Humu-2015-0326

Databases

dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Non-synonymous and Splice Site SNVs

Xiaoming Liu[1,2], Chunlei Wu[3], Chang Li[2] and Eric Boerwinkle[1,2,4]

[1]Human Genetics Center and [2]Department of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas, USA; [3]Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, California, USA; [4]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, USA.

***Correspondence to:** Xiaoming Liu, Ph.D., Human Genetics Center, The University of Texas Health Science Center at Houston, 1200 Pressler Street, E529, Houston, Texas 77030, USA; Phone: 1-713-500-9820;

E-mail: Xiaoming.Liu@uth.tmc.edu

Contract grant sponsor: National Institutes of Health; Contract grant number: 5RC2HL102419 and U54HG003273

ABSTRACT

The purpose of the dbNSFP is to provide a one-stop resource for functional predictions and annotations for human non-synonymous single-nucleotide variants (nsSNVs) and splice site variants (ssSNVs), and to facilitate the steps of filtering and prioritizing SNVs from a large list of SNVs discovered in an exome-sequencing study. A list of all potential nsSNVs and ssSNVs based on the human reference sequence were created, functional predictions and annotations were curated and compiled for each SNV. Here we report a recent major update of the database to version 3.0. The SNV list has been rebuilt based on GENCODE 22 and currently the database includes 82,832,027 nsSNVs and ssSNVs. An attached database dbscSNV, which compiled all potential human SNVs within splicing consensus regions and their deleteriousness predictions, add another 15,030,459 potentially functional SNVs. Eleven prediction scores (MetaSVM, MetaLR, CADD, VEST3, PROVEAN, 4× fitCons, fathmm-MKL and DANN) and allele frequencies from the UK10K cohorts and the Exome Aggregation Consortium (ExAC), among others, have been added. The original seven prediction scores in v2.0 (SIFT, 2× Polyphen2, LRT, MutationTaster, MutationAssessor and FATHMM) as well as many SNV and gene functional annotations have been updated. dbNSFP v3.0 is freely available at http://sites.google.com/site/jpopgen/dbNSFP.

Key Words: dbNSFP; dbscSNV, non-synonymous mutation; splice site mutation; functional prediction; database

INTRODUCTION

With the advancement of technologies and the drop of the associated expenses, DNA sequencing is increasingly used as a research as well as diagnostic tool for human diseases. Among all the sequencing strategies, whole exome sequencing (WES) is probably the most popular for identifying novel genes and mutations causing genetic diseases. Currently, the cost of WES is roughly on par with targeted sequencing of a few genes while delivering the genotypes of the whole exome. Compared to whole genome sequencing with the same depth, with only a fraction of the cost WES is able to discover some of the most important candidates for disease causing mutations, including presumably functional single-nucleotide variants (SNVs): stop-gain, stop-loss, missense, splice site, and those within splicing consensus regions (−3 to +8 at the 5’ splice site and −12 to +2 at the 3’ splice site).

The major aim of dbNSFP is to facilitate the process of filtering and prioritizing the above mentioned presumably functional SNVs from a long list of SNVs identified in a typical WES study. To make it truly scalable to large WES studies and avoid security concerns, dbNSFP was designed to work as a local and self-sustaining database without need for internet connection. This database compiled all potential non-synonymous SNVs (nsSNVs, including stop-gain, stop-loss and missense), splice site SNVs (ssSNVs) and SNVs in splicing consensus regions (scSNVs, via attached database dbscSNV; see below) based on a human reference sequence. Functional predictions and annotations for each SNV from many methods and resources were exhaustively curated. Searching the database using the companion Java program can be accomplished by a single command line call, therefore it is easy to operate for researchers with minimum bioinformatics training.

dbNSFP has expanded since its first release in 2011. dbNSFP v1.0 (Liu et al. 2011) was based on the human reference sequence version hg18 and the gene model of Consensus Coding Sequence (CCDS) version 20090327 (Pruitt et al. 2009). It included 75,931,005 nsSNVs and four functional prediction scores: SIFT (Ng and Henikoff 2001), Polyphen2 (Adzhubei et al. 2010), LRT (Chun and Fay 2009) and MutationTaster (Schwarz et al. 2010), and one conservation score: phyloP (Siepel et al. 2006) for each nsSNV. dbNSFP v2.0 (Liu et al. 2013) was rebuilt based on the human reference sequence version hg19 and the gene model of GENCODE 9 (Harrow et al. 2012). It compiled 87,347,043 nsSNVs and 2,270,742 ssSNVs. It added two functional prediction scores, MutationAssessor (Reva et al. 2011) and FATHMM (Shihab et al. 2013), two conservation scores, GERP++ (Davydov et al. 2010) and SiPhy (Garber et al. 2009; Lindblad-Toh et al. 2011), and allele frequencies from the 1000 Genomes Project phase 1 data (The 1000 Genomes Project Consortium 2012) and the NHLBI Exome Sequencing Project data (Fu et al. 2013). Rich functional annotations for human genes were also added to dbNSFP v2.0. dbNSFP has gained popularity among human geneticists and has been adopted by mainstream annotation tools/resources, including the UCSC Genome Browser's Variant Annotation Integrator (http://genome.ucsc.edu/cgi-bin/hgVai), Ensembl's Variant Effect Predictor (McLaren et al. 2010), ANNOVAR (Wang et al. 2010), SnpEff/SnpSift (Cingolani et al. 2012) and HGMD (Stenson et al. 2014), among others.

Here we report a recent major update of dbNSFP to v3.0. The core SNVs have been rebuilt based on the human reference sequence version hg38. It now includes 82,832,027 nsSNVs and ssSNVs. An attached database called dbscSNV (Jian et al. 2014) which compiled all potential human scSNVs (15,030,459 in total) is distributed along with dbNSFP, and can be searched using the same companion search program of dbNSFP. Compared to v2.0, the new version added eleven new prediction scores: MetaSVM and MetaLR (Dong et al. 2015), CADD (Kircher et al. 2014), VEST3 (Carter et al. 2013), PROVEAN (Choi et al. 2012), 4× fitCons scores (Gulko et al. 2015), fathmm-MKL (Shihab et al. 2015) and DANN (Quang et al. 2015), two conservation scores: 2× phastCons (Siepel et al. 2005), and allele frequencies from the UK10K cohorts (The UK10K Consortium 2015) and the Exome Aggregation Consortium (ExAC, http://exac.broadinstitute.org/), among others. Many prediction scores and resources have been updated. Details of the updates and preliminary analyses of the functional prediction scores and conservation scores are reported in the following sections.

NEW AND UPDATED FUNCTIONAL ANNOTATIONS

To keep up with the updates of new gene models, we have rebuilt our backbone nsSNVs and ssSNVs using the GENCODE 22, which is based on human reference sequence version hg38. As described previously (Liu et al. 2013), we artificially "mutated" each non-N reference allele to the three alternative alleles. Then we checked the "mutations" against the gene models and collected all those nsSNVs or ssSNVs (on the first two and last two nucleotide sites of an intron) into our database. To balance false positives and false negatives of the gene models, we included putative genes but excluded genes with incomplete 5' ends. Genes on the mitochondrial genome has been included for the first time. This resulted in 80,622,428 nsSNVs and 2,209,599 ssSNVs in the database. Genome positions were converted to corresponding coordinates in hg19 (no missing) and then in hg18 (0.09% missing) using the liftOver tool of the UCSC Genome Browser (Rosenbloom et al. 2015). Please note that there are a few SNVs whose coordinates in hg38 and hg19 (hg18) have inconsistent chromosome numbers.

Two new nsSNV-focused prediction scores, PROVEAN and VEST 3.0 have been added, which were kindly provided by Drs. Yongwook Choi and Rachel Karchin, respectively. PROVEAN scores range from -14 to 14 in dbNSFP, with a lower score indicating a higher likelihood to be deleterious. PROVEAN also provides binary predictions (*Neutral* versus *Damaging*) with a score cut-off of -2.5. Multiple scores and predictions corresponding to multiple transcripts of the same gene are separated by ";" and the transcript IDs are presented in the *Ensembl_proteinid* column. VEST 3.0 scores range from 0 to 1 with a higher score indicating a higher likelihood to be deleterious. VEST does not provide binary predictions. Multiple scores are separated by ";" and the corresponding transcript IDs are presented in the *Transcript_id_VEST3* column.

Recently, several "general" prediction scores have been proposed, which incorporated DNA/protein sequence features as well as epigenomic signals and provide deleteriousness predictions for any SNV in the human genome, coding or non-coding. Examples of such scores include CADD, fitCons, fathmm-MKL and DANN. Among them CADD, fathmm-MKL and DANN provide predictions for a SNV while fitCons is more coarse-grained and has predictions at the genome position level as a conservation score. We included the above mentioned four "general" prediction scores in dbNSFP v3.0 to provide more choices for our users. fathmm-MKL separated their scores for coding and non-coding SNVs and we included those designed for coding SNVs.

Although having more prediction scores for an nsSNV has an advantage of providing additional perspectives, sometime a consensus prediction is also useful in practice. We recently developed two ensemble scores, MetaLR and MetaSVM, based on 10 component scores (SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, GERP++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy, PhyloP) and the maximum frequency observed in the 1000 genomes populations (Dong et al. 2015). Based on our comparison, the two ensemble scores outperform all their component scores. MetaLR achieved the highest separation power (AUC = 0.92 and 0.94 for testing dataset I and II, respectively) followed by MetaSVM (AUC = 0.91 and 0.93 for testing dataset I and II, respectively).

To make the functional prediction scores and conservation scores in the dbNSFP more comparable to each other, we created a rank score for each of them. First, we converted scores if necessary to make them monotonic in the same direction (a higher score indicating more likely to be damaging, see Suppporting Information for details). Then for each type of score (such as a converted SIFT score) we ranked all the (converted SIFT) scores in the dbNSFP and the rank score is the ratio of the rank (or tied rank) of the (converted SIFT) score over the total number of (converted SIFT) scores in the dbNSFP. In the case when an nsSNV has multiple scores due to multiple transcripts, only the most deleterious one was used in ranking. Therefore, a rank score is always between 0 and 1 and a score of 0.9 means it is more likely to be damaging than 90% of all potential nsSNVs predicted by that method.

Many prediction scores and conservation scores have been updated from the dbNSFP v2.0 to v3.0: SIFT to the version based on Ensembl 66; MutationTaster to MutationTaster2 (Schwarz et al. 2014); FATHMM to v2.3; phyloP to phyloP7way_vertebrate and phyloP20way_mammalian (both based on hg38); phastCons to phastCons7way_vertebrate and phastCons20way_mammalian (both based on hg38). As many prediction scores provide multiple (often different) scores or predictions for the same nsSNV due to multiple transcripts of the same gene, we included those transcript specific predictions in this new version.

