CHARGE SNP Info v7 read me file

---

Table of Contents

---

---

Part 1. Illumina annotation

---

Index: serial number of all variants in the SNP Info file

Columns from Illumina annotation file HumanExome-12v1_A.csv (www.myillumina.com):

IlmnID: Illumina ID
Name: Offical SNP name used to identify the variant. Genotype data references the SNP Name.
IlmnStrand:
IlmnSNP:
AddressA_ID:
AlleleA_ProbeSeq:
AddressB_ID:
AlleleB_ProbeSeq:
GenomeBuild:
IlmnChr:
MapInfo: physical position on the chromosome as to hg19 (1-based coordinate)
Ploidy:
Species:
Source:
SourceVersion:
SourceStrand:
SourceSeq:
TopGenomicSeq:
BeadSetID:
Exp_Clusters:
IlmnRefStrand:

Appended columns:

v1: Site included on HumanExome BeadChip v1.0 array = 1
v1_1: Site included on HumanExome BeadChip v1.1 array = 1
v1_2: Site included on HumanExome BeadChip v1.2 array = 1

Flip_TOPtoFWD: If exome chip data was previously exported using the TOP strand, flip the alleles of variants = 1 to match CHARGE exome chip jointly called data which was exported using Illumina FWD. Strand flipping provided by Martina Mueller-Nurasyid.

RecodeALL_FlipFWDtoPLUS: If exome chip data was previously exported using the Illumina FWD strand, and recoded using the "recode_all.txt" file, the variants =1 would need to be flipped to match data referencing the HG19 PLUS strand. Strand flipping confirmed by VCF check and list provided by Gina Peloso, Josh Bis and Megan Grove.

SNP_list_to_be_flipped_KL_TW: If exome chip data was previously exported using the Illumina FWD strand, then the variants =1 would need to be flipped to match data referencing the HG19 PLUS strand. Strand flipping confirmed by VCF check and list provided by Ruth Loos and Kevin Lu.

Part 2. Variant selection

Column from annotatedList.txt (ftp://share.sph.umich.edu/exomeChip/IlluminaDesigns/):

VarCat: Variant selection category

Part 3. dbSNP rs ID

Columns from exome_annot_dsg.csv (provided by Borecki I)

dbSNPID: dbSNP ID available as of October 1, 2012

Blat_Flag: coded 1-4, see below for description of flags

PAR_Y: pseudoautosomal Y position

Table of BLAT RUN results:

GROUP	Not Run	No Match	One Match	PAR	Blat2	PosTOTAL
NO ISSUES	242,934	0	0	0	0	242,934
RS NAME MISSING	0	0	4,681	86	31	4,798
OTHERWISE FLAGGED	0	86	0	0	52	138
TOTAL	242,934	86	4,681	86	83	247,870

Table of BLAT FLAG results:

Group	Blat_Flag	Count
No ISSUES		242,934
Location Verified	1	4,681
Pseudoautosomal	2	86
Blat 2 Positions	3	83
No Match	4	86
Total		247,870

Comments from Boerecki I:
"Here is a summary of the rs-annotation progress we've made with the Exome chip variants. All but ~4,798 had an rs name in the file provided by Ben Neale. Of those, 86 were pseudo autosomal and 31 mapped to 2 locations (so suspect), leaving 4,681 SNPs. We verified the physical positions of these loci by BLAT using UCSC browser for hg19. All the SNPs that mapped to unique locations were verified as having the location reported by Illumina. 86 additional SNPs did not match with the hg19 map, but had Illumina-provided locations, and 52 others matched to two positions; while we left the Illumina locations, these are flagged as suspicious as they don't uniquely map."

Part 4. CHARGE Exome Chip Minor Allele Frequencies

Excluded the following samples before calculating MAF: all AGES samples, all HapMap controls, known duplicates (based on sample information provided in manifests from individual cohorts), p10GC<0.38, call rate < 0.97, or race was unknown or not provided.
MAFs not reported for 8,994 excluded SNPs.

CHARGE Exome Chip (EC) minor allele freq categories:
Alleles presented are based on the Illumina provided annotation of forward strand (abbreviated as Fwd).
"Fwd_A1" = the minor allele (aka coded allele in PLINK) for each race-specific category
"Fwd_A2" = the common allele (aka non-coded allele in PLINK) for each race-specific category
"ALL" = all CHARGE samples (across cohorts)
"AA" = African Americans (across cohorts)
"EA" = European Americans (across cohorts)
"HIS" = Hispanics (includes MESA participants only)
"ASI" = Asians (includes MESA and CHS participants only)
"CEU" = HapMap CEPH
"YRI" = HapMap Yoruban

**Fwd_A1_ALL should be used for analyses.
Download the "recode_all.txt" file from the wiki for a PLINK-ready text file to force standardized allele coding which is the same information presented here.

Fwd_A1_ALL:
Fwd_A2_ALL:
EC_ALL_MAF:
Fwd_A1_AA:
Fwd_A2_AA:
EC_AA_MAF:
Fwd_A1_EA:
Fwd_A2_EA:
EC_EA_MAF:
Fwd_A1_AA_EA:
Fwd_A2_AA_EA:
EC_AA_EA_MAF:
Fwd_A1_HIS:
Fwd_A2_HIS:

EC_HIS_MAF:
Fwd_A1_ASI:
Fwd_A2_ASI:
EC_ASI_MAF:

HapMap unrelated control samples (total n=96):

Fwd_A1_HapMap_CEU:
Fwd_A2_HapMap_CEU:
EC_HapMap_CEU_MAF:
Fwd_A1_HapMap_YRI:
Fwd_A2_HapMap_YRI:
EC_HapMap_YRI_MAF:

PLINK_file: Variants in main CHARGE PLINK file listed as "0" (n=247,039). Duplicate variants in CHARGE 1000 genomes PLINK file listed as "1" (n=831).

VarType: Variant type identified as follows if unique="Y".

VarType	Freq.
Indel	140
SNV	245,842
SNV;Duplicate;Complement	1,366
SNV;Duplicate;Identical	206
SNV;MT	226
SNV;Triallelic	90
Total	247,870

VarDup: Variants with same chr and position on the chip are identified as 0=unique, 1=first appearence of duplicated variant, and 2=second appearance of a duplicated variant. 831 duplicates identified as follows if unique="Y".

VarDup	Freq.
0	246,208
1	831
2	831
Total	247,870

Part 5. Annotation for analyses

The following functional classifications are based on the "**ANNOVAR_ucsc_precedent_consequence**" column.

sc_exonic: TRUE if variant is categorized as exonic, frameshift, ncRNA_exonic, nonsynonymous, stopgain, stoploss, synonymous, cRNA_splicing, or splicing.

sc_nonsynSplice: TRUE if variant is categorized as frameshift, nonsynonymous, stopgain, stoploss, or splicing.

sc_damaging: TRUE if variant is lof OR predicted damaging by at least 2 of the following methods: Polyphen, LRT, SIFT, Mutation Taster (including Polyphen 'P' [possibly damaging] or either Mutation Taster damaging category [A or D]).

sc_lof: TRUE if variant is categorized as splicing, stopgain, stoploss, or frameshift.

NS_strict: Based on the Purcell et al (PMID: 24463508) criteria. TRUE if a variant is stopgain, stoploss, frameshift, or predicted damaging by all 5 of the following algorithms: SIFT, mutationTaster category [A or D], LRT, PolyPhen_HDIV, and PolyPhen_HVAR.
NS_broad: Based on the Purcell et al (PMID: 24463508) criteria. TRUE if variant is stopgain, stoploss, frameshift, or predicted damaging by at least 1 of the following algorithms: SIFT, mutationTaster category [A or D], LRT, PolyPhen_HDIV, and PolyPhen_HVAR.
dmg_sift: TRUE if variant SIFT_score is not missing and less than 0.05.
sc_indel: TRUE if VarType is categorized as Indel.
sc_indel_coding: TRUE if variant is categorized as frameshift, nonframeshift, or splicing and VarType is Indel (sc_indel is TRUE).
sc_functional: TRUE if sc_indel_coding is TRUE or sc_nonsynSplice is TRUE.
SKATgene: If sc_exonic is TRUE or sc_indel_coding is TRUE, the value is ANNOVAR_ucsc_precedent_gene, otherwise the value is **Name**.

