AnnotSV Manual

Version 2.0
AnnotSV is a program for annotating and ranking structural variations from the human genome.

https://lbgi.fr/AnnotSV/

Please feel free to contact me for any suggestions or bug reports
email: veronique.geoffroy@inserm.fr

LEXIQUE

1000g: 1000 Genomes Project (phase 3)
ACMG: American College of Medical Genetics and Genomics
BED: Browser Extensible Data
bp: base pair
CDS: CoDing Sequence
CNV: Copy Number Variation
DDD: Deciphering Developmental Disorders
DECIPHER: DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources
DEL: Deletion
DGV: Database of Genomic Variants
DNA: DesoxyriboNucleic Acid
DUP: Duplication
ENCODE: Encyclopedia of DNA Elements
ExAC: Exome Aggregation Consortium
GRCh37: Genome Reference Consortium Human Build 37
GRCh38: Genome Reference Consortium Human Build 38
HI: Haploinsufficiency
hom: homozygous
htz: heterozygous
ID: Identifier
indel: Insertion/deletion
LoF: Loss of Function
misZ = Z score indicating gene intolerance to missense variation
NAHR: Non-Allelic Homologous Recombination
NM: RefSeq identifiers
OMIM: Online Mendelian Inheritance in Man
pLI = score computed by the ExAc consortium to indicate gene intolerance to a loss of function variation
SNV : Single Nucleotide Variation
SV: Structural Variations
synZ = Z score indicating gene intolerance to synonymous variation
TAD: Topologically Associating Domains
Tcl: Tool Command Language
TriS: Triplosensitivity
Tx: transcript
VCF: Variant Call Format

TABLE OF CONTENTS

1. INTRODUCTION

AnnotSV is a program designed for annotating and ranking Structural Variations (SV). This tool compiles functionally, regulatory and clinically relevant information and aims at providing annotations useful to i) **interpret SV potential pathogenicity** and ii) **filter out SV potential false positives**.

Different types of SV exist including deletions, duplications, insertions, inversions, translocations or more complex rearrangements. They can be either balanced or unbalanced. When unbalanced and resulting in a gain or loss of material, they are called Copy Number Variations (CNV). CNV can be described by coordinates on one chromosome, with the start and end positions of the SV (deletions, insertions, duplications). Complex rearrangements with several breakends can arbitrary be summarized as a set of novel adjacencies, as described in the Variant Call Format specification VCFv4.3 (Jul 2017).

AnnotSV takes as an input file a classical BED or VCF file describing the SV coordinates. The output file contains the overlaps of the SV with relevant genomic features where the genes refer to NCBI RefSeq genes. AnnotSV provides numerous additional relevant annotations:

- A genes-based annotation (OMIM, Gene intolerance, Haploinsufficiency…)
- An annotation with features overlapping the SV (DGV, 1000genomes…)
- An annotation with features overlapped with the SV (pathogenic SV from dbVar, promoters, TAD…)
- An annotation of the SV breakpoints (GC content, repeats…)

In addition to these annotations, AnnotSV also provide a systematic SV classification using the same type of categories delineated by the American College of Medical Genetics and Genomics (ACMG (; on behalf of the ACMG Laboratory Quality Assurance Committee et al., 2015)):

Class 1 = benign
Class 2 = likely benign
Class 3 = VOUS (variant of unknown significance)
Class 4 = likely pathogenic
Class 5 = pathogenic

2. INSTALLATION/REQUIREMENTS

a. Tcl (required)

The AnnotSV program is written in the Tcl language. Modern Unix systems have this scripting language already installed (otherwise it can be downloaded from http://www.tcl.tk/).

AnnotSV requires **the latest release of the Tcl distribution starting with version 8.6** as well as the following 2 packages "tar" and "csv" (used only when data sources are updated).

b. AnnotSV source code (required)

"AnnotSV sources" can be download at http://lbgi.fr/AnnotSV/downloads (under the GNU GPL license).

Install:

The sources .tar.gz should be extracted and uncompressed to any directory.
tar -xvf AnnotSV_latest.tar.gz

The installation requires simply to set the following environment variable:

$ANNOTSV : “AnnotSV installation directory”

and to save the settings in your .cshrc or .bashrc file.

- In csh, you can define it with the following command line:
  setenv ANNOTSV /path_of_AnnotSV_installation/bin
- In bash, you can define it with the following command line:
  export ANNOTSV=/path_of_AnnotSV_installation/bin

Make sure the program correctly finds the Tcl interpreter. By default, the best way to make a Tcl script executable is to put the following as the first line of the main script (which is already done in AnnotSV-main.tcl):
#!/usr/bin/env tclsh

It can be changed to any other path like:
#!/usr/local/ActiveTcl/bin tclsh

Typically, you can create an alias of the main Tcl script “sources/AnnotSV-main.tcl” for example to “AnnotSV”, place it in the “/bin” directory”(this is done by default already) and add the path to this in your $PATH.

AnnotSV installation directory:

By default, the AnnotSV installation directory looks like this:

```
AnnotSV                          #the program installation directory
 |
 |----- Annotations/             #where annotation files are stored (RefGene, OMIM, DGV…)
 |
 |----- bin/                     #where an alias is set to the main .tcl script
 |
 |----- changeLogs.txt           #description of AnnotSV changes
 |
 |----- configfile               #a configfile example that can be edited for modification purpose
 |
 |----- Example/                 #command/input/output example
 |
 |----- License.txt              #GNU GPL license
 |
 |----- README.AnnotSV_*.pdf     #this file
 |
 |----- Sources/                 #where the source .tcl files are stored
```

c. bedtools (required)

The **“bedtools”** toolset (developed by Quinlan AR) needs to be locally installed. Configuration requires to set the path to the bedtools executable in the AnnotSV configfile located in: $ANNOTSV/configfile.

Warning: the minimum version of bedtools compatible with AnnotSV is version 2.25.

3. ANNOTATION SOURCES (provided)

AnnotSV requires different data sources for the annotation of SV. **In order to provide a ready to start installation of AnnotSV, each annotation source listed below (that do not require a commercial license) is already provided with the AnnotSV sources.** The aim and update of each of these sources are explained below. Annotation can be performed using either the GRCh37 or GRCh38 build version of the human genome (user defined, see USAGE/OPTIONS), but depending on the availability of some data sources there might be some limitations. Some of the annotations are linked to the gene name and thus provided independently of the genome build.

a. Genes-based annotations

Gene annotations

The "Gene annotation" aims at providing information for the overlapping known genes with the SV in order to list the genes from the well annotated RefSeq database. These annotations include the definition of the genes and corresponding transcripts (RefSeq), the length of the CoDing Sequence (CDS) and of the transcript, the location of the SV in the gene (e.g. « txStart-exon3 ») and the coordinates of the intersection between the SV and the transcript.

Annotation columns:
Adds 7 annotation columns: "Gene name", "NM", "CDS length", "tx length", "location", "intersectStart", "intersectEnd".

Method:
For each gene, only a single transcript from all transcripts available in RefSeq for this gene is reported. In case of transcripts with different CDS length (considering the overlapping region with the SV), the transcript with the longest CDS is reported. Otherwise, if there is no differences in CDS length, the longest transcript is reported.

Updating the data source (if needed):

- Remove all the files in the "$ANNOTSV/Annotations/RefGene/GRCh37" and/or "$ANNOTSV/Annotations/RefGene/GRCh38" directories.
- Download and place the "refGene.txt.gz" file in the "$ANNOTSV/Annotations/RefGene/GRCh37" and/or "$ANNOTSV/Annotations/RefGene/GRCh38" directories. The latest update of this file is available for free download at:
  Genome build GRCh37:
  http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/refGene.txt.gz
  Genome build GRCh38:
  http://hgdownload.cse.ucsc.edu/goldenPath/hg38/database/refGene.txt.gz

After the update, this refGene.txt.gz file will be processed by AnnotSV during the first run (it will take longer than usual AnnotSV runtime).

It is to notice that the **promoter's annotations update** will be done at the same time (without supplementary update command).

DDD gene annotations

Aim:

The Deciphering Developmental Disorders (DDD) Study (Firth, et al., 2011) has recruited nearly 14,000 children with severe undiagnosed developmental disorders, and their parents from around the UK and Ireland. The patients have been deeply phenotyped by their referring clinician via DECIPHER using the Human Phenotype Ontology. The DNA from these children have been explored using high-resolution exon-arrayCGH and exome sequencing (trio) to investigate the genetic causes of their abnormal development. These annotations give additional information on each gene overlapped by a SV (independently of the genome build version).

Annotation columns:

Adds 5 annotation columns (only in the "split" lines): “DDD_status”, “DDD_mode”, “DDD_consequence”, “DDD_disease”, “DDD_pmids”.