Besides prediction scores and conservation scores, many annotation resources have been added or updated. Noticeably, allele frequencies from the UK10K cohorts and the Exome Aggregation Consortium (ExAC) have been added; those of human populations in the 1000 Genomes Project have been updated to the phase 3 data set. Clinvar (Landrum et al. 2014), dbSNP (Sherry et al. 2001) 142 and phenotypes of mouse and zebra fish homologs have been added. More details on the resources and their version in dbNSFP can be found in the Supporting Information and the readme file distributed with the database file.

The dbNSFP v3.0 is provided in two branches: v3.0a and v3.0c. The former includes all the prediction scores and annotation resources while the latter excludes prediction scores that require licenses for commercial usages, such as VEST, CADD and DANN. The whole database is in plain text format. No database management system is needed. A Java companion search program along with the database files are freely available at https://sites.google.com/site/jpopgen/dbNSFP. Alternatively, dbNSFP can be queried via MyVariant.info web service (http://myvariant.info/), either calling its API directly or using its Python client (Mark 2015a) and R client from Bioconductor (Mark 2015b).

A COMPARISON OF FUNCTIONAL PREDICTION SCORES AND CONSERVATION SCORES

We conducted some preliminary analyses comparing the 24 functional prediction scores and conservation scores based on the 80,622,428 nsSNVs in dbNSFP v3.0. A summary of the 24 scores is presented in Table 1. nsSNV-focused scores typically have a higher missingness percentage in the dbNSFP (a minimum of 2.15% for MutationTaster to a maximum of 16.68% for LRT) compared to “general” prediction scores or conservation scores (a minimum of <0.01%

for CADD to a maximum of 3.97% for fitCons), largely due to gene model inconsistency (Table 2). As to the distributions of the rank scores (Figure 1), while some rank scores are more or less evenly distributed, such as MutationAssessor, FATHMM, PROVEAN, VEST3, CADD, DANN, fathmm-MKL, MetaSVM, MetaLR, GERP++ and SiPhy, others are more sparse and have high spikes, suggesting a large amount of raw scores having tied ranks in the database.

Knowing the correlation between scores helps researchers to weight the predictions from multiple methods. For each pair of scores, we calculated the Pearson's correlation coefficient (r) between their rank scores as a measure of correlation (Table 3). Some of the highly correlated ($r > 0.7$) pairs either use the same training data or use the same method, such as Polyphen2-HVIR and Polyphen2-HVAR, MetaSVM and MetaLR, CADD and DANN, fitCons-i6 and fitCons-h1, phyloP7way_vertebrate and phyloP20way_mammalian, and phastCons7way_vertebrate and phastCons20way_mammalian. The others are less obvious, such as FATHMM and MetaLR (or MetaSVM), CADD and Polyphen2-HVAR (or Polyphen2-HDIV), CADD and VEST3, fathmm-MKL and GERP++, fathmm-MKL and SiPhy, and GERP++ and SiPhy. fitCons scores have low correlations ($r < 0.3$) with other scores. There is even a negative (though close to 0) correlation between fitCons-gm and FATHMM. To provide an entire perspective, we clustered the scores using UPGMA (Unweighted Pair Group Method with Arithmetic Mean) with 1-r as distance between scores (Figure 2). The scores fall in four clusters. The largest one includes LRT, MutationTaster, fathmm-MKL, GERP++, SiPhy, 2× phastCons scores and 2× phyloP scores. All conservation scores are in this cluster suggesting that the prediction scores in this cluster may put a heavy weight on conservation information. The second largest cluster includes SIFT, MutationAssessor, PROVEAN, VEST3, CADD, DANN and 2× Polyphen2 scores. The smallest cluster includes FATHMM, MetaSVM and MetaLR, which are highly correlated among themselves while having low to moderate correlation with other scores. Finally, the 4× fitCons scores form their own cluster and serve as an out group.

Among the 24 scores, eleven provide binary predictions (deleterious or tolerated) for nsSNVs. Comparison of their prediction agreement shows that majority of the pairs have low to moderate agreement rate (< 70%) (Table 3). The lowest agreement rate (40%) is between FATHMM and fathmm-MKL (coding score). The highest agreement is between MetaLR and MetaSVM (96%) followed by FATHMM and MetaLR (90%) and the two Polyphen2 scores (89%).

Finally, we compared the performance of the nsSNV prediction scores, "general" prediction scores and conservation scores in dbNSFP v3.0 using their rank scores. We re-used the testing dataset I and testing dataset II from Dong et al. (2015) after removing nsSNVs that causing different amino acid changes in different transcripts according to GENCODE 22, which resulting in 115 true positives and 117 true negatives in testing data set I (Supp. Table S1) and 5,979 true positives and 13,025 true negatives in testing data set II (Supp. Table S2), respectively. The performance of the scores was measured using receiver operating characteristic (ROC) curve and area under the curve (AUC) (Figure 3). We found that, the two ensemble rank scores, MetaSVM and MetaLR, achieved excellent prediction accuracy (AUC > 0.9) in both testing datasets. Two other scores that reached excellent prediction accuracy in either testing dataset include VEST3 (AUC=0.9294 in testing data set I) and FATHMM (AUC=0.912 in testing data set II). The results also showed that those recently proposed "general" scores did not stand out as to nsSNV deleteriousness prediction, although some of those, such as CADD, DANN and fathmm-MKL, showed comparable performance as popular nsSNV prediction scores Polyphen2 and SIFT.

ATTACHED DATABASE

Recently we developed a method for predicting the splice-altering effect of a scSNV (a SNV located within splicing consensus regions) (Jian et al. 2014). The resulting two ensemble prediction scores (ada_score and rf_score) and predictions were pre-computed for all potential scSNVs in the human genome based on RefSeq release 62 and Ensembl release 73. Those scores along with related annotations were compiled into a plain text database called dbscSNV and serves as an attached database for the dbNSFP. It is freely available for download at https://sites.google.com/site/jpopgen/dbNSFP. The companion Java search program distributed with dbNSFP v3.0 supports searching dbscSNV and SPIDEX (Xiong et al. 2015), another prediction tool for splice-altering SNVs, along with dbNSFP using the “-s” option.

ACKNOWLEDGMENTS

We thank Drs. Yongwook Choi, Rachel Karchin and Dominik Seelow for kindly providing the PROVEAN/SIFT, VEST3 and MutationTaster2 scores, respectively. We thank Dr. CS (Jonathan) Liu for providing hosting space. We thank Jason J. Corneveaux, Chris Gillies, Mihail Halachev, Jocob Hsu, James Ireland, Xueqiu Jian, Seung-Tae Lee, Alexander Li, John McGuigan, Zena Ng, Adam Novak, Kirill Prusov, and Lishuang Shen, for reporting bugs and providing suggestions. We thank Sara Barton for copy editing this manuscript. This project was supported by the US National Institutes of Health (5RC2HL102419 and U54HG003273). The authors declare that they have no conflict of interests.

REFERENCES

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat. Methods 7: 248–249.

Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. 2013. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics 14 Suppl 3: S3.

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. 2012. Predicting the functional effect of amino acid substitutions and indels. PloS One 7: e46688.

Chun S, Fay JC. 2009. Identification of deleterious mutations within three human genomes. Genome Res. 19: 1553 –1561.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6: 80–92.

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. PLoS Comput Biol 6: e1001025.

Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, Liu X. 2015. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum. Mol. Genet. 24: 2125–2137.

Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Rieder MJ, Altshuler D, Shendure J, Nickerson DA, Bamshad MJ, et al. 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493: 216–220.

Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. 2009. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics 25: i54–i62.

Gulko B, Hubisz MJ, Gronau I, Siepel A. 2015. A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat. Genet. 47: 276–283.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 22: 1760–1774.

Jian X, Boerwinkle E, Liu X. 2014. In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res. 42: 13534–13544.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. 46: 310–315.

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 42: D980–D985.

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, Ward LD, Lowe CB, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478: 476–482.

Liu X, Jian X, Boerwinkle E. 2011. dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. Hum. Mutat. 32: 894–899.

Liu X, Jian X, Boerwinkle E. 2013. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. Hum. Mutat. 34: E2393–2402.

Mark A. 2015a. MyVariant.py: MyVariant.info Python client. https://pypi.python.org/pypi/myvariant.

Mark A. 2015b. MyVariant.R: MyVariant.info R client. http://bioconductor.org/packages/myvariant/.

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26: 2069–2070.

Ng PC, Henikoff S. 2001. Predicting deleterious amino acid substitutions. Genome Res. 11: 863–874.

Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, Hart E, Suner M-M, et al. 2009. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Res. 19: 1316–1323.

Quang D, Chen Y, Xie X. 2015. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinforma. Oxf. Engl. 31: 761–763.

Reva B, Antipin Y, Sander C. 2011. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 39: e118.

Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, Harte RA, Heitner S, et al. 2015. The UCSC Genome Browser database: 2015 update. Nucleic Acids Res. 43: D670–681.

Schwarz JM, Cooper DN, Schuelke M, Seelow D. 2014. MutationTaster2: mutation prediction for the deep-sequencing age. Nat. Methods 11: 361–362.

Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. 2010. MutationTaster evaluates disease-causing potential of sequence alterations. Nat. Methods 7: 575–576.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 29: 308–311.

Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, Day INM, Gaunt TR. 2013. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum. Mutat. 34: 57–65.

Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day INM, Gaunt TR, Campbell C. 2015. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics 31: 1536–1543.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15: 1034 –1050.

Siepel A, Pollard KS, Haussler D. 2006. New methods for detecting lineage-specific selection. RECOMB 2006. LNCS (LNBI), vol. 3909, Heidelberg: Springer, p 190–205.

Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. 2014. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum. Genet. 133: 1–9.

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65.

The UK10K Consortium. 2015. The UK10K project identifies rare variants in health and disease. Nature advance online publication: 2015 Sep 14. doi: 10.1038/nature14962.

