TCAG SMALL VARIANT ANNOTATION PIPELINE

Creation date: 01 June 2022.

Pipeline version: 27.7

Table of contents

1. GENERAL RELEASE NOTES

Following information is valid for annotations generated by TCAG annotation pipeline rev27.7 and applicable to variants called by Dragen.(https://support-docs.illumina.com/SW/DRAGEN_v38/Content/SW/DRAGEN/GPipelineIntro_fDG.htm)

2. Overall Annotation Process:

The Variant Call Format is a standardized format to represent SNPs and indels. The TCAG annotation pipeline uses ANNOVAR to functionally annotate these small variants. In order to accurately add gene-based, region-based and filter-based annotations, the original VCF file generated by the variant caller needs to be converted into an ANNOVAR compatible format. The conversion involves variant decomposition (splitting of variants with two alternate alleles into multiple rows), left-aligning indels on the forward strand of the reference genome and normalization. For indels, the normalization sometimes shifts the position of the variants. The 'GT_PreNorm' captures the genotype before variant decomposition and 'Original_VCFKey' captures the position of the variant before normalization.

3. Detailed annotation field description

[coordinates]

CHROM: chromosome

start: start position (1-positional system)

end: end position (1-positional system)

Original_VCFKEY: unique variant identifier for original vcf file.

MULTI_ALLELIC: Flag to represent multi-allelic sites (ie 1=multi-allelic).

[sequence and ploidy]

<Sample-name>:zygosity: heterozygous-reference (ref-alt), homozygous-alternate (hom-alt) and heterozygous-alternate(alt-alt)

ref_allele: Reference allele

alt_allele: Alternate allele

<Sample-name>: Genotype: Genotype, represented as “reference allele | alternate allele” (ref-alt) or “alternate allele | alternate allele” (hom-alt) or “alternate allele 1 | alternate allele 2” (alt-alt). This is based on pre-normalized raw genotype (eg: A/C).

<Sample-name>: GT_PostNorm: Raw genotype after left-normalization (eg : 0/1)

<Sample-name>: GT_PreNorm: Raw genotype before left-normalization (eg : 1/2)

var_type: snp (single nucleotide variation, can be ref-alt), ins (insertion), del (deletion), mnp (multiple bases substitutions), complex (could be block substitution).

<Sample-name>: PRI: Physical phasing haplotype information, describing how the alternate alleles are phased in relation to one another.

<Sample-name>: PS: Physical phasing ID information, where each unique ID within a given sample (but not across samples) connects records within a phasing group.

[quality score and read counts]

FILTER: Filter status: PASS if this position has passed all filters.

DP: Approximate read depth for this variant.

FS: Phred-scaled p-value using Fisher's exact test to detect strand bias .

MQ: RMS Mapping Quality**.** This annotation provides an estimation of the overall mapping quality of reads supporting a variant call. It produces both raw data (sum of square and num of total reads) and the calculated root mean square **.**

MQRankSum: Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities .

ReadPosRankSum: Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias .

FractionInformativeReads: The fraction of informative reads out of the total reads.

<Sample-name>: AD_REF: Allelic depth for the ref allele.

<Sample-name>: AD_ALT: Allelic depth for the alt allele.

<Sample-name>: DP: Filtered basecall depth used for site genotyping.

<Sample-name>: GQ: Genotype quality.

<Sample-name>: PL: Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification .

<Sample-name>: GP: Phred-scaled posterior probabilities for genotypes as defined in the VCF specification.
<Sample-name>: MB: Per-sample component statistics to detect mate bias**.**
<Sample-name>: F1R2: Count of reads in F1R2 pair orientation supporting each allele.
<Sample-name>: F2R1: Count of reads in F2R1 pair orientation supporting each allele.