Part 6. Annotation

Variant annotation was completed using WGSA v055 with dbNSFP v2.9. References and sources are provided at the end of this section.

Note: Each SNP/indel may have multiple rows if annotated to multiple genes. Each SNP/indel only has one row with a 'Y' in the "unique_variant" column, which is determined by the most damaging functional annotation.

chr: chromosome number
pos: position (hg19)
ref: reference allele
alt: alternative allele
ANNOVAR_ensembl_summary: ANNOVAR consequence summary with Ensembl as gene model. Format: GeneID(total number of transcripts):consequence#1(number of transcripts affected):consequence#2(number of transcripts affected)… Multiple genes are separated by "|".
SnpEff_ensembl_summary: SnpEff consequence summary with Ensembl as gene model. Format: GeneID(total number of transcripts):consequence#1(number of transcripts affected):consequence#2(number of transcripts affected)… Multiple genes are separated by "|".
VEP_ensembl_summary: VEP consequence summary with Ensembl as gene model. Format: GeneID(total number of transcripts):consequence#1(number of transcripts affected):consequence#2(number of transcripts affected)... Multiple genes are separated by "|".
ANNOVAR_refseq_summary: SnpEff consequence summary with Refseq as gene model. Format: GeneID(total number of transcripts):consequence#1(number of transcripts affected): consequence#2(number of transcripts affected)... Multiple genes are separated by "|".
SnpEff_refseq_summary: SnpEff consequence summary with Refseq as gene model.

Format: GeneID(total number of transcripts):consequence#1(number of transcripts affected):consequence#2(number of transcripts affected)... Multiple genes are separated by "|".

VEP_refseq_summary: SnpEff consequence summary with Refseq as gene model.
Format: GeneID(total number of transcripts):consequence#1(number of transcripts affected):consequence#2(number of transcripts affected)... Multiple genes are separated by "|".

ANNOVAR_ucsc_summary: ANNOVAR consequence summary with UCSC knowgene as gene model.
Format: GeneID:consequence#1(number of transcripts affected):consequence#2(number of transcripts affected)... Multiple genes are separated by "|".

SnpEff_ensembl_LOF: SnpEff Loss-Of-Function summary with Ensembl as gene model.
Format: GeneID(total number of transcripts):consequence#1(percentage of transcripts affected*total number of coding transcripts):consequence#2(percentage of transcripts affected*total number of coding transcripts)...

SnpEff_refseq_LOF: SnpEff Loss-Of-Function summary with Refseq as gene model.
Format: GeneID(total number of transcripts):consequence#1(percentage of transcripts affected*total number of coding transcripts):consequence#2(percentage of transcripts affected*total number of coding transcripts)...

rs_dbSNP144: rs number from dbSNP144

sno_miRNA_name: the name of snoRNA or miRNA if the site is located within (from miRBase/snoRNABase)

sno_miRNA_type: the type of snoRNA or miRNA (from miRBase/snoRNABase)

UTR3_miRNA_target: the gene-miRNA pair, if the site is located within a predicted (conserved) target of conserved miRNA families (from TargetScan)

TargetScan_context++_score_percentile: context++ score is a mearsue of favarableness of the site for the miRNA family. The higer the percentile, the more favarable (from TargetScan)

splicing_consensus_ada_score: splicing-change prediction for splicing consensus SNPs based on adaboost. If the score >0.6, it predicts that the splicing will be changed, otherwise it predicts the splicing will not be changed.

splicing_consensus_rf_score: splicing-change prediction for splicing consensus SNPs based on random forest. If the score >0.6, it predicts that the splicing will be changed, otherwise it predicts the splicing will not be changed.

SPIDEX_dpsi_max_tissue: "This is the predicted change in percent-inclusion due to the variant, reported as the maximum across tissues (in percent)"

SPIDEX_dpsi_zscore: "This is the z-score of dpsi_max_tissue relative to the distribution of dPSI that are due to common SNP."

SPIDEX_gene: "The gene which is affected by the variant."

SPIDEX_transcript: "The RefSeq transcript affected."

SPIDEX_exon_number: "The exon for which percent inclusion is predicted."

SPIDEX_location: "Whether the variant is intronic or exonic."

SPIDEX_cds_type: "CDS type"

SPIDEX_ss_dist: "The distance of the variant to the splice site."

GWAS_catalog_rs: rs number according to GWAS catalog

GWAS_catalog_trait: associated trait according to GWAS catalog

GWAS_catalog_pubmedid: pubmedid of the paper describing the association

GRASP_rs: rs number by GRASP

GRASP_PMID: PMID number by GRASP

GRASP_p-value: p-value of the association test based on the SNP
GRASP_phenotype: phenotype the SNP associated with
GRASP_ancestry: population ancestry of the samples on which the association test was based
GRASP_platform: SNP platform on which the association test was based
clinvar_rs: rs number by clinvar
clinvar_clnsig: clinical significance by clinvar
2 - Benign, 3 - Likely benign, 4 - Likely pathogenic, 5 - Pathogenic, 6 - drug response, 7 - histocompatibility. A negative score means the score is for the ref allele
clinvar_trait: the trait/disease the clinvar_clnsig referring to
clinvar_golden_stars: ClinVar Review Status summary
0 - no assertion criteria provided, 1 - criteria provided, single submitter, 2 - criteria provided, multiple submitters, no conflicts, 3 - reviewed by expert panel, 4 - practice guideline
HGMD_ACC_NUM: HGMD acc number
HGMD_HGVS_cdna: Mutation in HGVS cDNA format
HGMD_HGVS_protein: Mutation in HGVS protein format
HGMD_disease: Disease caused by the mutation
HGMD_pmid: PubMed id reporting the mutation
HGMD_Variant_class: HGMD class tag:
DM - disease causing, DM? - disease causing with a degree of doubt, DP - disease associated polymorphism, DFP - disease associated polymorphism with additional supporting functional evidence, FP - functional polymorphism, R - retired record
COSMIC_ID: ID of the SNV at the COSMIC (Catalogue Of Somatic Mutations In Cancer) database
COSMIC_CNT: number of samples having this SNV in the COSMIC database
MAP20: average Duke mappability score based on 20bp read, 0-1, higher score means higher mappability
MAP35: average Duke mappability score based on 35bp read, 0-1, higher score means higher mappability
1000G_strict_masked: whether the site is within the 1000G strict masked region
Y (Yes) or N (No), Y means generally good mapping quality
RepeatMasker_masked: whether the site is masked by RepeatMasker
Y (Yes) or N (No), Y means generally lower mapping quality
phyloP46way_primate: a conservation score based on 46way alignment primate set, the higher the more conservative
phyloP46way_primate_rankscore: the rank of the phyloP46way_primate score among all phyloP46way_primate scores in genome
phyloP46way_placental: a conservation score based on 46way alignment placental set, the higher the more conservative
phyloP46way_placental_rankscore: the rank of the phyloP46way_placental score among all phyloP46way_placental scores in genome
phyloP100way_vertebrate: a conservation score based on 100way alignment vertebrate set, the higher the more conservative
phyloP100way_vertebrate_rankscore: the rank of the phyloP100way_vertebrate score among all phyloP100way_vertebrate scores in genome
phastCons46way_primate: a conservation score based on 46way alignment primate set, the higher the more conservative
phastCons46way_primate_rankscore: the rank of the phastCons46way_primate score among all phastCons46way_primate scores in genome

phastCons46way_placental: a conservation score based on 46way alignment placental set, the higher the more conservative

phastCons46way_placental_rankscore: the rank of the phastCons46way_placental score among all phastCons46way_placental scores in genome

phastCons100way_vertebrate: a conservation score based on 100way alignment vertebrate set, the higher the more conservative

phastCons100way_vertebrate_rankscore: the rank of the phastCons100way_vertebrate score among all phastCons100way_vertebrate scores in genome

