AnnotSV: Mouse Annotation Manual

Version 3.0
AnnotSV Mouse is a program for annotating structural variations from the mouse genome.

https://lbgi.fr/AnnotSV/

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

LEXIQUE

BED: Browser Extensible Data
bp: base pair
CDS: CoDing Sequence
CNV: Copy Number Variation
DEL: Deletion
DNA: DesoxyriboNucleic Acid
DUP: Duplication
ENST: ENSEMBL transcript identifier
GRCm38/mm10: Genome Reference Consortium Mouse Build 38
NCBI37/mm9: NCBI Mouse Build 37
hom: homozygous
htz: heterozygous
ID: Identifier
indel: Insertion/deletion
INS: Insertion
INV: Inversion
NAHR: Non-Allelic Homologous Recombination
NM: RefSeq identifiers
SNV: Single Nucleotide Variation
SV: Structural Variations
Tx: transcript
VCF: Variant Call Format

TABLE OF CONTENTS

1. INTRODUCTION

AnnotSV is a program designed for annotating 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**.

Mouse annotations are available in AnnotSV. Its installation and use are detailed in this manual.

2. MOUSE ANNOTATION INSTALLATION

The AnnotSV program needs to be installed before to add the Mouse annotation.
Please, see the AnnotSV README manual for more information on the installation/requirements.

3. ANNOTATION SOURCES

AnnotSV requires different data sources for the annotation of SV. **In order to provide a ready to start installation of AnnotSV, each Mouse annotation source listed below (that do not require a commercial license) is provided with the AnnotSV Mouse sources.** The aim and update of each of these sources are explained below. Annotation can be performed using either the **mm9 or mm10 build version** of the mouse 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. 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 8 annotation columns: “Gene_name”, “Tx”, “Overlapped_CDS_length”, “Overlapped_tx_length”, “Location”, “Location2”, “Intersect_start”, “Intersect_end”.

Method:
For each gene, only a single transcript from all transcripts available in RefSeq for this gene is reported in the following order of preference:
- The transcript selected by the user with the "-txFile" option is reported
- The transcript with the longest overlapped CDS is reported (considering the overlapping region with the SV)
- If there is no difference in CDS length, the longest transcript is reported.

Updating the data source (if needed):

- Remove all the files in the “$ANNOTSV/share/AnnotSV/Annotations_Mouse/Genes/mm9” and/or “$ANNOTSV/share/AnnotSV/Annotations_Mouse/Genes/mm10” directories.
- Download and place the “refGene.txt.gz” file in the

"$ANNOTSV/share/AnnotSV/Annotations_Mouse/Genes/mm9" and/or
"$ANNOTSV/share/AnnotSV/Annotations_Mouse/Genes/mm10" directories.
The latest update of this file is available for free download at:
Genome build mm9:
http://hgdownload.cse.ucsc.edu/goldenPath/mm9/database/refGene.txt.gz
Genome build mm10:
http://hgdownload.cse.ucsc.edu/goldenPath/mm10/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).

b. Annotations with features **overlapping** the SV

It is to notice that, for this type of annotations and only for this type, a reciprocal overlap can be used (see "reciprocal" option in USAGE/OPTIONS).

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 mouse 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 i) if the SV overlaps at least 70% of this promoter (user defined, see the "overlap" option in USAGE/OPTIONS) or ii) if the SV is an insertion included in the promoter.

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

d. Breakpoints annotations

GC content annotations

Aim:
GC content 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 mouse reference genome FASTA file to run the “bedtools nuc” command.

- Remove all the files in the “$ANNOTSV/share/AnnotSV/Annotations_Mouse/BreakpointsAnnotations/GCcontent/mm9” and/or “$ANNOTSV/share/AnnotSV/Annotations_Mouse/BreakpointsAnnotations/GCcontent/mm10” directories.
- Download and place the mouse reference genome FASTA file in the “$ANNOTSV/share/AnnotSV/Annotations_Mouse/BreakpointsAnnotations/GCcontent/mm9” and/or “$ANNOTSV/share/AnnotSV/Annotations_Mouse/BreakpointsAnnotations/GCcontent/mm10” directories.
  The latest update of this file is available for free download at:
  Genome build mm9:
  http://hgdownload.cse.ucsc.edu/goldenPath/mm9/bigZips/chromFa.tar.gz

  Genome build mm10:
  http://hgdownload.cse.ucsc.edu/goldenPath/mm10/bigZips/chromFa.tar.gz

