Published in final edited form as:
Hum Mutat. 2013 September ; 34(9): E2393–E2402. doi:10.1002/humu.22376.

dbNSFP v2.0: A Database of Human Non-synonymous SNVs and Their Functional Predictions and Annotations

Xiaoming Liu[*], **Xueqiu Jian**, and **Eric Boerwinkle**
Human Genetics Center, School of Public Health, University of Texas Health Science Center at Houston, Houston, Texas, USA

Abstract

dbNSFP is a database developed for functional prediction and annotation of all potential non-synonymous single-nucleotide variants (nsSNVs) in the human genome. This database significantly facilitates the process of querying predictions and annotations from different databases/web-servers for large amounts of nsSNVs discovered in exome-sequencing studies. Here we report a recent major update of the database to version 2.0. We have rebuilt the SNV collection based on GENCODE 9 and currently the database includes 87,347,043 nsSNVs and 2,270,742 essential splice site SNVs (an 18% increase compared to dbNSFP v1.0). For each nsSNV dbNSFP v2.0 has added two prediction scores (MutationAssessor and FATHMM) and two conservation scores (GERP++ and SiPhy). The original five prediction and conservation scores in v1.0 (SIFT, Polyphen2, LRT, MutationTaster and PhyloP) have been updated. Rich functional annotations for SNVs and genes have also been added into the new version, including allele frequencies observed in the 1000 Genomes Project phase 1 data and the NHLBI Exome Sequencing Project, various gene IDs from different databases, functional descriptions of genes, gene expression and gene interaction information, among others.

Keywords

dbNSFP; non-synonymous mutation; splice site mutation; functional prediction; database

Introduction

Exome sequencing has become a popular and effective strategy to identify variants causing Mendelian diseases or extreme phenotypes. Non-synonymous single nucleotide variants (nsSNVs) are the major candidate variants in such studies. Typically, exome sequencing will discover a large number of nsSNVs, among which many are novel (e.g. not reported in dbSNP). Researchers will rely on various functional predictions and annotations to filter and prioritize those nsSNVs to shorten the list for further (experimental) validation. To facilitate this process, we developed dbNSFP v1.0 (Liu et al. 2011).

[*]Correspondence to Xiaoming Liu, Ph.D., Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, 1200 Herman Pressler Street, E529, Houston, Texas 77030, USA; Phone: 1-713-500-9820; Xiaoming.Liu@uth.tmc.edu.

Based on the Consensus Coding Sequence (CCDS) version 20090327 (Pruitt et al. 2009), dbNSFP v1.0 compiled a collection of 75,931,005 nsSNVs in the human genome, including both previously reported and potentially novel ones. For each nsSNV, four functional prediction scores: SIFT (Kumar et al. 2009), 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), were collected. Since the publication of dbNSFP v1.0, it has attracted much attention and been broadly used by human geneticists and sequencing centers as well. It has been recommended by the Faculty of 1000 and adopted by various software (e.g. Lindenbaum et al. 2011; Li et al. 2012; San Lucas et al. 2012; Chang and Wang 2012; Sifrim et al. 2012; Zhang et al. 2013) and databases (e.g. Li et al. 2011).

To fulfill the increasing demand for better functional annotation for SNVs discovered in exome sequencing studies, we have upgraded the dbNSFP to v2.0. Now the database is separated into two parts, dbNSFP_variant and dbNSFP_gene. As their names indicate, the former focuses on variant annotations (including prediction scores and conservation scores), and the latter focuses on gene annotations. As to variant annotation, the database has expanded its SNV collections not only based on a more up-to-date GENCODE 9 annotation (Harrow et al. 2012), but also included all potential essential splice site SNVs (ssSNVs), which are another type of candidate variants in exome sequencing studies. Its core score collection has added two more functional prediction scores: MutationAssessor (Reva et al. 2011) and FATHMM (Shihab et al. 2013); and two more conservation scores: GERP++ (Davydov et al. 2010) and SiPhy (Garber et al. 2009; Lindblad-Toh et al. 2011). To facilitate filtering common SNVs observed in human populations, allele frequencies from the 1000 Genomes Project phase 1 data (Abecasis et al. 2012) and the NHLBI Exome Sequencing Project data (Fu et al. 2013) were added. As to gene annotation, dbNSFP v2.0 has collected rich functional annotations for all genes in the database, including various IDs for different databases, functional descriptions, and gene expression and gene interaction information, among others. Details of the upgrade and preliminary analyses of the core scores are reported in the following sections.

