AnnotSV Manual

Version 3.0.7
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
OMIM: Online Mendelian Inheritance in Man
pLI: score indicating 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
TS: 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**.

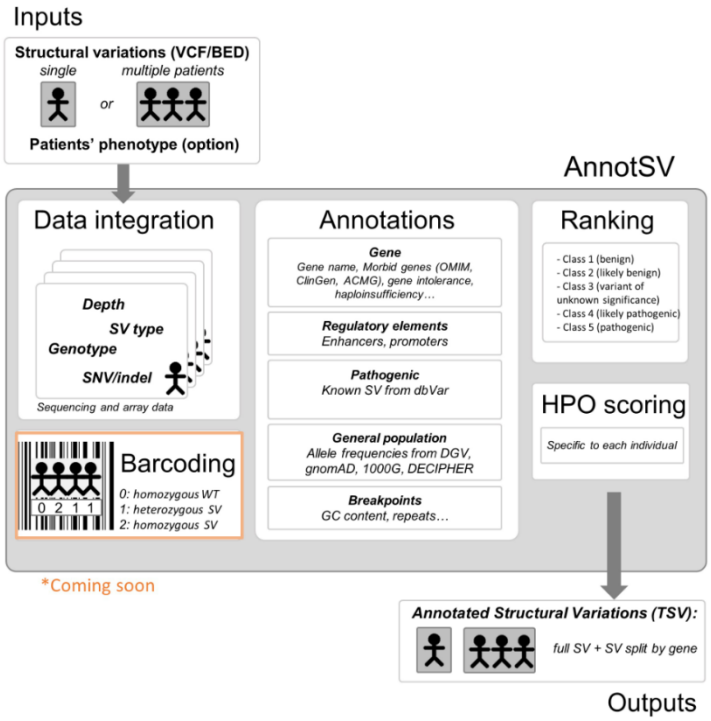
Different types of SV exist including deletions, duplications, insertions, inversions, translocations or more complex rearrangements. They can be either balanced or unbalanced. When unbalanced and resulting in a gain or loss of material, they are called Copy Number Variations (CNV). CNV can be described 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 (Jun 2020).

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:

- Gene-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) (Richards et al., 2015; Riggs et al., 2020).



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 to specifically report the information of the corresponding SV type.

Ex: A deletion of interest will be annotated with pathogenic SV using only the deletion data in details.

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 4 packages “http”, “json”, "tar" and "csv".
The “http” and the “json” packages are 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) Bcftools (required)

The **“bcftools”** toolset (Li, 2011) needs to be locally installed if using VCF input file(s).

Add the path of the bcftools 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 {$PATH}:/’somewhere’/bcftools-1.9/bin
- In bash, you can define it with the following command line:
  export PATH=$PATH:/’somewhere’/bcftools-1.9/bin

Warning: the minimum bcftools version compatible with AnnotSV is version 1.10. To check if bcftools exists and if the version is the good one, run:

bcftools --version

d) Java (optional)

In order to use the phenotype-driven analysis based on one Exomiser module, a minimal Java 8 installation is required.
Moreover, the Exomiser module writes in the /tmp/spring.log file that must, therefore, have write permissions.