GERP++_NR: GERP++ neutral rate

GERP++_RS: GERP++ RS score, the larger the score, the more conserved the site

GERP++_RS_rankscore: the rank of the GERP++_RS score among all GERP++_RS scores in genome

SiPhy_29way_logOdds: SiPhy score based on 29 mammals genomes. The larger the score, the more conserved the site

SiPhy_29way_logOdds_rankscore: the rank of the SiPhy_29way_logOdds score among all SiPhy_29way_logOdds scores in genome

integrated_fitCons_score: fitCons scores (i6) based on function evidence from multiple cell types, the higher the score the more potential for interesting genomic function

integrated_fitCons_rankscore: rank of the integrated_fitCons_score among all integrated_fitCons_scores in genome

integrated_confidence_value: confidence value for the integrated_fitCons_score:
0 - High confidence values (p<~.003), 1 - Likely Significant (p<.05),
2 - Likely Informative (p<.25), 3 - Best estimate (p>=.25)

GM12878_fitCons_score: fitCons scores (gm) based on function evidence from the GM12878 cell type, the higher the score the more potential for interesting genomic function

GM12878_fitCons_rankscore: rank of the GM12878_fitCons_score among all GM12878_fitCons_scores in genome

GM12878_confidence_value: confidence value for the GM12878_fitCons_score:
0 - High confidence values (p<~.003), 1 - Likely Significant (p<.05),
2 - Likely Informative (p<.25), 3 - Best estimate (p>=.25)

H1-hESC_fitCons_score: fitCons scores (h1) based on function evidence from the H1-hESC cell type, the higher the score the more potential for interesting genomic function

H1-hESC_fitCons_rankscore: rank of the H1-hESC_fitCons_score among all H1-hESC_fitCons_scores in genome

H1-hESC_confidence_value: confidence value for the H1-hESC_fitCons_score:
0 - High confidence values (p<~.003), 1 - Likely Significant (p<.05),
2 - Likely Informative (p<.25), 3 - Best estimate (p>=.25)

HUVEC_fitCons_score: fitCons scores (hu) based on function evidence from the HUVEC cell type, the higher the score the more potential for interesting genomic function

HUVEC_fitCons_rankscore: rank of the HUVEC_fitCons_score among all HUVEC_fitCons_scores in genome

HUVEC_confidence_value: confidence value for the HUVEC_fitCons_score:
0 - High confidence values (p<~.003), 1 - Likely Significant (p<.05),
2 - Likely Informative (p<.25), 3 - Best estimate (p>=.25)

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

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

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

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

1000Gp3_EUR_AC: Alternative allele counts in the 1000Gp3 European descendent samples.
1000Gp3_EUR_AF: Alternative allele frequency in the 1000Gp3 European descendent samples.
1000Gp3_AMR_AC: Alternative allele counts in the 1000Gp3 American descendent samples.
1000Gp3_AMR_AF: Alternative allele frequency in the 1000Gp3 American descendent samples.
1000Gp3_EAS_AC: Alternative allele counts in the 1000Gp3 East Asian descendent samples.
1000Gp3_EAS_AF: Alternative allele frequency in the 1000Gp3 East Asian descendent samples.
1000Gp3_SAS_AC: Alternative allele counts in the 1000Gp3 South Asian descendent samples.
1000Gp3_SAS_AF: Alternative allele frequency in the 1000Gp3 South Asian descendent samples.
TWINSUK_AC: Alternative allele count in called genotypes in UK10K TWINSUK cohort.
TWINSUK_AF: Alternative allele frequency in called genotypes in UK10K TWINSUK cohort.
ALSPAC_AC: Alternative allele count in called genotypes in UK10K TWINSUK cohort.
ALSPAC_AF: Alternative allele frequency in called genotypes in UK10K TWINSUK cohort.
ESP6500_AA_AC: Alternative allele counts in the African American samples of the NHLBI GO Exome Sequencing Project (ESP6500 data set).
ESP6500_AA_AF: Alternative allele frequency in the African American samples of the NHLBI GO Exome Sequencing Project (ESP6500 data set).
ESP6500_EA_AC: Alternative allele counts in the European American samples of the NHLBI GO Exome Sequencing Project (ESP6500 data set).
ESP6500_EA_AF: Alternative allele frequency in the European American samples of the NHLBI GO Exome Sequencing Project (ESP6500 data set).
ExAC_AC: Allele count in total ExAC samples (~60,706 unrelated individuals)
ExAC_AF: Allele frequency in total ExAC samples
ExAC_Adj_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in total ExAC samples
ExAC_Adj_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in total ExAC samples
ExAC_AFR_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in African & African American ExAC samples
ExAC_AFR_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in African & African American ExAC samples
ExAC_AMR_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in American ExAC samples
ExAC_AMR_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in American ExAC samples
ExAC_EAS_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in East Asian ExAC samples
ExAC_EAS_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in East Asian ExAC samples
ExAC_FIN_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in Finnish ExAC samples
ExAC_FIN_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in Finnish ExAC samples
ExAC_NFE_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in Non-Finnish European ExAC samples
ExAC_NFE_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in Non-Finnish European ExAC samples
ExAC_SAS_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in South Asian ExAC samples
ExAC_SAS_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in South Asian ExAC samples
RegulomeDB_motif: motif the SNP resides (from RegulomeDB)
RegulomeDB_score: categorical score from RegulomeDB. The smaller, the more likely the SNP affects binding
Motif_breaking: whether break a known motif (in-house script)
network_hub: whether the target gene is a network hub based on funseq-0.1
ENCODE_annotated: whether annotated by ENCODE based on funseq-0.1
sensitive: whether defined as sensitive region based on funseq-0.1

ultra_sensitive: whether defined as ultra-sensitive region based funseq-0.1
target_gene: target gene (for promoter, enhancer, etc.) based on funseq-0.1
funseq_noncoding_score: funseq-like noncoding score range 0-6, each of the previous 5 columns contribute 1 if "YES", or 0 if "NO"; the column Motif_breaking contribute 1 if it is not a "."
funseq2_noncoding_score: funseq2 noncoding score range 0-5.4 a weighted score designed for damaging prediction of cancer somatic SNPs
funseq2_noncoding_rankscore: the rank of the funseq2_noncoding_score among all funseq2_noncoding_scores in genome
CADD_raw: CADD raw score, the larger the number the more likely damaging
CADD_phred: CADD phred-like score, ranges 1-99, the larger the number the more likely damaging; score >10 means the variant in the top 10% (0.1) among the total 8.6 billion possible SNVs, >20 means in the top 1%, >30 means in the top 0.1%, etc. CADD suggests a cutoff between 10 and 20 (e.g. 15)
CADD_raw_rankscore: the rank of the CADD_raw score among all CADD_raw scores in genome
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)
DANN_rank_score: rank of the DANN_score among all DANN_scores
fathmm-MKL_non-coding_score: fathmm-MKL non-coding prediction probability, the larger the number the more likely damaging; the threshold separating deleterious prediction and neutral prediction is 0.5.
fathmm-MKL_non-coding_rankscore: the rank of the fathmm-MKL_non-coding_score among all fathmm-MKL_non-coding_scores in genome
fathmm-MKL_non-coding_pred: If a fathmm-MKL_non-coding_score is >0.5 the corresponding nsSNV is predicted as "D(AMAGING)"; otherwise it is predicted as "N(EUTRAL)".
fathmm-MKL_non-coding_group: fathmm-MKL non-coding group, the feature group used for the non-coding prediction fathmm-MKL_coding_score; fathmm-MKL coding prediction probability, the larger the number the more likely damaging the threshold separating deleterious prediction and neutral prediction is 0.5.
fathmm-MKL_coding_rankscore: the rank of the fathmm-MKL_coding_score among all fathmm-MKL_coding_scores in genome
fathmm-MKL_coding_pred: If a fathmm-MKL_coding_score is >0.5 the corresponding nsSNV is predicted as "D(AMAGING)"; otherwise it is predicted as "N(EUTRAL)".
fathmm-MKL_coding_group: fathmm-MKL coding group, the feature group used for the coding prediction.
ORegAnno_type: the type of regulatory region by ORegAnno
ORegAnno_PMID: the PMID of the paper describing the regulation
ENCODE_TFBS: name of the transcription factors (separated by ;) if the site is within a TFBS
ENCODE_TFBS_score: the higher the score the stronger the evidence of the TFBS
ENCODE_TFBS_cells: the cell lines (separated by ;) the TFBS was detected
ENCODE_Dnase_score: the higher the score the stronger the evidence of a DNase I hypersensitive site
ENCODE_Dnase_cells: number of cell lines supporting a DNase I hypersensitive site
Ensembl_Regulatory_Build_Overviews: genome segment prediction based on 17 cell types from ENCODE and Roadmap. Predicted states: ctcf - CTCF binding sites, distal - Predicted

enhancers open - Unannotated open chromatin regions, proximal - Predicted promoter flanking regions, tfbs - Unannotated transcription factor binding sites, tss - Predicted promoters