Updating the data source (if needed):

- Remove all the **DDG2P** files in the “$ANNOTSV/Annotations/Genes-based/DDD” directory.
- Download and place the “**DDG2P.csv.gz**” DECIPHER file in the “$ANNOTSV/Annotations/Genes-based/DDD” directory. The latest update of this file is available for free download at: http://www.ebi.ac.uk/gene2phenotype/downloads/DDG2P.csv.gz

This file will be computed the first time AnnotSV will be executed after the update.

Warning: This update requires the “csv” Tcl package.

OMIM annotations

Aim:

OMIM (Online Mendelian Inheritance in Man) (Hamosh, et al., 2000) focuses on the relationship between phenotype and genotype. These annotations give additional information on each gene overlapped by a SV (independently of the genome build version). Moreover, a morbid genes list is provided.

Annotation columns:

Add 2 annotation columns: “morbidGenes” and "morbidGenesCandidates".
Add 3 other annotation columns (only in the "split" lines): “Mim Number”, “Phenotypes” and “Inheritance”.

Update:

- Remove all the files in the “$ANNOTSV/Annotations/Genes-based/OMIM” directory.
- Download and place the “**genemap2.txt**” and “**morbidmap.txt**” OMIM files in the “$ANNOTSV/Annotations/Genes-based/OMIM” directory.
  The latest updates of these files are available for download following a registration and review process (https://omim.org/downloads/). “**genemap2.txt**” is a tab-delimited file containing OMIM's synopsis of the Human gene map including additional information such as genomic coordinates and inheritance. “**morbidmap.txt**” is a tab-delimited file of OMIM's Synopsis of the Human Gene Map (same as genemap.txt above) sorted alphabetically by disorder

Method:
The “morbidGenes” and "morbidGenesCandidates" are described in the “Disorder” column of the Gene Map file as follows:

- morbidGenes: the number in parentheses after the name of each disorder is set to (3) or (4):

(3) indicates that the molecular basis of the disorder is known; a mutation has been found in the gene.
(4) indicates that a contiguous gene deletion or duplication syndrome, multiple genes are deleted or duplicated causing the phenotype.

- morbidGenesCandidates: the symbol in front of the name of each disorder is set to "{ }" or ?:

"{ }", indicates mutations that contribute to susceptibility to multifactorial disorders (e.g., diabetes) or to susceptibility to infection (e.g., malaria).
"?", before the phenotype name indicates that the relationship between the phenotype and gene is provisional.

ACMG annotations

Aim:
The American College of Medical Genetics and Genomics has published recommendations for reporting incidental or secondary findings in genes with a medical benefit(; on behalf of the ACMG Laboratory Quality Assurance Committee et al., 2015). The most recent version of the recommendations is the ACMG SF v2.0 including 59 genes.

Annotation columns:
Add 1 annotation column (only in the "split" lines): “ACMG”.

Gene intolerance annotations (ExAC)

Aim:
Gene intolerance annotations from the ExAC (Lek, et al., 2016) give the significance deviation from the observed and the expected number of variants for each gene:

Column name	Constraint from ExAC	Score	Indication
synZ_ExAC	Synonymous	Z score	Positive Z scores indicate gene intolerance to synonymous variation.
misZ_ExAC	Missense	Z score	Positive Z scores indicate gene intolerance to missense variation.
pLI_ExAC	LoF (Nonsense, splice acceptor, and splice donor variants caused by SNV)	Score computed by the ExAC consortium	pLI indicates the probability that a gene is intolerant to a loss of function mutation. ExAC consider pLI >= 0.9 as an extremely LoF intolerant set of genes.
delZ_ExAC	Deletion	Z score	Higher positive values indicate greater intolerance (a lower than expected rate of CNVs for that gene).
dupZ_ExAC	Duplication	Z score	
cnvZ_ExAC	CNV	Z score	

These annotations give additional information on each gene overlapped by a SV (independently of the genome build version).

Annotation columns:

Adds 6 annotation columns: “synZ_ExAC”, “misZ_ExAC”, “pLI_ExAC”, "delZ_ExAC", "dupZ_ExAC" and “cnvZ_ExAC.

Updating the data source (if needed):

- Remove all the files in the “$ANNOTSV/Annotations/Genes-based/ExAC” directory.
- Download and place the “**fordist_cleaned_nonpsych_z_pli_rec_null_data.txt**” ExAC file in the “$ANNOTSV/Annotations/Genes-based/ExAC” directory. The latest update of this file is available for free download at:
  ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3.1/functional_gene_constraint/fordist_cleaned_nonpsych_z_pli_rec_null_data.txt

This file will be reprocessed the first time AnnotSV will be executed after the update.

Haploinsufficiency annotations (DDD)

Aim:

Haploinsufficiency, wherein a single functional copy of a gene is insufficient to maintain normal function, is a major cause of dominant disease. As detailed in DECIPHER, over 17,000 protein coding genes have been scored according to their predicted probability of exhibiting haploinsufficiency:

- High ranks (e.g. 0-10%) indicate a gene is more likely to exhibit haploinsufficiency
- Low ranks (e.g. 90-100%) indicate a gene is more likely to NOT exhibit haploinsufficiency.

This annotation give additional information on each gene overlapped by a SV (independently of the genome build version).

Annotation columns:

Add 1 annotation column: “HI_DDDpercent”.

Update:

- Remove all the files in the “$ANNOTSV/Annotations/Genes-based/DDD” directory.
- Download and place the “**HI_Predictions_Version3.bed.gz**” DECIPHER file in the “$ANNOTSV/Annotations/Genes-based/DDD” directory. The latest update of this file is available for free download at:
  https://decipher.sanger.ac.uk/about#downloads/data

This file will be computed the first time AnnotSV will be executed after the update.

Haploinsufficiency and triplosensitivity Scores annotations (ClinGen)

Aim:

The ClinGen Consortium Rating System is curating genes and regions of the genome to assess whether there is evidence to support that these genes/regions are dosage sensitive. Haploinsufficiency and triplosensitivity scorings are ranged as follow:

Score	Possible Clinical Interpretation
3	Sufficient evidence for dosage pathogenicity
2	Some evidence for dosage pathogenicity
1	Little evidence for dosage pathogenicity
0	No evidence for dosage pathogenicity
40	Evidence suggests the gene is not dosage sensitive
30	Gene associated with autosomal recessive phenotype

Annotation columns:
Add 2 annotation columns: "HI_CGscore" and "TriS_CGscore".
Concerning annotations on the **"full"** length of SV covering several genes, only the most pathogenic score is reported if any.

Update:

- Remove all the files in the "$ANNOTSV/Annotations/Genes-based/ClinGen" directory.
- Download and place the "**ClinGen_gene_curation_list_GRCh37.tsv**" ClinGen file in the "$ANNOTSV/Annotations/Genes-based/ClinGen/" directory. The latest update of this file is available for free download at:
  ftp://ftp.ncbi.nlm.nih.gov/pub/dbVar/clingen/ClinGen_gene_curation_list_GRCh37.tsv

This file will be computed the first time AnnotSV will be executed after the update. The annotations selected by AnnotSV are genome build independent, and only based on the gene name.

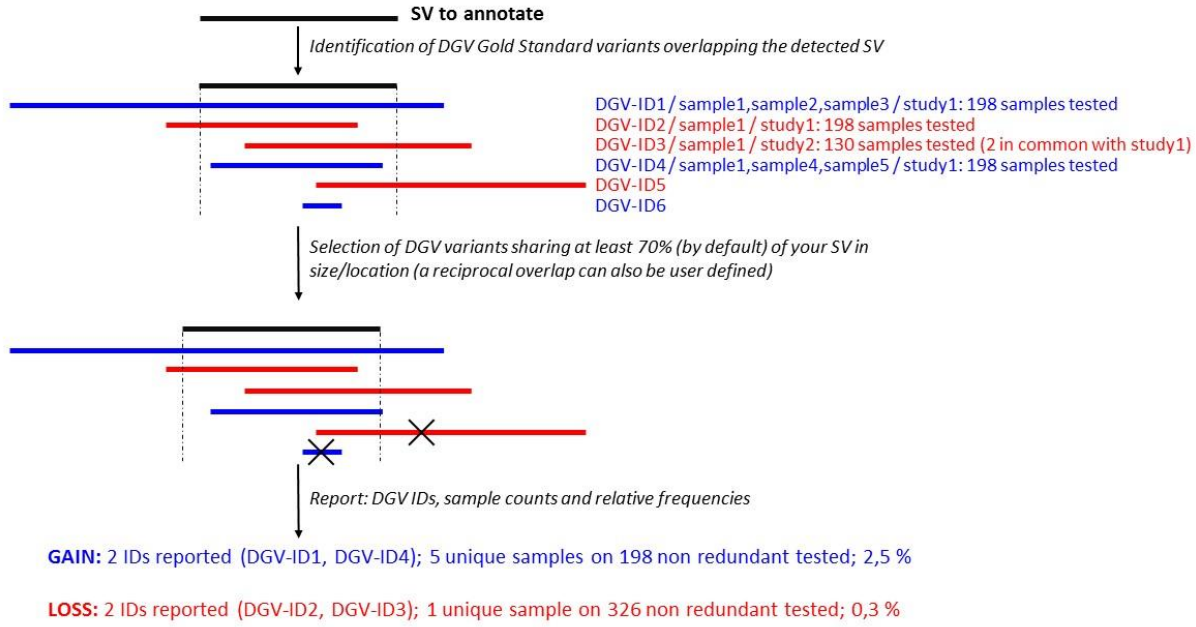
b. Annotations with features overlapping the SV

DGV Gold Standard annotations

Aim:
The Database of Genomic Variants (DGV) (MacDonald, et al., 2014) provides SV defined as DNA elements with a size >50 bp. The content of DGV is only representing SV identified in healthy control samples from large cohorts published and integrated by the DGV team. The annotations will give information about whether your SV is a rare or a benign common variant.