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

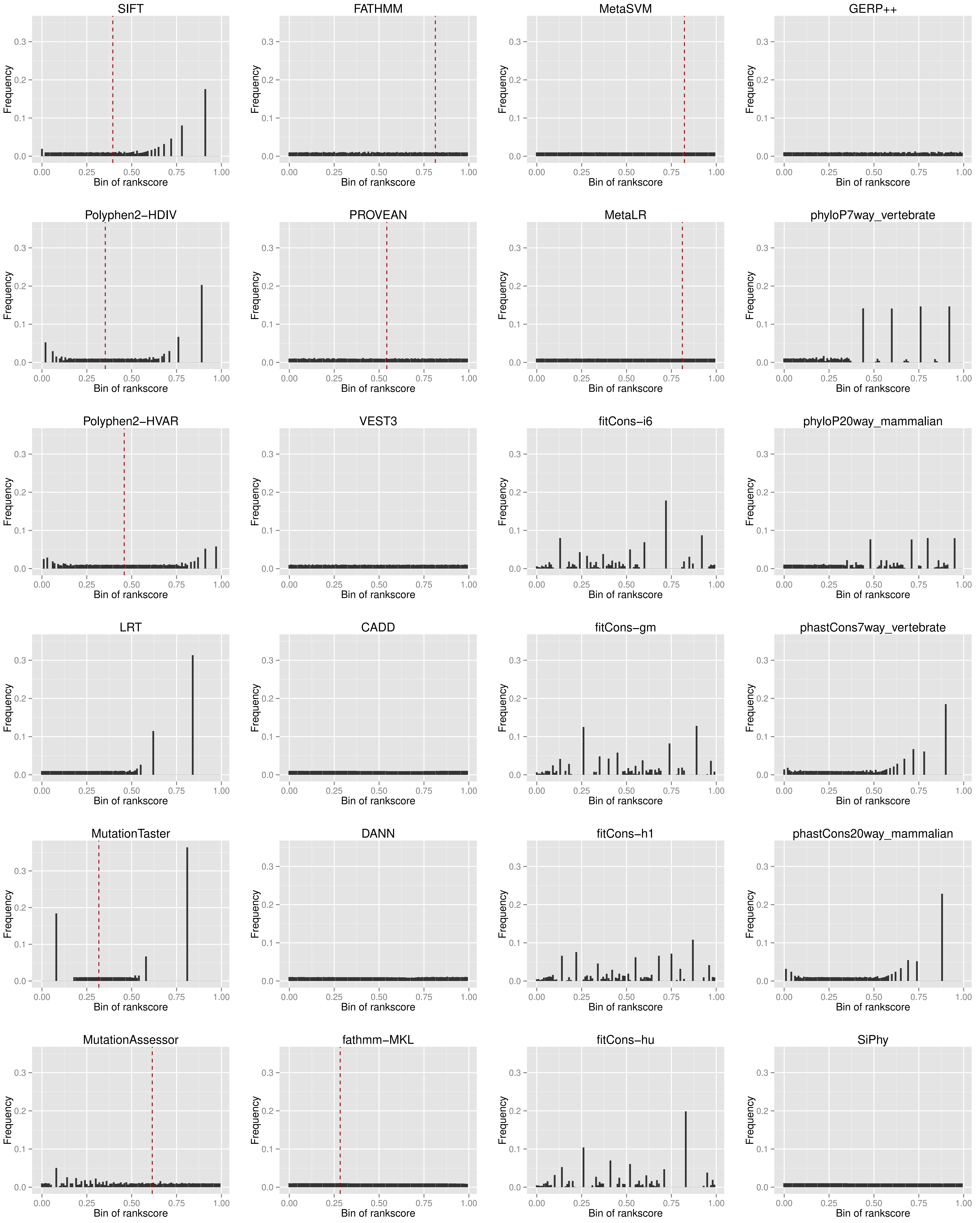
Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR, Morris Q, Barash Y, et al. 2015. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science 347: 1254806.

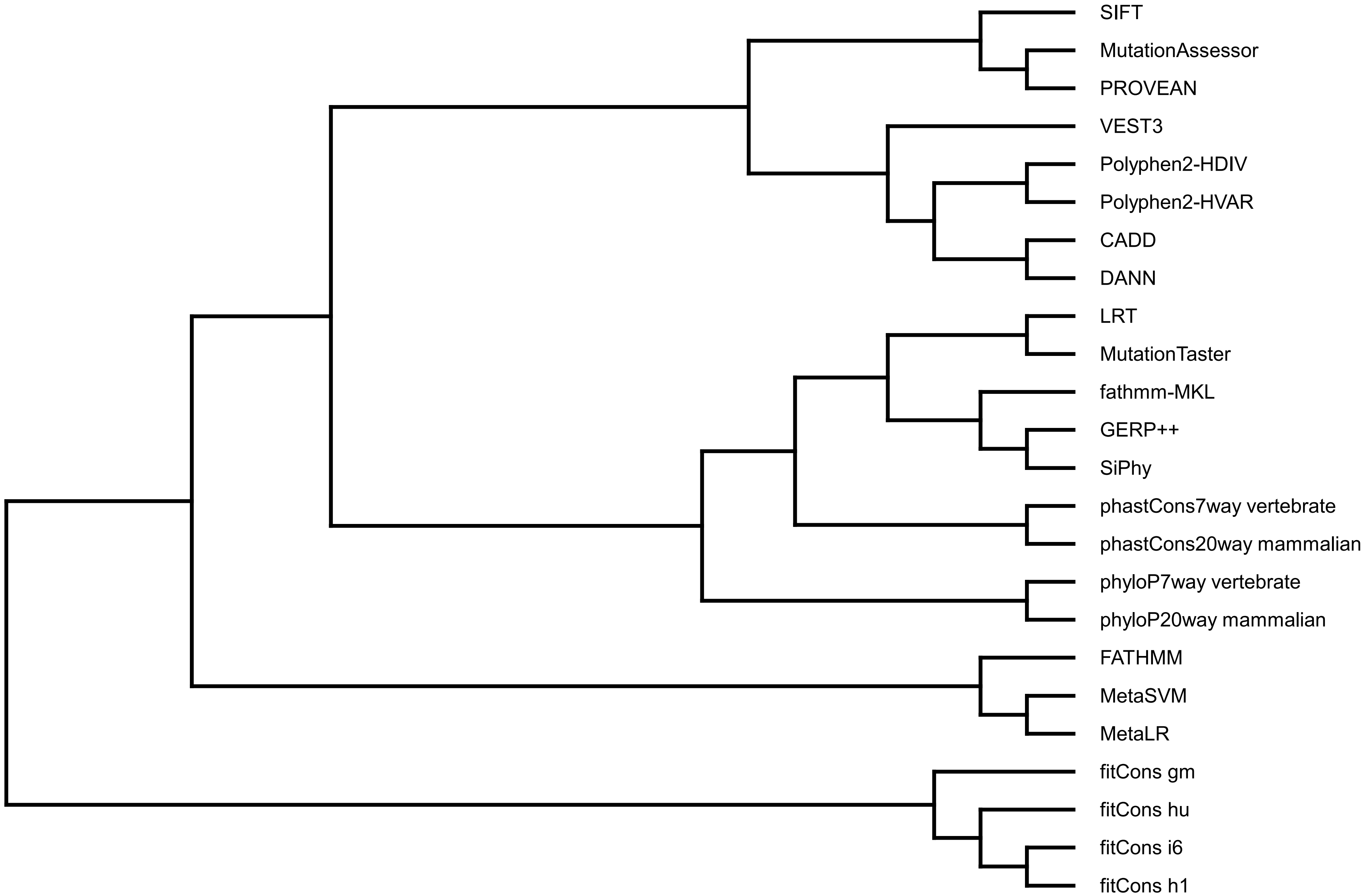
FIGURE LEGENDS

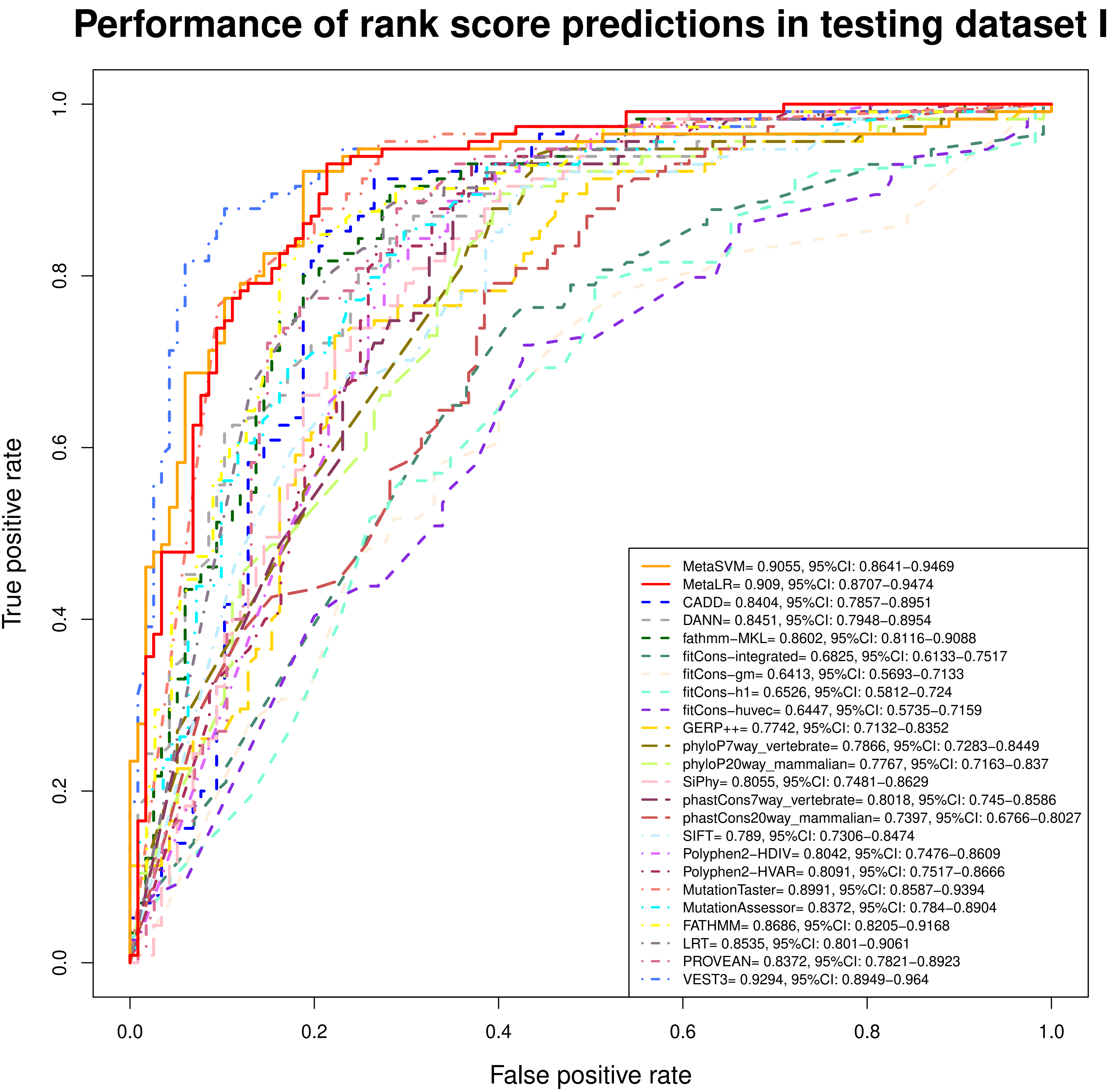
Figure 1: Distributions of the rank scores of the prediction and conservation scores based on 100 bins between 0 and 1. Dash lines indicate the cut-offs for binary predictions.

