CHARGE SNP Info v6 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 "func_region" 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.

Part 6. dbNSFP annotation

---

dbNSFP version 2.6
Release: July 26, 2014

Please cite:
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: dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. Human Mutation. 34:E2393-E2402.

Major sources:

Variant determination:

Gencode release 9/Ensembl 64, released May, 2011

Functional predictions:

SIFT Human_db_37_ensembl_63, released Auguest, 2011 http://sift.jcvi.org/
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, v1.0 http://cadd.gs.washington.edu/
VEST, v3.0 http://karchinlab.org/apps/appVest.html

Conservation scores:

phyloP46way_primate
http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phyloP46way/primates/
phyloP46way_placental
http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phyloP46way/placentalMammals/
phyloP100way_vertebrate
http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phyloP100way/hg19.100way.phyloP100way/
phastCons46way_primate
http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons46way/primates/
phastCons46way_placental
http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons46way/placentalMammals/

phastCons100way_vertebrate
http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons100way/hg19.100way.phastCons/
GERP++ http://mendel.stanford.edu/SidowLab/downloads/gerp/
SiPhy http://www.broadinstitute.org/mammals/2x/siphy_hg19/

Other variant annotation sources:

Interpro http://www.ebi.ac.uk/interpro/
SLR test statistics http://www.ebi.ac.uk/~greg/mammals/
UniSNP http://research.nhgri.nih.gov/tools/unisnp/
1000 Genomes project http://www.1000genomes.org/
ANNOVAR http://www.openbioinformatics.org/annovar/
ESP https://esp.gs.washington.edu/drupal/

Other gene annotation sources:

HGNC, downloaded on Oct. 1, 2013
Uniprot, relaesed Sept., 2013
IntAct, downloaded on Oct. 1, 2013
GWAS catalog, downloaded on Oct. 1, 2013
egenetics and GNF/Atlas expression data, downloaded from BioMart on Oct. 1, 2013
BioGRID, version 3.2.105
Haploinsufficiency probability data, from doi:10.1371/journal.pgen.1001154
Recessive probability data, from DOI:10.1126/science.1215040
GO Slim, GOC Validation Date: 09/27/2013
ConsensusPathDB, version 27
Essential genes, based on doi:10.1371/journal.pgen.1003484

SNP Exclusions:

MT SNPs not included in dbNSFP annotation.

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” column, which is determined by the most damaging functional annotation.

The following rules are applied:
1. If a SNP/indel is annotated for both a gene and a read-through RNA (if included). All rows for the read-through RNA will be ‘N’.
2. For the remaining rows, pick the row with the highest rank of damaging as ‘Y’ and others as ‘N’.
3. If there is a tie for the highest rank of damaging among some rows, pick one (the first one the program finds) as ‘Y’.

Damaging Ranking Table

Rank	Func_Region
1	stopgain
2	stoploss
3	splicing
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

All SNPs have the following columns:

chr: chromosome number (human)

pos: physical position on the chromosome as to hg19 (1-based coordinate). Indels noted in vcf style (-1 coordinate in MapInfo column above).

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

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

refstrand: updated IlmnStrand

snp: updated IlmnSNP

seq: updated SourceSeq

func_region: function annotation from ANNOVAR (based on RefSeq) for variants and indels only.

gene: gene name from ANNOVAR (based on RefSeq)

unique: whether the variant should be included in the "unique" SNP set (based on ANNOVAR)

rs_dbSNP138: rs number from dbSNP138

MAP20: average Duke mappability score based on 20bp read

MAP35: average Duke mappability score based on 35bp read

Ancestral_allele: Ancestral allele for SNPs based on 1000 genomes reference data. The following comes from its original README file:

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

actg - low-confindence 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

Ancestral allele for indels based on CADD annotation.

AltaiNeandertal: genotype of a deep sequenced Altai Neandertal

Denisova: genotype of a deep sequenced Denisova

phyloP46way_primate: a conservation score based on 46way alignment primate set, the higher the more conservative

phyloP46way_placental: a conservation score based on 46way alignment placental set, the higher the more conservative

phyloP100way_vertebrate: a conservation score based on 100way alignment vertebrate set, the higher the more conservative

phastCons46way_primate: a conservation score based on 46way alignment primate set, the higher the more conservative

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

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

GERP++_NR: GERP++ neutral rate

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

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

1000Gp1_AC: Alternative allele counts in the whole 1000 genomes phase 1 (1000Gp1) data.

1000Gp1_AF: Alternative allele frequency in the whole 1000Gp1 data.

1000Gp1_AFR_AC: Alternative allele counts in the 1000Gp1 African descendent samples.

1000Gp1_AFR_AF: Alternative allele frequency in the 1000Gp1 African descendent samples.

1000Gp1_EUR_AC: Alternative allele counts in the 1000Gp1 European descendent samples.

1000Gp1_EUR_AF: Alternative allele frequency in the 1000Gp1 European descendent samples.

1000Gp1_AMR_AC: Alternative allele counts in the 1000Gp1 American descendent samples.

1000Gp1_AMR_AF: Alternative allele frequency in the 1000Gp1 American descendent samples.

1000Gp1_ASN_AC: Alternative allele counts in the 1000Gp1 Asian descendent samples.

1000Gp1_ASN_AF: Alternative allele frequency in the 1000Gp1 Asian descendent samples.

1000Gp1_Fst: Fst calculated based on all 1000g phase 1 populations

ESP6500_AA_AF: Alternative allele frequency in the Afrian 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).

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 "."

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).

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

splicing_consensus_adaboost_score: splicing-change prediction for splicing consensus SNPs based on adaboost. If the score >0.5, 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 >0.5, it predicts that the splicing will be changed, otherwise it predicts the splicing will not be changed.

The following columns are unique to SNPs with an entry in dbNSFP v2.5:

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)

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

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

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

Uniprot_id: 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_feature: gene feature changes for indels only based on MutationTaster annotation

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 ";"

RadialSVM_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.

RadialSVM_rankscore: RadialSVM scores were ranked among all RadialSVM scores in dbNSFP. The rankscore is the ratio of the rank of the screo over the total number of RadialSVM scores in dbNSFP. The scores range from 0 to 1.

RadialSVM_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.

LR_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.

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

LR_pred: Prediction of our LR 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 RadialSVM and LR. Ranges from 1 to 10. As RadialSVM and LR 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.

The following columns are from dbNSFP gene annotation:

Gene_old_names: Old gene sybmol (from HGNC)
Gene_other_names: Other gene names (from HGNC)
Uniprot_acc(HGNC/Uniprot): Uniprot acc number (from HGNC and Uniprot)
Uniprot_id(HGNC/Uniprot): 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(s) the gene belongs to (from Uniprot)
Pathway(ConsensusPathDB): Pathway(s) the gene belongs to (from ConsensusPathDB)
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_Slim_biological_process: GO Slim terms for biological process
GO_Slim_cellular_component: GO Slim terms for cellular component
GO_Slim_molecular_function: GO Slim terms for molecular function
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): Other genes the gene interacted with (from IntAct) gene name followed by Pubmed id in "[]"
Interactions(BioGRID): Other genes the gene interacted with (from BioGRID) gene name followed by Pubmed id in "[]"
Interactions(ConsensusPathDB): Other genes the gene interacted with (from ConsensusPathDB) gene name followed by interaction confidence in "[]"
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)
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

Note 1: Missing data is designated as '.'.
Note 2: Mulitple annotations are separated by ';'

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_rev6.tsv.txt", header=TRUE, sep="\t", na.strings=c("NA","."), quote="", comment.char="", as.is=TRUE)

There are 203 variables and 267,389 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