Annotation columns:
Adds 8 annotation columns: respectively for GAIN and LOSS: "DGV_IDs", "n_samples_with_SV", "n_samples_tested" and "Frequency".

Method:
First, AnnotSV searches for DGV Gold Standard variants overlapping the SV to annotate. Second, only the DGV variants overlapping at least 70% of your SV in size/location are selected (default value, a different percentage or a reciprocal overlap can also be user defined with the "overlap" and "reciprocal" options). Third, the DGV IDs are reported. Then, all DGV samples information are merged: the counts of unique samples with gains and losses, the number of samples tested in the related studies (without redundancy) and subsequent relative frequencies are calculated and reported (genotype data are not considered).



Warning:

- **Exceptional overestimation of the relative frequencies**, can be observed in DGV Gold Standard (March 2016). ~10% of the supporting variants are not released with sample information preventing AnnotSV to properly differentiate whether some variation are redundant or not. Consequently, some relative frequencies can be exceptionally overestimated by AnnotSV.

- **The Gain/Loss status can be different for a same event.** A SV call in DGV can be relative to a specific reference sample, a pool of reference samples or relative to the reference assembly. Since different reference samples may have been used in different studies, what is called as a gain in one study may actually be called a loss in another.

Updating the data source (if needed):

- Remove all the files in the "$ANNOTSV/Annotations/SVincludedInFt/DGV/GRCh37" and/or "$ANNOTSV/Annotations/SVincludedInFt/DGV/GRCh38" directories.
- Download and place the 2 following DGV files in the "$ANNOTSV/Annotations/SVincludedInFt/DGV/GRCh37" and/or "$ANNOTSV/Annotations/SVincludedInFt/DGV/GRCh38" directories.

  Genome build GRCh37:
  The latest update of these 2 files are available for free download at http://dgv.tcag.ca/dgv/app/downloads
  - **DGV.GS.March2016.50percent.GainLossSep.Final.hg19.gff3** (see DGV Gold Standard Variants section)
  - **GRCh37_hg19_supportingvariants_2016-05-15.txt** (see Supporting Variants section)

  Genome build GRCh38:
  The dataset is not yet available from the DGV team.

To give access to the ranking of SV with GRCh38 coordinates, the GRCh37 DGV GS dataset has been lift over to GRCh38 with the UCSC web server and is provided by AnnotSV.

These 2 files will be computed the first time AnnotSV will be executed after the update.

DDD frequency annotations

Aim:
AnnotSV takes advantage of the DDD study (national blood service controls + generation Scotland controls), representing the 845 samples currently available (an update is planned in the near future).

Method:
By default, a DDD CNV is reported only if it overlaps at least 70% of the SV to annotate. Nevertheless, the user can modify the default behaviour by either use a different percentage or a reciprocal overlap (see "overlap" and "reciprocal" options in USAGE/OPTIONS).

Annotation columns:
Adds 5 annotation columns: “DDD_SV”, “DDD_DUP_n_samples_with_SV”, “DDD_DUP_Frequency”, “DDD_DEL_n_samples_with_SV”, “DDD_DEL_Frequency”.
Concerning the four last annotations, only 1 value is reported (the biggest one) n the **“full”** length lines.

Updating the data source (if needed):

- Remove all the files in the “$ANNOTSV/Annotations/SVincludedInFt/DDD/GRCh37” directory.
- Download and place the “**population_cnv.txt.gz**” DECIPHER files in the “$ANNOTSV/Annotations/SVincludedInFt/DDD/GRCh37” directory.
  Genome build GRCh37:
  The latest update of this file is available for free download at:
  https://decipher.sanger.ac.uk/files/downloads/population_cnv.txt.gz

  Genome build GRCh38:
  The dataset is not yet available from the DDD team.

This file will be computed the first time AnnotSV will be executed after the update.

1000 genomes frequency annotations

Aim:
The goal of the 1000 Genomes Project (Sudmant, et al., 2015) was to find most genetic variants with frequencies of at least 1% in the populations studied. Analyses were conducted looking at both the short variations (up to 50 base pairs in length) and the SV. These annotations give additional information on the SV allele frequencies from the 1000 genomes database overlapped by a SV to annotate.

Method:
By default, a 1000g SV is reported only if it overlaps at least 70% of the SV to annotate. Nevertheless, the user can modify the default behaviour by either use a different percentage or a reciprocal overlap (see "overlap" and "reciprocal" options in USAGE/OPTIONS).

Annotation columns:
Adds 3 annotation columns: “1000g_event”, “1000g_AF” and “1000g_max_AF”.

Concerning the frequencies, only 1 value is reported (the most frequent one) in the **"full"** length lines.

Updating the data source (if needed):

- Remove all the **1000g** files in the "$ANNOTSV/Annotations/SVincludedInFt/1000g/GRCh37" and/or "$ANNOTSV/Annotations/SVincludedInFt/1000g/GRCh38" directories.
- Download and place the VCF files in the "$ANNOTSV/Annotations/SVincludedInFt/1000g/GRCh37" and/or "$ANNOTSV/Annotations/SVincludedInFt/1000g/GRCh38" directories. The latest updates of these files are available for free download at:
  Genome build GRCh37:
  ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/ALL.wgs.mergedSV.v8.20130502.svs.genotypes.vcf.gz
  Genome build GRCh38:
  http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/supporting/GRCh38_positions/ALL.wgs.mergedSV.v8.20130502.svs.genotypes.GRCh38.vcf.gz

This file will be computed the first time AnnotSV will be executed after the update.

c. Annotations with features overlapped with the SV

Promoter annotations

Aim:
The contribution of SV affecting promoters to disease etiology is well established. Affecting possibly gene expression, understanding the consequences of these regulatory variants on the human transcriptome remains a major challenge. AnnotSV reports the list of the genes whose promoters are overlapped by the SV.

Annotation columns:
Adds 1 annotation column: "promoters"

Method:
Promoters are defined by default as 500 bp upstream from the transcription start sites (using the RefGene data). Nevertheless, the user can define a different bp size with the "promoterSize" option (see USAGE/OPTIONS). A promoter is reported only if the SV overlaps at least 70% of this promoter (user defined, see the "overlap" option in USAGE/OPTIONS).

Update:
The promoters' annotations update will be done at the same time as the Gene annotations update.

dbVar NR SV pathogenic annotations

Aim:
dbVar is the NCBI's database of genomic structural variation collecting insertion/deletion/duplications/mobile elements insertions/translocations data from large initiative including also medically relevant variations. A non-redundant version of the database, dbVar non-redundant SV (NR SV) datasets include more than 2.2 million deletions, 1.1 million insertions, and 300,000 duplications. These data are aggregated from over 150 studies including 1000 Genomes Phase 3, Simons Genome Diversity Project, ClinGen, ExAC, and others. By selecting pathogenic SV records from the dbVar NR SV database, AnnotSV obtained a clinically-relevant human SV dataset.

Method:
By default, a pathogenic SV is reported only if the SV overlaps at least 70% of this pathogenic SV. Nevertheless, the user can modify the default behaviour by either use a different percentage or a reciprocal overlap (see "overlap" and "reciprocal" options in USAGE/OPTIONS).

Annotation columns:
Adds 3 annotation columns: “dbvar_event”, “dbVar_variant” and “dbVar_status”.

Updating the data source (if needed):

- Remove all the files in the “$ANNOTSV/Annotations/FtIncludedInSV/dbVar_pathogenic_NR_SV/GRCh37” and/or “$ANNOTSV/Annotations/FtIncludedInSV/dbVar_pathogenic_NR_SV/GRCh38” directories.
- Download and place the 2 following dbVar files in the “$ANNOTSV/Annotations/FtIncludedInSV/dbVar/GRCh37” and/or “$ANNOTSV/Annotations/FtIncludedInSV/dbVar/GRCh38” directories.