Figure 2: UPGMA dendrogram of the prediction and conservation scores.

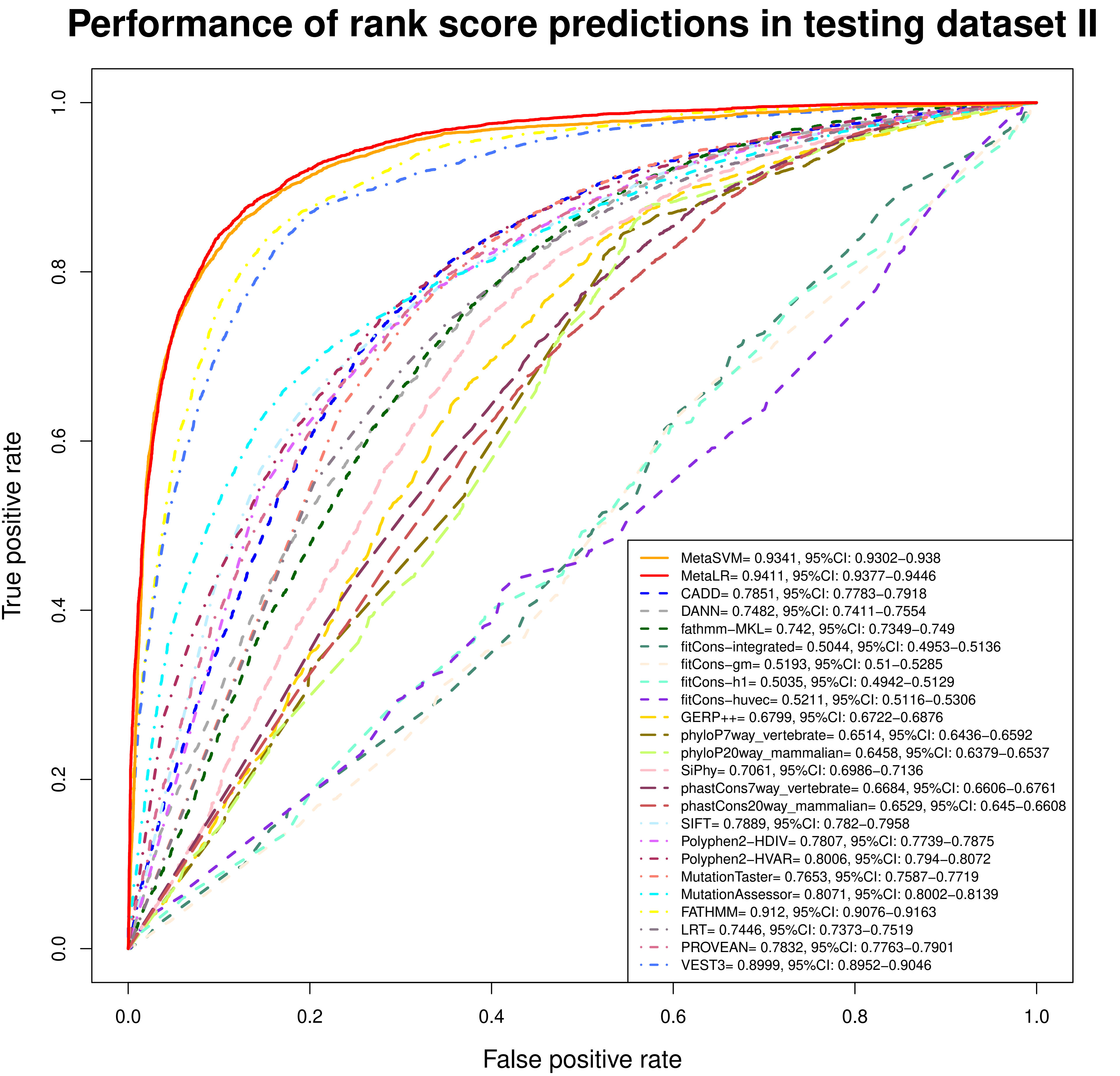
Figure 3: ROC curves for the functional prediction scores and conservation scores in dbNSFP v3.0 with testing data set I (A) and II (B).


SIFT
FATHMM
MetaSVM
GERP++
Polyphen2−HDIV
PROVEAN
MetaLR
phyloP7way_vertebrate
Polyphen2−HVAR
VEST3
fitCons−i6
phyloP20way_mammalian
LRT
CADD
fitCons−gm
phastCons7way_vertebrate
MutationTaster
DANN
fitCons−h1
phastCons20way_mammalian
MutationAssessor
fathmm−MKL
fitCons−hu
SiPhy
Frequency
Bin of rankscore
0.0
0.1
0.2
0.3
0.00
0.25
0.50
0.75
1.00


SIFT
MutationAssessor
PROVEAN
VEST3
Polyphen2-HDIV
Polyphen2-HVAR
CADD
DANN
LRT
MutationTaster
fathmm-MKL
GERP++
SiPhy
phastCons7way vertebrate
phastCons20way mammalian
phyloP7way vertebrate
phyloP20way mammalian
FATHMM
MetaSVM
MetaLR
fitCons gm
fitCons hu
fitCons i6
fitCons h1



Performance of rank score predictions in testing dataset I
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
MetaSVM= 0.9055, 95%CI: 0.8641–0.9469
MetaLR= 0.909, 95%CI: 0.8707–0.9474
CADD= 0.8404, 95%CI: 0.7857–0.8951
DANN= 0.8451, 95%CI: 0.7948–0.8954
fathmm-MKL= 0.8602, 95%CI: 0.8116–0.9088
fitCons-integrated= 0.6825, 95%CI: 0.6133–0.7517
fitCons-gm= 0.6413, 95%CI: 0.5693–0.7133
fitCons-h1= 0.6526, 95%CI: 0.5812–0.724
fitCons-huvec= 0.6447, 95%CI: 0.5735–0.7159
GERP++= 0.7742, 95%CI: 0.7132–0.8352
phyloP7way_vertebrate= 0.7866, 95%CI: 0.7283–0.8449
phyloP20way_mammalian= 0.7767, 95%CI: 0.7163–0.837
SiPhy= 0.8055, 95%CI: 0.7481–0.8629
phastCons7way_vertebrate= 0.8018, 95%CI: 0.745–0.8586
phastCons20way_mammalian= 0.7397, 95%CI: 0.6766–0.8027
SIFT= 0.789, 95%CI: 0.7306–0.8474
Polyphen2–HDIV= 0.8042, 95%CI: 0.7476–0.8609
Polyphen2–HVAR= 0.8091, 95%CI: 0.7517–0.8666
MutationTaster= 0.8991, 95%CI: 0.8587–0.9394
MutationAssessor= 0.8372, 95%CI: 0.784–0.8904
FATHMM= 0.8686, 95%CI: 0.8205–0.9168
LRT= 0.8535, 95%CI: 0.801–0.9061
PROVEAN= 0.8372, 95%CI: 0.7821–0.8923
VEST3= 0.9294, 95%CI: 0.8949–0.964



Performance of rank score predictions in testing dataset II
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
MetaSVM= 0.9341, 95%CI: 0.9302−0.938
MetaLR= 0.9411, 95%CI: 0.9377−0.9446
CADD= 0.7851, 95%CI: 0.7783−0.7918
DANN= 0.7482, 95%CI: 0.7411−0.7554
fathmm−MKL= 0.742, 95%CI: 0.7349−0.749
fitCons−integrated= 0.5044, 95%CI: 0.4953−0.5136
fitCons−gm= 0.5193, 95%CI: 0.51−0.5285
fitCons−h1= 0.5035, 95%CI: 0.4942−0.5129
fitCons−huvec= 0.5211, 95%CI: 0.5116−0.5306
GERP++= 0.6799, 95%CI: 0.6722−0.6876
phyloP7way_vertebrate= 0.6514, 95%CI: 0.6436−0.6592
phyloP20way_mammalian= 0.6458, 95%CI: 0.6379−0.6537
SiPhy= 0.7061, 95%CI: 0.6986−0.7136
phastCons7way_vertebrate= 0.6684, 95%CI: 0.6606−0.6761
phastCons20way_mammalian= 0.6529, 95%CI: 0.645−0.6608
SIFT= 0.7889, 95%CI: 0.782−0.7958
Polyphen2−HDIV= 0.7807, 95%CI: 0.7739−0.7875
Polyphen2−HVAR= 0.8006, 95%CI: 0.794−0.8072
MutationTaster= 0.7653, 95%CI: 0.7587−0.7719
MutationAssessor= 0.8071, 95%CI: 0.8002−0.8139
FATHMM= 0.912, 95%CI: 0.9076−0.9163
LRT= 0.7446, 95%CI: 0.7373−0.7519
PROVEAN= 0.7832, 95%CI: 0.7763−0.7901
VEST3= 0.8999, 95%CI: 0.8952−0.9046

Table 1: A summary of functional prediction scores and conservation scores.

Score	Training data	Information used	Prediction model
PolyPhen2-HDIV	5564 Mendelian disease mutations and 7539 divergence SNVs from close mammalian homolog proteins	eight sequence-based and three structure-based predictive features	naive Bayes classifier
PolyPhen2-HVAR	22196 disease associated SNVs and 21119 common SNVs	same as above	same as above
SIFT	1750 deleterious and 2254 tolerant nsSNVs of E. coli LacI gene	sequence homology based on PSI-BLAST	position specific scoring matrix
Mutation Taster	SNVs from 1000 G (1000 Genomes Project), HGMD	conservation, splice site, mRNA features, protein features; regulatory features	naive Bayes classifier
LRT	coding sequences of 32 vertebrate species	sequence homology	likelihood ratio test of codon neutrality
Mutation Assessor	SNVs from COSMIC database	sequence homology of protein families and sub-families within and between species	combinatorial entropy formalism
FATHMM	SNVs from HGMD and UniProt	sequence homology	hidden Markov models
PROVEAN	SNVs from UniProt/HUMSAVAR	sequence homology	Delta alignment score
VEST3	SNVs from HGMD and the Exome Sequencing Project	86 sequence features	Random Forest
fathmm-MKL coding	SNVs from HGMD and 1000G	conservation, epigenomic signals	multiple kernel learning
MetaSVM	36,192 SNVs from UnPprot	9 prediction scores and allele frequencies in 1000G	radial kernel support vector machine
MetaLR	same as above	same as above	logistic regression
CADD	16,627,775 "observed" variants and 49,407,057 "simulated" variants	63 annotations (949 features)	linear kernel support vector machine
DANN	same as above	same as above	deep neural network
fitCons-i6	genomes of 54 unrelated human individuals	epigenomic signals of GM12878, H1-hESC and HUVEC	INSIGHT (Inference of Natural Selection from Interspersed Genomically coHerent elemenTs)
fitCons-gm	same as above	epigenomic signals of GM12878	same as above
fitCons-h1	same as above	epigenomic signals of H1-hESC	same as above
fitCons-hu	same as above	epigenomic signals of HUVEC	same as above
SiPhy	genomes of 29 mammals	multiple alignments	inferring nucleotide substitution pattern per site
GERP++	genomes of 34 mammals	multiple alignments and phylogenetic tree	maximum likelihood evolutionary rate estimation
phyloP7way_vertebrate	genomes of 7 vertebrates	same as above	distributions of the number of substitutions based on a phylogenetic hidden Markov model
phyloP20way_mammalian	genomes of 20 mammals	same as above	same as above
phastCons7way_vertebrate	genomes of 7 vertebrates	same as above	two-state phylogenetic hidden Markov model
phastCons20way_mammalian	genomes of 20 mammals	same as above	same as above

Table 2: Number of nsSNVs in each chromosome and the percentages of missingness of functional prediction scores and conservation scores.

