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. A recurrent neural network that incorporates a window of +/- 1 codon around the affected codon was combined with 28 high-level annotation scores and allele frequency features to develop the two proposed models. 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 from a large list of candidate mutations in next-generation sequencing experiments 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 mutations that result in premature stop codons. Because experimentally validating the effects of these mutations is time-consuming and costly, computational predictive methods have been developed for this purpose[1-6]. 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 and score both nsSNVs and nfINDELs.

MetaRNN and MetaRNN-indel integrated information from 16 high-level pathogenicity prediction scores, such as CADD[1] and REVEL[2], 8 conservation scores and allele frequency information from the 1000 Genomes Project (1000GP)[7], ExAC[8], and gnomAD[9] (full list in **Supplementary Note**). The integrated information for all target mutations and their flanking base pairs, defined as a ±1 AA window around the affected codon(s), was extracted by a deep recurrent neural network (RNN) with multiple stacked bidirectional gated recurrent units[10] (GRUs) (**Supplementary Note**). The MetaRNN model takes as input a 11x28 matrix (11 base-pairs each with 28 features) and the MetaRNN-indel model takes as input a 58x28 matrix (58 base-pairs each with 28 features; details in **Supplementary Note**). We trained the RNN model with 26,517 rare nsSNVs (absent from at least one of the three population datasets) and 2,057 rare nfINDELs from ClinVar[11] 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 constructed from well-known variation-disease association databases, i.e. ClinVar[11] and HGMD[12], and a TP53 functional mutation dataset[13]. We observed that MetaRNN outperformed other competitors across all the test datasets, including two rare nsSNV test sets (RNTS) that was

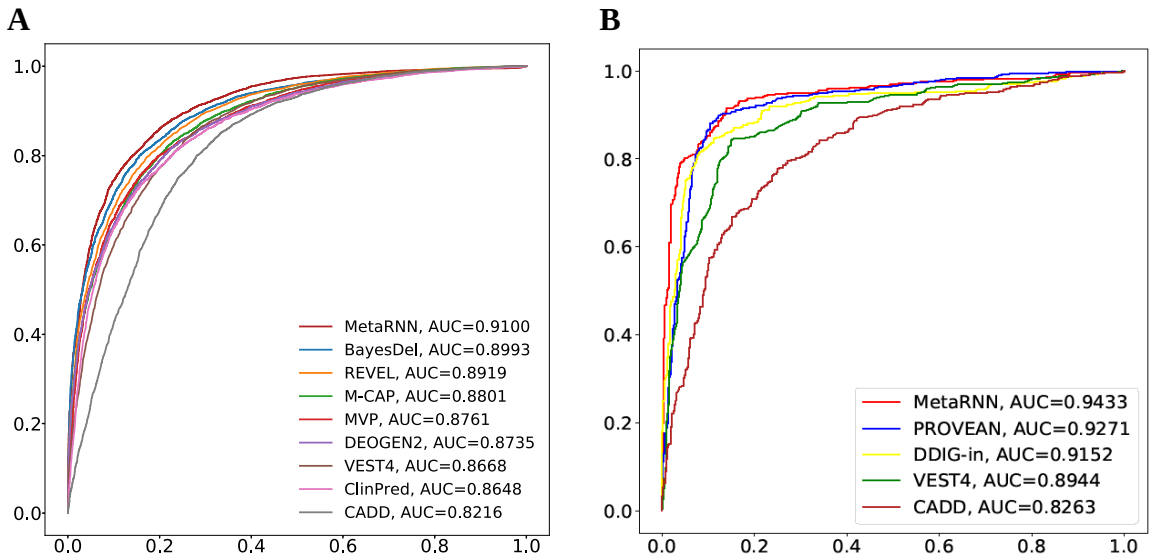

Figure 1. Comparisons of MetaRNN and MetaRNN-indel with other prediction tools. **A**: Performance comparison of MetaRNN and 8 other nsSNV prediction tools using the rare nsSNV test set (RNTS). **B**: Performance comparison of MetaRNN-indel and other nfINDEL prediction tools using ClinVar nfINDELs.

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**). To explore the generalizability of our model on disease-causing nsSNVs curated through other standards by another independent database, we retrieved disease-causing mutations (DM) from HGMD Professional v.2021.01 as true positives (n = 22,628) and matched rare nsSNVs observed in gnomAD v3 with allele frequencies between 0.01 and 0.0001 as true negatives (n = 22,628). Using this test dataset, MetaRNN still outperformed other competitors (**S. Figure 11**). 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 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[14] and CADD[1], that showed good performance for nsSNVs. The same findings were observed for nfINDELs retrieved from HGMD (**S. Figure 12**). 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[15], 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 conflicting group"), whereas 4,110 (20.2%) showed conflicting interpretations between pathogenic/likely pathogenic and unknown significance ("pathogenic conflicting group"). Based on the fact that the benign conflicting group had approximately 4 times more variants than the pathogenic conflicting 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[1], VEST[16], REVEL[2]) or potentially underestimated the proportion of VUSs at the extremes (BayesDel[5]), the proposed MetaRNN scores exhibited greater interpretability for sequencing data and VUSs. Additionally, we explored the score distributions for nsSNVs and nfINDELs used in testing the respective MetaRNN models. A clear bimodal distribution was observed for both MetaRNN and MetaRNN-indel predictions (**Figure 2B**). Based on a cut-off value of 0.5 as inherited by the sigmoid activation function used in the models, pathogenic nsSNVs and nfINDELs can be effectively separated from benign ones. These observations have two implications. First, using a cutoff of 0.5 is in accordance with interpretation of the scores as

probabilities where a score greater than 0.5 can be categorized as pathogenic and a score less than 0.5 can be categorized as benign. Second, with a shared cutoff value and similar distributions for nsSNV and nfINDEL scores across independent test datasets, we show that predictions from our two models, namely MetaRNN and MetaRNN-indel, are comparable. This feature can help increase the power of genome-wide association studies and related gene-set association analyses. 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. As illustrated previously, MetaRNN and MetaRNN-indel scores are well calibrated to be combined for an integrated (nsSNV+nfINDELs) rare-variant burden test for genotype-phenotype association. Such method has previously been shown to be able to increase power/reduce required sample size of the study[17].