New and Updated Functional Annotations

As CCDS might be over-conservative on gene annotation and human reference build NCBI36/hg18 has been replaced by GRCh37/hg19, we completely rebuilt the backbone SNVs of dbNSFP based on GENCODE 9. In short, at each coding site and essential splice site (defined as the first two and last two nucleotide sites of an intron) we arbitrarily “mutated” the reference allele to the other three alternative alleles and collected them into the database. The total backbone SNVs now reached 89,617,785 (including 87,347,043 nsSNVs (Table 1) and 2,270,742 ssSNVs), which is an 18% increase compared to dbNSFP v1.0. Among them, 89,572,881 contain hg18 coordinates, obtained via the liftOver tool from the UCSC Genome Browser (Meyer et al. 2012), to facilitate queries based on hg18.

Three functional predictions scores and one conservation score of dbNSFP v1.0 (SIFT, Polyphen2, MutationTaster and PhyloP) have been updated. To avoid confusion, the original scores from the algorithms (not the re-scaled scores as in v1.0) were used, except LRT, for which we believe our monotone re-scaled score is easier to interpret than its original score;

and missing scores were no longer imputed (as in v1.0) and were presented as a single dot (“.”). In the dbNSFP v2.0, we collected Polyphen2 version 2.2.2 scores based on both HDIV and HVAR training sets. As the same nsSNV may have multiple HDIV (or HVAR) predictions and scores according to different amino acid positions of different transcripts of the same Uniprot (The UniProt Consortium 2011) gene, all scores were included (separated by “;”) and their orders corresponded to the Uniprot accession numbers and amino acid positions in the Uniprot_acc column and the Uniprot_aapos column, respectively. As to Polyphen2 version 2.2.2, the score thresholds separating “probably damaging”, “possibly damaging” and “benign” predictions are 0.956 and 0.453 for HDIV, and 0.908 and 0.447 for HVAR, respectively. Throughout this paper, we regard score 0.5 as the threshold for Polyphen2's binary predictions (i.e. “deleterious” versus “tolerated”).

Two new functional predictions scores, MutationAssessor and FATHMM, have been added. Precomputed scores from MutationAssessor release 2 (hg19) were downloaded from http://mutationassessor.org/. We used its functional impact combined score as our MutationAssessor score, which ranges from -5.545 to 5.975; the larger the score the more likely it will be deleterious. MutationAssessor provides four types of predictions (to be functional): high, medium, low and neutral. To form binary predictions, we treat high and medium predictions as “deleterious” and low and neutral predictions as “tolerated”. FATHMM v2.1 database were downloaded from http://fathmm.biocompute.org.uk/ and installed on our local server, and its default scores (weighted for human inherited-disease mutations with Disease Ontology) for all the backbone SNVs were then retrieved. These score ranges from -18.09 to 11.0; the smaller the score the more likely it will be deleterious. Binary prediction is also provided and the threshold separating “deleterious” and “tolerated” is -1.5. In case there are more than one FATHMM scores for the same nsSNV due to isoforms, we took the smallest score (most deleterious) as our FATHMM score.

We have also added two new conservation scores: GERP++ and SiPhy. GERP++ base-wise scores were downloaded from http://mendel.stanford.edu/SidowLab/downloads/gerp/. Although GERP RS scores were typically used to measure the conservation of a nucleotide site in Mendelian disease studies (e.g. Cooper et al. 2010), an alternative measure might be a scaled RS score with the corresponding neutral rate (NR) of the site (i.e. RS/NR ratio). Therefore, both NR and RS scores were included in the database. SiPhy scores based on 29 mammalian genomes were downloaded from http://www.broadinstitute.org/mammals/2x/siphy_hg19/. We used its logOdds scores as our SiPhy scores. We also included the SiPhy estimated stationary distribution of A, C, G and T of the site (29way_pi) in our database to facilitate alternative conservation measures. For both the RS and logOdds scores, the larger the score the more conserved the site.