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

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

Repeated sequences annotations

Aim:
Repeated sequences 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 "$ANNOTSV/share/AnnotSV/Annotations_Mouse/BreakpointsAnnotations/Repeat/mm9" and/or "$ANNOTSV/share/AnnotSV/Annotations_Mouse/BreakpointsAnnotations/Repeat/mm10" 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:

  "Mouse" genome, "NCBI37/mm9" or "GRCm38/mm10" assembly, "Variations and 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 "$ANNOTSV/share/AnnotSV/Annotations_Mouse/BreakpointsAnnotations/Repeat/mm9" and/or "$ANNOTSV/share/AnnotSV/Annotations_Mouse/BreakpointsAnnotations/Repeat/mm10" directories.

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

e. External BED annotation files (optional)

Aim:
Several users might want to add their own private region 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 mm9:
➔ "$ANNOTSV/share/AnnotSV/Annotations_Mouse/Users/mm9/FtIncludedInSV" directory
or
➔ "$ANNOTSV/share/AnnotSV/Annotations_Mouse/Users/mm9/SVincludedInFt" directory

Genome build mm10:
➔ "$ANNOTSV/share/AnnotSV/Annotations_Mouse/Users/mm10/FtIncludedInSV" directory
or
➔ "$ANNOTSV/share/AnnotSV/Annotations_Mouse/Users/mm10/SVincludedInFt" directory

It is to notice that:

By placing the BED file in the "FtIncludedInSV" directory, only the features overlapped with the SV (>70% by default) will be reported

By placing the BED file in the "SVincludedInFt" directory, only the features overlapping the SV (>70% by default) will be reported. In this case, a reciprocal overlap can be used (see "reciprocal" option in USAGE/OPTIONS).

In both cases, the user can modify the default behaviour of the overlap by using a different percentage (see "overlap" option in USAGE/OPTIONS).

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

Examples:

- This first example has been set to provide the SV overlap with frequency (Freq) of internal cohort regions:

  'UserYYY'.bed file contains:

#Chrom	Start	End	Freq
1	2806107	107058351	0.0018
12	25687536	25699754	0.0023

The additional "Freq" annotation column is then made available in the output file (if "Freq" added in the configfile).

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

  'UserXXX'.bed file contains:

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

The additional "RoH" annotation column is then made available in the output file (if "RoH" added in the configfile).

f. 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/share/AnnotSV/Annotations_Mouse/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.

Moreover you need to use a configfile (located either in the same directory as your input file or directly in $ANNOTSV/etc/AnnotSV/configfile) and to define the output column names you want to be added.

4. VERSIONS OF THE ANNOTATION SOURCES

Annotations source	Version
...Gene annotations	
Gene annotations (RefSeq)	2020-08-22
…Regulatory Elements annotations	
Promoter data (RefSeq)	2020-08-22
...Breakpoints annotations	
GRCh37 FASTA genome	2007-07-25
GRCh38 FASTA genome	2012-02-09
Repeated sequences annotations	2020-07-16

5. 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 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.

Considering the "full" length annotation of one SV, AnnotSV does not report the genes-based annotation (value is set to empty), except for scores and percentages where AnnotSV reports the most pathogenic score or the maximal percentage.

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
Samples_ID	List of the samples ID for which the SV was called	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
Annotation_mode	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
Gene_count	Number of overlapped genes with the SV	X		X	X
Tx[1]	Transcript symbol		X	X	X
Tx_start	Starting position of the transcript		X	X	X
Tx_end	Ending position of the transcript		X	X	X
Overlapped_tx_length	Length of the transcript (bp) overlapping with the SV		X	X	X
Overlapped_CDS_length	Length of the CoDing Sequence (CDS) (bp) overlapping the SV		X	X	X
Overlapped_CDS_percent	Percent of the CoDing Sequence (CDS) (bp) overlapping the SV		X	X	X
Frameshift	Indicates if the CDS length is not divisible by three (yes or no)		X	X	X
Exon_count	Number of exons of the transcript		X	X	X
Location	SV location in the gene (e.g. « txStart-exon1 »)		X	X	X
Location2	SV location in the gene's coding regions (e.g. « 3'UTR-CDS »)		X	X	X
Dist_Nearest_SS	Absolute distance to nearest splice site after considering exonic and intronic SV breakpoints		X	X	X
Nearest_SS_type	Nearest splice site type: 5' (donor) or 3' (acceptor)		X	X	X
Intersect_start	Start position of the intersection between the SV and a transcript		X	X	X
Intersect_end	End position of the intersection between the SV and a transcript		X	X	X
RE_gene	List of the genes whose promoters are overlapped by the SV	X	X	X	X