[allele frequency]
A1000g_all: allele frequency in the full 1000 Genome data-set
A1000g_eur: allele frequency in the caucasian-european sub-set of 1000 Genome
A1000g_amr: allele frequency in the mixed-background latin americans (e.g. Mexicans, Puerto Ricans, Peruvians) sub-set of 1000 Genome
A1000g_eas: allele frequency in the east-asian sub-set of 1000 Genome
A1000g_sas: allele frequency in the south-asian sub-set of 1000 Genome
A1000g_afr: allele frequency in the black-african sub-set of 1000 Genome (note african here indicates the ethnic background, not the actual geographical location of the population sample)
gnomAD_exome211_FILTER: gnomAD v2.1.1 Filter status: PASS if this position has passed all filters. From gnomAD exome dataset.
gnomAD_exome211_AC: gnomAD v2.1.1 alternate allele count for samples. From gnomAD exome dataset.
gnomAD_exome211_AC_female: gnomAD v2.1.1 alternate allele count for female samples. From gnomAD exome dataset.
gnomAD_exome211_AC_male: gnomAD v2.1.1 alternate allele count for male samples. From gnomAD exome dataset.
gnomAD_exome211_AN: gnomAD v2.1.1 total number of alleles in all samples. From gnomAD exome dataset.
gnomAD_exome211_AN_female: gnomAD v2.1.1 total number of alleles in female samples. From gnomAD exome dataset.
gnomAD_exome211_AN_male: gnomAD v2.1.1 total number of alleles in male samples. From gnomAD exome dataset.
gnomAD_exome211_nhomalt: gnomAD v2.1.1 count of homozygous individuals in samples. From gnomAD exome dataset.
gnomAD_exome211_nhomalt_female: gnomAD v2.1.1 count of homozygous individuals in female samples. From gnomAD exome dataset.
gnomAD_exome211_nhomalt_male: gnomAD v2.1.1 count of homozygous individuals in male samples. From gnomAD exome dataset.
gnomAD_exome211_AF: gnomAD v2.1.1 allele frequencies (exomes).
gnomAD_exome211_AF_raw: gnomAD v2.1.1 raw allele frequencies (exomes) - global.
gnomAD_exome211_AF_afr: nomAD v2.1.1 allele frequencies (exomes) - African-American/African subset.
gnomAD_exome211_AF_amr: gnomAD v2.1.1 allele frequencies (exomes) - Latino/Admixed American subset.

gnomAD_exome211_AF_asj: gnomAD v2.1.1 allele frequencies (exomes) - Ashkenazi Jewish subset.
gnomAD_exome211_AF_eas: gnomAD v2.1.1 allele frequencies (exomes) - East Asian subset.
gnomAD_exome211_AF_nfe: gnomAD v2.1.1 allele frequencies (exomes) - Non-Finnish European subset.
gnomAD_exome211_AF_fin: gnomAD v2.1.1 allele frequencies (exomes) - Finnish subset.
gnomAD_exome211_AF_oth: gnomAD v2.1.1 allele frequencies (exomes) - others.
gnomAD_exome211_AF_sas: gnomAD v2.1.1 allele frequencies (exomes) - South Asian subset.
gnomAD_exome211_AF_female: gnomAD v2.1.1 alternate allele frequency in female samples (exomes) .
gnomAD_exome211_AF_male: gnomAD v2.1.1 alternate allele frequency in male samples (exomes) .
gnomAD_exome211_faf95_afr: Filtering allele frequency (using Poisson 95% CI) for samples of African-American/African ancestry (exomes).
gnomAD_exome211_faf95_amr: Filtering allele frequency (using Poisson 95% CI) for samples of Latino ancestry (exomes).
gnomAD_exome211_faf95_eas: Filtering allele frequency (using Poisson 95% CI) for samples of East Asian ancestry (exomes).
gnomAD_exome211_faf95_nfe: Filtering allele frequency (using Poisson 95% CI) for samples of Non-Finnish European ancestry (exomes).
gnomAD_exome211_faf95_sas: Filtering allele frequency (using Poisson 95% CI) for samples of South Asian ancestry (exomes).
gnomAD_genome31_FILTER: gnomAD v3.1.1 Filter status: PASS if this position has passed all filters. From the gnomAD genome dataset.

gnomAD_genome31_AC: gnomAD v3.1.1 alternate allele count for all samples. From the gnomAD genome dataset.

gnomAD_genome31_AC_XX: gnomAD v3.1.1 alternate allele count for female samples. From the gnomAD genome dataset.

gnomAD_genome31_AC_XY: gnomAD v3.1.1 alternate allele count for male samples. From the gnomAD genome dataset.

gnomAD_genome31_AN: gnomAD v3.1.1 total number of alleles in all samples. From the gnomAD genome dataset.

gnomAD_genome31_AN_XX: gnomAD v3.1.1 total number of alleles in female samples. From the gnomAD genome dataset.

gnomAD_genome31_AN_XY: gnomAD v3.1.1 total number of alleles in male samples. From the gnomAD genome dataset.

gnomAD_genome31_nhomalt: gnomAD v3.1.1 count of homozygous individuals in all samples. From the gnomAD genome dataset.

gnomAD_genome31_nhomalt_XX: gnomAD v3.1.1 count of homozygous individuals in female samples. From the gnomAD genome dataset.