Additional SNV annotations include: observed alternative allele frequencies in the 1000 Genomes Project phase 1 data; observed alternative allele frequencies in the NHLBI Exome Sequencing Project ESP6500 data set (from ANNOVAR (Wang et al. 2010)); rs numbers from a cleaned version of dbSNP build 129 (from UniSNP, http://research.nhgri.nih.gov/tools/unisnp/); ancestral allele (from the 1000 Genomes Project phase 1 data); reference amino acid (aaref); alternative amino acid (aaalt); coding sequence strand (+ or -); reference codon; SNV position in the codon (1, 2, or 3); codon degenerate type (0, 2 or 3); SLR test

statistic of the codon, which is a measure of natural selection acting on the codon (Massingham and Goldman 2005); the protein domain(s) the SNV resides on, according to the InterPro database (Hunter et al. 2012); and amino acid position(s) (aapos) as to Ensembl transcript(s). If there are multiple amino acid position(s) for the same SNV, the order of the positions corresponds to the multiple Ensembl transcript IDs in the Ensembl_transcriptid column. ssSNVs will have missing (".") in the aaref and aaalt columns and "-1" in the aapos column.

To facilitate gene-based SNV prioritization, we have strengthened the functional annotation of genes and added the following information: gene IDs from multiple databases (HGNC (Gray et al. 2013), Uniprot, Entrez Gene (Maglott et al. 2011), CCDS, Refseq (Pruitt et al. 2012), UCSC, and MIM (Amberger et al. 2011)); pathway information, function descriptions, disease descriptions and MIM phenotype IDs from Uniprot; trait association from the GWAS catalog (Hindorff et al. 2009); eGenetics (Kelso et al. 2003) and GNF/Atlas (Su et al. 2002) gene expression data from BioMart (Guberman et al. 2011); gene interaction data from IntAct (Kerrien et al. 2011) and BioGRID (Chatr-aryamontri et al. 2012); and estimated probability that the gene is haploinsufficient (Huang et al. 2010) and recessive (MacArthur et al. 2012).

A Summary of Functional Prediction Scores and Conservation Scores

As different functional prediction scores and conservation scores are derived from different methods using different information (summarized in Table 2), they have different ranges and distributions. Evenly dividing each score into 100 bins between its minimum and maximum, Figure 1 shows the frequency of each bin collected in dbNSFP v2.0. SIFT, Polyphen2 (both HDIV and HVAR) and MutationTaster have U-shape distributions with majority of the scores close to either 0 or 1. LRT (rescaled) presents an L-shape distribution with majority of its scores close to 0. FATHMM and GERP++ show skewed unimodal distributions, while MutationAssessor, PhyloP and SiPhy show complex multimodal distributions.

Understanding the correlation between the scores is important for SNV prioritization. Typically consensus prediction by multiple scores is considered more reliable than a prediction by a single score. However, some scores are more correlated than others; therefore an agreed prediction of two less correlated scores carries more weight than that of two highly correlated scores. Based on absolute Spearman's rank correlation coefficient (aRCC), correlation strengths between the scores are mostly low to moderate (0.25-0.65) (Table 3). Exceptions include high correlations between the two Polyphen2 scores (0.97) and the three conservation scores (0.78-0.90), and very low correlations between FATHMM and other scores (0.13-0.21) (Table 3). The pairwise agreement of the binary predictions between the scores ranges from 42.05% (between FATHMM and Polyphen2-HDIV) to 76.82% (between MutationTaster and LRT), when excluding the 88.97% agreement between the two Polyphen2 scores (Table 3). Figure 2 shows the UPGMA dendrogram of the scores when using 1-aRCC as a measure of distance between scores. It is worth notice that LRT and MutationTaster are more closely related to the conservation scores than other functional prediction scores. As to LRT, this observation is not surprising because it is based on the comparison of aligned coding sequences of 32 vertebrate species. As to

MutationTaster, on the other hand, this observation suggests that conservation information is heavily weighted in its prediction model. Another interesting observation from Table 3 and Figure 2 is that FATHMM has low correlation with any other score.

The numbers of missing data in the prediction scores and conservation scores per chromosome can be derived from Table 1. In general, conservation scores have low missing data rates (0.07%-1.50%), while prediction scores' missing rates range from 11.17% (SIFT) to 22.07% (LRT). The missing rates are higher than those of dbNSFP v1.0, partially due to the fact that the GENCODE annotation of a gene is less stringent than that of CCDS. Figure 3 shows the percentages of dbNSFP v2.0 entries having 0 to 6 (non-missing) prediction scores, and the percentages of those having 0 to 3 (non-missing) conservation scores.