  Genome build GRCh37:
  https://ftp.ncbi.nlm.nih.gov/pub/dbVar/sandbox/sv_datasets/nonredundant/deletions/GRCh37.nr_deletions.tsv.gz
  https://ftp.ncbi.nlm.nih.gov/pub/dbVar/sandbox/sv_datasets/nonredundant/duplications/GRCh37.nr_duplications.tsv.gz

  Genome build GRCh38:
  https://ftp.ncbi.nlm.nih.gov/pub/dbVar/sandbox/sv_datasets/nonredundant/deletions/GRCh38.nr_deletions.tsv.gz
  https://ftp.ncbi.nlm.nih.gov/pub/dbVar/sandbox/sv_datasets/nonredundant/duplications/GRCh38.nr_duplications.tsv.gz

These 2 files will be computed then removed the first time AnnotSV will be executed after the update.

TAD boundaries annotations

Aim:
The spatial organization of the human genome helps to accommodate the DNA in the nucleus of a cell and plays an important role in the control of the gene expression. In this non-random organization, topologically associating domains (TAD) emerge as a fundamental structural unit able to separate domains and define boundaries. Disruption of these structures especially by SV can result in gene misexpression (Lupianez, et al., 2016).

Method:
A TAD boundary is reported only if the SV overlaps at least 70% of this TAD boundary (user defined, see the "overlap" option in USAGE/OPTIONS).

Annotation columns:
Adds 2 annotation columns (“TADcoordinates”, “ENCODEexperiments”), containing i) the overlapping TAD coordinates with a SV and ii) the ENCODE experiments from which the TAD have been defined.

Very large SV (e.g. 30Mb) can sometime overlap too many TAD locations (e.g. more than 2600). It appears that depending on the visualisation program used (spreadsheet programs mostly) this annotation can be truncated.

In order to avoid such embarrassing glitch and maybe also because overlapping so many TAD is already a problem, AnnotSV restrict the number of overlapping reported TAD to 20 (including their associated ENCODE experiments).

Updating the data source (if needed):
AnnotSV needs ENCODE experiments in BED format for the TAD annotations.

- Remove all the files in the "$ANNOTSV/Annotations/FtIncludedInSV/TAD/GRCh37" and/or "$ANNOTSV/Annotations/TAD/GRCh38" directories.
- Download and place your ENCODE BED files in the "$ANNOTSV/Annotations/FtIncludedInSV/TAD/GRCh37" and/or "$ANNOTSV/Annotations/FtIncludedInSV/TAD/GRCh38" directories.
  These files (GRCh37 and GRCh38) are available for free download at:
  https://www.encodeproject.org/search/?type=Experiment&assay_title=Hi-C&files.file_type=bed+bed3%2B
  Click the "bed bed3+" button on your link (else the "file.txt" is blank). Then, click the "Download" button to download a "files.txt" file that contains a list of URLs. Keep only the *.bed URLs in your "files.txt". Then use the following command to download all the BED files in the list:
  xargs -n 1 curl -O -L < files.txt
  Finally, dispatch the downloaded files in either the GRCh37 or the GRCh38 directory.

These BED files will be reprocessed during the first time AnnotSV will be executed.

d. Breakpoints annotations

GC content annotations

Aim:
GC content (as well as repeated sequences, DNA sequence identity and concentration of the PRDM9 homologous recombination hotspot motif 5'-CCNCCNTNNCCNC-3') is positively correlated with the frequency of nonallelic homologous recombination (NAHR). Indeed, NAHR hot spots have a significantly higher GC content (Dittwald, et al., 2013). This information with others could help identifying a novel locus for recurrent NAHR-mediated SV.

Method:
The GC content is calculated using bedtools around each SV breakpoint (+/- 100bp) then reported.

Annotation columns:
Adds 2 annotation columns: "GCcontent_left", "GCcontent_right"

Updating the data source (if needed):
AnnotSV needs the human reference genome FASTA file to run the "bedtools nuc" command.

- Remove all the files in the "$ANNOTSV/Annotations/BreakpointsAnnotations/GCcontent/GRCh37" and/or "$ANNOTSV/Annotations/BreakpointsAnnotations/GCcontent/GRCh38" directories.
- Download and place the human reference genome FASTA file in the "$ANNOTSV/Annotations/BreakpointsAnnotations/GCcontent/GRCh37" and/or "$ANNOTSV/Annotations/BreakpointsAnnotations/GCcontent/GRCh38" directories.

The latest update of this file is available for free download at:
Genome build GRCh37:
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/chromFa.tar.gz

Genome build GRCh38:
http://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/hg38.chromFa.tar.gz

This FASTA file will be reprocessed during the first time AnnotSV will be executed after the update.

Warning: This update requires the "tar" Tcl package.

Repeated sequences annotations

Aim:
Repeated sequences (as well as GC content, DNA sequence identity and presence of the PRDM9 homologous recombination hotspot motif 5'-CCNCCNTNNCCNC-3') play a major role in the formation of structural variants.

Method:
The overlapping repeats are identified using bedtools at the SV breakpoint (+/- 100bp) and reported (coordinates and type).

Annotation columns:
Adds 2 annotation columns: "Repeats_coord" and "Repeats_type"

Updating the data source (if needed):
AnnotSV needs a UCSC Repeat BED file.

- Remove all the files in the "$ANNOTSV/Annotations/BreakpointsAnnotations/Repeat/GRCh37" and/or "$ANNOTSV/Annotations/BreakpointsAnnotations/Repeat/GRCh38" directories.
- You can freely download the BED file from the "http://genome.ucsc.edu/cgi-bin/hgTables". There are many output options, here are the changes that you'll need to make:

  "GRCh37" or "GRCh38" assembly, "Repeats" group and "Repeatmasker" track. Select output format as BED. Choose the following output filename: Repeat.bed. Then, click the get output button.

- Download and place the BED file in the "$ANNOTSV/Annotations/BreakpointsAnnotations/Repeat/GRCh37" and/or "$ANNOTSV/Annotations/BreakpointsAnnotations/Repeat/GRCh38" directories.

This BED file will be reprocessed during the first time AnnotSV will be executed after the update.

4. Versions of the annotations sources

Annotations source	Version
Gene annotations (refGene)	2018-11-25
DDD gene annotations	2018-12-10
OMIM annotation	2018-11-13
ACMG	ACMG SF v2.0
Gene intolerance annotations (ExAC)	2016-01-14
Haploinsufficiency annotations (DDD)	2018-12-18
Haploinsufficiency and triplosensitivity Scores annotations (ClinGen)	2018-12-17
DGV Gold Standard annotations	2016-05-15
DDD frequency annotations	2018-12-10
1000 genomes frequency annotations (GRCh37)	2017-05-19
1000 genomes frequency annotations (GRCh38)	2017-11-05
dbVar NR SV pathogenic annotations	2018-08-02
TAD boundaries annotations	2017-10-24
GRCh37 FASTA genome	2009-03-20
GRCh38 FASTA genome	2014-01-23
Repeated sequences annotations	2018-12-10

5. SV RANKING/CLASSIFICATION

In order to assist the clinical interpretation of SV, AnnotSV provides on top of the annotations a systematic classification of each SV into one of the 5 classes proposed by the ACMG guidelines using the following data and criteria:

Data used for the ranking:

- Benign SV from the DGV Gold Standard corresponding to a gain (the ones with DGV_GAIN_Frequency >1% and with DGV_GAIN_n_samples_tested >500 (default, see the -minTotalNumber option in USAGE/OPTIONS))
- Benign SV from the DGV Gold Standard corresponding to a loss (the ones with DGV_LOSS_Frequency >1% and with DGV_LOSS_n_samples_tested >500 (default, see the -minTotalNumber option in USAGE/OPTIONS))
- Pathogenic SV from the dbVar NR-SV dataset
- pLI scores of each genes from ExAC
- Haploinsufficiency (HI) and triplosensitivity (TriS) scores from ClinGen
- Morbid genes from OMIM
- Candidate morbid genes from OMIM
- Candidate genes provided by the user (see the -candidateGenesFile option in USAGE/OPTIONS)

Criteria:

- **Class 1 (benign):**
  The SV overlaps (>70%) with a benign SV with the same SV type
  AND the SV does not contain a CDS from a morbid gene
  AND the SV does not contain a CDS from morbid gene candidate

AND the SV does not contain a candidate gene

- **Class 2 (likely benign):**
  The SV has no overlap OR an overlap ≤70% with a benign SV
  AND the SV does not contain a CDS from a morbid gene
  AND the SV does not contain a CDS from morbid gene candidate
  AND the SV does not contain a candidate gene

- **Class 3 (variant of unknown significance):**
  The SV overlaps a CDS from a morbid gene candidate (with at least 1bp overlap)
  OR the SV overlaps a CDS from a candidate gene (with at least 1bp overlap)

- **Class 4 (likely pathogenic):**
  The SV overlaps a morbid gene (with at least 1bp)
  OR for a loss: the SV overlaps a gene with a pLI_ExAC > 0.9 or with a HI_CGscore value of 3 or 2
  OR for a gain: the SV overlaps a gene TriS_CGscore value of 3 or 2

- **Class 5 (pathogenic):**
  The SV overlaps a pathogenic SV (with at least 1bp) with the same SV type

Warning:
In order to be able to classify the SV, AnnotSV requires that the type of SV is provided (duplication, deletion...) in the input SV file (BED or VCF).

