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_egID: entrez gene IDs.

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.

exon_egID: official gene entrez gene IDs, exons overlap only.

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

cds_egID: official gene entrez IDs; 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: observed/expected ratio for loss of function from the genome aggregation database (gnomAD)

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

gnomAD_oe_mis: observed/expected missense ratio 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.

MPO_NervousSystem: (array of) entrez gene ID, gene symbol, and inheritance for genes in the MPO terms part of nervous system phenotype (for MPO description, see MPO_Other).

MPO_Growth: (array of) entrez gene ID, gene symbol, and inheritance for genes in the MPO terms part of growth phenotype (for MPO description, see MPO_Other).

MPO_Other: (array of) entrez gene ID, gene symbol, MPO term, and inheritance for genes in the MPO (Mammalian Phenotype Ontology) top level phenotype(s) excluding terms in nervous system or growth phenotypes, 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 confirmed on the MGI website.

HPO_NervousSystem: (array of) entrez gene ID, gene symbol and inheritance for genes in the HPO terms part of nervous system phenotype (for HPO description, see HPO_Other).

HPO_Growth: (array of) entrez gene ID, gene symbol and inheritance for genes in the HPO terms part of growth system phenotype (for HPO description, see HPO_Other).

HPO_Other: (array of) entrez gene ID, gene symbol, MPO term, and inheritance for genes in the HPO (Human Phenotype Ontology) top level phenotype(s), imported from HPO, excluding nervous system and growth phenotypes. The genotype-phenotype association is typically supported by an OMIM entry.

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.

ASDgenes: list of genes involved in ASD/DD, with their evidence source.

ISCA_region: Genomic disease region from ClinGen (ISCA).

CNV_ISCA_percOverlap: % length of CNV/SV overlapped by ClinGen (ISCA) region(s).

ISCA_CNV_percOverlap: (array of) % length of ClinGen (ISCA) region(s) overlapped by CNV/SV.

ISCA_CNV_percOverlap_max: highest % length of ClinGen (ISCA) region overlapped by CNV/SV.

ISCA_matchCNVmax_percOverlap: % length of CNV/SV overlapped by largest ClinGen (ISCA) region.

exon_symbol_ISCA: gene symbol(s) in the region covered by 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/.

CNV_decipher_percOverlap: % length of CNV/SV overlapped by Decipher region(s).

decipher_CNV_percOverlap: % length of Decipher region(s) overlapped by CNV.

decipher_CNV_percOverlap_max: highest % length of Decipher region(s) overlapped by CNV.

decipher_matchCNVmax_percOverlap: % length of CNV/SV overlapped by largest Decipher region.

exon_symbols_Decipher: gene symbol(s) in the region covered by Decipher.

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_N_studies_50percRecipOverlap: Number of studies in DGV where at least one subject in the study has a variant overlapping the CNV, restricted to 50% reciprocal overlap.

DGV_N_subjects_50percRecipOverlap: Number of subjects in DGV where the variant overlaps the CNV, restricted to 50% reciprocal overlap.

DGVpercFreq_subjects_allStudies_50percRecipOverlap: % frequency in DGV with at least 50% reciprocal overlap; all studies combined.

DGVpercFreq_subjects_coverageStudies_50percRecipOverlap: % frequency in DGV with at least 50% reciprocal overlap; only studies where at least one of the subjects had coverage.

DGV_percOverlap_any: (array of) % length of DGV region(s) overlapped by CNV (no cutoff used). The Database of Genomic Variants provides a comprehensive summary of structural variation in the human genome. For more information: http://dgv.tcag.ca/dgv/app/about?ref=GRCh37/hg19. The DVG was lifted over to obtain the corresponding intervals in the GRCh38 reference genome.

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.

CGparentalPercFreq_90percRecipOverlap: frequency based on internal database - parents sequenced by Complete Genomics, with at least 90% reciprocal overlap, matched by variant type; the type is ignored when complex.

otgMantaPercFreq_90percRecipOverlap: frequency based on the 1000G+ collection – parents sequenced by NovaSeq6000, called by Manta, with at least 90% reciprocal overlap, matched by variant type; the type is ignored when complex.

otgDellyPercFreq_90percRecipOverlap: frequency based on the 1000G+ collection – parents sequenced by NovaSeq6000, called by Delly, with at least 90% reciprocal overlap, matched by variant type; the type is ignored when complex.

svMantaXPercFreq_90percRecipOverlap: frequency based on internal database - parents sequenced by Illumina HiSeqX called by Manta, with at least 90% reciprocal overlap, matched by variant type; the type is ignored when complex.

svManta2PercFreq_90percRecipOverlap: frequency based on internal database - parents sequenced by Illumina HiSeq2000/2500 called by Manta, with at least 90% reciprocal overlap, matched by variant type; the type is ignored when complex.

svDellyXPercFreq_90percRecipOverlap: frequency based on internal database - parents sequenced by Illumina HiSeqX called by Delly, with at least 90% reciprocal overlap, matched by variant type; the type is ignored when complex.

svDelly2PercFreq_90percRecipOverlap: frequency based on internal database - parents sequenced by Illumina HiSeq2000/2500 called by Delly, with at least 90% reciprocal overlap, matched by variant type; the type is ignored when complex.

hsDragenPercFreq_90percRecipOverlap: frequency based on internal database – unrelated samples sequenced by Illumina NovaSeq 6000 with average coverage of 35x, called by Dragen 3.8.4, with at least 90% reciprocal overlap, matched by variant type; the type is ignored when complex.

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.

nearestLeftExonBoundary: gene with exon/intron junction closest to the left CNV/SV boundary.

nearestLeftExonDistance: distance of closest exon/intron junction to the left boundary; negative values indicate that the CNV/SV is upstream of the junction.

nearestRightExonBoundary: gene with exon/intron junction closest to the right CNV/SV boundary.

nearestRightExonDistance: distance of the closest exon/intron junction to the right boundary; negative values indicate that the CNV/SV is upstream of the junction.