gnomAD_genome31_nhomalt_XY: gnomAD v3.1.1 count of homozygous individuals in male samples. From the gnomAD genome dataset.

gnomAD_genome31_AF: gnomAD v3.1.1 allele frequencies (genomes).

gnomAD_genome31_AF_afr: gnomAD v3.1.1 allele frequencies (genomes) - African/African American subset.

gnomAD_genome31_AF_ami: gnomAD v3.1.1 allele frequencies (genomes) - Amish subset.

gnomAD_genome31_AF_amr: gnomAD v3.1.1 allele frequencies (genomes) -Latino/Admixed American subset.

gnomAD_genome31_AF_asj: gnomAD v3.1.1 allele frequencies (genomes) - Ashkenazi Jewish subset.

gnomAD_genome31_AF_eas: gnomAD v3.1.1 allele frequencies (genomes) - East Asian subset.

gnomAD_genome31_AF_nfe: gnomAD v3.1.1 allele frequencies (genomes) - Non-Finnish European subset.

gnomAD_genome31_AF_fin: gnomAD v3.1.1 allele frequencies (genomes) - Finnish subset.

gnomAD_genome31_AF_oth: gnomAD v3.1.1 allele frequencies (genomes) - others.

gnomAD_genome31_AF_raw: gnomAD v3.1.1 raw allele frequencies (genomes) - global.

gnomAD_genome31_AF_sas: gnomAD v3.1.1 allele frequencies (genomes) - South Asian subset.

gnomAD_genome31_AF_XX: gnomAD v3.1.1 alternate allele frequency in female samples (genomes) .

gnomAD_genome31_AF_XY: gnomAD v3.1.1 alternate allele frequency in male samples (genomes) .

gnomAD_genome31_faf95_afr: Filtering allele frequency (using Poisson 95% CI) for samples of African-American/African ancestry (genomes).

gnomAD_genome31_faf95_amr: Filtering allele frequency (using Poisson 95% CI) for samples of Latino ancestry (genomes).

gnomAD_genome31_faf95_eas: Filtering allele frequency (using Poisson 95% CI) for samples of East Asian ancestry (genomes).

gnomAD_genome31_faf95_nfe: Filtering allele frequency (using Poisson 95% CI) for samples of Non-Finnish European ancestry (genomes).

gnomAD_genome31_faf95_sas: Filtering allele frequency (using Poisson 95% CI) for samples of South Asian ancestry (genomes).

gnomAD_genome31_faf95_popmax: Filtering allele frequency (using Poisson 95% CI) for the population with the maximum allele frequency.

gnomAD_genome31_AF_popmax: Maximum allele frequency across populations**.**

freq_max: Maximum of GnomAD exome/GnomAD genome filtering allele frequencies(faf) .

A1000g_freq_max: Maximum of 1000 Genomes allele frequencies
gnomAD_exome_freq_max: Maximum filtering allele frequencies(faf) for gnomAD (exome).
gnomAD_genome_freq_max: Maximum filtering allele frequencies(faf) of gnomAD (genome).

[reference variant databases]

dbsnp: exact match (position, allele) to dbSNP.
dbsnp_common: exact match (position, allele) to common dbSNP track USCS; this is not meant as a replacement of frequency-based filtering, and is not practically used for hard filters.
dbsnp_region: overlap-based match for dbSNP; this is useful to look up variants that are split into multiple dbSNP entries, or otherwise only partially match to dbSNP entries.
Clinvar_SIG: Overall ClinVar significance code; “pathogenic” is the code of interest for rare disorders. Clinical significance values for all the individual submissions (SCVs) aggregated for the RCV record in ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/docs/clinsig/)
Clinvar_CLNREF: References includes PubMed ID
Clinvar_AlleleID: ClinVar alleleID. A unique integer identifier assigned to each individual variant in ClinVar.
Clinvar_ReviewStatus: The level of review supporting the assertion of clinical significance(https://www.ncbi.nlm.nih.gov/clinvar/docs/details/#review_status)
Clinvar_SIG_Simple: expected values 0,1 or -1; 0 = no current value of pathogenic; 1 = at least one record submitted with pathogenic/likely pathogenic; -1 = no values for clinical significance at all for this variant or set of variants. Used for the "included" variants that are only in ClinVar because they are included in a haplotype or genotype with an interpretation.
cosmic: exact match (position, allele) to the Cosmic database of somatic variants; useful only for cancer projects.
Clinvar_Disease: Clinvar interpreted condition**.**