Using a VCF containing SV as input file:
The INFO keys used for structural variants should follow at least the VCF version 4.2 specifications:
- The "SVTYPE" values should be one of DEL, INS, DUP, INV, CNV, BND, LINE1, SVA, ALU.
- The <CN0>, <CN2>, <CN3>... angle-bracketed ID from the "ALT" column should be used in case of SVTYPE=CNV in the INFO column.

Using a BED containing SV as input file:
The column number with the SV type information should be indicated (see the -svtBEDcol option). The "SVTYPE" values should be one of the following:
- Deletion: DEL, deletion, loss or <CN0>
- Insertion: INS or insertion
- Duplication: DUP, duplication, gain, <CN2>, <CN3>...
- Inversion: INV or inversion
- Breakend record: BND, breakpoint, breakend
-

6. INPUT

AnnotSV takes several arguments as input to the command line including options that are detailed in section 5 (“USAGE / OPTIONS”). The different arguments can be passed either on the command line or using a specific file named "configfile". The configfile file needs to be located in the AnnotSV installation directory. Four types of INPUT files are detailed below:

a. SV input file (required)

AnnotSV supports either the VCF (Variant Call Format) or the BED (Browser Extensible Data) formats as input files to describe the SV to annotate. It allows the program to be easily integrated into any bioinformatics pipeline dedicated to NGS analysis.

- **VCF format**:

It contains meta-information lines (prefixed with ”##”), a header line (prefixed with ”#”), and data lines each containing information about a position in the genome and genotype information on samples for each position (text fields separated by tabs). The specification are described at https://samtools.github.io/hts-specs/VCFv4.3.pdf. AnnotSV supports either native or gzipped VCF file.

By default, AnnotSV extracts and reports from the VCF input file the following informations:
- The REF, ALT, FORMAT and samples columns
- The SVTYPE value from the INFO column and only this one
- All other columns (QUAL, FILTER and INFO)

This report is user defined, see the "SVinputInfo" option in USAGE/OPTIONS.

Warning: AnnotSV will not report (and annotate) SV described with a non-official nomenclature.

- **BED format**.

Every single line of the BED file define a SV including the obligatory first 3 fields to describe its coordinates:

1. *chrom* - The name of the chromosome (e.g. 3, Y, …) - Preferred without “chr”.
2. *chromStart* - The starting position of the SV on the chromosome. According to the format, the base count starts at base ”0”.
3. *chromEnd* - The ending position of the SV on the chromosome. The *chromEnd* base is not included in the display of the feature. For example, the first 100 bases of a chromosome are defined as *chromStart*=0, *chromEnd*=100, and span the bases numbered 0-99.

Additional fields from the BED file are optional and can be reported in the AnnotSV output file (user defined). It can be used to store quality, read depth or other metrics produced by the SV caller. By default, AnnotSV reports the additional fields from the BED input file. This report is user defined, see the "SVinputInfo" option in USAGE/OPTIONS.

When the additional fields from the BED file are reported, the user can provide a BED of which the first line begins with a “#”, is tab separated and describe the columns header. The following example has been set to provide the SV coordinates associated to their SV type (DEL, DUP…) and score:

#Chrom	Start	End	SV type	Score
1	2806107	107058351	DEL	5.0256
12	25687536	25699754	DUP	1.3652

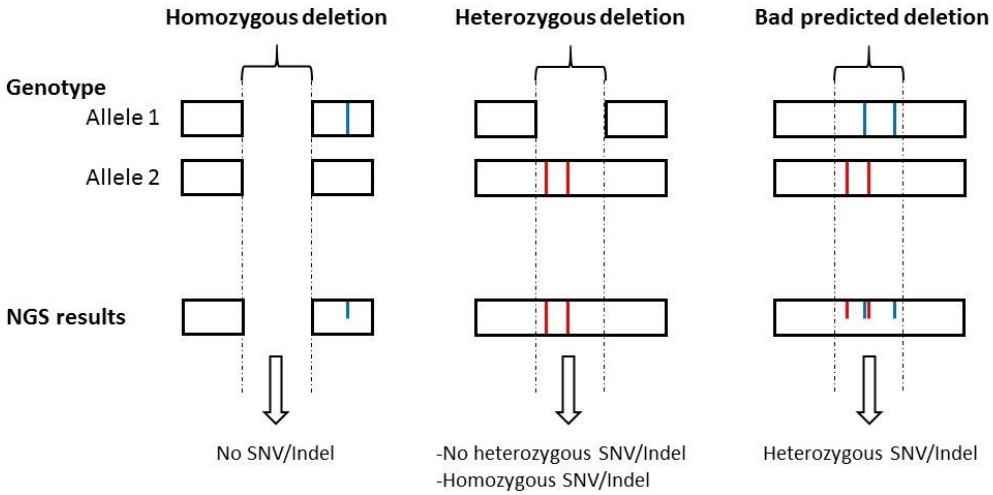
b. SNV/indel input files - for DELETION filtering (optional)

AnnotSV can take VCF file(s) with SNV/indel call from any sequencing experiment as input to the command line. These annotations report the counts of homozygous and heterozygous SNV/indel identified from the patients NGS data (user defined samples) and presents in the interval of the SV to annotate.

Annotation columns/Usage:

Add the “hom(sample)” and “htz(sample)” annotation columns.
The command line can be completed with the 2 following options: “-vcfFiles” and “-vcfSamples” (*cf* USAGE/OPTIONS).

Aim:
These annotations can be used by the user to filter out false positive SV calls or to confirm events as following:



-**Homozygous deletion** can be identified as a false positive by noting the presence of SNV/indel called at the predicted locus of the deletion in a sample.
-**Heterozygous deletion** can be identified as a false positive by noting the presence of heterozygous SNV/indel called at the predicted locus of the deletion in a sample. If no heterozygous SNV/indel are presents, the heterozygous deletion can be confirmed by reporting the presence of homozygous SNV/indel at that locus in the sample.

Warning:
In the VCF file(s), **the genotype of each variation should be indicated in the format field under the “GT” field**.

c. Filtered SNV/indel input files - for compound heterozygosity analysis (optional)

Aim:
AnnotSV can take a VCF file(s) with SNV/indel as input to the command line that is already filtered for genotype, frequency and effects on protein level. AnnotSV can report the heterozygous SNV/indel called (by any sequencing experiment) in the gene overlapped by the SV to annotate, as well in ‘healthy’ and ‘affected’ samples (user defined samples). AnnotSV offers an efficient way to highlight compound heterozygotes with one SNV/indel and one SV in the same gene. Indeed, in recessive genetic disorders, both copies of the gene are malfunctioning. This means that the maternally as well as the paternally inherited copy of an autosomal gene harbors a pathogenic variation. In addition, if the parents are non-consanguineous, compound heterozygosity is the best explanation for a recessive disease.

Usage:
To add the “**compound-htz**” annotation column**,** the command line can be completed with the 2 following options: “-filteredVCFfiles” and “-filteredVCFsamples” (cf USAGE/OPTIONS).

User challenge:

The user challenge in filtering variants for compound heterozygotes is to know whether the two heterozygous variants (the SNV/indel and the SV) are in *cis* or in *trans.* And when sequencing data of more than one family member is available, one can exclude certain variants based on rules of Mendelian inheritance (transmitted in a compound heterozygous mode from parents to the patient(s)).

Warning: In the VCF file(s), the genotype should be indicated in the format field as "GT".

d. External BED annotation files (optional)

Aim:
Several users might want to add their own private annotations to the one already provided by AnnotSV.

Inputs:
AnnotSV can integrate external annotations for specific regions that will be imported from a BED file into the output file.

Each external BED annotation file should be **copy or linked** in:

Genome build GRCh37:

➔ the "$ANNOTSV/Annotations/Users/GRCh37/FtIncludedInSV" directory

➔ the "$ANNOTSV/Annotations/Users/GRCh37/SVincludedInFt" directory

Genome build GRCh38:

➔ the "$ANNOTSV/Annotations/Users/GRCh38/FtIncludedInSV" directory

➔ the "$ANNOTSV/Annotations/Users/GRCh38/SVincludedInFt" directory

knowing that:

➔ FtIncludedInSV: Annotations will be done with features overlapped with the SV (>70% by default)

➔ SVincludedInFt: Annotations will be done with SV overlapped with the features (>70% by default)

Nevertheless, the user can modify the default behaviour by either use a different percentage or a reciprocal overlap (see "overlap" and "reciprocal" options in USAGE/OPTIONS).

Warning: After a formatting step, the copy and/or linked users file(s) will be deleted the first time AnnotSV will be executed after an update.

Header:
Each external BED annotation file (e.g. '*User'*.bed) can begin with a first line beginning with a "#" and describing the header of these new annotations.