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.

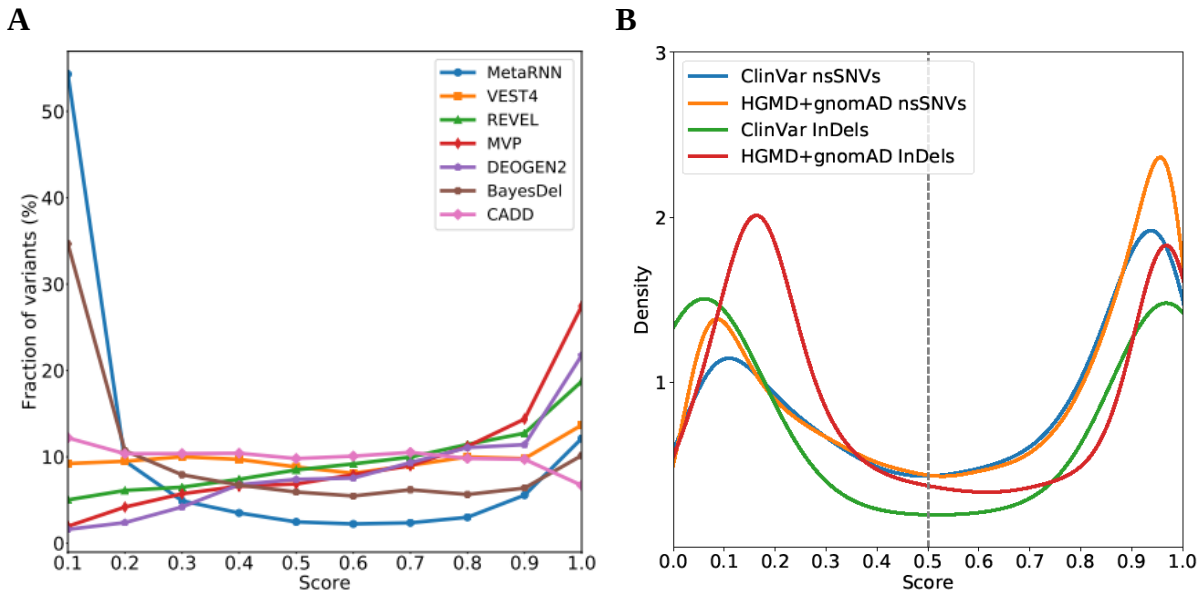

Figure 2. MetaRNN score distributions. **A**: Score distribution for ClinVar variants of unknown significance (VUSs). **B**: Score distributions of MetaRNN and MetaRNN-indel predictions on matched test datasets where the number of pathogenic and benign mutations are the same.

References

1 Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47**, D886-D894, doi:10.1093/nar/gky1016 (2019).

2 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).

3 Malhis, N., Jacobson, M., Jones, S. J. & Gsponer, J. LIST-S2: taxonomy based sorting of deleterious missense mutations across species. *Nucleic acids research* **48**, W154-W161 (2020).

4 Alirezaie, N., Kernohan, K. D., Hartley, T., Majewski, J. & Hocking, T. D. ClinPred: prediction tool to identify disease-relevant nonsynonymous single-nucleotide variants. *The American Journal of Human Genetics* **103**, 474-483 (2018).

5 Feng, B. J. PERCH: a unified framework for disease gene prioritization. *Human mutation* **38**, 243-251 (2017).

6 Qi, H. *et al.* MVP: predicting pathogenicity of missense variants by deep learning. *bioRxiv*, 259390 (2018).

7 in *Nature* Vol. 526 68-74 (2015).

8 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291, doi:10.1038/nature19057 (2016).

9 Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443, doi:10.1038/s41586-020-2308-7 (2020).

10 Cho, K., Bart, Bahdanau, D. & Bengio, Y. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *arXiv pre-print server*, doi:arxiv:1409.1259 (2014).

11 Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* **46**, D1062-D1067, doi:10.1093/nar/gkx1153 (2018).

12 Stenson, P. D. *et al.* The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting. *Human Genetics*, 1-11 (2020).

13 Leroy, B. *et al.* The TP53 website: an integrative resource centre for the TP53 mutation database and TP53 mutant analysis. *Nucleic Acids Res* **41**, D962-969, doi:10.1093/nar/gks1033 (2013).

14 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, doi:10.1002/humu.22911 (2016).

15 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. *Genetics in Medicine* **17**, 405-423, doi:10.1038/gim.2015.30 (2015).

16 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**, 1-16 (2013).

17 Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci U S A* **111**, E455-464, doi:10.1073/pnas.1322563111 (2014).

18 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 Med* **12**, 103, doi:10.1186/s13073-020-00803-9 (2020).

19 Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum Mutat* **32**, 894-899, doi:10.1002/humu.21517 (2011).

20 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164, doi:10.1093/nar/gkq603 (2010).