New Features of the Companion Search Program

The companion search program written in Java (search_dbNSFP) has also been upgraded. The human reference sequence build GRCh37/hg19 is now the default choice. To search SNVs based on hg18, users need to use the “-v hg18” option. The search program now supports vcf format for the input file. If the file name has an extension of “vcf”, the program will automatically query the database by the “chr pos ref alt” format. Alternatively, users can also search SNVs by “chr pos” and “chr pos ref alt refAA altAA” formats, by gene name or various database IDs, and by Uniprot ID/access number and protein position. By default all columns of dbNSFP2.0_variant and most columns of dbNSFP2.0_gene (except the first three columns) will be written to a user specified output file. Users can also specify the columns to output using the new “-w” option. All queries that do not have a match in the database will be written to an “.err” file. More details can be found in the readme file of the companion search program included in the database distribution package.

Suggested Usage

The main purpose of the dbNSFP is to facilitate the SNV filtering/prioritizing step in exome-sequencing based Mendelian disease studies. To filter/prioritize SNVs based on the contents of the database, we have the following suggestions: (1) filter in (i.e. retain) nsSNVs and ssSNVs from a list of SNVs discovered through sequencing by searching the dbNSFP using the companion search program; (2) filter out common nsSNVs and ssSNVs by removing any SNV with a MAF larger than a curtain threshold (e.g. 5%) in any of the 1000 Genomes Project populations and NHLBI Exome Sequencing Project populations; (3) prioritize (i.e. rank) SNVs by the number of prediction scores supporting a “deleterious” prediction (from all Polyphen2 scores, pick one score, e.g. the most deleterious one, to represent the Polyphen2 score); (4) prioritize SNVs by a conservation score (since the three conservation scores in the dbNSFP v2.0 are highly correlated); (5) filter out SNVs in genes that are not expressed in the disease related tissue(s); (6) highlight SNVs in genes that cause related MIM diseases or are associated with related phenotypes based on GWAS; (7) highlight SNVs in genes that interact with disease causing genes; and (8) highlight SNVs in genes that are in disease related pathways. At last, we want to give some friendly warnings to the users of dbNSFP or other functional prediction tools: all methods have their own

limitations; do not blindly trust any single method; using consensus prediction or majority vote might be a good practice in general but not a silver bullet.

Acknowledgments

Contract grant sponsor: This work was supported by the National Institutes of Health grant RC2 HL02419 and RC2 HL103010.

We thank Dr. Bo Peng for providing database storage and valuable comments during the process of the work. We thank Sara Barton for her help on polishing the writing. We thank the three anonymous reviewers for their comments and suggestions.

References

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491:56–65. [PubMed: 23128226]

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nature Methods. 2010; 7:248–249. [PubMed: 20354512]

Amberger J, Bocchini C, Hamosh A. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). Human Mutation. 2011; 32:564–567. [PubMed: 21472891]

Chang X, Wang K. wANNOVAR: annotating genetic variants for personal genomes via the web. Journal of Medical Genetics. 2012; 49:433–436. [PubMed: 22717648]

Chatr-aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, O'Donnell L, Reguly T, Breitkreutz A, et al. The BioGRID interaction database: 2013 update. Nucleic Acids Research. 2012; 41:D816–D823. [PubMed: 23203989]

Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Research. 2009; 19:1553–1561. [PubMed: 19602639]

Cooper GM, Goode DL, Ng SB, Sidow A, Bamshad MJ, Shendure J, Nickerson DA. Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. Nature Methods. 2010; 7:250–251. [PubMed: 20354513]

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++ PLoS Comput Biol. 2010; 6:e1001025. [PubMed: 21152010]

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. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013; 493:216–220. [PubMed: 23201682]

Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics. 2009; 25:i54–i62. [PubMed: 19478016]

Gray KA, Daugherty LC, Gordon SM, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2013. Nucleic Acids Research. 2013; 41:D545–552. [PubMed: 23161694]

Guberman JM, Ai J, Arnaiz O, Baran J, Blake A, Baldock R, Chelala C, Croft D, Cros A, Cutts RJ, Génova A, Di Forbes S, et al. BioMart Central Portal: an open database network for the biological community. Database: The Journal of Biological Databases and Curation 2011: bar041. 2011

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. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Research. 2012; 22:1760–1774. [PubMed: 22955987]

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America. 2009; 106:9362–9367. [PubMed: 19474294]