FAMTOM5_enhancer: whether the site is within a FAMTOM5 predicted enhancer region; Y (Yes) or N (No)

FAMTOM5_CAGE_peak: whether the site is within a FAMTOM5 Cap Analysis of Gene Expression (CAGE) peak. Y (Yes) or N (No). A CAGE peak generally suggests a promoter region

EnhancerFinder_general_developmental_enhancer: whether the site is within a predicted general developmental enhancer with 5% False Positive Rate; Y (Yes) or N (No)

EnhancerFinder_brain_enhancer: whether the site is within a predicted brain enhancer with 5% False Positive Rate; Y (Yes) or N (No)

EnhancerFinder_heart_enhancer: whether the site is within a predicted heart enhancer with 5% False Positive Rate; Y (Yes) or N (No)

EnhancerFinder_limb_enhancer: whether the site is within a predicted limb enhancer with 5% False Positive Rate; Y (Yes) or N (No)

Ensembl_Regulatory_Build_TFBS: TFBS from Ensembl Regulatory Build. Multiple TFBS separated by ";"

Ensembl_Regulatory_Build_TFBS_prob: the probability of observing TFBS binding. Multiple probabilities (corresponding to Ensembl_Regulatory_Build_TFBS) separated by ";"

The following columns are cell type specific:

Ensembl_A549_activity: A549 specific activity prediction from Ensembl Regulatory Build.
Predicted states: open - Unannotated active open chromatin regions
proximal - Predicted active promoter flanking regions/proximal enhancer
tss - Predicted active promoters
ctcf - Active CTCF binding sites
distal - Predicted active enhancers
tfbs - Unannotated active transcription factor binding sites

Ensembl_DND41_activity: DND41 specific activity prediction from Ensembl Regulatory Build.
Predicted states: open - Unannotated active open chromatin regions
proximal - Predicted active promoter flanking regions/proximal enhancer
tss - Predicted active promoters
ctcf - Active CTCF binding sites
distal - Predicted active enhancers
tfbs - Unannotated active transcription factor binding sites

Ensembl_GM12878_activity: GM12878 specific activity prediction from Ensembl Regulatory Build.
Predicted states: open - Unannotated active open chromatin regions
proximal - Predicted active promoter flanking regions/proximal enhancer
tss - Predicted active promoters
ctcf - Active CTCF binding sites
distal - Predicted active enhancers
tfbs - Unannotated active transcription factor binding sites

Ensembl_H1HESC_activity: H1HESC specific activity prediction from Ensembl Regulatory Build.
Predicted states: open - Unannotated active open chromatin regions
proximal - Predicted active promoter flanking regions/proximal enhancer
tss - Predicted active promoters
ctcf - Active CTCF binding sites

distal - Predicted active enhancers
tfbs - Unannotated active transcription factor binding sites

Ensembl_HELAS3_activity: HELAS3 specific activity prediction from Ensembl Regulatory Build.
Predicted states: open - Unannotated active open chromatin regions
proximal - Predicted active promoter flanking regions/proximal enhancer
tss - Predicted active promoters
ctcf - Active CTCF binding sites
distal - Predicted active enhancers
tfbs - Unannotated active transcription factor binding sites

Ensembl_HEPG2_activity: HEPG2 specific activity prediction from Ensembl Regulatory Build.
Predicted states: open - Unannotated active open chromatin regions
proximal - Predicted active promoter flanking regions/proximal enhancer
tss - Predicted active promoters
ctcf - Active CTCF binding sites
distal - Predicted active enhancers
tfbs - Unannotated active transcription factor binding sites

Ensembl_HMEC_activity: HMEC specific activity prediction from Ensembl Regulatory Build.
Predicted states: open - Unannotated active open chromatin regions
proximal - Predicted active promoter flanking regions/proximal enhancer
tss - Predicted active promoters
ctcf - Active CTCF binding sites
distal - Predicted active enhancers
tfbs - Unannotated active transcription factor binding sites

Ensembl_HSMM_activity: HSMM specific activity prediction from Ensembl Regulatory Build.
Predicted states: open - Unannotated active open chromatin regions
proximal - Predicted active promoter flanking regions/proximal enhancer
tss - Predicted active promoters
ctcf - Active CTCF binding sites
distal - Predicted active enhancers
tfbs - Unannotated active transcription factor binding sites

Ensembl_HSMMT_activity: HSMMT specific activity prediction from Ensembl Regulatory Build.
Predicted states: open - Unannotated active open chromatin regions
proximal - Predicted active promoter flanking regions/proximal enhancer
tss - Predicted active promoters
ctcf - Active CTCF binding sites
distal - Predicted active enhancers
tfbs - Unannotated active transcription factor binding sites

Ensembl_HUVEC_activity: HUVEC specific activity prediction from Ensembl Regulatory Build.
Predicted states: open - Unannotated active open chromatin regions
proximal - Predicted active promoter flanking regions/proximal enhancer
tss - Predicted active promoters
ctcf - Active CTCF binding sites
distal - Predicted active enhancers
tfbs - Unannotated active transcription factor binding sites

Ensembl_K562_activity: K562 specific activity prediction from Ensembl Regulatory Build.
Predicted states: open - Unannotated active open chromatin regions
proximal - Predicted active promoter flanking regions/proximal enhancer
tss - Predicted active promoters

ctcf - Active CTCF binding sites
distal - Predicted active enhancers
tfbs - Unannotated active transcription factor binding sites

Ensembl_MONO_activity: MONO specific activity prediction from Ensembl Regulatory Build.
Predicted states: open - Unannotated active open chromatin regions
proximal - Predicted active promoter flanking regions/proximal enhancer
tss - Predicted active promoters
ctcf - Active CTCF binding sites
distal - Predicted active enhancers
tfbs - Unannotated active transcription factor binding sites

Ensembl_NHA_activity: NHA specific activity prediction from Ensembl Regulatory Build.
Predicted states: open - Unannotated active open chromatin regions
proximal - Predicted active promoter flanking regions/proximal enhancer
tss - Predicted active promoters
ctcf - Active CTCF binding sites
distal - Predicted active enhancers
tfbs - Unannotated active transcription factor binding sites

Ensembl_NHDFAD_activity: NHDFAD specific activity prediction from Ensembl Regulatory Build.
Predicted states: open - Unannotated active open chromatin regions
proximal - Predicted active promoter flanking regions/proximal enhancer
tss - Predicted active promoters
ctcf - Active CTCF binding sites
distal - Predicted active enhancers
tfbs - Unannotated active transcription factor binding sites

Ensembl_NHEK_activity: NHEK specific activity prediction from Ensembl Regulatory Build.
Predicted states: open - Unannotated active open chromatin regions
proximal - Predicted active promoter flanking regions/proximal enhancer
tss - Predicted active promoters
ctcf - Active CTCF binding sites
distal - Predicted active enhancers
tfbs - Unannotated active transcription factor binding sites

Ensembl_NHLF_activity: NHLF specific activity prediction from Ensembl Regulatory Build.
Predicted states: open - Unannotated active open chromatin regions
proximal - Predicted active promoter flanking regions/proximal enhancer
tss - Predicted active promoters
ctcf - Active CTCF binding sites
distal - Predicted active enhancers
tfbs - Unannotated active transcription factor binding sites