Chr	nsSNV	SIFT	Poly phen2	LRT	Mutation Taster	Mutation Assessor	FATHMM	PROVEAN	VEST3	CADD	DANN	fathmm -MKL	MetaSVM MetaLR	fitCons	GERP++	phyloP 7way	phyloP 20way	Phast Cons 7way	Phast Cons 20way	SiPhy
M	23145	64.21	100.00	100.00	6.35	100.00	13.07	13.04	100.00	1.06	100.00	100.00	100.00	100.00	0.00	0.00	0.00	0.00	0.00	100.00
1	8085329	10.97	10.22	15.11	1.53	12.53	15.40	10.51	7.21	0.00	0.00	0.00	7.56	0.00	0.60	0.53	0.24	0.53	0.24	1.42
2	5960951	9.20	10.87	19.41	1.84	11.82	14.01	8.76	7.11	0.00	0.00	0.00	7.04	0.00	0.38	0.04	0.05	0.04	0.05	0.77
3	4647575	8.30	8.47	12.87	1.34	11.14	13.09	8.01	6.74	0.00	0.00	0.00	6.67	0.00	0.35	0.06	0.04	0.06	0.04	0.65
4	3238883	8.96	11.67	12.24	2.04	11.54	13.60	8.52	7.14	0.00	0.00	0.00	7.23	0.00	0.39	0.20	0.03	0.20	0.03	2.19
5	3718178	8.67	8.76	16.17	0.74	10.92	12.36	8.17	6.95	0.00	0.00	0.00	7.67	0.00	0.12	0.09	0.07	0.09	0.07	0.54
6	4123833	9.43	9.77	12.46	2.92	12.06	12.89	8.58	6.70	0.00	0.00	0.00	7.75	0.02	0.16	0.01	0.13	0.01	0.13	0.70
7	3797070	12.30	11.09	19.95	4.36	14.63	17.17	11.64	8.05	0.00	0.00	0.00	9.35	0.00	1.02	0.24	0.06	0.24	0.06	2.34
8	2706162	11.11	10.03	13.94	3.13	13.95	17.21	10.98	7.09	0.00	0.00	0.00	8.23	0.00	0.50	0.15	0.08	0.15	0.08	1.05
9	3168302	9.69	9.38	13.34	1.87	10.95	14.77	9.24	7.31	0.00	0.00	0.00	7.08	0.00	0.14	0.02	0.00	0.02	0.00	0.48
10	3114019	9.98	9.88	12.36	2.01	12.67	14.86	9.68	7.51	0.00	0.00	0.00	8.01	0.00	0.54	0.04	0.00	0.04	0.00	1.05
11	4735763	9.68	9.83	16.46	2.62	12.41	14.25	9.10	7.49	0.00	0.00	0.00	7.90	0.00	0.08	0.00	0.00	0.00	0.00	0.59
12	4223205	8.71	9.62	12.89	0.57	12.23	13.74	8.05	6.41	0.00	0.00	0.06	6.36	0.00	0.06	0.03	0.03	0.03	0.03	0.59
13	1470936	12.14	11.00	10.19	2.87	12.86	16.69	11.83	7.01	0.00	0.00	0.00	8.91	0.00	0.12	0.01	0.00	0.01	0.00	0.91
14	2546323	8.74	10.61	13.94	1.51	12.55	14.04	8.01	7.12	0.00	0.00	0.00	7.68	0.00	0.23	0.16	0.11	0.16	0.11	0.59
15	2790630	9.17	11.82	15.92	2.22	13.26	13.30	8.63	7.05	0.00	0.00	0.00	9.99	0.00	1.28	0.11	0.07	0.11	0.07	2.00
16	3434017	11.11	11.91	17.86	2.81	15.81	15.34	9.85	7.37	0.00	0.00	0.00	10.00	0.00	2.24	1.44	0.37	1.44	0.37	3.36
17	4608227	17.72	10.19	16.23	1.32	12.57	17.14	17.43	7.28	0.00	0.00	0.00	8.85	0.00	0.50	0.08	0.03	0.08	0.03	1.17
18	1286209	12.95	11.07	17.43	0.59	11.82	12.44	12.37	8.02	0.00	0.00	0.00	8.76	0.00	0.08	0.06	0.02	0.06	0.02	1.01
19	5373215	16.72	11.65	35.61	2.56	14.12	14.97	16.36	9.63	0.00	0.00	0.00	10.25	0.00	0.22	0.04	0.02	0.04	0.02	1.73
20	1930545	8.16	9.28	10.60	0.98	11.73	13.47	7.77	7.61	0.00	0.00	0.00	7.08	0.00	0.19	0.04	0.04	0.04	0.04	0.55
21	790792	10.06	9.03	16.10	0.46	13.34	12.77	8.05	7.21	0.00	0.00	0.00	6.59	0.00	0.24	0.11	0.22	0.11	0.22	1.35
22	1668348	10.95	9.63	13.39	1.41	12.17	17.38	10.76	6.91	0.00	0.00	0.00	6.64	0.00	0.46	0.14	0.03	0.14	0.03	1.25
X	3010269	11.94	11.13	16.72	1.27	13.59	14.48	11.70	7.03	0.00	0.00	0.00	7.39	100.00	0.42	0.12	0.06	0.12	0.06	2.76
Y	170502	15.38	24.85	96.11	100.00	22.95	20.52	13.93	11.34	0.00	0.00	0.00	88.33	100.00	38.26	3.78	2.81	3.78	2.81	100.00
Total	80622428	10.88	10.37	16.68	2.15	12.66	14.64	10.36	7.37	0.00	0.03	0.03	8.18	3.97	0.54	0.19	0.08	0.19	0.08	1.50

Table3: Pearson's correlation coefficients between rank scores (upper-triangle) and the ratio of binary predictions' agreement between scores (lower-triangle).

Score	SIFT	HDIV	HVAR	LRT	MT	MA	FAT	PROV	MKL	SVM	LR	VEST3	CADD	DANN	i6	gm	h1	hu	GERP	phP7	phP20	phC7	phC20	SiPhy
SIFT	-	0.63	0.63	0.36	0.36	0.59	0.14	0.63	0.36	0.40	0.41	0.53	0.63	0.50	0.02	0.01	0.02	0.02	0.28	0.25	0.22	0.21	0.18	0.31
HDIV	0.75	-	0.97	0.50	0.48	0.62	0.14	0.64	0.48	0.46	0.50	0.66	0.72	0.62	0.07	0.05	0.07	0.06	0.42	0.31	0.28	0.31	0.28	0.46
HVAR	0.74	0.89	-	0.53	0.51	0.64	0.16	0.66	0.51	0.48	0.53	0.68	0.73	0.62	0.07	0.06	0.08	0.06	0.43	0.32	0.30	0.34	0.31	0.48
LRT	0.66	0.71	0.72	-	0.66	0.44	0.18	0.46	0.68	0.37	0.42	0.64	0.55	0.50	0.22	0.19	0.22	0.21	0.55	0.38	0.36	0.59	0.53	0.59
MT	0.66	0.72	0.70	0.80	-	0.43	0.22	0.45	0.70	0.38	0.45	0.65	0.66	0.54	0.24	0.18	0.23	0.21	0.59	0.41	0.41	0.58	0.55	0.62
MA	0.68	0.66	0.72	0.65	0.61	-	0.16	0.69	0.43	0.49	0.53	0.57	0.60	0.50	0.05	0.05	0.05	0.05	0.31	0.28	0.24	0.25	0.20	0.36
FAT	0.46	0.43	0.50	0.48	0.44	0.61	-	0.16	0.22	0.71	0.85	0.22	0.19	0.16	0.02	-0.01	0.02	0.01	0.16	0.12	0.11	0.18	0.16	0.18
PROV	0.73	0.70	0.74	0.67	0.65	0.76	0.56	-	0.46	0.43	0.46	0.64	0.66	0.50	0.07	0.06	0.07	0.06	0.35	0.32	0.28	0.28	0.24	0.39
MKL	0.65	0.71	0.68	0.76	0.86	0.56	0.40	0.61	-	0.39	0.47	0.67	0.62	0.56	0.30	0.26	0.31	0.29	0.76	0.56	0.55	0.63	0.60	0.76
SVM	0.53	0.51	0.60	0.55	0.50	0.71	0.88	0.65	0.44	-	0.87	0.46	0.47	0.37	0.04	0.02	0.04	0.04	0.32	0.24	0.21	0.27	0.23	0.37
LR	0.52	0.50	0.58	0.53	0.48	0.69	0.90	0.63	0.44	0.96	-	0.51	0.52	0.44	0.08	0.04	0.08	0.07	0.39	0.29	0.27	0.34	0.30	0.43
VEST3	-	-	-	-	-	-	-	-	-	-	-	-	0.73	0.57	0.18	0.14	0.18	0.16	0.60	0.45	0.44	0.51	0.47	0.61
CADD	-	-	-	-	-	-	-	-	-	-	-	-	-	0.74	0.19	0.15	0.18	0.16	0.57	0.37	0.38	0.48	0.49	0.58
DANN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.19	0.15	0.18	0.16	0.52	0.34	0.36	0.45	0.46	0.53
i6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.67	0.74	0.68	0.23	0.15	0.17	0.26	0.27	0.22
gm	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.57	0.60	0.20	0.13	0.14	0.21	0.22	0.18
h1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.57	0.23	0.15	0.17	0.25	0.25	0.22
hu	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.22	0.14	0.16	0.24	0.24	0.21
GERP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.61	0.64	0.57	0.55	0.80
phP7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.74	0.51	0.44	0.43
phP20	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.44	0.50	0.43
phC7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.82	0.54
phC20	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.50

HDIV: Polyphen2_HDIV; HVAR: Polyphen2_HVAR; MT: MutationTaster; MA: MutationAssessor; FAT: FATHMM; PROV: PROVEAN; MKL: fathmm-MKL; SVM: MetaSVM; LR: MetaLR; i6: fitCons-i6; gm: fitCons-gm; h1: fitCons-h1; hu: fitCons-hu; GERP: GERP++; phP7: phyloP7way_vertebrate; phP20: phyloP20way_mammalian; phC7: phastCons7way_vertebrate; phC20: phastCons20way_mammalian

Supporting Information

1. Column description for variant files

1 chr: chromosome number

2 pos(1-based): physical position on the chromosome as to hg38 (1-based coordinate).

