AnnotSV Manual

Version 2.3
AnnotSV is a program for annotating and ranking structural variations from genomes of several organisms.
This README version is dedicated to 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
GH: GeneHancer
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
INS: Insertion
INV: Inversion
LoF: Loss of Function
MCNV: multiallelic CNV
MEI: Mobile Element Insertion
misZ = Z scoreindicating 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 VCF v4.3 (Jul 2017).

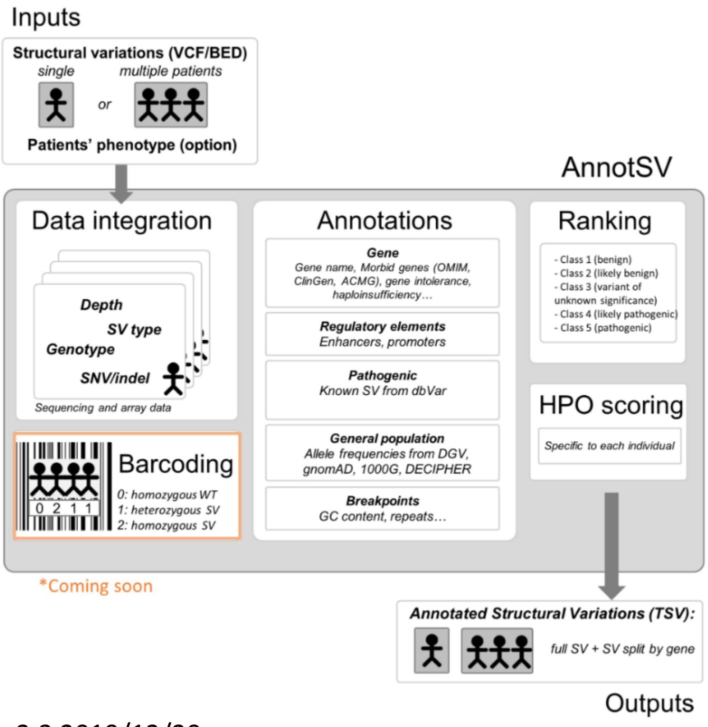
a. Overview

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

- Genes-based annotations (OMIM, Gene intolerance, Haploinsufficiency…)
- Annotations with features overlapping the SV (DGV, 1000genomes…)
- Annotations with features overlapped with the SV (pathogenic SV from dbVar, promoters, enhancers, TAD…)
- Annotations of the SV breakpoints (GC content, repeats…)

In addition to these annotations, AnnotSV also provide a systematic SV classification/ranking 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


Inputs
Structural variations (VCF/BED)
single
or
multiple patients
Patients' phenotype (option)
AnnotSV
Data integration
Depth
SV type
Genotype
SNV/indel
Sequencing and array data
Barcoding
0 2 1 1
0: homozygous WT
1: heterozygous SV
2: homozygous SV
*Coming soon
Annotations
Gene
Gene name, Morbid genes (OMIM, ClinGen, ACMG), gene intolerance, haploinsufficiency…
Regulatory elements
Enhancers, promoters
Pathogenic
Known SV from dbVar
General population
Allele frequencies from DGV, gnomAD, 1000G, DECIPHER
Breakpoints
GC content, repeats…
Ranking
- Class 1 (benign)
- Class 2 (likely benign)
- Class 3 (variant of unknown significance)
- Class 4 (likely pathogenic)
- Class 5 (pathogenic)
HPO scoring
Specific to each individual
Annotated Structural Variations (TSV):
full SV + SV split by gene
Outputs

It is important to notice that, in order to reduce or at least not to expand too much the list of annotation columns, we have decided for the new and upcoming annotations (gnomAD, IMH) to specifically report the information of the corresponding SV type.
Ex: A deletion of interest will be annotated with gnomAD using only the deletion data in details. However, events of different SV type (such as duplication, inversion…) overlapping our initial SV will be reported using only their identifiers.

b. Supported organisms

AnnotSV is mainly dedicated for the annotation and ranking of structural variations from human genomes. However, since version 2.2 AnnotSV supports also the mouse genome. If you are interested, please see the specific mouse README file.

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 https://www.activestate.com/activetcl/downloads).

AnnotSV requires **the latest release of the Tcl distribution starting with version 8.5** as well as the following 3 packages “http”, "tar" and "csv".
The “http” package is used for the phenotype-driven analysis.
The "tar" and "csv" packages are used only when data sources are updated.

b. bedtools (required)

The **“bedtools”** toolset (developed by Quinlan AR) needs to be locally installed.

Add the path of the bedtools bin directory to your PATH and save the settings in your .cshrc or .bashrc file:

- In csh, you can define it with the following command line:
```
setenv PATH {$PATH}:/’somewhere’/bedtools-2.25.0/bin
```
- In bash, you can define it with the following command line:
```
export PATH=$PATH:/’somewhere’/bedtools-2.25.0/bin
```

Warning: the minimum bedtools version compatible with AnnotSV is version 2.25. To check if bedtools exists and if the version is the good one, run:

```
bedtools --version
```

c. Java (optional)

In order to use the phenotype-driven analysis based on one Exomiser module, a minimal Java 8 installation is required.

d. AnnotSV source code (required)

Since the 2.3 version, **"AnnotSV source code"** is only downloadable on GitHub at the following address (under the GNU GPL license):
https://github.com/lgmgeo/AnnotSV

Install:
The sources can be cloned to any directory:

```
cd /'somewhere'/
git clone https://github.com/lgmgeo/AnnotSV
```

Then, the user can choose either to easily set the install by default in /usr/local:

```
make install
```

or to define $PREFIX as a specific installation directory:

```
make PREFIX=/'somewhere_else'/AnnotSV_'version'/ install
```

or to define $PREFIX as the actual directory:

```
make PREFIX=. install
```

The AnnotSV installation directory (/path_of_AnnotSV_installation) will be either set to:
/usr/local
or: /*'somewhere_else'*/AnnotSV_***'version'***/
or: /*'somewhere'*/AnnotSV_***'version'***/
Thus, the AnnotSV executable will be located in:
/path_of_AnnotSV_installation/bin/AnnotSV

Then, the annotations requested by the user (human, mouse or both) need to be installed with the following command lines:

```
make PREFIX=… install-human-annotation
make PREFIX=… install-mouse-annotation
make PREFIX=… install-mouse-annotation install-human-annotation
make PREFIX=… install-all-annotations
```

Finally, 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/
  ```
- In bash, you can define it with the following command line:
  ```
  export ANNOTSV=/path_of_AnnotSV_installation/
  ```

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 (already done in the AnnotSV executable):

```
#!/usr/bin/env tclsh
```

It can be changed to any other path like:

```
#!/usr/local/ActiveTcl/bin/tclsh
```

e. Filesystem Hierarchy Standard (FHS)

AnnotSV follows the Filesystem Hierarchy Standard (FHS) that defines the directory structure and directory contents in Linux distributions.

AnnotSV installation directory:
By default, the AnnotSV installation directory looks like this:

```
${DESTDIR}${PREFIX}                     #the program installation directory (default = /usr/local)
|
|----- bin/                             #where the executable script is stored
|
|----- etc/AnnotSV/                     #where a configfile example is stored, that can be copied to any
|                                       #analysis directory for modification purpose
|----- Makefile
|
|----- share/                           #Architecture-independent (shared) data
|        |----- AnnotSV                 #where annotation files are stored (RefGene, OMIM…)
|        |   |---- Annotations_Exomiser
|        |   |---- Annotations_Human
|        |   |---- Annotations_Mouse
|        |   |---- jar
|        |
|        |----- bash                    #where bash files are stored
|        |----- doc/AnnotSV/
|        |   |----- Example             #command/input/output examples
|        |   |----- changeLog.txt       #description of AnnotSV changes
|        |   |----- commandLineOptions.txt  #command line usage
|        |   |----- License.txt         #GNU GPL license
|        |   |----- README.AnnotSV_*.pdf  #this file
|        |
|        |----- tcl*/AnnotSV/           #where the procedures .tcl files are stored
```

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 annotation source listed below (that do not require a commercial license) is automatically downloaded during the installation. One exception is the GeneHancer source for which a licence is required (request to the GeneCards team).** The aim and update of each of these sources are explained below. Annotation can be performed using either the GRCh37 or GRCh38 build 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. Gene-based annotations

Each gene overlapped by the SV to annotate is reported (even with 1bp overlap).

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", "NM", "CDS length", "tx length", "location", "location2", "intersectStart", "intersectEnd".

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 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 "$ANNOTSV/share/AnnotSV/Annotations_Human/RefGene/GRCh37" and/or "$ANNOTSV/share/AnnotSV/Annotations_Human/RefGene/GRCh38" directories.
- Download and place the "refGene.txt.gz" file in the "$ANNOTSV/share/AnnotSV/Annotations_Human/RefGene/GRCh37" and/or "$ANNOTSV/share/AnnotSV/Annotations_Human/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-array CGH 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 "$ANNOTSV/share/AnnotSV/Annotations_Human/Genes-based/DDD" directory.
- Download and place the "**DDG2P.csv.gz**" DECIPHER file in the "$ANNOTSV/share/AnnotSV/Annotations_Human/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 is 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 "$ANNOTSV/share/AnnotSV/Annotations_Human/Genes-based/OMIM" directory.
- Download and place the "**genemap2.txt**" and "**morbidmap.txt**" OMIM files in the "$ANNOTSV/share/AnnotSV/Annotations_Human/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)	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:
Adds6 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/share/AnnotSV/Annotations_Human/Genes-based/ExAC” directory.

- Download and place the “**fordist_cleaned_nonpsych_z_pli_rec_null_data.txt**” and the “**exac-final-cnv.gene.scores071316**” ExAC files in the “$ANNOTSV/share/AnnotSV/Annotations_Human/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
  ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3.1/cnv/exac-final-cnv.gene.scores071316

This file will be reprocessed the first time AnnotSV is 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 the “*_HI.tsv.gz“ file in the “$ANNOTSV/share/AnnotSV/Annotations_Human/Genes-based/DDD” directory.
- Download and place the “**HI_Predictions_Version3.bed.gz**” DECIPHER file in the “$ANNOTSV/share/AnnotSV/Annotations_Human/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 is 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 "$ANNOTSV/share/AnnotSV/Annotations_Human/Genes-based/ClinGen/" directory.
- Download and place the "**ClinGen_gene_curation_list_GRCh37.tsv**" ClinGen file in the "$ANNOTSV/share/AnnotSV/Annotations_Human/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 is executed after the update. The annotations selected by AnnotSV are genome build independent, and only based on the gene name.

Phenotype-driven analysis extracted from Exomiser

Aim:

To score genes overlapped with a SV on biological relevance to the individual phenotype, AnnotSV takes use of Exomiser (Smedley et al., 2015) and HPO (Köhler et al., 2019).

For a given phenotype, a HPO-based score corresponding to a damaging probability is provided for each gene overlapped with an SV so that:

- Genes previously associated with disease can be highlighted easily
- Genes not previously associated with disease can be highlighted
- Genes associated with diseases that have little or no similarity to the observed phenotypes can be removed along

HPO:

AnnotSV uses the Human Phenotype Ontology (version reported in the AnnotSV output).
Find out more at http://www.human-phenotype-ontology.org.

human
phenotype
ontology

Please cite the 3 following articles if you use these data in your work:

- AnnotSV: An integrated tool for Structural Variations annotation. Geoffroy V., *et al*, Bioinformatics (2018) doi: doi:10.1093/bioinformatics/bty304
- Next-generation diagnostics and disease-gene discovery with the Exomiser. Smedley D., *et al*, Nature Protocols (2015) doi:10.1038/nprot.2015.124
- Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Köhler S., *et al*, Nucleic Acids Research (2019) doi: 10.1093/nar/gky1105