[gene mapping]

refseq_id: combined Annovar output on coding sequence mapping and effect, composed of: (a) for coding exotic changes ("typeset" exonic): gene official symbol : RefSeq ID : position in the coding sequence : amino acid change, [idem]; (b) for core splice site changes (typeseq "exonic"): gene official symbol (RefSeq ID : exon number : coding sequence position and change, [idem])
typeseq: type of sequence overlapped, with respect to known genes/transcripts and their coding / noncoding status: (a) “exonic” represents coding exons, (b) “splicing” represents core splicing site (by default, 15 bp on the intron side of intron-exon and exon-intron junctions), (c)

"ncRNA_exonic" represents exons of non-coding RNA genes, (d) "ncRNA_splicing" represents core splicing sites of non-coding RNA genes , (e) “UTR5” represents 5’ untranslated region, (f) “UTR3” represents 3’ untranslated region, (g) “upstream” represents 1kb upstream of TSS, (h) “downstream” represents 1kb downstream of TSS and (i) “intergenic” represents intergenic regions (beyond upstream/downstream threshold(1kb)). For variants with multiple sequence overlaps (eg, exonic for one transcript and intronic for other), all possible typseq values will be listed in semicolon-delimited format (eg: exonic;intronic).

typeseq_priority: Prioritized sequence overlap for multi-sequence overlap variants. Implementation of the Annovar prioritization scheme (http://annovar.openbioinformatics.org/en/latest/user-guide/gene/).

typeseq_RefseqSelect: Type of sequence overlapped, with respect to RefSeq Select transcript (https://www.ncbi.nlm.nih.gov/refseq/refseq_select/).

effect: type of effect on the coding sequence: (a) “synonymous SNV”, (b) "nonsynonymous SNV", (c) "stopgain SNV", (d) "frameshift deletion", (e) "frameshift insertion", (f) "frameshift substitution", (g) "nonframeshift deletion", (h) "nonframeshift insertion", (i) "nonframeshift substitution" , (j) “stoploss SNV”. For variants with multiple effects, all possible values will be represented in comma-separated fashion.

effect_priority: Prioritized effects for coding variants with multiple effects (http://annovar.openbioinformatics.org/en/latest/user-guide/gene/).

effect_RefseqSelect: Effect on the coding sequence with respect to RefSeq Select transcript (https://www.ncbi.nlm.nih.gov/refseq/refseq_select/).

aa_flag: this flag is set to 1 if more than one distinct amino acid change is reported in the “refseq-id” field.

distance_spliceJunction: A positive integer represents the distance from the nearest exon boundary, in genomic level.

gene_symbol: official gene symbol.

entrez.id: entrez-gene id.

gene_desc: full gene name.

omim_id: OMIM gene accession id.

omim_Phenotype: OMIM disorder/disease description when available for the corresponding omim gene accession.

MPO: (array of) MPO (Mammalian Phenotype Ontology) top level phenotype(s), imported from MGI and mapped from an orthologous mouse gene; the genotype-phenotype association is typically supported by a heterozygous/homozygous knock-out or other transgenic experiment, sometimes involving more than one gene: these details are exported by TCAG from MGI, but not included in this annotation field, so they should be looked up on the mgi website.

HPO: (array of) HPO (Human Phenotype Ontology) top level phenotype(s), imported from HPO; the genotype-phenotype association is typically supported by an OMIM entry; modes of inheritance are also exported by TCAG from HPO, but not included in this annotation field, so they should be looked up on the HPO website.
CGD_disease: the Clinical Genomics Database is compiled by curators and maintained by the NHGRI (National Human Genome Research Institute); for every gene in the database, the CGD provides a list of one or more genetic disorders and a mode of inheritance; this field reports the genetic disorder(s) .
CGD_inheritance: the Clinical Genomics Database is compiled by curators and maintained by the NHGRI (National Human Genome Research Institute); for every gene in the database, the CGD provides a list of one or more genetic disorders and a mode of inheritance; this field reports the mode of inheritance (AD, AR, AD/AR, XL, more complex modes); since the CGD mode of inheritance is directly added by a curator and it's tied to specific genetic disorder(s), it could be considered more accurate than the mode of inheritance for top-level HPO phenotypes.
gnomAD_oe_lof: LOF observed/expected (oe) metric from gnomAD constraint matrix.
gnomAD_oe_lof_upper: LOF observed/expected (oe) metric - CI upper bound
gnomAD_oe_mis: Missense observed/expected (oe) metric from gnomAD constraint matrix.
gnomAD_oe_mis_upper: missense observed/expected (oe) metric - CI upper bound.
gnomAD_pLI: Probability of being loss-of-function intolerant.
gnomAD_pRec: Probability of being intolerant of homozygous but not heterozygous LOF variants.
gnomAD_mis_z: GnomAD missense Z score**.**
ACMG_disease: Any (exonic, intronic or splice) variants in genes in ACMG (v3.0) published recommendations for reporting incidental findings. (https://www.nature.com/articles/s41436-021-01172-3)