For mitochondrial SNV, this position refers to the rCRS (GenBank: NC_012920).

3 ref: reference nucleotide allele (as on the + strand)

4 alt: alternative nucleotide allele (as on the + strand)

5 aaref: reference amino acid

"." if the variant is a splicing site SNP (2bp on each end of an intron)

6 aaalt: alternative amino acid

"." if the variant is a splicing site SNP (2bp on each end of an intron)

7 rs_dbSNP142: rs number from dbSNP 142

8 hg19_chr: chromosome as to hg19, "." means missing

9 hg19_pos(1-based): physical position on the chromosome as to hg19 (1-based coordinate).

For mitochondrial SNV, this position refers to a YRI sequence (GenBank: AF347015)

10 hg18_chr: chromosome as to hg18, "." means missing

11 hg18_pos(1-based): physical position on the chromosome as to hg18 (1-based coordinate)

For mitochondrial SNV, this position refers to a YRI sequence (GenBank: AF347015)

12 genename: gene name; if the nsSNV can be assigned to multiple genes, gene names are

separated by ";"

13 cds_strand: coding sequence (CDS) strand (+ or -)

14 refcodon: reference codon

15 codonpos: position on the codon (1, 2 or 3)

16 codon_degeneracy: degenerate type (0, 2 or 3)

17 Ancestral_allele: the ancestral allele.

Ancestral alleles of the mitochondrial genome are from RSRS.

Ancestral alleles of autosomes and X/Y chromosomes are provided by VEP based on Ensembl 71. The following comes from its original README file:

ACTG - high-confidence call, ancestral state supported by the other two sequences

actg - low-confidence call, ancestral state supported by one sequence only

N - failure, the ancestral state is not supported by any other sequence

\- - the extant species contains an insertion at this position

. - no coverage in the alignment

18 AltaiNeandertal: genotype of a deep sequenced Altai Neanderthal

19 Denisova: genotype of a deep sequenced Denisova

20 Ensembl_geneid: Ensembl gene id

21 Ensembl_transcriptid: Ensembl transcript ids (Multiple entries separated by ";")

22 Ensembl_proteinid: Ensembl protein ids

Multiple entries separated by ";", corresponding to Ensembl_transcriptids

23 aapos: amino acid position as to the protein.
"-1" if the variant is a splicing site SNP (2bp on each end of an intron).
Multiple entries separated by ";", corresponding to Ensembl_proteinid

24 SIFT_score: SIFT score (SIFTori). Scores range from 0 to 1. The smaller the score the more likely the SNP has damaging effect.
Multiple scores separated by ";", corresponding to Ensembl_proteinid.

25 SIFT_converted_rankscore: SIFTori scores were first converted to SIFTnew=1-SIFTori, then ranked among all SIFTnew scores in dbNSFP. The rankscore is the ratio of the rank the SIFTnew score over the total number of SIFTnew scores in dbNSFP.
If there are multiple scores, only the most damaging (largest) rankscore is presented.
The rankscores range from 0.00963 to 0.91219.

26 SIFT_pred: If SIFTori is smaller than 0.05 (rankscore>0.395) the corresponding nsSNV is predicted as "D(amaging)"; otherwise it is predicted as "T(olerated)".
Multiple predictions separated by ";"

27 Uniprot_acc_Polyphen2: Uniprot accession number provided by Polyphen2.
Multiple entries separated by ";".

28 Uniprot_id_Polyphen2: Uniprot ID numbers corresponding to Uniprot_acc_Polyphen2.
Multiple entries separated by ";".

29 Uniprot_aapos_Polyphen2: amino acid position as to Uniprot_acc_Polyphen2.
Multiple entries separated by ";".

30 Polyphen2_HDIV_score: Polyphen2 score based on HumDiv, i.e. hdiv_prob.
The score ranges from 0 to 1.
Multiple entries separated by ";", corresponding to Uniprot_acc_Polyphen2.

31 Polyphen2_HDIV_rankscore: Polyphen2 HDIV scores were first ranked among all HDIV scores in dbNSFP. The rankscore is the ratio of the rank the score over the total number of the scores in dbNSFP. If there are multiple scores, only the most damaging (largest) rankscore is presented. The scores range from 0.02634 to 0.89865.

32 Polyphen2_HDIV_pred: Polyphen2 prediction based on HumDiv, "D" ("probably damaging", HDIV score in [0.957,1] or rankscore in [0.52844,0.89865]), "P" ("possibly damaging", HDIV score in [0.453,0.956] or rankscore in [0.34282,0.52689]) and "B" ("benign", HDIV score in [0,0.452] or rankscore in [0.02634,0.34268]). Score cutoff for binary classification is 0.5 for HDIV score or 0.3528 for rankscore, i.e. the prediction is "neutral" if the HDIV score is smaller than 0.5 (rankscore is smaller than 0.3528), and "deleterious" if the HDIV score is larger than 0.5 (rankscore is larger than 0.3528). Multiple entries are separated by ";".

33 Polyphen2_HVAR_score: Polyphen2 score based on HumVar, i.e. hvar_prob.
The score ranges from 0 to 1.
Multiple entries separated by ";", corresponding to Uniprot_acc_Polyphen2.

34 Polyphen2_HVAR_rankscore: Polyphen2 HVAR scores were first ranked among all HVAR scores in dbNSFP. The rankscore is the ratio of the rank the score over the total number of

the scores in dbNSFP. If there are multiple scores, only the most damaging (largest) rankscore is presented. The scores range from 0.01257 to 0.97092.

35 Polyphen2_HVAR_pred: Polyphen2 prediction based on HumVar, "D" ("probably damaging", HVAR score in [0.909,1] or rankscore in [0.62797,0.97092]), "P" ("possibly damaging", HVAR in [0.447,0.908] or rankscore in [0.44195,0.62727]) and "B" ("benign", HVAR score in [0,0.446] or rankscore in [0.01257,0.44151]). Score cutoff for binary classification is 0.5 for HVAR score or 0.45833 for rankscore, i.e. the prediction is "neutral" if the HVAR score is smaller than 0.5 (rankscore is smaller than 0.45833), and "deleterious" if the HVAR score is larger than 0.5 (rankscore is larger than 0.45833). Multiple entries are separated by ";".

36 LRT_score: The original LRT two-sided p-value (LRTori), ranges from 0 to 1.

37 LRT_converted_rankscore: LRTori scores were first converted as LRTnew=1-LRTori*0.5 if Omega<1, or LRTnew=LRTori*0.5 if Omega>=1. Then LRTnew scores were ranked among all LRTnew scores in dbNSFP. The rankscore is the ratio of the rank over the total number of the scores in dbNSFP. The scores range from 0.00162 to 0.84324.

38 LRT_pred: LRT prediction, D(eleterious), N(eutral) or U(nknown), which is not solely determined by the score.

39 LRT_Omega: estimated nonsynonymous-to-synonymous-rate ratio (Omega, reported by LRT)

40 MutationTaster_score: MutationTaster p-value (MTori), ranges from 0 to 1. Multiple scores are separated by ";". Information on corresponding transcript(s) can

be found by querying http://www.mutationtaster.org/ChrPos.html

41 MutationTaster_converted_rankscore: The MTori scores were first converted: if the prediction is "A" or "D" MTnew=MTori; if the prediction is "N" or "P", MTnew=1-MTori. Then MTnew scores were ranked among all MTnew scores in dbNSFP. If there are multiple scores of a SNV, only the largest MTnew was used in ranking. The rankscore is the ratio of the rank of the score over the total number of MTnew scores in dbNSFP. The scores range from 0.08977 to 0.81031.

42 MutationTaster_pred: MutationTaster prediction, "A" ("disease_causing_automatic"), "D" ("disease_causing"), "N" ("polymorphism") or "P" ("polymorphism_automatic"). The score cutoff between "D" and "N" is 0.5 for MTnew and 0.31709 for the rankscore.

43 MutationTaster_model: MutationTaster prediction models.

44 MutationTaster_AAE: MutationTaster predicted amino acid change.

45 Uniprot_id_MutationAssessor: Uniprot ID number provided by MutationAssessor.

46 Uniprot_variant_MutationAssessor: AA variant as to Uniprot_id_MutationAssessor.

47 MutationAssessor_score: MutationAssessor functional impact combined score (MAori). The score ranges from -5.545 to 5.975 in dbNSFP.

48 MutationAssessor_rankscore: MAori scores were ranked among all MAori scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of MAori scores in dbNSFP. The scores range from 0 to 1.

49 MutationAssessor_pred: MutationAssessor's functional impact of a variant :

predicted functional, i.e. high ("H") or medium ("M"), or predicted non-functional, i.e. low ("L") or neutral ("N"). The MAori score cutoffs between "H" and "M", "M" and "L", and "L" and "N", are 3.5, 1.9 and 0.8, respectively. The rankscore cutoffs between "H" and "M", "M" and "L", and "L" and "N", are 0.941, 0.61456 and 0.26284, respectively.

50 FATHMM_score: FATHMM default score (weighted for human inherited-disease mutations with Disease Ontology) (FATHMMori). Scores range from -16.13 to 10.64. The smaller the score the more likely the SNP has damaging effect.

Multiple scores separated by ";", corresponding to Ensembl_proteinid.

51 FATHMM_converted_rankscore: FATHMMori scores were first converted to FATHMMnew=1-(FATHMMori+16.13)/26.77, then ranked among all FATHMMnew scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of FATHMMnew scores in dbNSFP. If there are multiple scores, only the most damaging (largest) rankscore is presented. The scores range from 0 to 1.

52 FATHMM_pred: If a FATHMMori score is <=-1.5 (or rankscore >=0.81332) the corresponding nsSNV is predicted as "D(AMAGING)"; otherwise it is predicted as "T(OLERATED)".

Multiple predictions separated by ";", corresponding to Ensembl_proteinid.

53 PROVEAN_score: PROVEAN score (PROVEANori). Scores range from -14 to 14. The smaller the score the more likely the SNP has damaging effect.

Multiple scores separated by ";", corresponding to Ensembl_proteinid.

54 PROVEAN_converted_rankscore: PROVEANori were first converted to PROVEANnew=1-(PROVEANori+14)/28, then ranked among all PROVEANnew scores in dbNSFP. The rankscore is the ratio of the rank the PROVEANnew score over the total number of PROVEANnew scores in dbNSFP. If there are multiple scores, only the most damaging (largest) rankscore is presented. The scores range from 0 to 1.