The following example has been set to provide the SV overlap with Regions of Homozygosity (RoH) of 2 individuals (sample1 and sample2):

'User'.bed file contains:

#Chrom	Start	End	RoH
1	2806107	107058351	sample1, sample2
12	25687536	25699754	sample2

"RoH" annotation column is then available in the output file.

e. External gene annotation files (optional)

In order to further enrich the annotation for each SV gene, AnnotSV can integrate external annotations imported from tab separated values file(s) into the output file. The first line should be a header including a column entitled "genes". The following example has been set to provide annotation for the interacting partners of a gene.

genes	Interacting genes
BBS1	BBS7, TTC8, BBS5, BBS4, BBS9, ARL6, BBS2, RAB3IP, BBS12, BBS10

"Interacting genes" annotation column is then available in the output file.

Each external gene annotation file (*.tsv) should be located in the "$ANNOTSV/Annotations/Users/" directory. It is to notice that these files should not contain any of these 2 specific characters "{" and "}" (that would be replaced by "(" and ")"). AnnotSV supports either native or gzipped tsv file.

7. OUTPUT

a. Output format

Giving a SV input file, AnnotSV produces a tab-separated values file that can be easily integrated in bioinformatics pipelines or directly read in a spreadsheet program.

b. Output file path and name

Two options (-outputDir and -outputFile) can be used to specify the output directory and/or file name. The output file extension should be ".tsv" (tab separated values).
By default, an output directory is created where AnnotSV is run ('YYYYMMDD'_AnnotSV). As an example, an input SV file named "mySVinputFile.vcf" will produce by default an output file named "20180320_AnnotSV/mySVinputFile.annotated.tsv".

c. "AnnotSV type" column

A typical AnnotSV use would be to first look at the annotation and ranking of each SV as a whole (i.e. "full") and then focus on the content of that SV. This is possible thanks to the way AnnotSV can present the data. Indeed, there are 2 types of lines produced by AnnotSV (*cf* the "AnnotSV type" output column):

- An annotation on the **"full"** length of the SV. Every SV are reported, even those not covering a gene. This type of annotation gives an estimate of the SV event itself.

- An annotation of the SV "**split"** by gene. This type of annotation gives an opportunity to focus on each gene overlapped by the SV. Thus, when a SV spans over several genes, the output will contain as many annotations lines as covered genes (*cf* example in FAQ). This latter annotation is extremely powerful to shorten the identification of mutation implicated in a specific gene.

-typeOfAnnotation
Indiquer cas particulier des lignes full : la valeur la plus grande ou la plus patho

d. Annotation columns available in the output file

In the following table, we describe the annotations that are available in the AnnotSV output file. It is to notice that, since AnnotSV can be configured to output the annotations using 2 different modes (full or split), in some cases specific gene annotations are only present while using one of the two modes.

Column name	Annotation	Full	Split	BED input	VCF input
AnnotSV ID	AnnotSV ID	X	X	X	X
SV chrom	Name of the chromosome	X	X	X	X
SV start	Starting position of the SV in the chromosome	X	X	X	X
SV end	Ending position of the SV in the chromosome	X	X	X	X
SV length	Length of the SV (bp)	X	X	X	X
SV type	Type of the SV (DEL, DUP, …)	X	X	X	X
REF	Nucleotide sequence in the reference genome (extracted only from a VCF input file)	X	X		X
ALT	Alternate nucleotide sequence (extracted only from a VCF input file)	X	X		X
FORMAT	The FORMAT column from a VCF file	X	X		X
Sample ID	The sample ID column from a VCF file	X	X		X
AnnotSV type	Indicate the type of annotation generated: - annotation on the SV full length (“full”) - annotation on each gene overlapped by the SV (“split”)	X	X	X	X
Gene name	Gene symbol	X	X	X	X
NM	Transcript symbol[1]		X	X	X
CDS length	Length of the CoDing Sequence (CDS) (bp) overlapping the SV		X	X	X
tx length	Length of the transcript (bp) overlapping with the SV		X	X	X
location	SV location in the gene (e.g. « txStart-exon1 »)		X	X	X
intersectStart	Start position of the intersection between the SV and a transcript		X	X	X
intersectEnd	End position of the intersection between the SV and a transcript		X	X	X
promoters	List of the genes whose promoters are overlapped by the SV	X	X	X	X
DGV_GAIN_IDs	DGV Gold Standard GAIN IDs overlapping the annotated SV	X	X	X	X
DGV_GAIN_n_samples_with_SV	Number of individuals with a shared DGV_GAIN_ID	X	X	X	X
DGV_GAIN_n_samples_tested	Number of individuals tested	X	X	X	X
DGV_GAIN_Frequency	Relative GAIN frequency = DGV_GAIN_n_samples_with_SV/DGV_GAIN_n_samples_tested	X	X	X	X
DGV_LOSS_IDs	DGV Gold Standard LOSS IDs overlapping the annotated SV	X	X	X	X
DGV_LOSS_n_samples_with_SV	Number of individuals with a shared DGV_LOSS_ID	X	X	X	X
DGV_LOSS_n_samples_tested	Number of individuals tested	X	X	X	X
DGV_LOSS_Frequency	Relative LOSS frequency = DGV_LOSS_n_samples_with_SV/DGV_LOSS_n_samples_tested	X	X	X	X
DDD_SV	List of the DDD SV coordinates from the DDD study (data control sets) overlapping the annotated SV	X	X	X	X
DDD_DUP_n_samples_with_SV	Maximum number of individuals with a shared DDD_DUP (among the DDD_SV)	X	X	X	X
DDD_DUP_Frequency	Maximum DUP Frequency (among the DDD_SV)	X	X	X	X
DDD_DEL_n_samples_with_SV	Maximum number of individuals with a shared DDD_DEL (among the DDD_SV)	X	X	X	X

DDD_DEL_Frequency	Maximum DEL Frequency (among the DDD_SV)	X	X	X	X
DDD_status	DDD category e.g. confirmed, probable, possible, …		X	X	X
DDD_mode	DDD allelic requirement: e.g. biallelic, hemizygous …		X	X	X
DDD_consequence	DDD mutation consequence: e.g. "loss of function", uncertain …		X	X	X
DDD_disease	DDD disease name: e.g. "OCULOAURICULAR SYNDROME"		X	X	X
DDD_pmids	DDD Pubmed Ids		X	X	X
1000g_event	List of the 1000 genomes event types (e.g. DEL, DUP, <CN3>...)	X	X	X	X
1000g_AF	Maximum of the 1000 genomes allele frequency among the 1000g_event	X	X	X	X
1000g_max_AF	Highest observed allele frequency across the 1000g populations	X	X	X	X
dbVar_event	List of the dbVar NR SV event types (e.g. deletion, duplication…) overlapped with the annotated SV	X	X	X	X
dbVar_variant	List of the dbVar NR SV accession (e.g. nssv1415016) overlapped with the annotated SV	X	X	X	X
dbVar_status	dbVar NR SV clinical assertion (e.g. pathogenic, likely pathogenic)	X	X	X	X
ACMG	ACMG genes		X	X	X
HI_CGscore	ClinGen Haploinsufficiency Score	X	X	X	X
TriS_CGscore	ClinGen Triplosensitivity Score	X	X	X	X
synZ_ExAC	Positive synZ_ExAC (Z score) from ExAC indicate gene intolerance to synonymous variation	X	X	X	X
misZ_ExAC	Positive misZ_ExAC (Z score) from ExAC indicate gene intolerance to missense variation	X	X	X	X
pLI_ExAC	Score computed by ExAC indicating the probability that a gene is intolerant to a loss of function variation (Nonsense, splice acceptor/donor variants due to SNV/indel). ExAC consider pLI>=0.9 as an extremely LoF intolerant gene	X	X	X	X
delZ_ExAC	Positive delZ_ExAC (Z score) from ExAC indicate gene intolerance to deletion	X	X	X	X
dupZ_ExAC	Positive dupZ_ExAC (Z score) from ExAC indicate gene intolerance to duplication	X	X	X	X
cnvZ_ExAC	Positive cnvZ_ExAC (Z score) from ExAC indicate gene intolerance to CNV	X	X	X	X
HI_DDDpercent	Haploinsufficiency ranks from DDD	X	X	X	X
Mim Number	OMIM unique six-digit identifier		X	X	X
Phenotypes	e.g. Charcot-Marie-Tooth disease		X	X	X
Inheritance	e.g. AD (= "Autosomal dominant")[2]		X	X	X
morbidGenes	Set to “yes” if the SV overlaps an OMIM morbid gene	X	X	X	X
morbidGenesCandidates	Set to “yes” if the SV overlaps an OMIM morbid gene candidate	X	X	X	X
GCcontent_left	GC content around the left SV breakpoint (+/- 100bp)	X		X	X
GCcontent_right	GC content around the right SV breakpoint (+/- 100bp)	X		X	X
Repeats_coord_left	Repeats coordinates around the left SV breakpoint (+/- 100bp)	X		X	X
Repeats_type_left	Repeats type around the left SV breakpoint (+/- 100bp) e.g. AluSp, L2b, L1PA2, LTR12C, SVA_D, …	X		X	X
Repeats_coord_right	Repeats coordinates around the right SV breakpoint (+/- 100bp)	X		X	X
Repeats_type_right	Repeats type around the right SV breakpoint (+/- 100bp) e.g. AluSp, L2b, L1PA2, LTR12C, SVA_D, …	X		X	X
TADcoordinates	Coordinates of the TAD whose boundaries overlapped with the annotated SV (boundaries included in the coordinates)	X		X	X
ENCODEexperiments	ENCODE experiments used to define the TAD	X		X	X