Huang N, Lee I, Marcotte EM, Hurles ME. Characterising and predicting haploinsufficiency in the human genome. PLoS Genetics. 2010; 6:e1001154. [PubMed: 20976243]

Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, de Castro E, Coggill P, et al. InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Research. 2012; 40:D306–312. [PubMed: 22096229]

Kelso J, Visagie J, Theiler G, Christoffels A, Bardien S, Smedley D, Otgaar D, Greyling G, Jongeneel CV, McCarthy MI, Hide T, Hide W. eVOC: a controlled vocabulary for unifying gene expression data. Genome Research. 2003; 13:1222–1230. [PubMed: 12799354]

Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U, Jandrasits C, Jimenez RC, et al. The IntAct molecular interaction database in 2012. Nucleic Acids Research. 2011; 40:D841–D846. [PubMed: 22121220]

Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protocols. 2009; 4:1073–1081.

Li MJ, Wang P, Liu X, Lim EL, Wang Z, Yeager M, Wong MP, Sham PC, Chanock SJ, Wang J. GWASdb: a database for human genetic variants identified by genome-wide association studies. Nucleic Acids Research. 2011; 40:D1047–D1054. [PubMed: 22139925]

Li MX, Gui HS, Kwan JSH, Bao SY, Sham PC. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucleic Acids Research. 2012; 40:e53–e53. [PubMed: 22241780]

Li MX, Kwan JSH, Bao SY, Yang W, Ho SL, Song YQ, Sham PC. Predicting Mendelian Disease-Causing Non-Synonymous Single Nucleotide Variants in Exome Sequencing Studies. PLoS Genetics. 2013; 9:e1003143. [PubMed: 23341771]

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. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011; 478:476–482. [PubMed: 21993624]

Lindenbaum P, Scouarnec SL, Portero V, Redon R. Knime4Bio: a set of custom nodes for the interpretation of next-generation sequencing data with KNIME. Bioinformatics. 2011; 27:3200–3201. [PubMed: 21984761]

Liu X, Jian X, Boerwinkle E. dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. Human Mutation. 2011; 32:894–899. [PubMed: 21520341]

MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers CA, Zhang ZD, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012; 335:823–828. [PubMed: 22344438]

Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Research. 2011; 39:D52–57. [PubMed: 21115458]

Massingham T, Goldman N. Detecting amino acid sites under positive selection and purifying selection. Genetics. 2005; 169:1753–1762. [PubMed: 15654091]

Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, Raney BJ, Pohl A, et al. The UCSC Genome Browser database: Extensions and updates 2013. Nucleic Acids Research. 2012; 41:D64–D69. [PubMed: 23155063]

Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, Hart E, Suner MM, et al. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Research. 2009; 19:1316–1323. [PubMed: 19498102]

Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): Current status, new features and genome annotation policy. Nucleic Acids Research. 2012; 40:D130–135. [PubMed: 22121212]

Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: Application to cancer genomics. Nucleic Acids Research. 2011; 39:e118. [PubMed: 21727090]

San Lucas FA, Wang G, Scheet P, Peng B. Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools. Bioinformatics. 2012; 28:421–422. [PubMed: 22138362]

Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nature Methods. 2010; 7:575–576. [PubMed: 20676075]

Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, Day INM, Gaunt TR. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Human Mutation. 2013; 34:57–65. [PubMed: 23033316]

Sifrim A, Van Houdt JK, Tranchevent LC, Nowakowska B, Sakai R, Pavlopoulos GA, Devriendt K, Vermeesch JR, Moreau Y, Aerts J. Annotate-it: a Swiss-knife approach to annotation, analysis and interpretation of single nucleotide variation in human disease. Genome Medicine. 2012; 4:73. [PubMed: 23013645]

Siepel, A.; Pollard, KS.; Haussler, D. New methods for detecting lineage-specific selection. Proceedings of the 10th International Conference on Research in Computational Molecular Biology (RECOMB2006); 2006. p. 190-205.

Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, Patapoutian A, Hampton GM, et al. Large-scale analysis of the human and mouse transcriptomes. Proceedings of the National Academy of Sciences of the United States of America. 2002; 99:4465–4470. [PubMed: 11904358]

The UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Research. 2011; 40:D71–D75. [PubMed: 22102590]

Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research. 2010; 38:e164. [PubMed: 20601685]

Zhang L, Zhang J, Yang J, Ying D, lung Lau Y, Yang W. PriVar: a toolkit for prioritizing SNVs and indels from next-generation sequencing data. Bioinformatics. 2013; 29:124–125. [PubMed: 23104884]

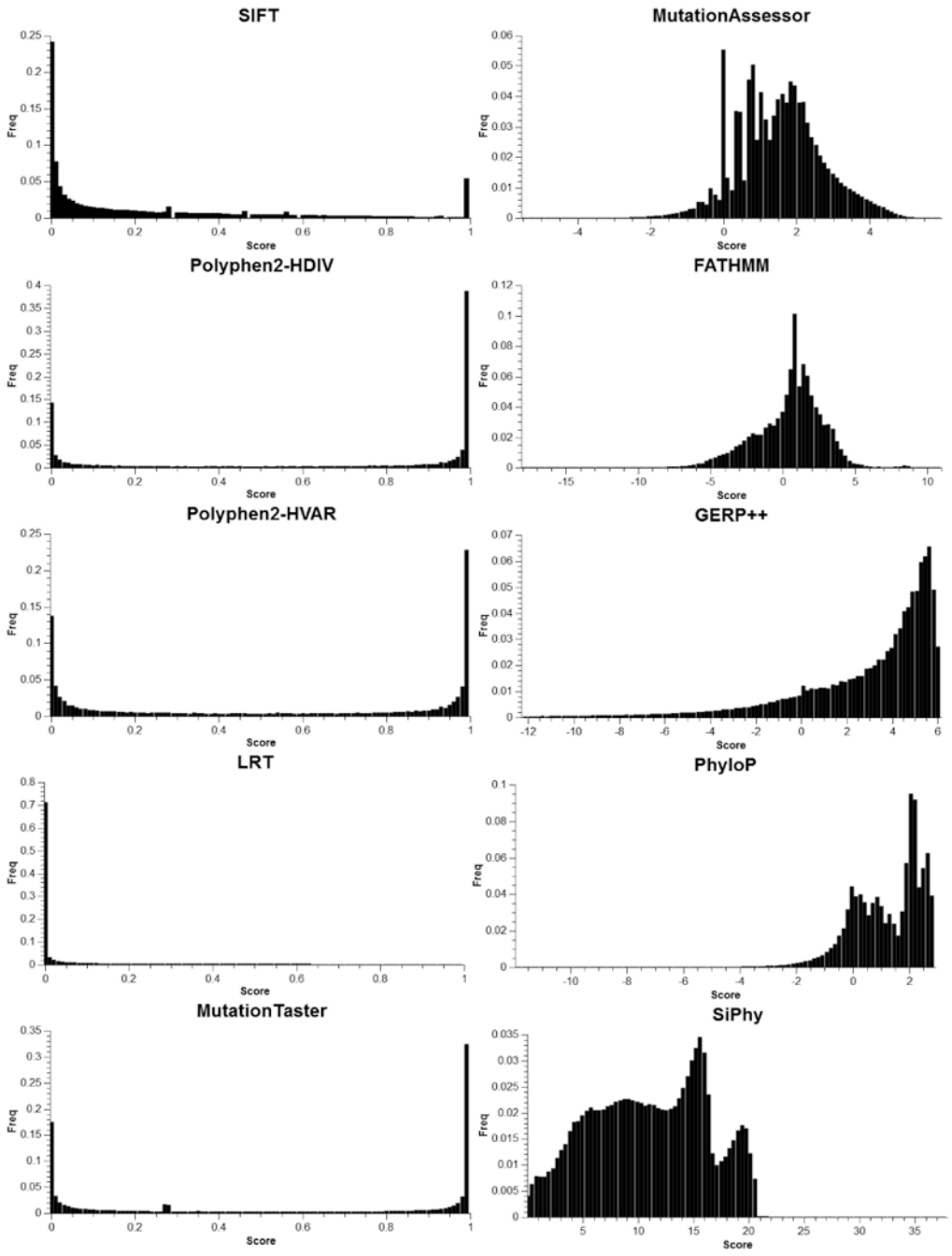

Figure 1.
Distributions of the prediction and conservation scores.