55 PROVEAN_pred: If PROVEANori <= -2.5 (rankscore>=0.543) the corresponding nsSNV is predicted as "D(amaging)"; otherwise it is predicted as "N(eutral)". Multiple predictions separated by ";", corresponding to Ensembl_proteinid.

56 Transcript_id_VEST3: Transcript id provided by VEST3.

57 Transcript_var_VEST3: amino acid change as to Transcript_id_VEST3.

58 VEST3_score: VEST 3.0 score. Score ranges from 0 to 1. The larger the score the more likely the mutation may cause functional change. Multiple scores separated by ";", corresponding to Transcript_id_VEST3. Please note this score is free for non-commercial use. For more details please refer to http://wiki.chasmsoftware.org/index.php/SoftwareLicense. Commercial users should contact the Johns Hopkins Technology Transfer office.

59 VEST3_rankscore: VEST3 scores were ranked among all VEST3 scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of VEST3 scores in dbNSFP. In case there are multiple scores for the same variant, the largest score (most damaging) is presented. The scores range from 0 to 1.

Please note VEST score is free for non-commercial use. For more details please refer to http://wiki.chasmsoftware.org/index.php/SoftwareLicense. Commercial users should contact the Johns Hopkins Technology Transfer office.

60 CADD_raw: CADD raw score for functional prediction of a SNP. Please refer to Kircher et al. (2014) Nature Genetics 46(3):310-5 for details. The larger the score the more likely the SNP has damaging effect. Scores range from -7.535037 to 35.788538 in dbNSFP. Please note the following copyright statement for CADD:
"CADD scores (http://cadd.gs.washington.edu/) are Copyright 2013 University of Washington and Hudson-Alpha Institute for Biotechnology (all rights reserved) but are freely available for all academic, non-commercial applications. For commercial licensing information contact Jennifer McCullar (mccullaj@uw.edu)."

61 CADD_raw_rankscore: CADD raw scores were ranked among all CADD raw scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of CADD raw scores in dbNSFP. Please note the following copyright statement for CADD: "CADD scores (http://cadd.gs.washington.edu/) are Copyright 2013 University of Washington and Hudson-Alpha Institute for Biotechnology (all rights reserved) but are freely available for all academic, non-commercial applications. For commercial licensing information contact Jennifer McCullar (mccullaj@uw.edu)."

62 CADD_phred: CADD phred-like score. This is phred-like rank score based on whole genome CADD raw scores. Please refer to Kircher et al. (2014) Nature Genetics 46(3):310-5

for details. The larger the score the more likely the SNP has damaging effect. Please note the following copyright statement for CADD: "CADD scores (http://cadd.gs.washington.edu/) are Copyright 2013 University of Washington and Hudson-Alpha Institute for Biotechnology (all rights reserved) but are freely available for all academic, non-commercial applications. For commercial licensing information contact Jennifer McCullar (mccullaj@uw.edu)."

63 DANN_score: DANN is a functional prediction score retrained based on the training data of CADD using deep neural network. Scores range from 0 to 1. A larger number indicate a higher probability to be damaging. More information of this score can be found in doi: 10.1093/bioinformatics/btu703. For commercial application of DANN, please contact Daniel Quang (dxquang@uci.edu)

64 DANN_rankscore: DANN scores were ranked among all DANN scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of DANN scores in dbNSFP.

65 fathmm-MKL_coding_score: fathmm-MKL p-values. Scores range from 0 to 1. SNVs with scores >0.5 are predicted to be deleterious, and those <0.5 are predicted to be neutral or benign. Scores close to 0 or 1 are with the highest-confidence. Coding scores are trained using 10 groups of features. More details of the score can be found in doi: 10.1093/bioinformatics/btv009.

66 fathmm-MKL_coding_rankscore: fathmm-MKL coding scores were ranked among all fathmm-MKL coding scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number

of fathmm-MKL coding scores in dbNSFP.

67 fathmm-MKL_coding_pred: If a fathmm-MKL_coding_score is >0.5 (or rankscore >0.28317) the corresponding nsSNV is predicted as "D(AMAGING)"; otherwise it is predicted as "N(EUTRAL)".

68 fathmm-MKL_coding_group: the groups of features (labeled A-J) used to obtained the score. More details can be found in doi: 10.1093/bioinformatics/btv009.

69 MetaSVM_score: Our support vector machine (SVM) based ensemble prediction score, which incorporated 10 scores (SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, GERP++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy, PhyloP) and the maximum frequency observed in the 1000 genomes populations. Larger value means the SNV is more likely to be damaging. Scores range from -2 to 3 in dbNSFP.

70 MetaSVM_rankscore: MetaSVM scores were ranked among all MetaSVM scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of MetaSVM scores in dbNSFP. The scores range from 0 to 1.

71 MetaSVM_pred: Prediction of our SVM based ensemble prediction score,"T(olerated)" or "D(amaging)". The score cutoff between "D" and "T" is 0. The rankscore cutoff between "D" and "T" is 0.82268.

72 MetaLR_score: Our logistic regression (LR) based ensemble prediction score, which incorporated 10 scores (SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, GERP++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy, PhyloP) and the maximum frequency observed in the 1000 genomes populations. Larger value means the SNV is more likely to be damaging.

Scores range from 0 to 1.

73 MetaLR_rankscore: MetaLR scores were ranked among all MetaLR scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of MetaLR scores in dbNSFP. The scores range from 0 to 1.

74 MetaLR_pred: Prediction of our MetaLR based ensemble prediction score,"T(olerated)" or "D(amaging)". The score cutoff between "D" and "T" is 0.5. The rankscore cutoff between "D" and "T" is 0.81113.

75 Reliability_index: Number of observed component scores (except the maximum frequency in the 1000 genomes populations) for MetaSVM and MetaLR. Ranges from 1 to 10. As MetaSVM and MetaLR scores are calculated based on imputed data, the less missing component scores, the higher the reliability of the scores and predictions.

76 integrated_fitCons_score: fitCons score predicts the fraction of genomic positions belonging to a specific function class (defined by epigenomic "fingerprint") that are under selective pressure. Scores range from 0 to 1, with a larger score indicating a higher proportion of nucleic sites of the functional class the genomic position belong to are under selective pressure, therefore more likely to be functional important. Integrated (i6) scores are integrated across three cell types (GM12878, H1-hESC and HUVEC). More details can be found in doi:10.1038/ng.3196.

77 integrated_fitCons_rankscore: integrated fitCons scores were ranked among all integrated fitCons scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number

of integrated fitCons coding scores in dbNSFP.

78 integrated_confidence_value: 0 - highly significant scores (approx. p<.003); 1 - significant scores (approx. p<.05); 2 - informative scores (approx. p<.25); 3 - other scores (approx. p>=.25).

79 GM12878_fitCons_score: fitCons score predicts the fraction of genomic positions belonging to a specific function class (defined by epigenomic "fingerprint") that are under selective pressure. Scores range from 0 to 1, with a larger score indicating a higher proportion of nucleic sites of the functional class the genomic position belong to are under selective pressure, therefore more likely to be functional important. GM12878 fitCons scores are based on cell type GM12878. More details can be found in doi:10.1038/ng.3196.

80 GM12878_fitCons_rankscore: GM12878 fitCons scores were ranked among all GM12878 fitCons scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of GM12878 fitCons coding scores in dbNSFP.

81 GM12878_confidence_value: 0 - highly significant scores (approx. p<.003); 1 - significant scores (approx. p<.05); 2 - informative scores (approx. p<.25); 3 - other scores (approx. p>=.25).

82 H1-hESC_fitCons_score: fitCons score predicts the fraction of genomic positions belonging to a specific function class (defined by epigenomic "fingerprint") that are under selective pressure. Scores range from 0 to 1, with a larger score indicating a higher proportion of nucleic sites of the functional class the genomic position belong to are under selective pressure, therefore more likely to be functional important. GM12878 fitCons scores are based on cell type H1-hESC. More details can be found in doi:10.1038/ng.3196.

83 H1-hESC_fitCons_rankscore: H1-hESC fitCons scores were ranked among all H1-hESC fitCons scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of H1-hESC fitCons coding scores in dbNSFP.

84 H1-hESC_confidence_value: 0 - highly significant scores (approx. p<.003); 1 - significant scores (approx. p<.05); 2 - informative scores (approx. p<.25); 3 - other scores (approx. p>=.25).

85 HUVEC_fitCons_score: fitCons score predicts the fraction of genomic positions belonging to a specific function class (defined by epigenomic "fingerprint") that are under selective pressure. Scores range from 0 to 1, with a larger score indicating a higher proportion of nucleic sites of the functional class the genomic position belong to are under selective pressure, therefore more likely to be functional important. GM12878 fitCons scores are based on cell type HUVEC. More details can be found in doi:10.1038/ng.3196.

86 HUVEC_fitCons_rankscore: HUVEC fitCons scores were ranked among all HUVEC fitCons scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of HUVEC fitCons coding scores in dbNSFP.

87 HUVEC_confidence_value: 0 - highly significant scores (approx. p<.003); 1 - significant scores (approx. p<.05); 2 - informative scores (approx. p<.25); 3 - other scores (approx. p>=.25).

88 GERP++_NR: GERP++ neutral rate

89 GERP++_RS: GERP++ RS score, the larger the score, the more conserved the site. Scores range from -12.3 to 6.17.

90 GERP++_RS_rankscore: GERP++ RS scores were ranked among all GERP++ RS scores in dbNSFP.

The rankscore is the ratio of the rank of the score over the total number of GERP++ RS scores in dbNSFP.

91 phyloP7way_vertebrate: phyloP (phylogenetic p-values) conservation score based on the multiple alignments of 7 vertebrate genomes (including human). The larger the score, the more conserved the site. Scores range from -5.172 to 1.062 in dbNSFP.

92 phyloP7way_vertebrate_rankscore: phyloP7way_vertebrate scores were ranked among all phyloP7way_vertebrate scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of phyloP7way_vertebrate scores in dbNSFP.

93 phyloP20way_mammalian: phyloP (phylogenetic p-values) conservation score based on the multiple alignments of 20 mammalian genomes (including human). The larger the score, the more conserved the site. Scores range from -13.282 to 1.199 in dbNSFP.

94 phyloP20way_mammalian_rankscore: phyloP20way_mammalian scores were ranked among all phyloP20way_mammalian scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of phyloP20way_mammalian scores in dbNSFP.

95 phastCons7way_vertebrate: phastCons conservation score based on the multiple alignments of 7 vertebrate genomes (including human). The larger the score, the more conserved the site. Scores range from 0 to 1.

96 phastCons7way_vertebrate_rankscore: phastCons7way_vertebrate scores were ranked among all phastCons7way_vertebrate scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of phastCons7way_vertebrate scores in dbNSFP.