e) 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 git@github.com:lgmgeo/AnnotSV.git
```

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

f) 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 (Genes, 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. Two exceptions need to be noticed with specific licence required in case**

the GeneHancer and/or the COSMIC resources are of any interest to you. 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.

IMPORTANT NOTE: To update the data sources, please download the latest available files (not the ones given as an example in the README).

a) Gene annotations

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

Aim:
The "Gene annotation" aims at providing information for the overlapping known genes with the SV. This will result in gene list from the well annotated RefSeq or ENSEMBL databases. These annotations include the definition of the genes and corresponding RefSeq transcripts from NCBI (default value). Transcripts from ENSEMBL can be user defined with the "-transcript" option, (see in USAGE/OPTIONS). This will also integrate 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:
Add 15 annotation columns: "Gene_name", "Gene_count", "Tx", "Tx_start", "Tx_end", "Overlapped_tx_length", "Overlapped_CDS_length", "Frameshift", "Exon_count", "Location", "Location2", "Dist_nearest_SS", "Nearest_SS_type", "Intersect_start", "Intersect_end".

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

Updating the data source from RefSeq (if needed):

- Remove the "genes.RefSeq.sorted.bed" file in the "$ANNOTSV/share/AnnotSV/Annotations_Human/Genes/GRCh37" and/or "$ANNOTSV/share/AnnotSV/Annotations_Human/Genes/GRCh38" directories.
- Download and place the "refGene.txt.gz" file in the "$ANNOTSV/share/AnnotSV/Annotations_Human/Genes/GRCh37" and/or "$ANNOTSV/share/AnnotSV/Annotations_Human/Genes/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).

Updating the data source from ENSEMBL (if needed):

- Remove the “genes.ENSEMBL.sorted.bed” file in the “$ANNOTSV/share/AnnotSV/Annotations_Human/Genes/GRCh37” and/or “$ANNOTSV/share/AnnotSV/Annotations_Human/Genes/GRCh38” directories.
  Genome build GRCh37:
  bash
  cd $ANNOTSV/share/AnnotSV/Annotations_Human/Genes/GRCh37/
  wget http://ftp.ensembl.org/pub/release-75/gtf/homo_sapiens/Homo_sapiens.GRCh37.75.gtf.gz
  wget http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/gtfToGenePred
  chmod +x gtfToGenePred
  gunzip Homo_sapiens.GRCh37.75.gtf.gz
  ./gtfToGenePred -genePredExt -geneNameAsName2 Homo_sapiens.GRCh37.75.gtf refGene.txt
  for i in 1 10 11 12 13 14 15 16 17 18 19 2 20 21 22 3 4 5 6 7 8 9 M MT X Y; do \
  awk -v chr=$i '$2 ==chr {print $2"\t"$4"\t"$5"\t"$3"\t"$12"\t"$1"\t"$6"\t"$7"\t"$9"\t"$10}' \
  refGene.txt | sed 's/^MT/M/' | sort -k1,1 -k2,2n -k3,3n >> genes.ENSEMBL.sorted.bed; done
  rm gtfToGenePred Homo_sapiens.GRCh37.75.gtf refGene.txt

  Genome build GRCh38:
  cd $ANNOTSV/share/AnnotSV/Annotations_Human/Genes/GRCh38/
  wget http://ftp.ensembl.org/pub/current_gtf/homo_sapiens/Homo_sapiens.GRCh38.102.chr.gtf.gz
  wget http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/gtfToGenePred
  chmod +x gtfToGenePred
  gunzip Homo_sapiens.GRCh38.100.gtf.gz
  ./gtfToGenePred -genePredExt -geneNameAsName2 Homo_sapiens.GRCh38.100.gtf refGene.txt
  for i in 1 10 11 12 13 14 15 16 17 18 19 2 20 21 22 3 4 5 6 7 8 9 M MT X Y; do \
  awk -v chr=$i '$2 ==chr {print $2"\t"$4"\t"$5"\t"$3"\t"$12"\t"$1"\t"$6"\t"$7"\t"$9"\t"$10}' \
  refGene.txt | sed 's/^MT/M/' | sort -k1,1 -k2,2n -k3,3n >> genes.ENSEMBL.sorted.bed; done
  rm gtfToGenePred Homo_sapiens.GRCh38.100.gtf.gz refGene.txt

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

b) Regulatory Elements annotations

Aim:
The contribution of SV affecting promoters/enhancers 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.

Method:
AnnotSV reports the list of the genes whose promoters/enhancers are overlapped (even with 1bp overlap) by the SV. When available, the regulated gene name is detailed with associated haploinsufficiency (HI), triplosensitivity (TS) and exomiser (EX) scores as well as OMIM morbid annotation.

Sources:

- **Gene data (RefSeq or ENSEMBL):** Promoters are defined by default as 500 bp upstream from the transcription start sites of the RefSeq or ENSEMBL databases (see the "-transcript" option in

USAGE/OPTIONS). Nevertheless, the user can define a different bp size with the "promoterSize" option (see USAGE/OPTIONS).

- **EnhancerAtlas 2.0:** Enhancers are reported from EnhancerAtlas 2.0. Based on the enhancer consensus and gene expression data, EnhancerAtlas 2.0 predicted the target genes of enhancers for many tissue/cell types in human.
- **GeneHancer** (Fishilevich et al., 2017)**:** Promoters and enhancers are reported from four different databases: the Encyclopedia of DNA Elements (ENCODE), the Ensembl regulatory build, the functional annotation of the mammalian genome (FANTOM) project and the VISTA Enhancer Browser.

WARNING:

GeneHancer data is under a specific licence that prevent the systematic availability in 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

Annotation columns:

Add 1 annotation column (only in the "full" lines): "RE_gene".

NOTE:

- Depending on the "tx" option setting, either RefSeq (default) or ENSEMBL gene name are reported.
- To come back to the regulatory elements coordinates and sources, the user can set the "REreport" option to "yes" (default = "no") which will allow to report this information in an "*.SV_RE_intersect.tmp" output file.

Updating the data source (if needed):

- *Remove all the files* in the "$ANNOTSV/share/AnnotSV/Annotations_Human/FtIncludedInSV/RegulatoryElements/GRCh37" and/or "$ANNOTSV/share/AnnotSV/Annotations_Human/FtIncludedInSV/RegulatoryElements/GRCh38" directories.
- *Gene data:* Promoters will be automatically updated by using the Gene annotations.
- *Download EnhancerAtlas files:* You can freely download the GRCh37 EnhancerAtlas TXT files from http://www.enhanceratlas.org/downloadv2.php. Click the "Download enhancer-gene interactions" section. Download the 114 human files (*_EP.txt) in the "$ANNOTSV/share/AnnotSV/Annotations_Human/FtIncludedInSV/RegulatoryElements/GRCh37".
  To do this:

```
tclsh

foreach tissuOrCell {A375 A549 AML_blast Astrocyte BJ Bronchia_epithelial
Caco-2 Calu-3 CD14+ CD19+ CD20+ CD34+ CD36+ CD4+ CD8+ Cerebellum CUTLL1 DOHH2
ECC-1 ESC_neuron Esophagus Fetal_heart Fetal_kidney Fetal_muscle_leg
Fetal_placenta Fetal_small_intestine Fetal_spinal_cord Fetal_stomach
Fetal_thymus FT246 FT33 GM10847 GM12878 GM12891 GM12892 GM18505 GM18526
GM18951 GM19099 GM19193 GM19238 GM19239 GM19240 H1 H9 HCC1954 HCT116 HEK293T
HEK293 Hela-S3 Hela HepG2 HFF HL-60 hMADS-3 HMEC hNCC HSMM HT1080 HT29 HUVEC
IMR90 Jurkat K562 Kasumi-1 KB Keratinocyte Left_ventricle LHCN-M2 Liver LNCaP-
abl LNCaP Lung MCF-7 MCF10A ME-1 Melanocyte melanoma Mesendoderm MS1 Myotube
Namalwa NB4 NHDF NHEK NHLF NKC OCI-Ly7 Osteoblast Ovary PANC-1 Pancreas
Pancreatic_islet PBMC PC3 PrEC SGBS_adipocyte SK-N-SH SK-N-SH_RA
```

```
Skeletal_muscle Small_intestine Sperm Spleen T47D T98G th1 Thymus U2OS VCaP
ZR75-30} {
   catch                    {eval                    exec                    wget
http://www.enhanceratlas.org/data/AllEPs/hs/${tissuOrCell}_EP.txt}
}
```

These GRCh37 files will be computed then removed the first time AnnotSV is executed after the update. After processing, you need to lift over the resulting GRCh37 file to GRCh38 with the UCSC web server and to move it in the "$ANNOTSV/share/AnnotSV/Annotations_Human/FtIncludedInSV/RegulatoryElements/GRCh38" directory.

- GeneHancer data: Put the "GeneHancer_<version>_for_annotsv.zip" file in the following directory: "$ANNOTSV/share/AnnotSV/Annotations_Human/FtIncludedInSV/RegulatoryElements"

  Unzip this GRCh37/GRCh38 file:

```
cd $ANNOTSV/share/AnnotSV/Annotations_Human/
cd FtIncludedInSV/RegulatoryElements/
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 computed then removed the first time AnnotSV is executed after the update.

c) Gene-based annotations

These annotations are linked to the **gene name** and thus provided independently of the genome build.