Ensembl_OSTEO_activity: OSTEO specific activity prediction from Ensembl Regulatory Build.
Predicted states: open - Unannotated active open chromatin regions
proximal - Predicted active promoter flanking regions/proximal enhancer
tss - Predicted active promoters
ctcf - Active CTCF binding sites
distal - Predicted active enhancers
tfbs - Unannotated active transcription factor binding sites

Ensembl_A549_segmentation: A549 specific genome segment prediction from Ensembl
Regulatory
Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated
tss - Active promoter
ctcf - Distal CTCF
weak - Weak signal
distal - Distal enhancer
dead - Polycomb repressed

Ensembl_DND41_segmentation: DND41 specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer
gene - Transcription associated
tss - Active promoter
ctcf - Distal CTCF
weak - Weak signal
distal - Distal enhancer
dead - Polycomb repressed

Ensembl_GM12878_segmentation: GM12878 specific genome segment prediction from Ensembl Regulatory
Build. Predicted states: proximal - Proximal enhancer
gene - Transcription associated
tss - Active promoter
ctcf - Distal CTCF
weak - Weak signal
distal - Distal enhancer
dead - Polycomb repressed

Ensembl_H1HESC_segmentation: H1HESC specific genome segment prediction from Ensembl Regulatory
Build. Predicted states: proximal - Proximal enhancer
gene - Transcription associated
tss - Active promoter
ctcf - Distal CTCF
weak - Weak signal
distal - Distal enhancer
dead - Polycomb repressed

Ensembl_HELAS3_segmentation: HELAS3 specific genome segment prediction from Ensembl Regulatory
Build. Predicted states: proximal - Proximal enhancer
gene - Transcription associated
tss - Active promoter
ctcf - Distal CTCF
weak - Weak signal
distal - Distal enhancer
dead - Polycomb repressed

Ensembl_HEPG2_segmentation: HEPG2 specific genome segment prediction from Ensembl Regulatory
Build. Predicted states: proximal - Proximal enhancer
gene - Transcription associated
tss - Active promoter
ctcf - Distal CTCF
weak - Weak signal

distal - Distal enhancer
dead - Polycomb repressed

Ensembl_HMEC_segmentation: HMEC specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer
gene - Transcription associated
tss - Active promoter
ctcf - Distal CTCF
weak - Weak signal
distal - Distal enhancer
dead - Polycomb repressed

Ensembl_HSMM_segmentation: HSMM specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer
gene - Transcription associated
tss - Active promoter
ctcf - Distal CTCF
weak - Weak signal
distal - Distal enhancer
dead - Polycomb repressed

Ensembl_HSMMT_segmentation: HSMMT specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer
gene - Transcription associated
tss - Active promoter
ctcf - Distal CTCF
weak - Weak signal
distal - Distal enhancer
dead - Polycomb repressed

Ensembl_HUVEC_segmentation: HUVEC specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer
gene - Transcription associated
tss - Active promoter
ctcf - Distal CTCF
weak - Weak signal
distal - Distal enhancer
dead - Polycomb repressed

Ensembl_K562_segmentation: K562 specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer
gene - Transcription associated
tss - Active promoter
ctcf - Distal CTCF
weak - Weak signal
distal - Distal enhancer
dead - Polycomb repressed

Ensembl_MONO_segmentation: MONO specific genome segment prediction from Ensembl
Regulatory
Build. Predicted states: proximal - Proximal enhancer
gene - Transcription associated
tss - Active promoter
ctcf - Distal CTCF
weak - Weak signal
distal - Distal enhancer
dead - Polycomb repressed

Ensembl_NHA_segmentation: NHA specific genome segment prediction from Ensembl
Regulatory
Build. Predicted states: proximal - Proximal enhancer
gene - Transcription associated
tss - Active promoter
ctcf - Distal CTCF
weak - Weak signal
distal - Distal enhancer
dead - Polycomb repressed

Ensembl_NHDFAD_segmentation: NHDFAD specific genome segment prediction from Ensembl
Regulatory
Build. Predicted states: proximal - Proximal enhancer
gene - Transcription associated
tss - Active promoter
ctcf - Distal CTCF
weak - Weak signal
distal - Distal enhancer
dead - Polycomb repressed

Ensembl_NHEK_segmentation: NHEK specific genome segment prediction from Ensembl
Regulatory
Build. Predicted states: proximal - Proximal enhancer
gene - Transcription associated
tss - Active promoter
ctcf - Distal CTCF
weak - Weak signal
distal - Distal enhancer
dead - Polycomb repressed

Ensembl_NHLF_segmentation: NHLF specific genome segment prediction from Ensembl
Regulatory
Build. Predicted states: proximal - Proximal enhancer
gene - Transcription associated
tss - Active promoter
ctcf - Distal CTCF
weak - Weak signal
distal - Distal enhancer
dead - Polycomb repressed

Ensembl_OSTEO_segmentation: OSTEO specific genome segment prediction from Ensembl
Regulatory
Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated
tss - Active promoter
ctcf - Distal CTCF
weak - Weak signal
distal - Distal enhancer
dead - Polycomb repressed

ENCODE_Gm12878_segmentation: the genome segmentation of the cell line Gm12878 using two different unsupervised machine learning techniques (ChromHMM and Segway). TSS - Predicted promoter region including TSS, PF - Predicted promoter flanking region, E - Predicted enhancer, WE - Predicted weak enhancer or open chromatin cis regulatory element, CTCF - CTCF enriched element, T - Predicted transcribed region, R - Predicted Repressed or Low Activity region

ENCODE_H1hesc_segmentation: the genome segmentation of the cell line H1hesc using two different unsupervised machine learning techniques (ChromHMM and Segway). TSS - Predicted promoter region including TSS, PF - Predicted promoter flanking region, E - Predicted enhancer, WE - Predicted weak enhancer or open chromatin cis regulatory element, CTCF - CTCF enriched element, T - Predicted transcribed region, R - Predicted Repressed or Low Activity region

ENCODE_Helas3_segmentation: the genome segmentation of the cell line Helas3 using two different unsupervised machine learning techniques (ChromHMM and Segway). TSS - Predicted promoter region including TSS, PF - Predicted promoter flanking region, E - Predicted enhancer, WE - Predicted weak enhancer or open chromatin cis regulatory element, CTCF - CTCF enriched element, T - Predicted transcribed region, R - Predicted Repressed or Low Activity region

ENCODE_Hepg2_segmentation: the genome segmentation of the cell line Hepg2 using two different unsupervised machine learning techniques (ChromHMM and Segway). TSS - Predicted promoter region including TSS, PF - Predicted promoter flanking region, E - Predicted enhancer, WE - Predicted weak enhancer or open chromatin cis regulatory element, CTCF - CTCF enriched element, T - Predicted transcribed region, R - Predicted Repressed or Low Activity region

ENCODE_Huvec_segmentation: the genome segmentation of the cell line Huvec using two different unsupervised machine learning techniques (ChromHMM and Segway). TSS - Predicted promoter region including TSS, PF - Predicted promoter flanking region, E - Predicted enhancer, WE - Predicted weak enhancer or open chromatin cis regulatory element, CTCF - CTCF enriched element, T - Predicted transcribed region, R - Predicted Repressed or Low Activity region

ENCODE_K562_segmentation: the genome segmentation of the cell line K562 using two different unsupervised machine learning techniques (ChromHMM and Segway). TSS - Predicted promoter region including TSS, PF - Predicted promoter flanking region, E - Predicted enhancer, WE - Predicted weak enhancer or open chromatin cis regulatory element, CTCF - CTCF enriched element, T - Predicted transcribed region, R - Predicted Repressed or Low Activity region

The following columns are based on the ANNOVAR_ucsc_summary:

ANNOVAR_ucsc_precedent_consequence: the most "damaging" consequence based on ANNOVAR annotation with UCSC knowngene gene models. The rank of damaging applied:

1 stopgain
2 splicing

3 stoploss
4 frameshift
5 nonframeshift
6 nonsynonymous
7 synonymous
8 exonic
9 UTR5
10 UTR3
11 ncRNA_splicing
12 ncRNA_exonic
13 upstream
14 intronic
15 ncRNA_intronic
16 downstream
17 intergenic
18 unknown