GC_content_left	GC content around the left SV breakpoint (+/- 100bp)	X		X	X
GC_content_right	GC content around the right SV breakpoint (+/- 100bp)	X		X	X
Repeat_coord_left	Repeats coordinates around the left SV breakpoint (+/- 100bp)	X		X	X
Repeat_type_left	Repeats type around the left SV breakpoint (+/- 100bp) e.g. AluSp, L2b, L1PA2, LTR12C, SVA_D, …	X		X	X
Repeat_coord_right	Repeats coordinates around the right SV breakpoint (+/- 100bp)	X		X	X
Repeat_type_right	Repeats type around the right SV breakpoint (+/- 100bp) e.g. AluSp, L2b, L1PA2, LTR12C, SVA_D, …	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
#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
#htz/allHom(sample)	Ratio for QC filtering: #htz(sample)/#allHom(sample)5	X	X		X
#htz/total(cohort)	Ratio for QC filtering: #htz(sample)/#total(cohort)	X	X		X
#total(cohort)	Total count of SNV/indel called from all the samples of the cohort and present in the interval of the deletion	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

[1]*Given one gene, only a single transcript from all transcripts available in RefSeq is reported. The transcript selected by the user with the "-txFile" option is firstly reported. Else, 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.*

e. User selection of the annotation columns

Users can disable the default annotation columns provided by AnnotSV and selects only the one of interest for its analysis. This could especially help in reducing the size of the output file and the time of the annotation.

This setting can be easily done in a configfile located in the same directory as the INPUT file (an example of configfile is provided in the AnnotSV installation directory), the user can comment column names with a hash character («#»).

6. USAGE / OPTIONS

To run AnnotSV, the default command line is the following:
$ANNOTSV/bin/AnnotSV/AnnotSV.tcl -SvinputFile '/Path/Of/Your/VCF/or/BED/Input/File' **–genomeBuild 'Your genome build'** >&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/AnnotSV.tcl -help
or
$ANNOTSV/bin/AnnotSV/AnnotSV.tcl

OPTIONS:

-annotationsDir: Path of the annotations directory

-bcftools: Path of the bcftools local installation

-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).

-candidateGenesFiltering: To select only the SV “split” annotations overlapping a gene from the "candidateGenesFile"
Values: no (default) or yes

-candidateSnvIndelFiles: 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

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

-genomeBuild: Genome build used
Values: GRCh37 (default) or GRCh38 or mm9 or mm10

-help: More information on the arguments

-hpo: HPO terms list describing the phenotype of the individual being investigated.
Values: use comma, semicolon or space separated class values,
Default = "" (e.g.: "HP:0001156,HP:0001363,HP:0011304")

-includeCI: To expand the "start" and "end" SV positions with the VCF confidence intervals (CIPOS, CIEND) around the breakpoints
Values: yes (default) or no

-metrics: Changing numerical values from frequencies to us or fr metrics (e.g. 0.2 or 0,2).
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 user features (User BED) and the annotated SV to be reported
Range values: [0-100], default = 100

-overwrite: To overwrite existing output results.
Values: yes (default) or no

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

-rankFiltering: To select the SV of a user-defined specific class (from 1 to 5)
Values: use comma separated class values, or use a dash to denote a range of values
(e.g.: "3,4,5" or "3-5"), default = "1-5"

-reciprocal: Use of a reciprocal overlap between SV and user features (only for annotations with features overlapping the SV)
Values: no (default) or yes

- REreport: Create a report to link the annotated SV and the overlapped regulatory elements (coordinates and sources)
Values: no (default) or yes

-samplesidBEDcol: Number of the column reporting the samples ID for which the SV was called (if the input SV file is a BED)
Range values: [4-[, default = -1 (value not given)
(Samples ID should be comma or space separated)

-snvIndelFiles: 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

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

-snvIndelSamples: To specify the sample names from the VCF files defined from the -snvIndelFiles option
Default: use all samples from the VCF files

-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 outputfile
Values: 1 (default) or 0

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

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

-tx: Origin of the transcripts (RefSeq or ENSEMBL)
Values: RefSeq (default) or ENSEMBL

-txFile:	Path of a file containing a list of preferred genes transcripts to be used in priority during the annotation (Preferred genes transcripts names should be tab or space separated)
-annotationMode:	Description of the types of lines produced by AnnotSV Values: both (default), full or split