[conservation and predicted impact]
sift_score: SIFT score for predicted protein impact, values <= 0.05 correspond to damaging (can be interpreted as a p-value); note that SIFT is based on amino acid conservation / substitution rates inferred from protein sequence alignments.
polyphen_score: Polyphen2 scores for predicted protein impact, values >= 0.95 correspond to damaging; note that Polyphen2 is based on a set of sequence and structural attributes, combined in a machine classifier trained on a data-set of positive cases, and thus it partially is complementary and partially correlated to SIFT and MA.
ma_score: mutation assessor scores for predicted protein impact, values >= 2 correspond to damaging; note that MA is based on amino acid conservation / substitution rates inferred from protein sequence alignments, additionally modeling protein family groupings, thus can be

regarded as an improved version of SIFT (improved performance has been shown particularly for somatic variants), although to this date SIFT remains more popular / broadly used.

mt_score: Mutationtaster prediction scores.

MPC_score: MPC (**M**issense badness, **P**olyPhen-2, and **C**onstraint) score. MPC is a missense deleteriousness metric based on the analysis of genic regions depleted of missense mutations in the Exome Aggregation Consortium (ExAC) data.

PROVEAN_score: Amino acid substitution or indel prediction score from Provean software (http://provean.jcvi.org/index.php). Values < -2.5 corresponds to a damaging variant.

REVEL_score: Rare Exome Variant Ensemble Learner (REVEL) scores for predicting the deleteriousness of each nucleotide change. The REVEL score for an individual missense variant can range from 0 to 1, with higher scores reflecting greater likelihood that the variant is disease-causing.

phylopMam: value array of PhyloP nucleotide-level conservation inferred from the Placental Mammal genome group. PhyloP is based uniquely on nucleotide substitutions and does not factor structural variation.

phylopMam_avg: Average value of PhyloP nucleotide-level conservation inferred from the Placental Mammal genome group. This field is suitable for hard filters, especially for missense variants and non-frameshift substitutions; note that PhyloP is based uniquely on nucleotide substitutions and does not factor structural variation.

phylopVert100: Value array of PhyloP nucleotide-level conservation inferred from the 100 Vertebrate genome group; note that PhyloP is based uniquely on nucleotide substitutions and does not factor structural variation.

phylopVert100_avg: Average value of PhyloP nucleotide-level conservation inferred from the 100 Vertebrate genome group; this field is suitable to hard filter, especially for missense variants and non-frameshift substitutions; note that PhyloP is based uniquely on nucleotide substitutions and does not factor structural variation.

phastCons_placental: PhastCons score for the Placental Mammal genome group, based on UCSC track (calculated by HMM scan of PhyloP values to infer conserved status); this is useful to assess conservation at the *regional level* (rather than PhyloP's *nucleotide level*); this field can be used as an evidence *suggestive* of functional relevance of a sequence (especially if non-coding), and thus could be used to hard filter non-coding variation.

pfam_annovar: overlap with coding sequence matching to a PFAM protein domain; can help further assess protein impact, but not meant for hard-filter.

per_cds_affected: percentage of coding exonic sequence affected, should be used to prioritize loss-of-function variation (stopgain, frameshift, splicing), but not meant for hard filters.

per_transcripts_affected: percentage of transcripts with variant overlapping them, should be used to prioritize variation, but not meant for hard filters.