ANNOVAR_ucsc_precedent_gene: gene name associated with ANNOVAR_ucsc_precedent_consequence
unique_variant: "Y" for the most "damaging" consequence/gene of the variant; "N" for other consequences/genes

The following columns are based on ANNOVAR_ucsc_precedent_gene and from dbNSFP3.0_gene:

Gene_old_names: Old gene symbol (from HGNC)
Gene_other_names: Other gene names (from HGNC)
Uniprot_acc: Uniprot acc number (from HGNC and Uniprot)
Uniprot_id: Uniprot id (from HGNC and Uniprot)
Entrez_gene_id: Entrez gene id (from HGNC)
CCDS_id: CCDS id (from HGNC)
Refseq_id: Refseq gene id (from HGNC)
ucsc_id: UCSC gene id (from HGNC)
MIM_id: MIM gene id (from HGNC)
Gene_full_name: Gene full name (from HGNC)
Pathway(Uniprot): Pathway description from Uniprot
Pathway(BioCarta)_short: Short name of the Pathway(s) the gene belongs to (from BioCarta)
Pathway(BioCarta)_full: Full name(s) of the Pathway(s) the gene belongs to (from BioCarta)
Pathway(ConsensusPathDB): Pathway(s) the gene belongs to (from ConsensusPathDB)
Pathway(KEGG)_id: ID(s) of the Pathway(s) the gene belongs to (from KEGG)
Pathway(KEGG)_full: Full name(s) of the Pathway(s) the gene belongs to (from KEGG)
Function_description: Function description of the gene (from Uniprot)
Disease_description: Disease(s) the gene caused or associated with (from Uniprot)
MIM_phenotype_id: MIM id(s) of the phenotype the gene caused or associated with (from Uniprot)
MIM_disease: MIM disease name(s) with MIM id(s) in "[]" (from Uniprot)
Trait_association(GWAS): Trait(s) the gene associated with (from GWAS catalog)
GO_biological_process: GO terms for biological process
GO_cellular_component: GO terms for cellular component
GO_molecular_function: GO terms for molecular function

Tissue_specificity(Uniprot): Tissue specificity description from Uniprot
Expression(egenetics): Tissues/organs the gene expressed in (egenetics data from BioMart)
Expression(GNF/Atlas): Tissues/organs the gene expressed in (GNF/Atlas data from BioMart)
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
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
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
P(HI): Estimated probability of haploinsufficiency of the gene (from doi:10.1371/journal.pgen.1001154)
P(rec): Estimated probability that gene is a recessive disease gene (from DOI:10.1126/science.1215040)
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)
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
RVIS_percentile: The percentile rank of the gene based on RVIS, the higher the percentile the more tolerant to mutational burden the gene is.
Essential_gene: Essential ("E") or Non-essential phenotype-changing ("N") based on Mouse Genome Informatics database. from doi:10.1371/journal.pgen.1003484
MGI_mouse_gene: Homolog mouse gene name from MGI
MGI_mouse_phenotype: Phenotype description for the homolog mouse gene from MGI
ZFIN_zebrafish_gene: Homolog zebrafish gene name from ZFIN
ZFIN_zebrafish_structure: Affected structure of the homolog zebrafish gene from ZFIN
ZFIN_zebrafish_phenotype_quality: Phenotype description for the homolog zebrafish gene from ZFIN
ZFIN_zebrafish_phenotype_tag: Phenotype tag for the homolog zebrafish gene from ZFIN
Ancestral_allele: Ancestral allele (based on the EPO pipeline). 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
AltaiNeandertal: genotype of a deep sequenced Altai Neandertal
Denisova: genotype of a deep sequenced Denisova

The following columns are nonsynonymous or splicing SNPs with entries in dbNSFP v2.9 (multiple entries separated by "|"):

aaref: reference amino acid
"-" if the variant is a splicing site SNP (2bp on each end of an intron)

aaalt: alternative amino acid
"-" if the variant is a splicing site SNP (2bp on each end of an intron)

genename: gene name; if the NScan be assigned to multiple genes, gene names are separated by ";"

Uniprot_acc_1: Uniprot accession number. Multiple entries separated by ";".

Uniprot_id_1: Uniprot ID number. Multiple entries separated by ";".

Uniprot_aapos: amino acid position as to Uniprot. Multiple entries separated by ";".

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 ";".

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

refcodon: reference codon

SLR_test_statistic: SLR test statistic for testing natural selection on codons.
A negative value indicates negative selection, and a positive value indicates positive selection. Larger magnitude of the value suggests stronger evidence.

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

fold-degenerate: degenerate type (0, 2 or 3)

Ensembl_geneid: Ensembl gene id

Ensembl_transcriptid: Ensembl transcript ids (separated by ";")

aapos: amino acid position as to the protein
"-1" if the variant is a splicing site SNP (2bp on each end of an intron)

aapos_SIFT: ENSP id and amino acid positions corresponding to SIFT scores. Multiple entries separated by ";"

aapos_FATHMM: ENSP id and amino acid positions corresponding to FATHMM scores. Multiple entries separated by ";"

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 ";".

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 0.02654 to 0.87932.

SIFT_pred: If SIFTori is smaller than 0.05 (rankscore>0.55) the corresponding NS is predicted as "D(amaging)"; otherwise it is predicted as "T(olerated)". Multiple predictions separated by ";"

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

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 0.02656 to 0.89917.

Polyphen2_HDIV_pred: Polyphen2 prediction based on HumDiv, "D" ("probably damaging", HDIV score in [0.957,1] or rankscore in [0.52996,0.89917]), "P" ("possibly damaging", HDIV score in [0.453,0.956] or rankscore in [0.34412,0.52842]) and "B" ("benign", HDIV score in [0,0.452] or rankscore in [0.02656,0.34399]). Score cutoff for binary classification is 0.5 for HDIV score or 0.35411 for rankscore, i.e. the prediction is "neutral" if the HDIV score is smaller than 0.5 (rankscore is smaller than 0.35411), and "deleterious" if the HDIV score is larger than 0.5 (rankscore is larger than 0.35411). Multiple entries are separated by ";".

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

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 0.01281 to 0.9711.

Polyphen2_HVAR_pred: Polyphen2 prediction based on HumVar, "D" ("probably damaging", HVAR score in [0.909,1] or rankscore in [0.62955,0.9711]), "P" ("possibly damaging", HVAR in [0.447,0.908] or rankscore in [0.44359,0.62885]) and "B" ("benign", HVAR score in [0,0.446] or rankscore in [0.01281,0.44315]). Score cutoff for binary classification is 0.5 for HVAR score or 0.45998 for rankscore, i.e. the prediction is "neutral" if the HVAR score is smaller than 0.5 (rankscore is smaller than 0.45998), and "deleterious" if the HVAR score is larger than 0.5 (rankscore is larger than 0.45998). Multiple entries are separated by ";".

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

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 0.00166 to 0.85682.

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

MutationTaster_score: MutationTaster p-value (MTori), ranges from 0 to 1.

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. The rankscore is the ratio of the rank of the score over the total number of MTnew scores in dbNSFP. The scores range from 0.0931 to 0.80722.

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 MTori and 0.328 for the rankscore.

MutationAssessor_score: MutationAssessor functional impact combined score (MAori). The score ranges from -5.545 to 5.975 in dbNSFP. Please refer to Reva et al. (2011) Nucl. Acids Res. 39(17):e118 for details.

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.

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 0.9416, 0.61387 and 0.26162, respectively.

FATHMM_score: FATHMM default score (weighted for human inherited-disease mutations with Disease Ontology) (FATHMMori). Scores range from -18.09 to 11.0. Multiple scores separated by ";" Please refer to Shihab et al. (2013) Human Mutation 34(1):57-65 for details.

FATHMM_rankscore: FATHMMori scores were ranked among all FATHMMori scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of FATHMMori scores in dbNSFP. If there are multiple scores, only the most damaging (largest) rankscore is presented. The scores range from 0 to 1.