compound-htz(sample)	List of heterozygous SNV/indel (reported with "chrom_position") presents in the gene overlapped by the annotated SV	X	X	X	X
hom(sample)	Number of homozygous variants (extracted from VCF input file) in the individual "sample" which are presents: - in the SV ("full" annotation) - between intersectStart and intersectEnd ("split" annotation)	X	X	X	X
htz(sample)	Number of heterozygous variants (extracted from VCF input file) in the individual "sample" which are presents: - in the SV ("full" annotation) - between intersectStart and intersectEnd ("split" annotation)	X	X	X	X
AnnotSV ranking	SV ranking into 1 of 5: class 1 (benign) class 2 (likely benign) class 3 (variant of unknown significance) class 4 (likely pathogenic) class 5 (pathogenic)	X	X	X	X

[1]Given one gene, only a single transcript from all transcripts available in RefSeq is reported. In case of transcripts with different CDS length (considering the overlapping region with the SV), the transcript with the longest CDS is reported. Otherwise, if there is no differences in CDS length, the longest transcript is reported.
[2]Detailed in the FAQ

8. USAGE / OPTIONS

To run AnnotSV, the default command line is the following:
$ANNOTSV/bin/AnnotSV -SVinputFile '/Path/Of/Your/VCF/or/BED/Input/File' >& AnnotSV.log &

The command line can be completed by the list of options described below or modified in the configfile. To show the options simply type:
$ANNOTSV/bin/AnnotSV –help or $ANNOTSV/bin/AnnotSV

OPTIONS:

-bedtools: Path of the bedtools local installation

-candidateGenesFile: Path of a file containing the candidate genes of the user (gene names can be space-separated, tabulation-separated, or line-break-separated).

-overlap: Minimum overlap (%) between the features (DGV, DDD, promoter, TAD…) and the annotated SV to be reported
Range values: [0-100], default = 70

-filteredVCFfiles: Path of the filtered VCF input file(s) with SNV/indel coordinates for compound heterozygotes report (optional)
Gzipped VCF files are supported as well as regular expression

-filteredVCFsamples: To specifiy the sample names from the VCF files defined from the -filtereVCFfiles option
Default: use all samples from the filtered VCF files

-genomeBuild: Human genome build used
Values: GRCh37 (default) or GRCh38

-help: More information on the arguments

-metrics: Changing numerical values from frequencies to us or fr metrics (e.g. 0.2 or 0,2).
Range values: us (default) or fr

-minTotalNumber: Minimum number of individuals tested to consider a benign SV for the ranking
Range values: [100-1000], default = 500

-outputDir: Output path name

-outputFile: Output path and file name

-overlap: Minimum overlap (%) between the features (DGV, DDD, promoter, TAD…) and the annotated SV to be reported
Range values: [0-100], default = 70

-promoterSize: Number of bases upstream from the transcription start site
Default = 500

-SVinputFile: Path of the input file (VCF or BED) with SV coordinates
Gzipped VCF file is supported

-SVinputInfo: To extract the additional SV input fields and insert the data in the output file
Range values: 0 (default) or 1

-SVminSize: SV minimum size (in bp)
Default = 50

-reciprocal: Use of a reciprocal overlap between SV and features.
Values: no (default) or yes

-svtBEDcol: Number of the column describing the SV type (DEL, DUP)
Range values: [4-[, default = -1 (value not given)

-typeOfAnnotation: Description of the types of lines produced by AnnotSV
Values: both (default), full or split

-vcfFiles: Path of the VCF input file(s) with SNV/indel coordinates used for false positive discovery
Use counts of the homozygous and heterozygous variants
Gzipped VCF files are supported as well as regular expression

-vcfPASS: Boolean. To only use variants from VCF input files that passed all filters during the calling (FILTER column value equal to PASS)
Range values: 0 (default) or 1

-vcfSamples: To specifiy the sample names from the VCF files defined from the -vcfFiles option
Default: use all samples from the VCF files

9. Test

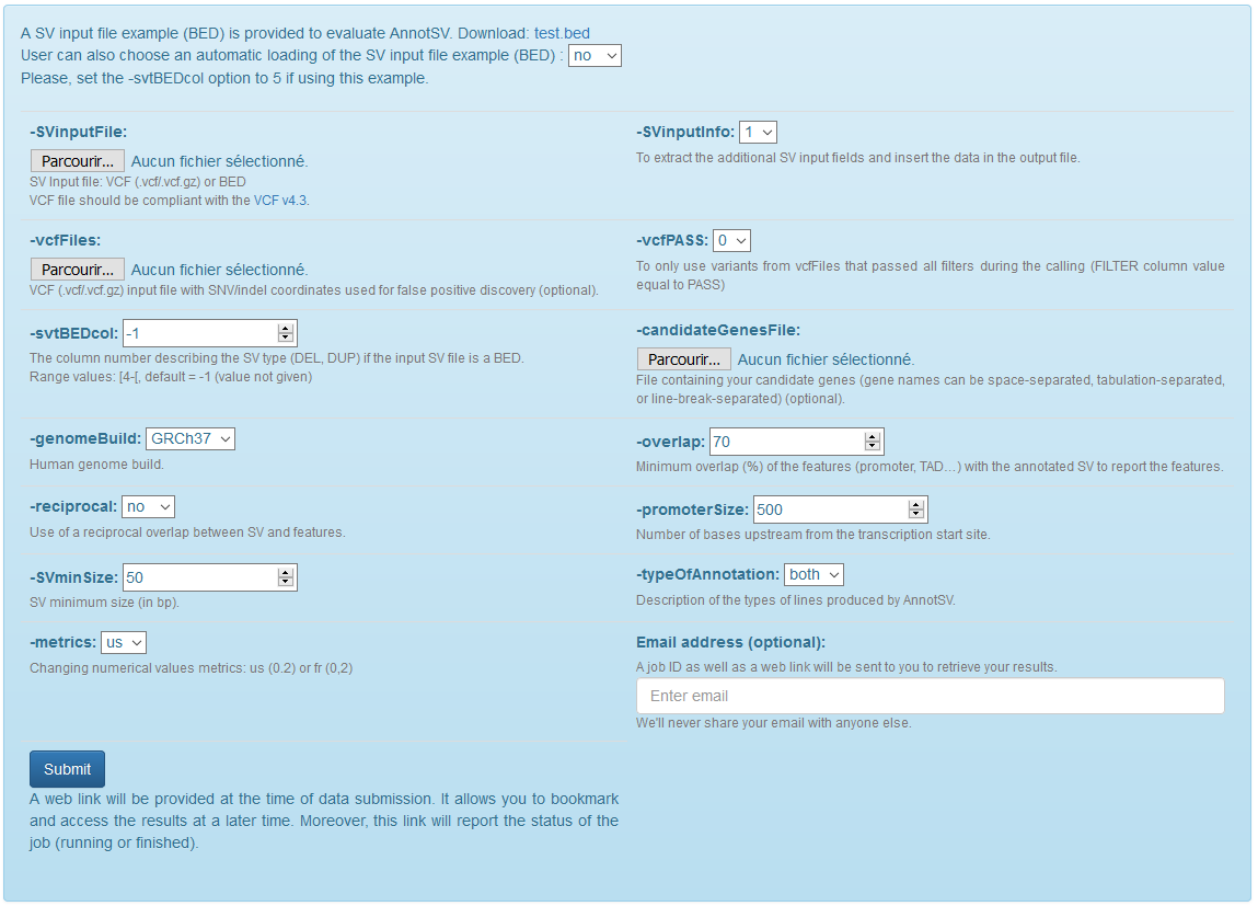
In order to validate the AnnotSV installation and its functioning, an example is available in the "$ANNOTSV/Example" directory. Command lines examples are available in the following file "$ANNOTSV/Example/commands.README".

Moreover, an input/output example (the HG00096 individual from the 1000 Genomes project) is available on the AnnotSV website.