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:

Add 5 annotation columns (only in the "split" lines): "DDD_status", "DDD_mode", "DDD_consequence", "DDD_disease", "DDD_pmid".

Updating the data source (if needed):

- Remove all the ***DDG2P*** files in the "$ANNOTSV/share/AnnotSV/Annotations_Human/Gene-based/DDD" directory.
- Download and place the "**DDG2P.csv.gz**" DECIPHER file in the "$ANNOTSV/share/AnnotSV/Annotations_Human/Gene-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 3 annotation columns: “OMIM_ID”, “OMIM_morbid” and "OMIM_morbid_candidate".
Add 2 other annotation columns (only in the "split" lines): “OMIM_phenotype” and “OMIM_inheritance”.

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 number in parentheses after the name of each disorder is set to (3) or (4) AND 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.

Updating the data source (if needed):

- Remove all the files in the “$ANNOTSV/share/AnnotSV/Annotations_Human/Gene-based/OMIM” directory.
- Download and place the “**genemap2.txt**” and “**morbidmap.txt**” OMIM files in the “$ANNOTSV/share/AnnotSV/Annotations_Human/Gene-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

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 (Richards 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 (gnomAD)

Aim:
Gene intolerance annotations from the gnomAD dataset give the significant deviation from the observed and the expected number of variants for each gene.

pLI is a score indicating the probability that a gene is intolerant to a loss of function variation (Nonsense, splice acceptor/donor variants due to SNV/indel). A gene with pLI>=0.9 is considered as an extremely LoF intolerant gene.

LOEUF stands for the "loss-of-function observed/expected upper bound fraction."

- Low LOEUF scores (e.g. 0) indicate strong selection against predicted loss-of-function (pLoF) variation in a given gene
- High LOEUF scores (e.g. 9) suggest a relatively higher tolerance to inactivation.

LOEUF advantage over pLI is that it can be used as a continuous value rather than a dichotomous scale (e.g. pLI > 0.9) - if such a single cutoff is still desired, pLI is a perfectly fine metric to use. At large sample sizes, the observed/expected ratio will be a more appropriate measure for selection, but at the moment, LOEUF provides a good compromise of point estimate and significance measure.

Annotation columns:
Add 3 annotation columns: “LOEUF_bin”, “GnomAD_pLI” and “ExAC_pLI”.

Updating the data source (if needed):

- Remove the file in the “$ANNOTSV/share/AnnotSV/Annotations_Human/Gene-based/gnomAD” directory.
- Download, uncompress and place the “**gnomad.v2.1.1.lof_metrics.by_gene.txt.bgz**” gnomAD file in the “$ANNOTSV/share/AnnotSV/Annotations_Human/Gene-based/gnomAD” directory. The latest update of this file is available for free download in the "pLoF Metrics by Gene TSV" section at:
  https://gnomad.broadinstitute.org/downloads#v2-constraint
  This file will be computed the first time AnnotSV is executed after the update.

This annotation is genome build independent, and only based on the gene name.
For genes with several transcripts, the maximal “LOEUF_bin” score is reported.

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.
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:
Add 5 annotation columns: “ExAC_synZ”, “ExAC_misZ”, "ExAC_delZ", "ExAC_dupZ" and “ExAC_cnvZ”.

Updating the data source (if needed):

- Remove all the files in the “$ANNOTSV/share/AnnotSV/Annotations_Human/Gene-based/ExAC” directory.
- Download and place the “**fordist_cleaned_nonpsych_z_pli_rec_null_data.txt**” and the “**exac-final-cnv.gene.scores071316**” ExAC files in the “$ANNOTSV/share/AnnotSV/Annotations_Human/Gene-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: “DDD_HI_percent”.

Updating the data source (if needed):

- Remove the “*_HI.tsv.gz“ file in the “$ANNOTSV/share/AnnotSV/Annotations_Human/Gene-based/DDD” directory.
- Download and place the “**HI_Predictions_Version3.bed.gz**” DECIPHER file in the “$ANNOTSV/share/AnnotSV/Annotations_Human/Gene-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" and "TS".
Concerning annotations on the **"full"** length of SV covering several genes, only the most pathogenic score is reported if any.

