Published in final edited form as:

J Med Genet. 2017 February ; 54(2): 134–144. doi:10.1136/jmedgenet-2016-104369.

The performance of deleteriousness prediction scores for rare non-protein-changing single nucleotide variants in human genes

Xiaoming Liu¹, Chang Li¹, and Eric Boerwinkle^{1,2}

¹Human Genetics Center and Department of Epidemiology, Human Genetics and Environmental Sciences, UTHealth School of Public Health, Houston, Texas, USA

²Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, USA

In recent years, whole genome sequencing has increasingly been used as a replacement of whole exome sequencing for identifying causal variants of human diseases. In response to this trend, several “genome-level” deleteriousness prediction scores have been proposed to implement the scores designed specifically for missense or splicing variants. The aim of this study was to investigate the prediction accuracy of those genome-level scores for rare non-protein-changing SNVs (npcSNVs) in and near human genes. We compared 15 genome-level deleteriousness prediction scores and 8 conservation scores using receiver operating characteristic (ROC) and Area Under Curve (AUC). We found fathmm-MKL coding score [1] was the best score for npcSNVs (AUC=0.875), outperforming other genome-level deleteriousness prediction scores and conservation scores.

As the cost of whole genome and exome sequencing has reduced considerably, clinical use of sequencing data is becoming more popular. Even though more candidate single nucleotide variants (SNVs) were identified and reported, interpretation of these variants accurately in a clinical context remains a challenge. To facilitate and standardize the interpretation of sequence variants, the American College of Medical Genetics and Genomics (ACMG) have developed a new five-tier, evidence based guideline [2]. As a major component of this guideline, in silico prediction of variant deleteriousness has been widely used in screening and prioritizing candidate variants from a large number of background variants.

Multiple algorithms exist to predict variant deleteriousness based on different properties of the variant. Previously most algorithms focused on SNVs altering amino acid or splicing. In recent years a new group of algorithms have been proposed for predicting the deleteriousness of a variant in genes or intergenic regions, such as CADD [3], DeepSEA [4], Eigen/Eigen-PC [5], fitCons [6] and deltaSVM [7]. How those genome-level deleteriousness prediction algorithms perform for npcSNVs, especially compared to conservation scores, are challenging questions faced by medical geneticists who have begun applying whole genome sequencing to clinical diagnosis. To meet the growing needs of SNV functional prediction at

Correspondence: Xiaoming Liu (Xiaoming.Liu@uth.tmc.edu).

Contributors X.L. designed the study and collected the annotation resources. C.L. performed the comparison. E.B. and X.L. supervised the study. X.L. and E.B. wrote the draft manuscript.

Competing interests The authors declare that they have no competing interests.

both clinical and research settings, we conducted a comprehensive comparison of the predictive performance of genome-level deleteriousness prediction scores using high quality testing data sets in and near genes (from 5 kb upstream to 5 kb downstream).