97 phastCons20way_mammalian: phastCons conservation score based on the multiple alignments of 20 mammalian genomes (including human). The larger the score, the more conserved the site. Scores range from 0 to 1.

98 phastCons20way_mammalian_rankscore: phastCons20way_mammalian scores were ranked among all phastCons20way_mammalian scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of phastCons20way_mammalian scores in dbNSFP.

99 SiPhy_29way_pi: The estimated stationary distribution of A, C, G and T at the site, using SiPhy algorithm based on 29 mammals genomes.

100 SiPhy_29way_logOdds: SiPhy score based on 29 mammals genomes. The larger the score, the more conserved the site. Scores range from 0 to 37.9718 in dbNSFP.

101 SiPhy_29way_logOdds_rankscore: SiPhy_29way_logOdds scores were ranked among all SiPhy_29way_logOdds scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of SiPhy_29way_logOdds scores in dbNSFP.

102 1000Gp3_AC: Alternative allele counts in the whole 1000 genomes phase 3 (1000Gp3) data.

103 1000Gp3_AF: Alternative allele frequency in the whole 1000Gp3 data.

104 1000Gp3_AFR_AC: Alternative allele counts in the 1000Gp3 African descendent samples.

105 1000Gp3_AFR_AF: Alternative allele frequency in the 1000Gp3 African descendent samples.

106 1000Gp3_EUR_AC: Alternative allele counts in the 1000Gp3 European descendent samples.

107 1000Gp3_EUR_AF: Alternative allele frequency in the 1000Gp3 European descendent samples.

108 1000Gp3_AMR_AC: Alternative allele counts in the 1000Gp3 American descendent samples.

109 1000Gp3_AMR_AF: Alternative allele frequency in the 1000Gp3 American descendent samples.

110 1000Gp3_EAS_AC: Alternative allele counts in the 1000Gp3 East Asian descendent samples.

111 1000Gp3_EAS_AF: Alternative allele frequency in the 1000Gp3 East Asian descendent samples.

112 1000Gp3_SAS_AC: Alternative allele counts in the 1000Gp3 South Asian descendent samples.

113 1000Gp3_SAS_AF: Alternative allele frequency in the 1000Gp3 South Asian descendent samples.

114 TWINSUK_AC: Alternative allele count in called genotypes in UK10K TWINSUK cohort.

115 TWINSUK_AF: Alternative allele frequency in called genotypes in UK10K TWINSUK cohort.

116 ALSPAC_AC: Alternative allele count in called genotypes in UK10K TWINSUK cohort.

117 ALSPAC_AF: Alternative allele frequency in called genotypes in UK10K TWINSUK cohort.

118 ESP6500_AA_AC: Alternative allele count in the African American samples of the NHLBI GO Exome Sequencing Project (ESP6500 data set).

119 ESP6500_AA_AF: Alternative allele frequency in the African American samples of the NHLBI GO Exome Sequencing Project (ESP6500 data set).

120 ESP6500_EA_AC: Alternative allele count in the European American samples of the NHLBI GO Exome Sequencing Project (ESP6500 data set).

121 ESP6500_EA_AF: Alternative allele frequency in the European American samples of the NHLBI GO Exome Sequencing Project (ESP6500 data set).

122 ExAC_AC: Allele count in total ExAC samples (~60,706 unrelated individuals)

123 ExAC_AF: Allele frequency in total ExAC samples

124 ExAC_Adj_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in total ExAC samples

125 ExAC_Adj_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in total ExAC samples

126 ExAC_AFR_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in African & African American ExAC samples

127 ExAC_AFR_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in African & African American ExAC samples

128 ExAC_AMR_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in American ExAC samples

129 ExAC_AMR_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in American ExAC samples

130 ExAC_EAS_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in East Asian ExAC samples

131 ExAC_EAS_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in East Asian ExAC samples

132 ExAC_FIN_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in Finnish ExAC samples

133 ExAC_FIN_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in Finnish ExAC samples

134 ExAC_NFE_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in Non-Finnish European ExAC samples

135 ExAC_NFE_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in Non-Finnish European ExAC samples

136 ExAC_SAS_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in South Asian ExAC samples

137 ExAC_SAS_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in South Asian ExAC samples

138 clinvar_rs: rs number from the clinvar data set

139 clinvar_clnsig: clinical significance as to the clinvar data set

2 - Benign, 3 - Likely benign, 4 - Likely pathogenic, 5 - Pathogenic, 6 - drug response,

7 - histocompatibility. A negative score means the the score is for the ref allele

140 clinvar_trait: the trait/disease the clinvar_clnsig referring to

141 Interpro_domain: domain or conserved site on which the variant locates. Domain annotations come from Interpro database. The number in the brackets following a specific domain is the count of times Interpro assigns the variant position to that domain, typically coming from different predicting databases. Multiple entries separated by ";".

2. Column description for gene annotation file

1 Gene_name: Gene symbol from HGNC

2 Ensembl_gene: Ensembl gene id (from HGNC)

3 chr: Chromosome number (from HGNC)

4 Gene_old_names: Old gene symbol (from HGNC)

5 Gene_other_names: Other gene names (from HGNC)

6 Uniprot_acc(HGNC/Uniprot): Uniprot acc number (from HGNC and Uniprot)

7 Uniprot_id(HGNC/Uniprot): Uniprot id (from HGNC and Uniprot)

8 Entrez_gene_id: Entrez gene id (from HGNC)

9 CCDS_id: CCDS id (from HGNC)

10 Refseq_id: Refseq gene id (from HGNC)

11 ucsc_id: UCSC gene id (from HGNC)

12 MIM_id: MIM gene id (from HGNC)

13 Gene_full_name: Gene full name (from HGNC)

14 Pathway(Uniprot): Pathway description from Uniprot

15 Pathway(BioCarta)_short: Short name of the Pathway(s) the gene belongs to (from BioCarta)

16 Pathway(BioCarta)_full: Full name(s) of the Pathway(s) the gene belongs to (from BioCarta)

17 Pathway(ConsensusPathDB): Pathway(s) the gene belongs to (from ConsensusPathDB)

18 Pathway(KEGG)_id: ID(s) of the Pathway(s) the gene belongs to (from KEGG)

19 Pathway(KEGG)_full: Full name(s) of the Pathway(s) the gene belongs to (from KEGG)

20 Function_description: Function description of the gene (from Uniprot)

21 Disease_description: Disease(s) the gene caused or associated with (from Uniprot)

22 MIM_phenotype_id: MIM id(s) of the phenotype the gene caused or associated with (from Uniprot)

23 MIM_disease: MIM disease name(s) with MIM id(s) in "[]" (from Uniprot)

24 Trait_association(GWAS): Trait(s) the gene associated with (from GWAS catalog)

25 GO_biological_process: GO terms for biological process

26 GO_cellular_component: GO terms for cellular component

27 GO_molecular_function: GO terms for molecular function

28 Tissue_specificity(Uniprot): Tissue specificity description from Uniprot

29 Expression(egenetics): Tissues/organs the gene expressed in (egenetics data from BioMart)

30 Expression(GNF/Atlas): Tissues/organs the gene expressed in (GNF/Atlas data from BioMart)

31 Interactions(IntAct): The number of other genes this gene interacting with (from IntAct).

Full information (gene name followed by Pubmed id in "[]") can be found in the ".complete" table

32 Interactions(BioGRID): The number of other genes this gene interacting with (from BioGRID)
Full information (gene name followed by Pubmed id in "[]") can be found in the ".complete" table

33 Interactions(ConsensusPathDB): The number of other genes this gene interacting with (from ConsensusPathDB). Full information (gene name followed by Pubmed id in "[]") can be found in the ".complete" table

34 P(HI): Estimated probability of haploinsufficiency of the gene (from doi:10.1371/journal.pgen.1001154)

35 P(rec): Estimated probability that gene is a recessive disease gene (from DOI:10.1126/science.1215040)

36 Known_rec_info: Known recessive status of the gene (from DOI:10.1126/science.1215040)
"lof-tolerant = seen in homozygous state in at least one 1000G individual"
"recessive = known OMIM recessive disease"
(original annotations from DOI:10.1126/science.1215040)

37 RVIS: Residual Variation Intolerance Score, a measure of intolerance of mutational burden, the higher the score the more tolerant to mutational burden the gene is.
from doi:10.1371/journal.pgen.1003709

38 RVIS_percentile: The percentile rank of the gene based on RVIS, the higher the percentile the more tolerant to mutational burden the gene is.

39 Essential_gene: Essential ("E") or Non-essential phenotype-changing ("N") based on Mouse Genome Informatics database. from doi:10.1371/journal.pgen.1003484

40 MGI_mouse_gene: Homolog mouse gene name from MGI

41 MGI_mouse_phenotype: Phenotype description for the homolog mouse gene from MGI

42 ZFIN_zebrafish_gene: Homolog zebrafish gene name from ZFIN

43 ZFIN_zebrafish_structure: Affected structure of the homolog zebrafish gene from ZFIN

44 ZFIN_zebrafish_phenotype_quality: Phenotype description for the homolog zebrafish gene from ZFIN

45 ZFIN_zebrafish_phenotype_tag: Phenotype tag for the homolog zebrafish gene from ZFIN

3. Column description for dbscSNV files

1 chr: chromosome number

2 pos: physical position on the chromosome as to hg19 (1-based coordinate)

3 ref: reference nucleotide allele (as on the + strand)

4 alt: alternative nucleotide allele (as on the + strand)

5 hg38_chr: chromosome number as to hg38

6 hg38_pos: physical position on the chromosome as to hg38 (1-based coordinate)

7 RefSeq?: whether the SNV is a scSNV according to RefSeq

8 Ensembl?: whether the SNV is a scSNV according to Ensembl

9 RefSeq_region: functional region the SNV located according to RefSeq

10 RefSeq_gene: gene name according to RefSeq

11 RefSeq_functional_consequence: functional consequence of the SNV according to RefSeq

12 RefSeq_id_c.change_p.change: SNV in format of c.change and p.change according to RefSeq

13 Ensembl_region: functional region the SNV located according to Ensembl

14 Ensembl_gene: gene id according to Ensembl

15 Ensembl_functional_consequence: functional consequence of the SNV according to Ensembl

16 Ensembl_id_c.change_p.change: SNV in format of c.change and p.change according to Ensembl

17 ada_score: ensemble prediction score based on ada-boost. Ranges 0 to 1. The larger the score the higher probability the scSNV will affect splicing. The suggested cutoff for a binary prediction (affecting splicing vs. not affecting splicing) is 0.6.

18 rf_score: ensemble prediction score based on random forests. Ranges 0 to 1. The larger the score the higher probability the scSNV will affect splicing. The suggested cutoff for a binary prediction (affecting splicing vs. not affecting splicing) is 0.6.