Annotation columns:

Add 4 annotation columns: "EXOMISER_GENE_PHENO_SCORE", "HUMAN_PHENO_EVIDENCE", "MOUSE_PHENO_EVIDENCE" and "FISH_PHENO_EVIDENCE"

Usage:

The user enters a human phenotype as a list of HPO terms (see "hpo" option in USAGE/OPTIONS). The HPO terms needs to be as specific as possible.

According to our own experience (limited), a gene with an EXOMISER_GENE_PHENO_SCORE >= 0.7 can be considered to be associated with the disease. For a gene that has not been previously associated with a disease, the threshold can be lowered to 0.5.

Updating the data source (if needed):

AnnotSV needs matching between the “HGNC symbols” and “NCBI gene ID”.

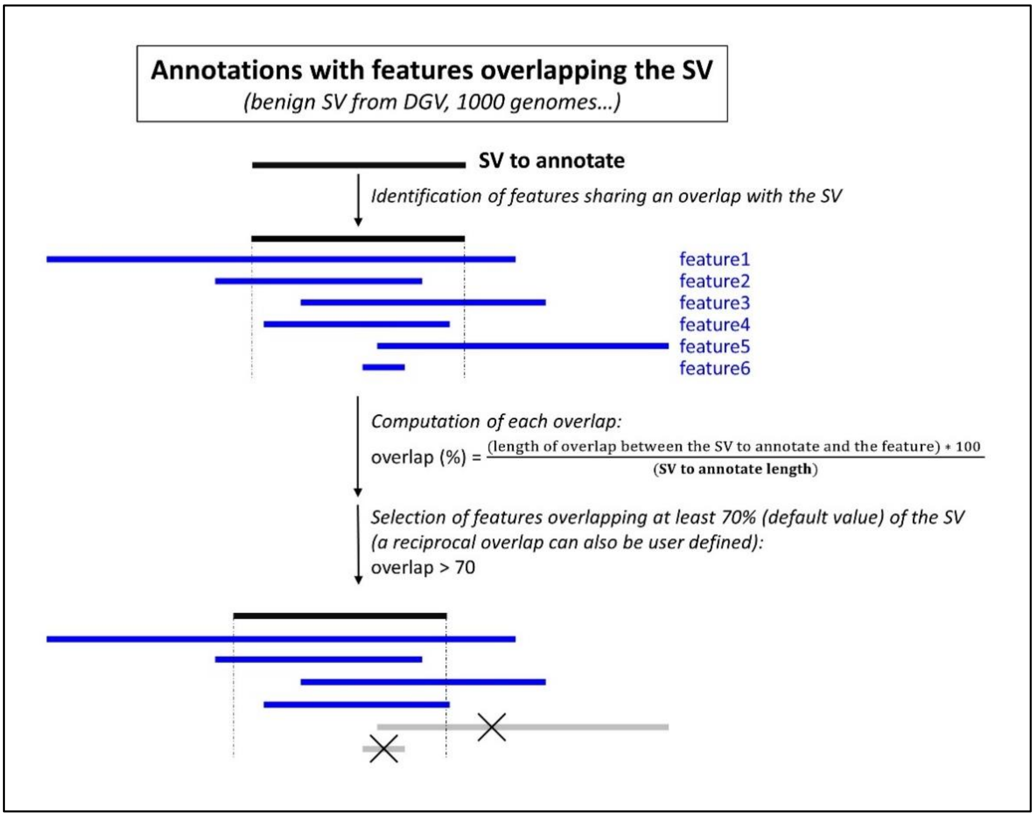
- Remove all the files in the “$ANNOTSV/share/AnnotSV/Annotations_Human/Genes-based/NCBIgeneID/” directory.
- Download and place your NCBI gene ID file (“results.txt”) in the “$ANNOTSV/share/AnnotSV/Annotations_Human/Genes-based/NCBIgeneID/” directory
  This file is available for free download at:
  https://biomart.genenames.org/martform/#!/default/HGNC?datasets=hgnc_gene_mart
  In the “Attributes” section:
  - Select only the “Approved symbol”, the “Alias symbol” and the “Previous symbol”.
  In the “Gene resources” section:
  - Select only the “NCBI gene ID”.
  Click the "Go >>" button.
  Then, click the “Download data” button to download the “results.txt” file.

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

First, AnnotSV searches for features sharing an overlap with the SV to annotate. Second, only the features overlapping at least 70% of the SV in size/location are selected (default value, a different percentage can also be user defined with the "overlap" option).

Interest of this computation:
For example, AnnotSV considers that a benign SV is informative enough only if > 70% length of the SV to annotate is overlapped with this benign SV. So, and only then, the SV to annotate can be considered as benign.

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).


Annotations with features overlapping the SV
(benign SV from DGV, 1000 genomes...)
SV to annotate
Identification of features sharing an overlap with the SV
feature1
feature2
feature3
feature4
feature5
feature6
Computation of each overlap:
overlap (%) = (length of overlap between the SV to annotate and the feature) * 100 / (SV to annotate length)
Selection of features overlapping at least 70% (default value) of the SV
(a reciprocal overlap can also be user defined):
overlap > 70

DGV Gold Standard annotations

Aim:

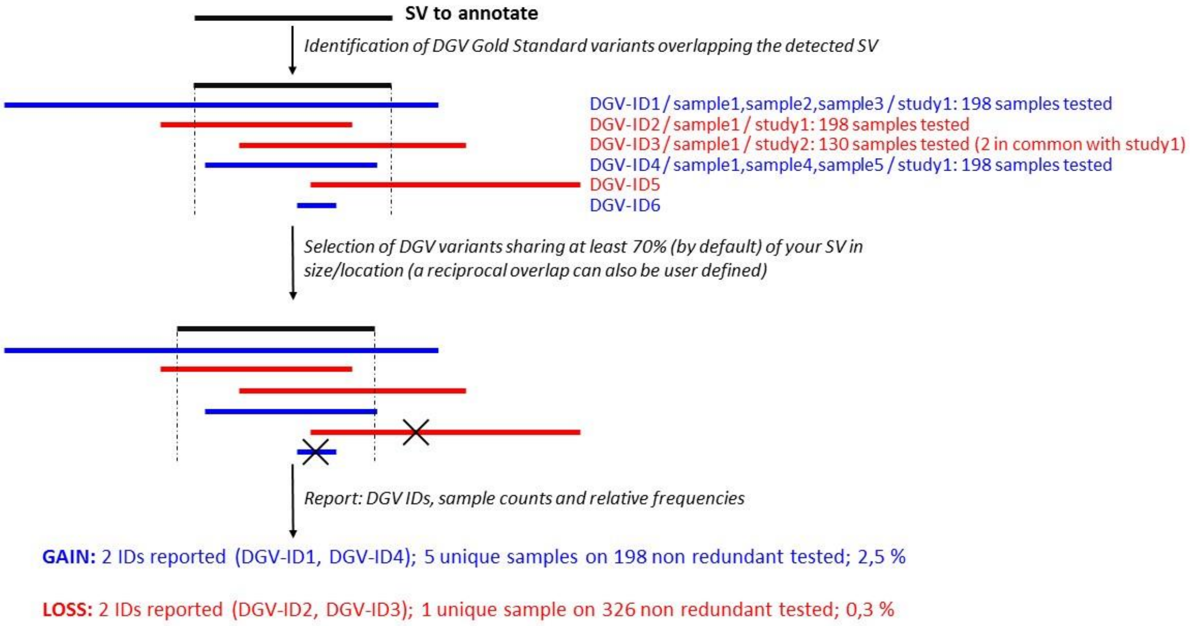
The Database of Genomic Variants (DGV) (MacDonald, et al., 2014) provides SV defined as DNA elements with a size >50 bp. The content of DGV is only representing SV identified in healthy control samples from large cohorts published and integrated by the DGV team. The annotations will give information 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% (default) of your SV in size/location are selected. 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).


SV to annotate
Identification of DGV Gold Standard variants overlapping the detected SV
DGV-ID1 / sample1,sample2,sample3 / study1: 198 samples tested
DGV-ID2 / sample1 / study1: 198 samples tested
DGV-ID3 / sample1 / study2: 130 samples tested (2 in common with study1)
DGV-ID4 / sample1,sample4,sample5 / study1: 198 samples tested
DGV-ID5
DGV-ID6
Selection of DGV variants sharing at least 70% (by default) of your SV in size/location (a reciprocal overlap can also be user defined)
Report: DGV IDs, sample counts and relative frequencies
GAIN: 2 IDs reported (DGV-ID1, DGV-ID4); 5 unique samples on 198 non redundant tested; 2,5 %
LOSS: 2 IDs reported (DGV-ID2, DGV-ID3); 1 unique sample on 326 non redundant tested; 0,3 %

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 "$ANNOTSV/share/AnnotSV/Annotations_Human/SVincludedInFt/DGV/GRCh37" and/or "$ANNOTSV/share/AnnotSV/Annotations_Human/SVincludedInFt/DGV/GRCh38" directories.
- Download and place the 2 following DGV files in the "$ANNOTSV/share/AnnotSV/Annotations_Human/SVincludedInFt/DGV/GRCh37" and/or "$ANNOTSV/share/AnnotSV/Annotations_Human/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 is 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).

Annotation columns:

Adds5 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) in the **"full"** length lines.

Updating the data source (if needed):

- Remove all the files in the "$ANNOTSV/share/AnnotSV/Annotations_Human/SVincludedInFt/DDD/GRCh37" directory.
- Download and place the "**population_cnv.txt.gz**" DECIPHER files in the "$ANNOTSV/share/AnnotSV/Annotations_Human/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 is 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.

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/share/AnnotSV/Annotations_Human/SVincludedInFt/1000g/GRCh37" and/or

“$ANNOTSV/share/AnnotSV/Annotations_Human/SVincludedInFt/1000g/GRCh38” directories.

- Download and place the VCF files in the
“$ANNOTSV/share/AnnotSV/Annotations_Human/SVincludedInFt/1000g/GRCh37”
and/or
“$ANNOTSV/share/AnnotSV/Annotations_Human/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 is executed after the update.

gnomAD SV frequency annotations

Aim:
A reference atlas of SV from deep WGS of 14,891 individuals across diverse global populations has been constructed as a component of the gnomAD database (Collins et al., 2019). The publicly available SV data represents a relatively diverse collection of unrelated individuals that should have rates of most severe diseases equivalent to, if not lower than, the general population.

Data sources:
Genome build GRCh37:
The gnomAD data are based on the genome build GRCh37/hg19. They can be freely downloaded at:
https://storage.googleapis.com/gnomad-public/papers/2019-sv/gnomad_v2_sv.sites.bed.gz

Genome build GRCh38:
The GRCh38 gnomAD SV dataset is not yet available from the gnomAD team.
However, the GRCh37 gnomAD SV dataset has been lifted over to GRCh38 with the UCSC web server and is provided as is by AnnotSV.

Method:
The DUP, DEL, INV and INS from gnomAD are reported.

Annotation columns:
Adds 7 annotation columns: “GD_ID”, “GD_AN”, “GD_N_HET”, “GD_N_HOMALT”, “GD_AF”, “GD_POPMAX” and “GD_ID_others”.

Concerning the 6 first columns, only the gnomAD SV with the same type as the SV to annotate are reported.
If no SVtype is provided for the SV to annotate, no gnomAD annotation is reported.

Concerning the frequencies (“GD_AF” and “GD_POPMAX”), only 1 value is reported (the most frequent one).

Ira M. Hall’s lab SV frequency annotations

Aim:

Ira M. Hall’s lab characterized SV in 17,795 deeply sequenced human genomes from common disease trait mapping studies (Abel et al., 2018). They publicly released SV frequency annotations to guide SV analysis and interpretation in the era of WGS.

Data sources:

Supplementary files 1 and 2 from (Abel et al., 2018) was downloaded. Outer breakpoints of duplications, deletions, inversions and mobile element insertions are used in AnnotSV annotations with GRCh37 and GRCh38 coordinates.

Method:

The DUP, DEL, INV and MEI from the IMH (Ira M. Hall’s lab) are reported.

Annotation columns:

Adds 3 annotation columns: “IMH_ID”, “IMH_AF” and “IMH_ID_others”.

Concerning the 2 first columns, only the IMH SV with the same type as the SV to annotate are reported.
If no SVtype is provided for the SV to annotate, no IMH annotation is reported.

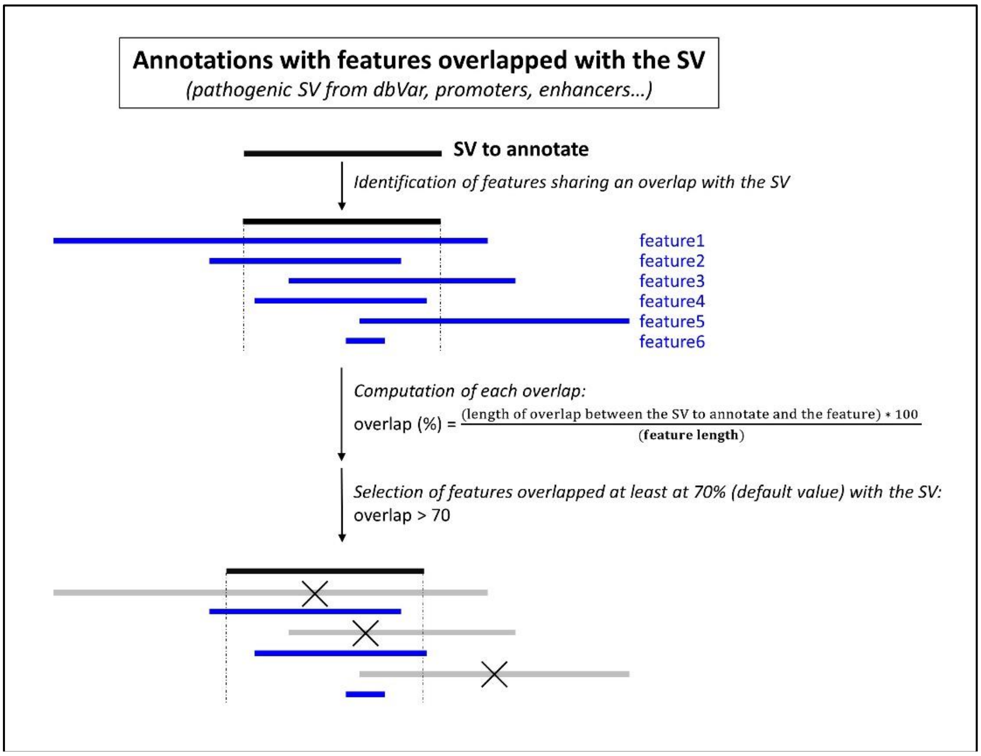
Concerning the “IMH_AF” frequencies, only 1 value is reported (the most frequent one).

c. Annotations with features **overlapped** with the SV

First, AnnotSV searches for features sharing an overlap with the SV to annotate. Second, only the features overlapped at least at 70% with the SV are selected (default value, a different percentage can also be user defined with the "overlap" option).

Interest of this computation:
For example, AnnotSV considers that a pathogenic SV is informative enough only if > 70% length of the pathogenic SV is overlapped with the SV to annotate. So, and only then, the SV to annotate can be considered as pathogenic.

It is to notice that, for this type of annotations, a reciprocal overlap cannot be used.


Annotations with features overlapped with the SV
(pathogenic SV from dbVar, promoters, enhancers…)
SV to annotate
Identification of features sharing an overlap with the SV
feature1
feature2
feature3
feature4
feature5
feature6
Computation of each overlap:
overlap (%) = (length of overlap between the SV to annotate and the feature) * 100 / (feature length)
Selection of features overlapped at least at 70% (default value) with the SV:
overlap > 70

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 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.

dbVarNR 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 (user defined, see the "overlap" option 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 “$ANNOTSV/share/AnnotSV/Annotations_Human/FtIncludedInSV/dbVar_pathogenic_NR_SV/GRCh37” and/or “$ANNOTSV/share/AnnotSV/Annotations_Human/FtIncludedInSV/dbVar_pathogenic_NR_SV/GRCh38” directories.
- Download and place the 2 following dbVar files in the “$ANNOTSV/share/AnnotSV/Annotations_Human/FtIncludedInSV/dbVar_pathogenic_NR_SV/GRCh37” and/or “$ANNOTSV/share/AnnotSV/Annotations_Human/FtIncludedInSV/dbVar_pathogenic_NR_SV/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 is 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 if i) the SV overlaps at least 70% of this TAD boundary (user defined, see the "overlap" option in USAGE/OPTIONS) or ii) if the SV is an insertion included in the TAD.

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 “$ANNOTSV/share/AnnotSV/Annotations_Human/FtIncludedInSV/TAD/GRCh37” and/or “$ANNOTSV/share/AnnotSV/Annotations_Human/TAD/GRCh38” directories.
- Download and place your ENCODE BED files in the “$ANNOTSV/share/AnnotSV/Annotations_Human/FtIncludedInSV/TAD/GRCh37” and/or “$ANNOTSV/share/AnnotSV/Annotations_Human/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 is executed.

GeneHancer annotations (not distributed)

Aim:

Enhancer and promoter genomic aberrations have been reported to underlie genetic diseases that represent a current challenge. For this, we include GeneHancer (Fishilevich et al., 2017),an integrated compendium of human promoters, enhancers and their inferred target genes.

WARNING:

GeneHancer data, as part of the GeneCards Suite, cannot be redistributed. Thus, GeneHancer annotation cannot be supplied as part of the AnnotSV sources. Users need to request the up-to-date GeneHancer data dedicated to AnnotSV (“GeneHancer_<version>_for_annotsv.zip“) by contacting directly the GeneCards team:

- Academic users: genecards@weizmann.ac.il
- Commercial users: support@lifemapsc.com

Method:

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

Annotation columns:

Adds 6 annotation columns (“GHid_elite”, “GHid_not_elite”, “GHtype”, “GHgene_elite”, “GHgene_not_elite” and “GHtissue”).

Installing the data source:

AnnotSV needs the “GeneHancer_<version>_for_annotsv.zip” file.

- Put the “GeneHancer_<version>_for_annotsv.zip” file in the following directory :
  “$ANNOTSV/share/AnnotSV/Annotations_Human/FtIncludedInSV/GeneHancer/
- Unzip this file:

```
cd “$ANNOTSV/share/AnnotSV/Annotations_Human/FtIncludedInSV/GeneHancer/
unzip GeneHancer_<version>_for_annotsv.zip
        Archive:  GeneHancer_<version>_for_annotsv.zip
        inflating: ReadMe.txt
        inflating: GeneHancer_elements.txt
        inflating: GeneHancer_gene_associations_scores.txt
        inflating: GeneHancer_hg19.txt
        inflating: GeneHancer_tissues.txt
```

  These files will be reprocessed and then removed the first time AnnotSV is 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 hot spot motif 5’-CCNCCNTNNCCNC-3’) is positively correlated with the frequency of non allelic 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 “$ANNOTSV/share/AnnotSV/Annotations_Human/BreakpointsAnnotations/GCcontent/GRCh37” and/or “$ANNOTSV/share/AnnotSV/Annotations_Human/BreakpointsAnnotations/GCcontent/GRCh38” directories.
- Download and place the human reference genome FASTA file in the “$ANNOTSV/share/AnnotSV/Annotations_Human/BreakpointsAnnotations/GCcontent/GRCh37”

and/or
“$ANNOTSV/share/AnnotSV/Annotations_Human/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 is 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 “$ANNOTSV/share/AnnotSV/Annotations_Human/BreakpointsAnnotations/Repeat/GRCh37” and/or “$ANNOTSV/share/AnnotSV/Annotations_Human/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 “$ANNOTSV/share/AnnotSV/Annotations_Human/BreakpointsAnnotations/Repeat/GRCh37” and/or “$ANNOTSV/share/AnnotSV/Annotations_Human/BreakpointsAnnotations/Repeat/GRCh38” directories.

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

4. Versions of the annotations sources

Annotations source	Version
...refGene annotations	
Gene annotations (refGene)	2019-12-19
ACMG	ACMG SF v2.0
...Genes-based annotations	
Haploinsufficiency and triplosensitivity Scores annotations (ClinGen)	2019-12-19
DDD gene annotations	2019-12-19
Haploinsufficiency annotations (DDD)	2019-12-19
Gene intolerance annotations (ExAC)	2016-01-14
Morbid genes annotations (OMIM)	2019-12-16
OMIM annotation	2019-12-16
Exomiser	2019-09-17
...Annotations with features overlapping the SV	
DGV Gold Standard annotations	2016-05-15
gnomAD (GRCh37)	2019-03-14
DDD frequency annotations	2019-03-18
1000 genomes frequency annotations (GRCh37)	2017-05-19
1000 genomes frequency annotations (GRCh38)	2017-11-05
Ira M. Hall's lab annotations	2018-12-31
...Annotations with features overlapped with the SV	
dbVar NR SV pathogenic annotations (GRCh37, GRCh38)	2019-12-19
GeneHancer annotations	Downloaded by the user
TAD boundaries annotations	2017-10-24
...Breakpoints annotations	
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 5classes proposed by the ACMG guidelines using the following data and criteria:

Data used for the ranking:

- Frequent SV from gnomAD (the ones with a GD_POPMAX_AF > 1%)
- 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)
- Enhancer and promoter elements from GeneHancer

Criteria:

- **Class 1 (benign):**
  The SV overlaps (>70%) with a frequent SV with the same SV type
  AND the SV does not overlap with a morbid gene (or its enhancer/promoter)
  AND the SV does not overlap with morbid gene candidate (or its enhancer/promoter)
  AND the SV does not overlap a candidate gene (or its enhancer/promoter)

- **Class 2 (likely benign):**
  The SV has no overlap OR an overlap≤70% with a benign SV
  AND the SV does not overlap with a morbid gene (or its enhancer/promoter)
  AND the SV does not overlap with a morbid gene candidate (or its enhancer/promoter)
  AND the SV does not overlap with a candidate gene (or its enhancer/promoter)

- **Class 3 (variant of unknown significance):**
  The SV overlaps a morbid gene candidate (or its enhancer/promoter) (with at least 1bp overlap)
  OR the SV overlaps a candidate gene (or its enhancer/promoter) (with at least 1bp overlap)

- **Class 4 (likely pathogenic):**
  The SV overlaps a morbid gene (or its enhancer/promoter) (with at least 1bp)
  OR for a loss: the SV overlaps a gene (or its enhancer/promoter) 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 (or its enhancer/promoter) with a 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

6. SV Type

In order to be able to classify the SV and to provide relevant annotations, 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>
- Duplication: DUP, duplication, gain, MCNV, <CN2>, <CN3>...
- Insertion: INS, insertion, ALU, LINE, SVA or MEI
- Inversion: INV or inversion
- Breakend record: BND, breakpoint, breakend

7. INPUT

AnnotSV takes several arguments as input including options that are detailed in section 5 ("USAGE / OPTIONS"). The different arguments can be passed either on the command line (priority) or using a specific file named "configfile". This configfile file needs to be located in the same directory as the INPUT file, an example of configfile is provided in the AnnotSV installation directory. Five 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 information:
- 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 outputfile (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 calls from any sequencing experiment as input to the command line. These annotations report the counts and ratio of homozygous and heterozygous SNV/indel identified from the patients NGS data (user defined samples) and presents in the interval of the **deletion** to annotate.

Usage:

The command line can be completed with the 2 following options: "-snvIndelFiles" and "-snvIndelSamples" (*cf* USAGE/OPTIONS).

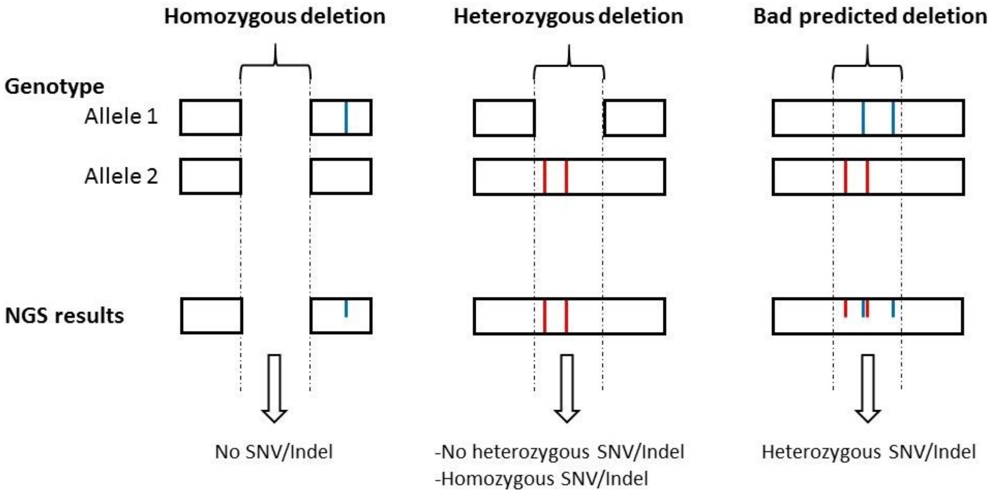
Annotation columns:

Add the "#hom(sample)", "#htz(sample)", "#htz/allHom(sample)", "#htz/total(cohort)" and "#total(cohort)" annotation columns.

- **#hom(sample):** Count of homozygous SNV/indel called from the sample and present in the interval of the deletion
- **#htz(sample):** Count of heterozygous SNV/indel called from the sample and present in the interval of the deletion
- **#allHom(sample):** Count of homozygous SNV/indel called from the sample, including homozygous WT SNV/indel (extracted from VCF input file, GT=0/0), and present in the interval of the deletion
- **#total(cohort):** Total count of SNV/indel called from the sample and present in the interval of the deletion

Aim:

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



- **Homozygous deletion:** No SNV/indel is expected in the region. 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. So we expect a zero “#htz/allHom(sample)” and “#htz/total(cohort)” ratio.

- **Heterozygous deletion:** All SNV/indel are expected to be homozygous. 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. So we expect small “#htz/allHom(sample)” and “#htz/total(cohort)” ratio. However, threshold for these ratio are dependant on sequencing protocols and calling/filtering strategies and can not be determined as a standard.

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

A deletion QC can be performed by checking both ratio, ONLY if:
- analysing a cohort VCF where all samples have been jointly called.
- there is a minimum number of SNV/indel located in the SV. So, AnnotSV reports these ratio only if #total(cohort) > 50 ; otherwise the ratio will be set to "NA" (not applicable).

The deletion QC do not apply to standard VCF for single sample, since homozygous reference positions are not usually reported.

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: “-candidateSnvIndelFiles” and “-candidateSnvIndelSamples” (*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.* Especially, when sequencing data of more than one family member is available, one can exclude certain variants based on the expected Mendelian inheritance (transmitted in a compound heterozygous mode from parents to the patient(s)). A specific feature (barcode) will be implemented soon for this.

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 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 GRCh37:

➔ "$ANNOTSV/share/AnnotSV/Annotations_Human/Users/GRCh37/FtIncludedInSV" directory
or
➔ "$ANNOTSV/share/AnnotSV/Annotations_Human/Users/GRCh37/SVincludedInFt" directory

Genome build GRCh38:

➔ "$ANNOTSV/share/AnnotSV/Annotations_Human/Users/GRCh38/FtIncludedInSV" directory
or
➔ "$ANNOTSV/share/AnnotSV/Annotations_Human/Users/GRCh38/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.