10. Web server

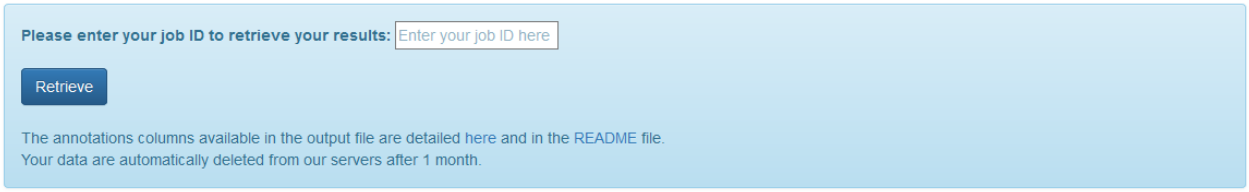
AnnotSV annotation and ranking of your SV are available online. A web server is freely available at:
https://lbgi.fr/AnnotSV/runjob

User can so operate through a web browser, which can be used to select the parameters, run the program, and retrieve the results:



A web link is provided at the time of data submission. It allows user to bookmark and access the results at a later time. Moreover, this link will report the status of the job (running or finished).

Moreover, a job ID is also provided to retrieve the results at:
https://lbgi.fr/AnnotSV/retrievejob



User data are automatically deleted from our servers after 1 month.

11. FAQ

Q: What are Structural Variations (SV)?
SV are generally defined as variation in a DNA region that vary in length from ~50 base pairs to many megabases and include several classes such as translocations, inversions, insertions, deletions.

Q: What are Copy Number Variations (CNV)?
CNV are deletions and duplications in the genome (unbalanced SV) that vary in length from ~50 base pairs to many megabases.

Q: What are the differences between SV and CNV?
CNV are unbalanced SV with gain or loss of genomic material. For example, a heterozygous duplication as a CNV will be characterized with the start and end coordinates and the number of copies which is 3.

Q: Can AnnotSV annotate every type of SV?
AnnotSV supports as well VCF or BED format in input.
- VCF format supports complex rearrangements with breakends, that can arbitrary be summarized as a set of novel adjacencies, as described in the Variant Call Format Specification VCFv4.3 (Jul 2017).
- BED format does not allow inter-chromosomal feature definitions (e.g. inter-chromosomal translocation). A new file format (BEDPE) is proposed in order to concisely describe disjoint genome features but it is not yet supported by AnnotSV.

Q: I would like to annotate my SV with new annotation sources but I don't know how to do that…
No problem. AnnotSV is under active and continuous development. You can email me with a detailed request and I will answer as quickly as possible.

Q: I have just updated AnnotSV or the annotations sources and the annotation process is longer than usual, is it normal?
After an update of AnnotSV sources, some files will be reprocessed and thus taking several additional time. Further use of AnnotSV will be quicker!

Q: How to cite AnnotSV in my work?
If you are using AnnotSV, please cite our work using the following reference:
AnnotSV: An integrated tool for Structural Variations annotation. Geoffroy V, Herenger Y, Kress A, Stoetzel C, Piton A, Dollfus H, Muller J. Bioinformatics. 2018 Apr 14. doi: 10.1093/bioinformatics/bty304

Q: What are the WARNINGs that AnnotSV mention while running?
AnnotSV writes to the standard output progress of the analysis including warnings about issues or missing information that can be either blocking or simply informative.

Q: Why are some values empty in the output files?
When no information is available for a specific type of annotation, then the value is empty.

Q: Why some SV have empty gene annotation in the output file?
If a SV is located in an intergenic region and so does not cover a gene, then the SV is reported in the output file but without gene annotation.

Q: Why can we have several gene annotations for one SV?
In some cases, one SV overlaps a large portion of the genome including several genes. In these cases, the annotation of the SV is split on several lines.

Annotation example for the deletion 1:16892807-17087595
AnnotSV keep all gene annotations, with only one transcript annotation for each gene:

1	16892807	17087595	DEL	CROCCP2	NR_026752	1	12652	txStart-txEnd
1	16892807	17087595	DEL	ESPNP	NR_026567	1	28941	txStart-txEnd
1	16892807	17087595	DEL	FAM231A	NM_001282321	511	511	txStart-txEnd
1	16892807	17087595	DEL	FAM231C	NM_001310138	511	656	txStart-txEnd
1	16892807	17087595	DEL	LOC102724562	NR_135824	1	2998	txStart-txEnd
1	16892807	17087595	DEL	MIR3675	NR_037446	1	75	txStart-txEnd
1	16892807	17087595	DEL	MST1L	NM_001271733	2015	6468	txStart-exon14
1	16892807	17087595	DEL	MST1P2	NR_027504	1	4848	txStart-txEnd
1	16892807	17087595	DEL	NBPF1	NM_017940	2912	47294	intron3-txEnd

Q: I am confused by the difference between the 'full' and the 'split' AnnotSV type mode. CNVs have been split into several lines, but each line get different DB annotation (DGV, 1000g…). I thought that same region should have the same annotations (excluding gene/transcript)?
AnnotSV builds 2 types of annotations, one based on the full-length SV (corresponding to the AnnotSV type = "full") and one based on each gene within the SV (corresponding to the AnnotSV type = "split"). Thus you will have access to:

- all the overlapped genes information (ID, OMIM...)
- the SV location within each overlapped gene (e.g. "exon3-intron11", "txStart-intron19", ...)

Be careful: the first 3 columns (SV chrom, SV start and SV end) remains the same despite being in "full" or in "split" type.

Regarding these "split" lines,

- DGV and 1000g SV overlaps are examined with regards to these gene coordinates. So, each "split" line get different DB annotation (DGV, 1000g...).
- 2 more annotation columns (intersectStart and intersectEnd) providing the intersection coordinates between the SV and the gene transcript.

Q: What do the OMIM Inheritance annotations mean?
AD = "Autosomal dominant"
AR = "Autosomal recessive"
XLD = "X-linked dominant"
XLR = "X-linked recessive"
YLD = "Y-linked dominant"
YLR = "Y-linked recessive"
XL = "X-linked"
YL = "Y-linked"

Q: Why do I get this error message: "Feature (10:134136286-134136486) beyond the length of 10 size (133797422 bp). Skipping."
One possibility is that you are using the bad "-genomeBuild" option. For example, you are using a bedfile in input with the SV coordinates on GRCh37 but with the "-genomeBuild GRCh38" option.

Q: How to interpret the presence of my SV in DGV or DDD databases?
DGV is populated with healthy samples whereas DDD is presenting affecting patients. The presence of a SV from your sample in DGV or DDD does not necessarily imply a disease-causing event. Healthy carriers of pathogenic SV do exist in either databases. When available allele frequency can be helpful to decide on the status.

Q: Is AnnotSV available for other organisms?
The main objective of AnnotSV is to annotate SV information from human data. All the annotations are based on human specific databases. Nevertheless, some files can be modified with the proper dataset but this is not currently supported.

Q: Is there an option to just generate SV "split" by gene?
You can choose to keep only the split annotation lines thanks to the "-typeOfAnnotation" option.

Q: I am unable to run the code on the input files provided. It crashes on the Repeat annotation step due to a bad_alloc error. Do you have any ideas on why this is happening?
AnnotSV needs to be run with an appropriate RAM (depending of the annotations used). Setting your system to allocate 10 Go should solve the problem.

Q: I'm getting the error: "ANNOTSV environment variable not specified. Please define it before running AnnotSV. Exit". How can I fix this problem?
ANNOTSV is the environment variable defining the installation path of the software.

- In csh, you can define it with the following command line:
  setenv ANNOTSV /path_of_AnnotSV_installation/bin
- In bash, you can define it with the following command line:
  export ANNOTSV=/path_of_AnnotSV_installation/bin

I advise you to save the good command in your .cshrc or .bashrc file.

Q: My annotated SV is intersecting both a benign SV and a pathogenic SV. How can I explain that?
Several possible explanations can be considered:

- The pathogenicity can concern a recessive disease. So the pathogenic SV can be present in the heterozygous state in the healthy population (with a DGV low frequency)
- The pathogenic region of the dbVar SV is not overlapping the DGV SV

12. REFERENCES

; on behalf of the ACMG Laboratory Quality Assurance Committee, Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., et al. (2015). **Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology**. Genetics in Medicine *17*, 405–423.

Dittwald, P., et al. **NAHR-mediated copy-number variants in a clinical population: mechanistic insights into both genomic disorders and Mendelizing traits**. Genome research 2013;23(9):1395-1409.

Firth, H.V., Wright, C.F. and Study, D.D.D. **The Deciphering Developmental Disorders (DDD) study**. Developmental medicine and child neurology 2011;53(8):702-703.

Hamosh, A., et al. **Online Mendelian Inheritance in Man (OMIM)**. Human mutation 2000;15(1):57-61.

Lek, M., et al. **Analysis of protein-coding genetic variation in 60,706 humans**. Nature 2016;536(7616):285-291.

Lupianez, D.G., Spielmann, M. and Mundlos, S. Breaking **TADs: How Alterations of Chromatin Domains Result in Disease**. Trends in genetics : TIG 2016;32(4):225-237.

MacDonald, J.R., et al. **The Database of Genomic Variants: a curated collection of structural variation in the human genome.** Nucleic acids research 2014;42(Database issue):D986-992.

Sudmant, P.H., et al. **An integrated map of structural variation in 2,504 human genomes.** Nature 2015;526(7571):75-81.