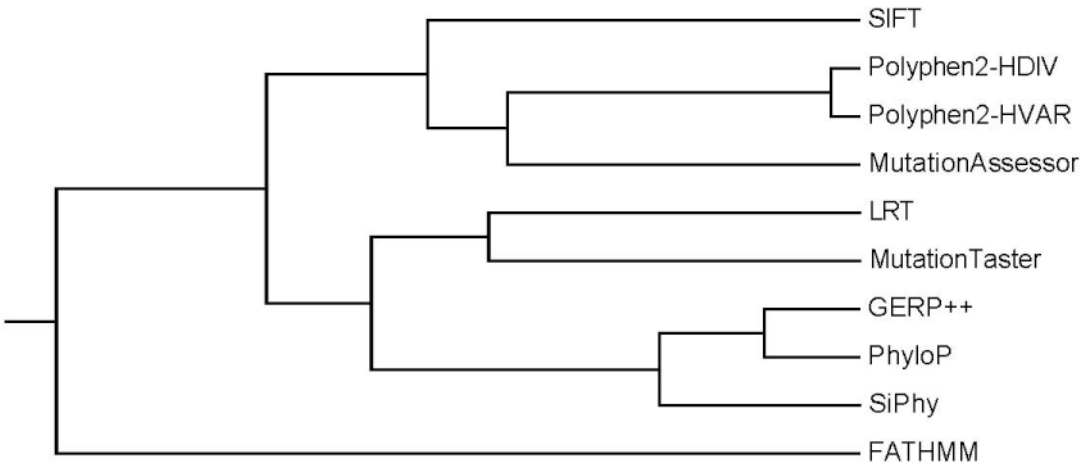

Figure 2.
UPGMA dendrogram of the prediction and conservation scores.

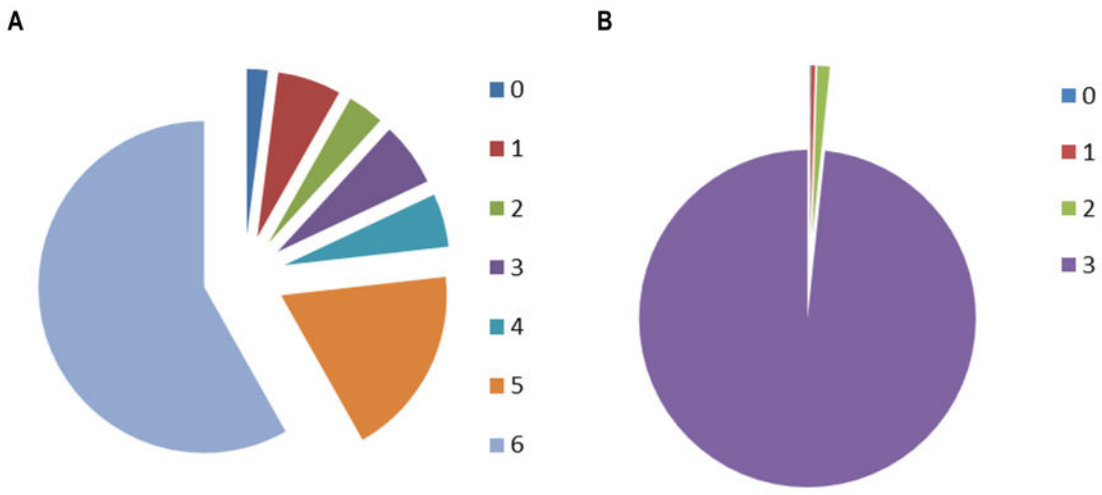

Figure 3.
Percentages of the dbNSFP v2.0 entries having 0 to 6 (non-missing) prediction scores (A), and the percentages of those having 0 to 3 (non-missing) conservation scores (B).