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.

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:

The*'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):

The *'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).

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/share/AnnotSV/Annotations_Human/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.

8. 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(s) and name(s)

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 "*'date'*_AnnotSV/mySVinputFile.annotated.tsv".

AnnotSV can create two other output files:

- A report of unannotated variants (".unannotated.tsv" file)
  Indeed, AnnotSV does not annotate variants from a VCF input file:
  - If the variant is an indel (variant length < SVminSize)
  - If the SV is not well formatted
  - If the "END" of the SV is not defined
- A report of the decisions that explain the ranking of each SV (see the "-rankOutput" option in USAGE/OPTIONS)

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 are2 types of lines provided 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
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")	X	X	X	X

	- annotation on each gene overlapped by the SV (“split”)				
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’s (e.g. « txStart-exon1 »)		X	X	X
location2	SV location in the gene’s coding regions (e.g. « 3’UTR-CDS »)		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
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
GD_ID	gnomAD IDs overlapping the annotated SV with the same SV type	X	X	X	X
GD_AN	gnomAD total number of alleles genotyped (for biallelic sites) or individuals with copy-state estimates (for multiallelic sites)	X	X	X	X
GD_N_HET	gnomAD number of individuals with heterozygous genotypes	X	X	X	X
GD_N_HOMALT	gnomAD number of individuals with homozygous alternate genotypes	X	X	X	X
GD_AF	Maximum of the gnomAD allele frequency (for biallelic sites) and copy-state frequency (for multiallelic sites)	X	X	X	X
GD_POPMAX	Maximum of the gnomAD maximum allele frequency across any population	X	X	X	X
GD_ID_others	Other gnomAD IDs overlapping the annotated SV (with a different SV type)	X	X	X	X
1000g_event	List of the 1000 genomes event types (e.g. DEL, DUP, <CN3>...)	X	X	X	X
1000g_AF	Estimated global allele frequency among the 1000g_event	X	X	X	X
1000g_max_AF	Highest observed allele frequency across all the 1000g populations	X	X	X	X
IMH_ID	Ira M. Hall’s lab IDs overlapping the annotated SV	X	X	X	X
IMH_AF	IMH Allele Frequency	X	X	X	X
IMH_ID_others	Other IMH IDs overlapping the annotated SV (with a different SV type)	X	X	X	X

promoters	List of the genes whose promoters are overlapped by the SV	X	X	X	X
dbVar_event					
dbVar_variant					
dbVar_status					
GHid_elite[3,4]	List of the GeneHancer (GH) IDs for each "elite" element overlapped with the annotated SV	X	X	X	X
GHid_not_elite[3,4]	List of the GeneHancer (GH) IDs for each "not elite" element overlapped with the annotated SV	X	X	X	X
GHtype[4]	Type of the overlapped GH element(s) (Enhancer or Promoter)	X	X	X	X
GHgene_elite[3,4]	List of the genes for which an "elite" element-gene relation was identified	X	X	X	X
GHgene_not_elite[3,4]	List of the genes for which a "not elite" element-gene relation was identified	X	X	X	X
GHtissue[3,4]	List of the tissues in which elements were identified	X	X	X	X
TADcoordinates[3]	Coordinates of the TAD whose boundaries overlapped with the annotated SV (boundaries included in the coordinates)	X		X	X
ENCODEexperiments[3]	ENCODE experiments used to define the TAD	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
ACMG	ACMG genes		X	X	X
HI_CGscore	ClinGen Haploinsufficiency Score	X	X	X	X
TriS_CGscore	ClinGen Triplosensitivity Score	X	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 Id		X	X	X
HI_DDDpercent	Haploinsufficiency ranks from DDD	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 considers 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
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

Mim Number	OMIM unique six-digit identifier	X	X	X	X
Phenotypes	e.g. Charcot-Marie-Tooth disease		X	X	X
Inheritance	e.g. AD (= "Autosomal dominant")[2]		X	X	X
EXOMISER_GENE_PHENO_SCORE	Exomiser score for how close each overlapped gene is to the phenotype	X	X	X	X
HUMAN_PHENO_EVIDENCE	Phenotypic evidence from Human model		X	X	X
MOUSE_PHENO_EVIDENCE	Phenotypic evidence from Mouse model		X	X	X
FISH_PHENO_EVIDENCE	Phenotypic evidence from Fish model		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 SNV/indel (extracted from VCF input file) in the individual “sample” which are presents: - in the deletion SV (“full” annotation) - between intersectStart and intersectEnd (“split” annotation)	X	X	X	X
#htz(sample)	Number of heterozygous SNV/indel (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/allHom(sample)	Ratio for QC filtering: #htz(sample)/#allHom(sample)[5]				
#htz/total(cohort)	Ratio for QC filtering: #htz(sample)/#total(cohort)				
#total(cohort)	Total count of SNV/indel called from the sample and present in the interval of the deletion				

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. The transcript selected by the user with the "-txFile" option is firstly 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*

[3] *Very large SV (e.g. 30Mb) can sometime overlap too many features locations. 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 features is already a problem, AnnotSV restrict the number of overlapping reported features to 20.*

[4] *GeneHancer data, as part of the GeneCards Suite, cannot be redistributed. Thus, GeneHancer annotation cannot be supplied as part of the AnnotSV sources. Users need to request the up-to-date GeneHancer data dedicated to AnnotSV by contacting the GeneCards team (see “GeneHancer annotations”)*

[5] *#allHom(sample): Count of homozygous SNV/indel called from the sample, including* homozygous WT SNV/indel (extracted from VCF input file, GT=0/0), and present in the interval of the deletion

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 («#»).

9. 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:

-annotationsDir: Path of the annotations directory

-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")

-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

-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"

-rankOutput: To save in an output file the decisions that explain the rank of each SV
Values: no (default) or yes

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

-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)
Range 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
Range values: 1 (default) or 0

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

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

-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)

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

10. Test

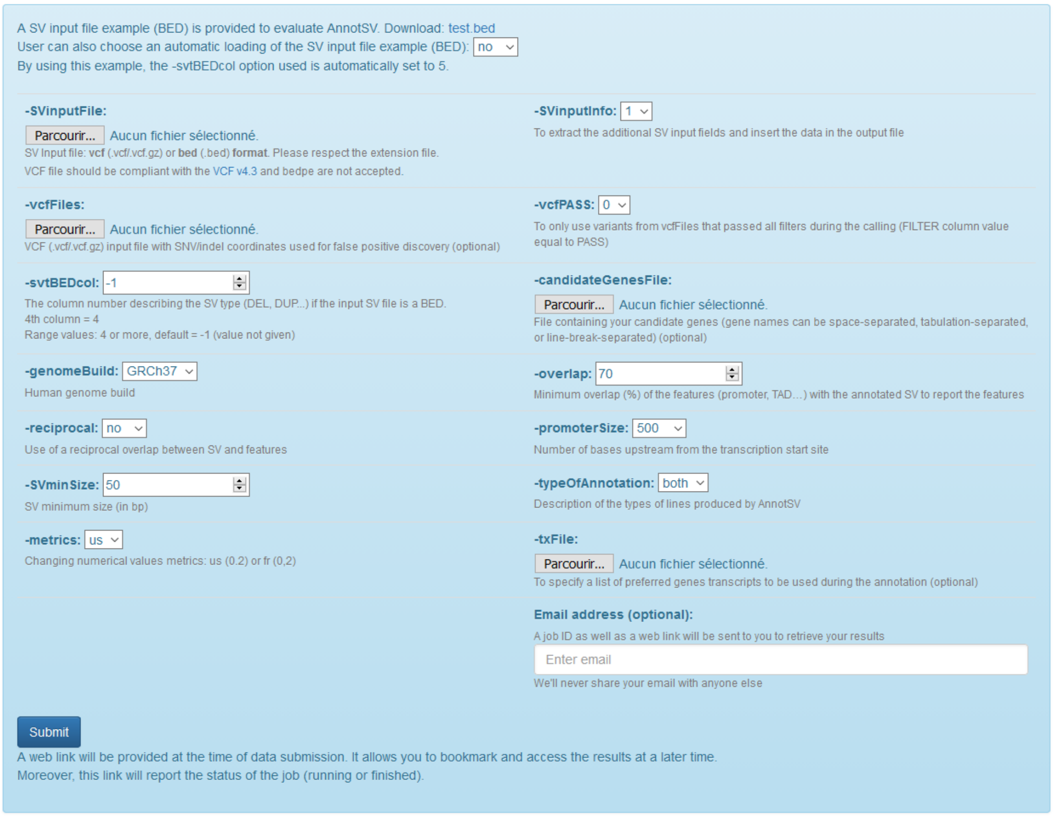
In order to validate the AnnotSV installation and its functioning, an example is available in the "$ANNOTSV/share/doc/AnnotSV/Example" directory. Command lines examples are available in the following file "$ANNOTSV/share/doc/AnnotSV/commands.README".

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

11. 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 SV input file example (BED) is provided to evaluate AnnotSV. Download: test.bed
User can also choose an automatic loading of the SV input file example (BED): no
By using this example, the -svtBEDcol option used is automatically set to 5.

-SVinputFile: Aucun fichier sélectionné.
SV Input file: **vcf** (.vcf/.vcf.gz) or **bed** (.bed) **format**. Please respect the extension file.
VCF file should be compliant with the VCF v4.3 and bedpe are not accepted.

-vcfFiles: Aucun fichier sélectionné.
VCF (.vcf/.vcf.gz) input file with SNV/indel coordinates used for false positive discovery (optional)

-svtBEDcol: -1
The column number describing the SV type (DEL, DUP...) if the input SV file is a BED.
4th column = 4
Range values: 4 or more, default = -1 (value not given)

-genomeBuild: GRCh37
Human genome build

-reciprocal: no
Use of a reciprocal overlap between SV and features

-SVminSize: 50
SV minimum size (in bp)

-metrics: us