We collected disease-causing npcSNVs from the professional version of the HGMD database [8] version 2015.4 as our “deleterious” groups (see online supplementary notes for details). As there are very few (< 10) intergenic SNVs (> 5 kb from any coding or non-coding genes) reported as disease-causing, we excluded those SNVs and focused on SNVs in or near a gene (from 5 kb upstream to 5 kb downstream). We collected likely benign singleton SNVs observed in healthy cohort samples from the UK10K [9] for npcSNVs (see online supplementary notes). We used only singleton SNVs in our “benign” control group because (1) mismatching the allele frequencies of the disease-causing SNVs in the “deleterious” and “benign” groups may introduce comparison bias; (2) the ability to separate very rare disease-causing SNVs from very rare benign SNVs is critical in narrowing down the candidates of novel disease-causing SNVs for Mendelian disease studies. To avoid bias caused by different numbers of SNVs from different sets of genes presented in the “deleterious” and “benign” groups, we matched the number and the functional category of the “benign” SNVs to those of the “deleterious” SNVs for each gene (see online supplementary notes for details). The SNV lists of HGMD (hg38 coordinates) and UK10K (hg19 coordinates) were annotated using the WGS [10] pipeline. The SNVs were filtered to remove splicing site SNVs, stopgain SNVs, stoploss SNVs and SNVs causing different amino acid changes in different transcripts. Any “deleterious” npcSNVs that are observed in UK10K data were removed from further analysis. This resulted in 2,578 deleterious npcSNVs and 2,578 benign npcSNVs as testing set I (see online supplementary table S1). We obtained 23 deleteriousness prediction scores and conservation scores (Table 1) (CADD, DANN, DeepSEA, DeepSEA HGMD probability, deltaSVM_gm12878weights, deltaSVM_hepg2weights, deltaSVM_k562weights, Eigen, Eigen-PC, fathmm-MKL-coding, fathmm-MKL-noncoding, fitCons_integrated, funseq2_noncoding, GenoCanyon, REMM, GERP++, phyloP100way Vertebrate, phyloP46way Placental, phyloP46way Primate, phastCons100way Vertebrate, phastCons46way Placental, phastCons46way Primate, and SiPhy_29way_logOdds, see online supplementary notes for complete references) for the testing sets, using the WGS annotation pipeline. In case for a SNV there are multiple scores from the same algorithm, the most deleterious one was used. Because some of the other algorithms may have used an earlier version the HGMD data for training, we also constructed a subset of testing set I as testing set II (196 deleterious SNVs and 196 benign SNVs, online supplementary table S2) using deleterious SNVs reported in year 2015 and matched benign SNVs from UK10K singletons. We compared the prediction accuracy of those scores using receiver operating characteristic (ROC) curves and the area under the ROC curves (AUC).

The comparison results suggest that for npcSNVs, fathmm-MKL coding score topped the list with AUC of 0.875 (95% CI: 0.8654–0.8847) followed by fathmm-MKL noncoding score (AUC of 0.8671) (Figure 1A). The difference between fathmm-MKL coding score and noncoding score is that the former was based on 10 groups of conservation and epigenomic features and trained with coding disease-causing variants, while the latter was based on a subset of 4 features and trained with noncoding disease-causing variants [1]. The fathmm-

MKL coding score significantly outperforms other genome-level deleteriousness prediction scores and conservation scores. For testing set II, the two fathmm-MKL scores also outperformed other scores but not significantly better than some of the conservation scores (Figure 1B). The performance of other genome-level deleteriousness prediction scores is just on a par with conservation scores (Figure 1).

Readers shall note that there are limitations of our comparison: (1) we do not have sufficient data to compare the genome-level prediction scores for intergenic npcSNVs; (2) our “deleterious” testing data are from disease-causing SNVs, which are certainly biased towards SNVs with large effects on conserved sites, therefore may in favor of methods training on similar data, such as the fathmm-MKL scores; (3) although we have taken precautionary steps to collect our testing data, false positive and false negative variants may still exist. However, considering all the results, we remain optimistic that using genome-level scores, such as the fathmm-MKL coding score, can be a reasonable choice for analyses that prefer a one-size-fit-all score, such as a sliding-window based genotype-phenotype association analysis. Finally, as recommended by ACMG, deleteriousness prediction shall only be considered as one line of supporting evidence [2]. The result of these predictive algorithms shall to be further evaluated in combination with other lines of evidence such as the variant segregating pattern in pedigrees [2].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Dr. Alanna Morrison for providing the HGMD database.

Funding This study was supported by the US National Institutes of Health (5RC2HL102419 and U54HG003273).