Table 1

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

Chr.	nsSNV	SIFT	Polyphen2	LRT	Mutation Taster	Mutation Assessor	FAT HMM	GERP ++	PhyloP	SiPhy
1	9017117	11.55	18.10	22.23	18.07	15.15	20.33	1.08	0.13	2.15
2	6515708	11.38	18.27	25.76	25.85	14.09	20.06	0.38	0.02	0.87
3	5174843	14.00	17.74	21.11	24.65	13.78	21.02	0.33	0.01	0.78
4	3627319	13.90	21.06	20.97	20.66	14.23	21.04	0.82	0.00	2.15
5	3982618	11.36	15.28	21.27	15.45	12.94	17.59	0.27	0.09	0.82
6	4423891	10.43	15.99	17.71	17.53	14.12	18.25	0.20	0.10	0.74
7	4096944	11.80	18.80	24.98	22.38	16.06	21.27	0.85	0.04	2.06
8	3076675	13.97	20.80	23.26	24.75	17.14	24.24	0.19	0.04	1.32
9	3585198	11.16	17.30	21.21	20.63	14.43	20.25	0.40	0.03	0.84
10	3505818	11.27	17.76	18.71	21.09	16.32	20.99	0.89	0.00	1.31
11	5146903	12.76	16.99	21.94	19.18	14.71	19.39	0.14	0.01	0.84
12	4646355	14.95	17.39	20.24	22.08	14.89	19.23	0.09	0.01	0.73
13	1544883	13.43	14.85	14.12	15.12	12.92	17.55	0.10	0.01	0.91
14	2855480	15.01	20.01	22.62	22.41	15.13	23.83	0.16	0.01	0.76
15	2849761	3.85	12.63	16.29	19.80	13.14	15.04	1.51	0.08	1.85
16	3538182	7.25	14.24	19.31	20.75	16.06	19.74	2.67	0.74	3.83
17	4773221	6.51	13.46	18.53	13.98	13.51	19.07	0.58	0.05	1.32
18	1299988	7.02	12.09	18.27	12.88	11.60	14.95	0.36	0.01	1.13
19	5454137	5.09	12.45	35.74	24.59	13.47	16.97	0.20	0.02	1.53
20	2105920	10.85	16.53	17.13	19.10	13.92	19.47	0.17	0.00	0.73
21	898748	15.43	19.69	25.84	21.65	17.15	22.04	0.34	0.00	1.97
22	1889471	11.49	18.38	22.61	22.78	15.50	22.42	0.88	0.05	2.21
X	3304139	14.18	18.87	22.91	17.61	16.23	20.73	0.47	0.08	3.15
Y	33724	100.00	54.29	75.60	60.67	37.07	27.58	52.70	3.59	100.00
Total	87347043	11.17	16.96	22.07	20.47	14.62	19.81	0.61	0.07	1.50

Table 2

A summary of functional prediction scores and conservation scores.

Score	Training data	Information used	Prediction model
PolyPhen-2	UniProtKB/UniRef100; PDB/DSSP; UCSC alignments of 45 vertebrate genomes	eight sequence-based and three structure-based predictive features	naive Bayes classifier
SIFT	SWISS-PROT/TrEMBL	sequence homology based on PSI-BLAST	position specific scoring matrix
Mutation Taster	UniProt; homologous genes in humans and 10 other species; dbSNP; HapMap	conservation, splice site, mRNA features, protein features;	naive Bayes classifier
LRT	coding sequences of 32 vertebrate species	sequence homology	likelihood ratio test of codon neutrality
Mutation Assessor	homologous sequences from Uniprot identified by BLAST	sequence homology of protein families and sub-families within and between species	combinatorial entropy formalism
FATHMM	homologous sequences from UniRef90, SUPERFAMILY and Pfam	sequence homology	hidden Markov models
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
PhyloP	genomes of 33 placental mammals	multiple alignments and phylogenetic tree	distributions of the number of substitutions based on phylogenetic hidden Markov model

Table 3

Pair-wise prediction agreement percentages (upper-right triangle) and Spearman's rank correlation coefficients (lower-left triangle).

Score	SIFT	Polyphen2-HDIV	Polyphen2-HVAR	LRT	Mutation Taster	Mutation Assessor	FAT HMM	GERP++	PhyloP
SIFT	-	65.62	69.50	60.42	58.13	72.66	57.56	-	-
Polyphen2-HDIV	-0.54	-	88.97	71.22	67.42	66.09	42.05	-	-
Polyphen2-HVAR	-0.54	0.97	-	72.12	69.15	71.97	49.56	-	-
LRT	0.28	-0.48	-0.52	-	76.82	64.93	47.65	-	-
MutationTaster	-0.20	0.47	0.50	-0.61	-	65.69	51.85	-	-
MutationAssessor	-0.56	0.62	0.64	-0.43	0.48	-	60.57	-	-
FATHMM	0.14	-0.13	-0.15	0.16	-0.21	-0.15	-	-	-
GERP++	-0.19	0.42	0.43	-0.53	0.45	0.31	-0.15	-	-
PhyloP	-0.18	0.41	0.42	-0.46	0.41	0.29	-0.14	0.90	-
SiPhy	-0.22	0.46	0.48	-0.57	0.50	0.36	-0.17	0.80	0.78

Hum Mutat. Author manuscript; available in PMC 2014 July 24.