FATHMM_pred: If a FATHMMori score is <=-1.5 (or rankscore <=0.81415) the corresponding NS is predicted as "D(AMAGING)"; otherwise it is predicted as "T(OLERATED)". Multiple predictions separated by ";"

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.

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.

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 0.83357.

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.

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.

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 0.82268.

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.

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. In case there are multiple scores for the same variant, the largest score (most damaging) is presented. Please refer to Carter et al., (2013) BMC Genomics. 14(3) 1-16 for details. 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.

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. 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.

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 ";". Details can be found in DOI: 10.1371/journal.pone.0046688

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.

PROVEAN_pred: If PROVEANori <= -2.5 (rankscore>=0.59) the corresponding NS is predicted as "D(amaging)"; otherwise it is predicted as "N(eutral)". Multiple predictions separated by ";"

The following columns are based on per site/SNV annotations with a number within {} presenting the count of that annotation:

indel_focal_length: length of the focal region, i.e. the size of indel (see details in doi:10.1136/jmedgenet-2015-103423)

focal_snv_number: number of "SNVs" created by the indel within the focal region (see details in doi:10.1136/jmedgenet-2015-103423)

dbNSFP version 2.9 Resources

Released: February 3, 2015

Major sources:

Variant determination:

Gencode release 9/Ensembl 64, released May, 2011

Funtional predictions:

SIFT ensembl 66, released Jan, 2015 http://provean.jcvi.org/index.php
PROVEAN 1.1 ensembl 66, released Jan, 2015 http://provean.jcvi.org/index.php
Polyphen-2 v2.2.2, released Feb, 2012 http://genetics.bwh.harvard.edu/pph2/
LRT, released November, 2009 http://www.genetics.wustl.edu/jflab/lrt_query.html
MutationTaster, data retrieved in 2013 http://www.mutationtaster.org/
MutationAssessor, release 2 http://mutationassessor.org/
FATHMM, v2.3 http://fathmm.biocompute.org.uk
CADD, v1.2 http://cadd.gs.washington.edu/
VEST, v3.0 http://karchinlab.org/apps/appVest.html

Citations:

Liu X, Jian X, and Boerwinkle E. 2011. dbNSFP: a lightweight database of human non-synonymous SNPs and their functional predictions. Human Mutation. 32:894-899.

Liu X, Jian X, and Boerwinkle E. 2013. dbNSFP v2.0: A Database of Human Non-synonymous SNVs and Their Functional Predictions and Annotations. Human Mutation. 34:E2393-E2402.

Liu X, White S, Peng B, Johnson AD, Brody JA, Li AH, Huang Z, Carroll A, Wei P, Gibbs R, Klein RJ and Boerwinkle E. (2016) WGSA: an annotation pipeline for human genome sequencing studies. Journal of Medical Genetics 53:111-112.

WGSA v0.55 Resources

List of resources (WGSA v0.55)

Resource	**Brief Description**	**Ref.**	**URL**
Functional annotation for missense and splicing SNVs & gene-centric annotation			
dbNSFP (v2.9)	An integrated functional annotation database for missense	[1,2]	https://sites.google.com/site/jpopgen/dbNSFP

	SNVs and splicing SNVs		
dbscSNV (v1.0)	A deleteriousness prediction score for SNVs within splicing consensus regions (scSNVs)	3	https://sites.google.com/site/jpopgen/dbNSFP
SPIDEX (free non-commercial v1)	A deleteriousness prediction score for SNVs near splicing sites. (note: independent license/download required)	4	http://www.deepgenomics.com/spidex/
Functional prediction scores for non-coding SNVs			
CADD (v1.3)	A genome-wide deleteriousness prediction score for DNA variants based on 63 sequence features (only SNV annotations are in WGSA)	5	http://cadd.gs.washington.edu/
DANN	A functional prediction score retrained based on the training data of CADD using deep neural network.	6	https://cbcl.ics.uci.edu/public_data/DANN/
FATHMM-MKL	A genome-wide deleteriousness prediction score for SNVs based on 10 feature groups	7	http://fathmm.biocompute.org.uk/fathmmMKL.htm
fitCons	A genome-wide deleteriousness measure for genomic positions based on functional assays and selective pressure estimation.	8	http://compgen.cshl.edu/fitCons/
Funseq	A genome-wide categorical deleteriousness prediction score for DNA variants (only non-coding SNV annotations are in WGSA)	9	http://funseq.gersteinlab.org/
Funseq2	A genome-wide	10	http://funseq2.gersteinlab.org/

	deleteriousness prediction score designed for non-coding somatic SNVs		
RegulomeDB (v1.0)	A genome-wide categorical functional prediction score for SNVs based on ENCODE annotation	11	http://regulomedb.org/
Allele frequencies (SNVs and indels)			
1000G	Whole genome allele frequencies from the 1000 Genomes Project phase 3 data	12	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/
ESP6500	Exome allele frequencies from the Exome Variant Server ESP6500SI-V2 release	13	http://evs.gs.washington.edu/EVS/
ExAC (r0.3)	Exome allele frequencies from the Exome Aggregation Consortium		http://exac.broadinstitute.org/
UK10K	Whole genome allele frequencies from TWINSUK cohort		http://www.uk10k.org/studies/cohorts.html
Disease-related variants (SNVs and indels)			
ClinVar (2015/09/01)	DNA variants related to human diseases/phenotypes	14	http://www.ncbi.nlm.nih.gov/clinvar/
COSMIC (v71)	Somatic variants discovered in cancer	15	http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/
GWAS catalog (2015/03/05)	DNA variants associated with human diseases/phenotypes discovered in GWAS studies (downloaded 03/05/2015)	16	http://www.genome.gov/gwastudies/
GRASP 2.0	DNA variants associated with human diseases/phenotypes discovered in GWAS studies, including eQTLs and other quantitative trait scans	17	http://apps.nhlbi.nih.gov/Grasp/

Conservation scores			
GERP++	A conservation score measured by "Rejected Substitutions"	18	http://mendel.stanford.edu/SidowLab/downloads/gerp/
phastCons46way primate	A conservation score based on 46way alignment primate set	19	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons46way/primates/
phastCons46way placental	A conservation score based on 46way alignment placental set	19	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons46way/placentalMammals/
phastCons100way vertebrate	A conservation score based on 100way alignment vertebrate set	19	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons100way/hg19.100way.phastCons/
phyloP46way primate	A conservation score based on 46way alignment primate set	20	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phyloP46way/primates/
phyloP46way placental	A conservation score based on 46way alignment placental set	20	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phyloP46way/placentalMammals/
phyloP100way vertebrate	A conservation score based on 100way alignment vertebrate set	20	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phyloP100way/hg19.100way.phyloP100way/
SiPhy	A conservation score based on 29 mammals genomes	21	http://www.broadinstitute.org/mammals/2x/siphy_hg19/
Epigenomics			
ENCODE	DNase clusters (across 125 cell types), uniform TFBS	22	http://genome.ucsc.edu/ENCODE/downloads.html
EnhancerFinder	Predicted enhancers based on VISTA enhancer data set	23	http://genome-mirror.bscb.cornell.edu/cgi-bin/hgTrackUi?g=disc
FANTOM5	Predicted enhancers, CAGE peaks including TSS (promoter)	24	http://fantom.gsc.riken.jp/data/
Roadmap +ENCODE	Regulatory segmentations , TFBS binding probability	25	http://ngs.sanger.ac.uk/production/ensembl/regulation/hg19/
Ancestral information			
Ancestral allele	Ancestral allele inferred via 6 primates EPO +	26,27	ftp://ftp.ebi.ac.uk/pub/databases/ensembl/ancestral_alleles/