References

1. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day INM, Gaunt TR, Campbell C. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*. 2015; 31:1536–43. [PubMed: 25583119]
2. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL, ACMG Laboratory Quality Assurance Committee. 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*. 2015; 17:405–24. [PubMed: 25741868]
3. 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; 46:310–5. [PubMed: 24487276]
4. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods*. 2015; 12:931–4. [PubMed: 26301843]
5. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet*. 2016; 48:214–20. [PubMed: 26727659]
6. Gulko B, Hubisz MJ, Gronau I, Siepel A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat Genet*. 2015; 47:276–83. [PubMed: 25599402]

7. Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, Beer MA. A method to predict the impact of regulatory variants from DNA sequence. *Nat Genet.* 2015; 47:955–61. [PubMed: 26075791]
8. Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet.* 2014; 133:1–9. [PubMed: 24077912]
9. The UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature.* 2015; 526:82–90. [PubMed: 26367797]
10. Liu X, White S, Peng B, Johnson AD, Brody JA, Li AH, Huang Z, Carroll A, Wei P, Gibbs R, Klein RJ, Boerwinkle E. WGS: an annotation pipeline for human genome sequencing studies. *J Med Genet.* 2016; 53:111–2. [PubMed: 26395054]

A

B

Figure 1. Comparison of deleteriousness predictions for non-protein-changing SNVs. (A) Comparison of 23 deleteriousness prediction scores or conservation scores for non-protein-changing SNVs using testing set I (2,578 deleterious SNVs and 2,578 benign SNVs). (B) Comparison of 23 deleteriousness prediction scores or conservation scores for non-protein-changing SNVs using testing set II (196 deleterious SNVs and 196 benign SNVs). Legends of the scores are ranked according to AUC.

Table 1

A list of prediction scores and conservation scores used in this comparison.

Score	Brief Description	URL
CADD (v1.3)	A genome-wide deleteriousness prediction score for DNA variants based on 63 sequence features.	http://cadd.gs.washington.edu/
DANN	A functional prediction score retrained based on the training data of CADD using deep neural network.	https://cbcl.ics.uci.edu/public_data/DANN/
FATHMM-MKL	A genome-wide deleteriousness prediction score for SNVs based on 10 feature groups (for coding variants) or on 4 feature groups (for noncoding variants).	http://fathmm.biocompute.org.uk/fathmmMKL.htm
GenoCanyon	A genome-wide functional prediction score based on 22 computational and experimental annotations using an unsupervised statistical learning.	http://genocanyon.med.yale.edu/
Eigen & EigenPC	A genome-wide functional prediction score using an unsupervised statistical learning.	http://www.columbia.edu/~ii2135/eigen.html
fitCons	A genome-wide deleteriousness measure for genomic positions based on functional assays and selective pressure estimation.	http://compgen.cshl.edu/fitCons/
deltaSVM	A genome-wide deleteriousness prediction score for regulatory DNA variants. Scores weighted with GM12878 DHS, K562 DHS + H3K4me1 and HepG2 DHS + H3K4me1.	http://www.beerlab.org/deltasvm/
Funseq2	A genome-wide deleteriousness	http://funseq2.gersteinlab.org/

Score	Brief Description	URL
	prediction score designed for non-coding somatic SNVs in cancer.	
GERP++	A conservation score measured by “Rejected Substitutions”.	http://mende1.stanford.edu/SidowLab/downloads/gerp/
phastCons46way primate	A conservation score based on 46way alignment primate set.	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons46way/primates/
phastCons46way placental	A conservation score based on 46way alignment placental set.	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons46way/placentalMammals/
phastCons100way vertebrate	A conservation score based on 100way alignment vertebrate set.	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons100way/hg19.100way.phastCons/
phyloP46way primate	A conservation score based on 46way alignment primate set.	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phyloP46way/primates/
phyloP46way placental	A conservation score based on 46way alignment placental set.	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phyloP46way/placentalMammals/
phyloP100way vertebrate	A conservation score based on 100way alignment vertebrate set.	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phyloP100way/hg19.100way.phyloP100way/
SiPhy	A conservation score based on 29 mammals genomes.	http://www.broadinstitute.org/mammals/2x/siphy_hg19/