SV Annotation Table

DESCRIPTION OF EACH COLUMN

SV_Name: name of the SV

chrAnn: Chromosome (chr1 to chr22, plus chrX, chrY, and chrM).

startAnn: CNV/SV start.

endAnn: CNV/SV end.

variantTypeAnn: Variant type as interpreted by the annotation pipeline; for CNVs, deletions and duplications. For CNVs, complex is used when the calling software assigns a type which cannot be ascribed as either duplication or deletion (e.g. translocations, inversions, and other events); in this case the annotation pipeline ignores the type in the database when matching it with the variant. SV types recognized by the pipeline are: duplications, deletions, inversions, and insertions. Insertions of size 1 are padded (+/- 50bp) prior to frequency calculations. BND are currently ignored.

sizeAnn: CNV/SV size as determined by the pipeline.

GC_content_perc: GC content based on the UCSC genome reference, hg37 or hg38.

cytobandAnn: Cytoband.

numberOfGeneSymbols: number of official gene symbol(s) for genes spanned by the CNV/SV based on the UCSC RefSeq gene definitions.

gene_symbol: official gene symbol(s) for genes spanned by the CNV/SV based on the UCSC RefSeq gene definitions.

gene_symbol_CNVstart: official gene symbol(s); overlap on CNV/SV start coordinate.

gene_symbol_CNVend: official gene symbol(s); overlap on CNV/SV end coordinate.

exon_symbol: official gene symbol(s); exons overlap only.

cds_symbol: official gene symbol(s); coding exons overlap.

ISCA_haploinsufficient: (array of) entrez gene ID, gene symbol, and score from the dosage sensitivity map, haploinsufficient phenotype defined in ClinGen (ISCA). ClinGen is a National Institutes of Health (NIH)-funded resource dedicated to building an authoritative central resource that defines the clinical relevance of genes and variants for use in precision medicine and research.

ISCA_triplosensitive: (array of) entrez gene ID, gene symbol, and score for dosage sensitivity map, triplosensensitive phenotype defined in ClinGen (ISCA).

ExAC_pLI: probability of being loss-of-function intolerant; for more information of ExAc functional constraint scores, see Samocha *et al*. - Nature Genetics 2014 (http://www.ncbi.nlm.nih.gov/pubmed/25086666).

gnomAD_oe_lof_upper: observed/expected upper bound loss of function from the genome aggregation database (gnomAD)

gnomAD_oe_mis_upper: observed/expected upper bound missense from the genome aggregation database (gnomAD)

gnomAD_pLI: probability that a gene falls into the class of intollerant of a single LoF gene (LoF-haploinsufficient intolerant genes), from the genome aggregation database (gnomAD)

gnomAD_pRec: probability that a gene falls into the class of intolerant of two LoF genes (recessive genes), from the genome aggregation database (gnomAD)

repeatMasker_percOverlap: percent overlap with repeat regions (RepeatMasker annotation from UCSC).

dirtyRegion_percOverlap: percent overlap with gaps (including centromeres and telomeres), and segmental duplications.

chrRegion: telomere/centromere tag.

CGD: (array of) entrez gene ID, gene symbol, disease name(s), and inheritance found in the Clinical Genomics Database; it 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 (AD, AR, AD/AR, XL, more complex modes); since the CGD mode of inheritance is directly added by a curator and is tied to specific genetic disorder(s), it could be considered more accurate than the mode of inheritance for top-level HPO phenotypes.

OMIM_MorbidMap: (array of) entrez gene ID, gene symbol, and disorder/disease name(s) found in OMIM.

ISCA_region: Genomic disease region from ClinGen (ISCA).

decipher_region: Genomic disease region from Decipher; the DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources is an interactive web-based database which incorporates a suite of tools designed to aid the interpretation of genomic variants. For more information, see: https://decipher.sanger.ac.uk/.

gnomAD_commonSV: % overlap with the common (greater than 1%) features in the structural variants genome aggregation database (gnomAD)

gnomAD_rareSV: % overlap with the rare (less than or equal to 1%) features in the structural variants genome aggregation database (gnomAD)

DGV_50percRecipOverlap: % length covered by merged variants in DGV, restricted to those with at least 50% reciprocal overlap.

DGV_commonPerc: % overlap with CNVs with a frequency higher than 1% in DGV.

pacBioPercFreq_90percRecipOverlap: frequency based on internal database – unrelated 18 samples called by pbsv 2.6.2, with at least 90% reciprocal overlap, matched by variant type; the type is ignored when complex.