CADD_Raw: Raw score from CADD(Combined Annotation Dependent Depletion).
CADD_phred: PHRED-like c-score ranking from CADD.
spx_gene: The RefSeq gene affected by the variant.
spx_transcript: The RefSeq transcript affected by the variant
spx_exonN: The exon for which percent inclusion is predicted
spx_spliceDist: The distance from the variant to the splice site
spx_dpsi: Splice site prediction score from SPIDEX. The change in percentage exon inclusion reported as the maximum across tissues (https://www.ncbi.nlm.nih.gov/pubmed/25525159)
dbscSNV_ADA_SCORE, dbscSNV_RF_SCORE: Splice site prediction scores from dbscSNV (https://www.ncbi.nlm.nih.gov/pubmed/26555599)
spliceAI_SYMBOL: Gene symbol from SpliceAI (https://pypi.org/project/spliceai/)
spliceAI_STRAND: Strand from SpliceAI.
spliceAI_TYPE: Sequence overlap for this variant (I=intronic; E=exonic) from SpliceAI.
spliceAI_DIST: Distance from exon/intron junction from SpliceAI.
spliceAI_DS_AG: SpliceAI Delta score, the probability of the variant being splice-altering (acceptor gain). To filter use 0.5 (recommended), for higher recall 0.2 can be used.
spliceAI_DS_AL: SpliceAI Delta score, the probability of the variant being splice-altering Delta score (acceptor loss). To filter use 0.5 (recommended), for higher recall 0.2 can be used.
spliceAI_DS_DG: SpliceAI Delta score, the probability of the variant being splice-altering Delta score (donor gain). To filter use 0.5 (recommended), for higher recall 0.2 can be used.
spliceAI_DS_DL: SpliceAI Delta score, the probability of the variant being splice-altering Delta score (donor loss). To filter use 0.5 (recommended), for higher recall 0.2 can be used.
spliceAI_DP_AG: SpliceAI Delta position (acceptor gain).
spliceAI_DP_AL: SpliceAI Delta position (acceptor loss).
spliceAI_DP_DG: SpliceAI Delta position (donor gain).
spliceAI_DP_DL: SpliceAI Delta position (donor loss).

[other]
SegDup: overlap with segmental duplications, often useful to explain why ploidy is not 2 outside male sex chromosomes.
Repeat: Repeatmasker annotation from UCSC.

4 .Database and tool versions

Genome build : hg38
Annovar : October 2019 version

Gene-type : refGene
Gene_Updated: 07-FEB-2022
OMIM: version: 07-MAR-2022
HPO: version: 07-MAR-2022
MPO: version: 07-MAR-2022
gnomAD Constraint matrices: Downloaded August 2019
CGD: version: 07-MAR-2022
ACMG: v0.3, MAY-2021

dbsnp: dbsnp150, Downloaded from annovar, updated on 29-SEP-2017
dbsnpCommon: dbsnp151 common, Download from NCBI on 10-APR-2020
clinvar: Downloaded from NCBI and curated , 08-MAR-2022 version
cosmic: v94, Download from Cosmic , v95, MAR-2022 version
1000g: Downloaded from annovar, 24-AUG-2015 version
gnomAD : Version 3.1.1, Downloaded from gnomAD

pfam: Downloaded from UCSC on MAR 2022
SPIDEX: splicing predictions v1.1
segdup:genomicSuperDups, Downloaded from UCSC , Oct 2014 version

phastcon: Downloaded from UCSC, SEPI 2015 version
phylop_placental: phyloP30way Downloaded from UCSC. Nov 2017 version
phylop_100way: phyloP100way, Downloaded from UCSC May 2015
CADD: v1.6, Updated AUG-2020
dbnsfp: v3.5 Downloaded from annovar, 21-FEB-2017 version
dbscsnv: Downloaded from annovar, 18-DEC-2015 version
SpliceAI: Downloaded from Illumina, 08-OCT-2019

5. File Descriptions

*_FINAL_rev27.7.tsv.gz: Complete set of annotated variants.
*_FINAL_rev27.7.tsv.gz.tbi: Tabix index file for above.
*_SUBSET_rev27.7.tsv: Subset of annotated variants from above file. This subset consists of variants that satisfy following conditions:

rare (<5% for maxImum allele frequency (freq_max column))

AND

typeseq column contains at least one of the following terms; exonic, splicing, downstream,upstream or UTR

OR
have significant splicing prediction score (2 < spx_dpsi < -2 , dbscSNV_ADA_SCORE > 0.6, dbscSNV_RF_SCORE > 0.6 , (spliceAI_DS_AG > 0.2 and abs (spliceAI_DP_AG) <= 50), (spliceAI_DS_AL > 0.2 and abs (spliceAI_DP_AL) <= 50), (spliceAI_DS_DG > 0.2 and abs (spliceAI_DP_DG) <= 50) or (spliceAI_DS_DL > 0.2 & abs (spliceAI_DP_DL) <= 50))
OR
distance_spliceJunction <= 100.

Use the column named “SUBSET_filter” to identify the criteria satisfied for each variant in SUBSET file.