Changing numerical values metrics: us (0.2) or fr (0,2)

-SVinputInfo: 1
To extract the additional SV input fields and insert the data in the output file

-vcfPASS: 0
To only use variants from vcfFiles that passed all filters during the calling (FILTER column value equal to PASS)

-candidateGenesFile: Aucun fichier sélectionné.
File containing your candidate genes (gene names can be space-separated, tabulation-separated, or line-break-separated) (optional)

-overlap: 70
Minimum overlap (%) of the features (promoter, TAD...) with the annotated SV to report the features

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

-typeOfAnnotation: both
Description of the types of lines produced by AnnotSV

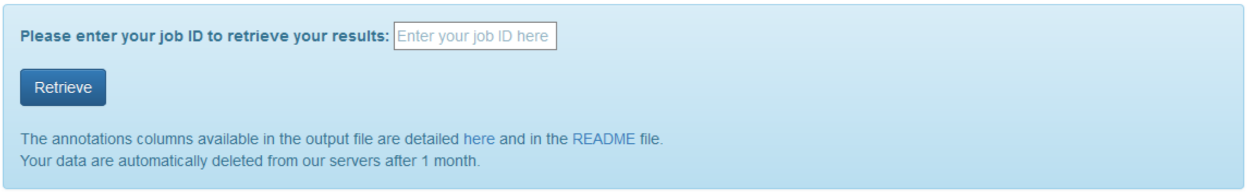
-txFile: Aucun fichier sélectionné.
To specify a list of preferred genes transcripts to be used during the annotation (optional)

Email address (optional):
A job ID as well as a web link will be sent to you to retrieve your results
Enter email
We'll never share your email with anyone else

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

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



Please enter your job ID to retrieve your results: Enter your job ID here

The annotations columns available in the output file are detailed here and in the README file.
Your data are automatically deleted from our servers after 1 month.

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

12. 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 format 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

And if you use the phenotype-driven analysis in your work, please cite also the following articles:

- Next-generation diagnostics and disease-gene discovery with the Exomiser. Smedley D., *et al*, Nature Protocols (2015) doi:10.1038/nprot.2015.124
- Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Köhler S., *et al*, Nucleic Acids Research (2019) doi: 10.1093/nar/gky1105

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 or set to -1 in the output files?
When no information is available for a specific type of annotation, then the value is empty. Regarding the frequencies, the default is set to -1.

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: Why does AnnotSV only report overlapping SV (from gnomAD, IHM…) with the same type?
Because reporting more and more columns is problematic, we decided to report more precisely the information of the same type of SV as the one in question (e.g. a duplication with a duplication, a deletion with a deletion …). However, to keep the user aware with different type of rearrangements overlapping the SV to annotate, the ID of such events are reported in a specific annotation column (e.g. GD_ID_others, IMH_ID_others…)

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 exclude/imply a disease-causing event. Healthy carriers of pathogenic SV do exist in either databases. 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. By default, all the annotations are based on human specific databases. Nevertheless, some additional annotation files can be added for mouse. If you are interested, please see the specific mouse README file.

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 am 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

Q: I am getting the error: "-- max size for a Tcl value (2147483647 bytes) exceeded". How can I fix this problem?

You are probably using AnnotSV to annotate a very large SV input file (from a large cohort). Thus you are facing a memory issue either caused by the current machine specification or the programming language used for AnnotSV (Tcl). To solve this you can split your input file into smaller files, run AnnotSV and then later merge them into a single output file. This will be fixed in a future release.

Q: For a VCF with only "BND" events, which refers to breakpoints, how are these being shown in the AnnotSV output when SVminSize is set to 50bp? Since a breakpoint start and stop positions only differ by 1bp, I am wondering why these are not filtered out by AnnotSV.

AnnotSV is designed to annotate SV and not SNV/indel from a VCF, which is the aim of the "SVminSize" option. Actually, SV can be described in three different ways in a VCF file:
- Type1: ref="G" and alt="ACTGCTAACGATCCGTTTGCTGCTAACGATCTAACGATCGGGATTGCTAACGATCTCGGG" (length >SVminSize)
- Type2: alt="<INS>", "<DEL>", "<BND>"...
- Type3: complex rearrangements with breakends: alt="G]17:1584563]"

The "SVminSize" parameter is only used to exclude SNV/indel from the SV of Type1.

Q: How is calculated the "SV length" annotation?

- AnnotSV reports the "SVLEN" value if given in a VCF input file.
- Nevertheless, when it is not provided, AnnotSV calculates the SV length (with "alt length" - "ref length") depending on the description of it in a VCF input file: ref="G" and alt="ACTGCTAACGATCCGTTTGCTGCTAACGATCTAACGATCGGGATTGCTAATCTCGGG"
- Else, AnnotSV calculates the SV length only for deletion, duplication and inversion (with "SVend - SVstart", and with a negative value for deletion). Indeed, this calculation cannot be done for insertion, breakend, translocation…
- Else, the SV length is blank.

Q: What does the candidateGenesFile parameter refer to?

The candidateGenesFile contains the candidate genes of the user. This information is used:
- To improve the ranking of the SV (see the "SV RANKING/CLASSIFICATION" section)
- To filter out the SV annotations that do not overlap a candidate gene (-candidateGenesFiltering yes). In this configuration, only "split" annotations can be reported.

Q: My input bed file contains ~10000 SV, but only ~2000 SV are annotated. Why?

AnnotSV does not annotate:
- the SNV/indel (size<50bp)
- the SV in a bad format
- the SV for which the "END" is not defined

If you want to annotate SNV/indel, please set the -SVminSize to 1.

Q: How overlaps (%) are calculated?

AnnotSV provides different types of annotations:

- An annotation with features **overlapping** the SV (DGV, 1000 genomes…):

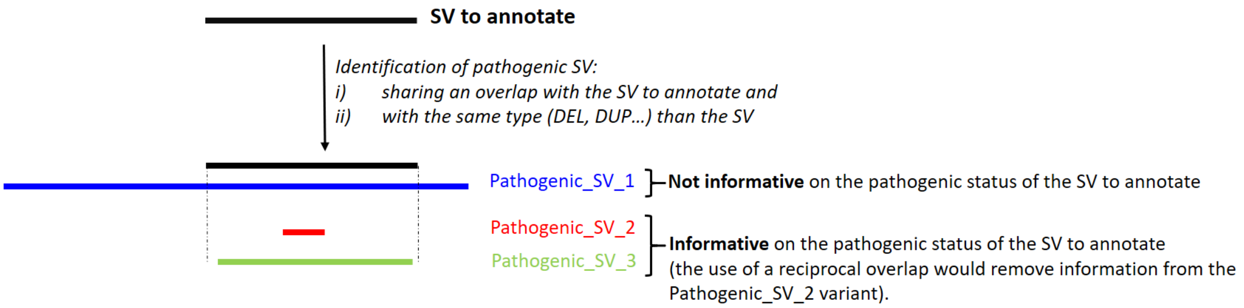
$$\text{overlap (\%)} = \frac{(\text{length of overlap between the SV to annotate and the feature}) * 100}{(\text{SV to annotate length})}$$

- An annotation with features **overlapped** with the SV (pathogenic SV from dbVar, promoters, enhancers…):

$$\text{overlap (\%)} = \frac{(\text{length of overlap between the SV to annotate and the feature}) * 100}{(\text{feature length})}$$

- A gene-based annotations
Each gene overlapped by the SV to annotate is reported (even with 1bp overlap).

Q: Why not to use a reciprocal overlap with features overlapped with the SV to annotate?
Let’s take the example of pathogenic SV as features.


SV to annotate
Identification of pathogenic SV:
i) sharing an overlap with the SV to annotate and
ii) with the same type (DEL, DUP…) than the SV
Pathogenic_SV_1
Not informative on the pathogenic status of the SV to annotate
Pathogenic_SV_2
Pathogenic_SV_3
Informative on the pathogenic status of the SV to annotate
(the use of a reciprocal overlap would remove information from the Pathogenic_SV_2 variant).

=> AnnotSV would lose some information if using a reciprocal overlap.

Q: What are the minimal info/headers needed in a VCF input file to run AnnotSV?
AnnotSV is using the VCF format following official specification VCF v4.3 (Jul 2017). Nevertheless, some flexibility is allowed:
- No meta-information line (prefixed with “##”) is required
But the following is mandatory:
- A header line (prefixed with “#CHROM”)
- The following INFO keys are required: GT, SVLEN, END and SVTYPE.

In order to be able to classify the SV, the "SVTYPE" values should be one of DEL, INS, DUP, INV, CNV, BND, LINE1, SVA, ALU. In addition, the <CN0>, <CN2>, <CN3>... angle-bracketed ID from the "ALT" column should be used in case of SVTYPE=CNV in the INFO column.

In order to use the “snvIndelPASS” option (using of the variants only if they passed all filters during the calling), the FILTER column value is mandatory.

Q: I’m getting the error: “ERROR: chromosome sort ordering for file … is inconsistent with other files”. How can I fix this problem?

The locale specified by your environment can affect the traditional “sort” order that uses native byte values.
Please, set LC_ALL=C.
In csh, you can define it with the following command line:

```
setenv LC_ALL C
```

In bash:

```
export LC_ALL=C
```

13. 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.

Abel, H.J., Larson, D.E., Chiang, C., Das, I., Kanchi, K.L., Layer, R.M., Neale, B.M., Salerno, W.J., Reeves, C., Buyske, S., et al. (2018). Mapping and characterization of structural variation in 17,795 deeply sequenced human genomes. BioRxiv 508515.

Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alföldi, J., Khera, A.V., Francioli, L.C., Gauthier, L.D., Wang, H., Watts, N.A., et al. (2019). An open resource of structural variation for medical and population genetics. BioRxiv 578674.

Dittwald, P., Gambin, T., Szafranski, P., Li, J., Amato, S., Divon, M.Y., Rodríguez Rojas, L.X., Elton, L.E., Scott, D.A., Schaaf, C.P., et al. (2013). NAHR-mediated copy-number variants in a clinical population: mechanistic insights into both genomic disorders and Mendelizing traits. Genome Res. *23*, 1395–1409.

Firth, H.V., Wright, C.F., and DDD Study (2011). The Deciphering Developmental Disorders (DDD) study. Dev Med Child Neurol *53*, 702–703.

Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., Rosen, N., Kohn, A., Twik, M., Safran, M., et al. (2017). GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database (Oxford) *2017*.

Hamosh, A., Scott, A.F., Amberger, J., Valle, D., and McKusick, V.A. (2000). Online Mendelian Inheritance in Man (OMIM). Hum. Mutat. *15*, 57–61.

Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J.O.B., Danis, D., Gourdine, J.-P., Gargano, M., Harris, N.L., Matentzoglu, N., McMurry, J.A., et al. (2019). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res. *47*, D1018–D1027.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Lupiáñez, D.G., Spielmann, M., and Mundlos, S. (2016). Breaking TADs: How Alterations of Chromatin Domains Result in Disease. Trends Genet. *32*, 225–237.

MacDonald, J.R., Ziman, R., Yuen, R.K.C., Feuk, L., and Scherer, S.W. (2014). The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. *42*, D986-992.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, 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. Genet. Med. *17*, 405–424.

Smedley, D., Jacobsen, J.O.B., Jager, M., Köhler, S., Holtgrewe, M., Schubach, M., Siragusa, E., Zemojtel, T., Buske, O.J., Washington, N.L., et al. (2015). Next-generation diagnostics and disease-gene discovery with the Exomiser. Nat Protoc *10*, 2004–2015.

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H.-Y., et al. (2015). An integrated map of structural variation in 2,504 human genomes. Nature *526*, 75–81.