	RSRS allele (for mitochondrial variants)		
AltaiNeandertal	Genotype of a deep sequenced Altai Neandertal	28	http://cdna.eva.mpg.de/neandertal/altai/AltaiNeandertal/VCF/
Denisova	Genotype of a deep sequenced Denisova	29	http://www.eva.mpg.de/denisova
Read mappability / genome accessibility			
MAP20	Average Duke mappability score based on 20bp read	22	http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeMapability
MAP35	Average Duke mappability score based on 35bp read	22	http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeMapability
1000G strict mask	Regions which are considered callable by the 1000 Genomes Project when analyzed with a stricter stringency (20120824_strict_mask)	12	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/accessible_genome_masks/
RepeatMasker	Regions masked by RepeatMasker		http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=rmsk
Other annotations			
dbSNP	rs number from dbSNP 144	30	http://www.ncbi.nlm.nih.gov/SNP/
snoRNA/miRNA	snoRNA and miRNA in human genome collected in miRBase r21 and snoRNABase v3	31,32	http://www.mirbase.org/ https://www-snorna.biotoul.fr/
miRNA target	3'UTR miRNA target in human genome predicted by TargetScan v7	33	http://www.targetscan.org/
ORegAnno	Known regulatory elements in human genome	34	http://www.oreganno.org/oregano/

References:

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

2. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum. Mutat.* **34,** E2393–2402 (2013).
3. Jian, X., Boerwinkle, E. & Liu, X. In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Res.* **42,** 13534–13544 (2014).
4. Xiong, H. Y. *et al.* RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347,** 1254806 (2015).
5. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46,** 310–315 (2014).
6. Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinforma. Oxf. Engl.* **31,** 761–763 (2015).
7. Shihab, H. A. *et al.* An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* **31,** 1536–1543 (2015).
8. Gulko, B., Hubisz, M. J., Gronau, I. & Siepel, A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.* **47,** 276–283 (2015).
9. Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342,** 1235587 (2013).
10. Fu, Y. *et al.* FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.* **15,** 480 (2014).
11. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22,** 1790–1797 (2012).
12. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
13. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493,** 216–220 (2013).
14. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42,** D980–D985 (2014).
15. Forbes, S. A. *et al.* COSMIC: exploring the world’s knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43,** D805–D811 (2015).
16. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** D1001–D1006 (2014).
17. Leslie, R., O’Donnell, C. J. & Johnson, A. D. GRASP: analysis of genotype–phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics* **30,** i185–i194 (2014).
18. Davydov, E. V. *et al.* Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Comput Biol* **6,** e1001025 (2010).
19. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034 –1050 (2005).
20. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15,** 901–913 (2005).
21. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478,** 476–482 (2011).

22. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
23. Erwin, G. D. *et al.* Integrating Diverse Datasets Improves Developmental Enhancer Prediction. *PLoS Comput Biol* **10,** e1003677 (2014).
24. FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. *Nature* **507,** 462–470 (2014).
25. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).
26. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26,** 2069–2070 (2010).
27. Behar, D. M. *et al.* A 'Copernican' Reassessment of the Human Mitochondrial DNA Tree from its Root. *Am. J. Hum. Genet.* **90,** 675–684 (2012).
28. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505,** 43–49 (2014).
29. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338,** 222–226 (2012).
30. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311 (2001).
31. Lestrade, L. & Weber, M. J. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res.* **34,** D158–162 (2006).
32. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42,** D68–D73 (2014).
33. Friedman, R. C., Farh, K. K.-H., Burge, C. B. & Bartel, D. P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* **19,** 92–105 (2009).
34. Griffith, O. L. *et al.* ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res.* **36,** D107–D113 (2008).

Part 7. Importing SNP Info file into Excel and R

Viewing in Excel: use text import wizard feature, tab delimiter, and specify text format for columns gene, Gene_old_names, and Gene_other_names. This will prevent Excel from using an automated formatting feature which will convert gene names to dates (e.g., DEC12).

Use the following exemplary command to load into R:

```
anno <-read.table ("/path/to/SNPInfo_HumanExome-12v1_rev7.tsv.txt", header=TRUE, sep="\t",
na.strings=c("NA","."), quote="", comment.char="", as.is=TRUE)
```

There are 368 columns and 292,329 rows in this file.

Part 8. SNP info file version update log

Version 1 - SNPInfo_HumanExome-12v1_rev1.csv
posted 11/15/12
notes: original file with missing MAFs

Version 2 - SNPInfo_HumanExome-12v1_rev2.tsv
posted 12/21/12
notes: added MAFs for 247,039 SNPs in main PLINK files, corrected comma-delimited formatting issue by converting to tab-delimited format

Version 3 - SNPInfo_HumanExome-12v1_rev3.tsv
posted 1/2/13
notes:

- added single gene info for most damaging variant
- corrected Excel automated date formatting issue of gene names (ex: DEC12)

Version 4 - SNPInfo_HumanExome-12v1_rev4.tsv.txt
posted 1/17/13
notes:

- re-annotated 9 variants in or nearby MIR548H3 with newest RefSeq release (Dec 30 2012); updated single_gene and SKATgene columns; was previously mapped to 2 chromosomes
- corrected formatting issue in 2 genes (MARC1 and MARC2); updated single_gene and SKATgene columns
- PPP2R3B variant on chrom XY (rs6603251) confirmed as PAR SNP and left as is, other variants in PPP2R3B were only mapped to chrom X
- updated "sc_exonic", "sc_nonsynSplice", "sc_lof", and "sc_damaging" TRUE/FALSE categorization based on "single_func_region" column (provided by Jen Brody)
- added "Fwd_A1" and "Fwd_A2" allele information for each race-specific minor allele freq calculation; minor allele noted in "Fwd_A1"
- replaced "NA" notation in "single_gene" column with "." so that individuals would not consider "NA" a gene
- added strand flipping notation ("Flip_TOPtoFWD" = 1) for TOP to FWD strands based on Illumina reference (provided by Martina Mueller-Nurasyid)
- added notation of which SNPs were packaged in main PLINK file ("PLINK_file" = 0) or duplicate PLINK file ("PLINK_file" = 1)
- added SNP info file version update log to track revisions
- clarified description and use of "single_gene" column

Version 5 - SNPInfo_HumanExome-12v1_rev5.tsv.txt
posted 2/7/13
notes:

- updated annotation using dbNSFP for 2,017 variants previously described as "exonic" in the single_func_region" column
  *According to Kai Wang, author of ANNOVAR, the “exonic” only case occurs when the gene does not have a complete ORF so the exact amino acid change

cannot be inferred correctly. dbNSFP was used to further categorize these variants.

- replaced "." notation in "single_gene" column with "NA" for intergenic variants
- replaced "." notation in "single_gene" and "single_func_region" column with no value (empty cells) for variants not annotated (indels and mitochondrial)
- updated "PLINKgene", "sc_exonic", "sc_nonsynSplice", "sc_lof", and "sc_damaging" TRUE/FALSE categorization based on revised "single_func_region" column (provided by Jen Brody)

Version 6 - SNPInfo_HumanExome-12v1_rev6.tsv.txt
posted 11/7/14
notes:

- new annotation using dbNSFP v2.6
- all possible annotations provided, therefore some variants listed more than once
- variant annotated with highest damaging rank identified by unique='Y'
- updated annotation using dbNSFP for 1,980 variants previously described as "exonic" in the "func_region" column
  *According to Kai Wang, author of ANNOVAR, the “exonic” only case occurs when the gene does not have a complete ORF so the exact amino acid change cannot be inferred correctly. dbNSFP was used to further categorize these variants.
- updated "sc_exonic", "sc_nonsynSplice", "sc_lof", and "sc_damaging" TRUE/FALSE categorization based on revised "func_region" column (provided by Heather Highland)
- included Purcell (PMID: 24463508) criteria as "NS_strict" and "NS_broad"
- identified variants on the HumanExome BeadChip v1.2 array
- categorized variant types in "VarType"
- identified duplicate/triallelic variants on the chip in "VarDup"
- added instructions for importing SNP info file into Excel and R

Version 7 - SNPInfo_HumanExome-12v1_rev7.tsv.txt
posted 10/6/2016
notes:

- For variant annotation, WGSA v055 with dbNSFP v2.9 based on hg19 was used
- For gene annotation, dbNSFP 3 was used as the content is updated and independent of a human reference version
- updated TRUE/FALSE categorization based on revised "**ANNOVAR_ucsc_precedent_consequence**" column in “**Part 5. Annotation for analyses**”
- variant annotated with highest damaging rank identified by unique_variant='Y' N=247,870
- updated Y chromosome variant allele frequencies based on the Y Gen Consortium recalling effort