Updating the data source (if needed):

- Remove all the files in the "$ANNOTSV/share/AnnotSV/Annotations_Human/Gene-based/ClinGen/" directory.
- Download and place the "**ClinGen_gene_curation_list_GRCh37.tsv**" ClinGen file in the "$ANNOTSV/share/AnnotSV/Annotations_Human/Gene-based/ClinGen/" directory. The latest update of this file is available for free download at: ftp://ftp.clinicalgenome.org/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 powered by Exomiser

Aim:

To score genes overlapped with a SV on biological relevance to the individual phenotype, AnnotSV rely on 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.

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 need to be as specific as possible.

According to our own (limited) experience, a known disease 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.

If not provided, the Exomiser_gene_pheno_score is set to "-1.0". For SV overlapping several genes, the highest Exomiser_gene_pheno_score is reported in the full annotation.

Updating the data source (if needed):

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

- Remove all the files in the "$ANNOTSV/share/AnnotSV/Annotations_Human/Gene-based/NCBIgeneID/" directory.
- Download and place your NCBI gene ID file ("results.txt") in the "$ANNOTSV/share/AnnotSV/Annotations_Human/Gene-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" / "HGNC data" section:
  - Select only the "Approved symbol", the "Alias symbol" and the "Previous symbol".

  In the "Attributes"/"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.

Exomiser data can be updated (e.g. with the 2003 version):

```
cd $ANNOTSV/share/AnnotSV/Annotations_Exomiser/
mkdir 2003/2003_hg19
cd 2003/2003_hg19
cp $ANNOTSV/share/AnnotSV/Annotations_Exomiser/1902/1902_hg19/1902_hg19_genome.h2.db 2003_hg19_genome.h2.db
cp $ANNOTSV/share/AnnotSV/Annotations_Exomiser/1902/1902_hg19/1902_hg19_transcripts_ensembl.ser 2003_hg19_transcripts_ensembl.ser
cp $ANNOTSV/share/AnnotSV/Annotations_Exomiser/1902/1902_hg19/1902_hg19_variants.mv.db 2003_hg19_variants.mv.db
cd ..
```

wget https://data.monarchinitiative.org/exomiser/data/2003_phenotype.zip
unzip 2003_phenotype.zip
rm 2003_phenotype.zip 2003_phenotype/2003_phenotype.sha256

Then, check the $ANNOTSV/etc/AnnotSV/application.properties file are pointing to the correct versions:
exomiser.phenotype.data-version=2003
exomiser.hg19.data-version=2003

d) Known pathogenic genes or genomic regions annotation

AnnotSV searches for known pathogenic genes or genomic regions completely overlapped with the SV to annotate.

Aim:
According to the ACMG technical standards (Riggs et al., 2020), a SV completely overlapping an established pathogenic CNV region would be classified as pathogenic (if sharing the same SV type).

Annotation columns:
Add 12 annotation columns:
"P_gain_phen", "P_gain_hpo", "P_gain_source", "P_gain_coord",
"P_loss_phen", "P_loss_hpo", "P_loss_source", "P_loss_coord",
"P_ins_phen", "P_ins_hpo", "P_ins_source", "P_ins_coord",

Pathogenic dataset creation:
For each SV type (Loss, Gain, Ins, Inv), different sources of pathogenic genes or genomic regions have been merged in AnnotSV:

- ClinVar
- ClinGen
- dbVar
- OMIM

Updating the data source (if needed):

- Remove all the files in the following directories:
  "$ANNOTSV/share/AnnotSV/Annotations_Human/FtIncludedInSV/PathogenicSV/GRCh37"
  "$ANNOTSV/share/AnnotSV/Annotations_Human/FtIncludedInSV/PathogenicSV/GRCh38"
- Download and place the files of the different sources in the following directories:
  "$ANNOTSV/share/AnnotSV/Annotations_Human/FtIncludedInSV/PathogenicSV/GRCh37"
  "$ANNOTSV/share/AnnotSV/Annotations_Human/FtIncludedInSV/PathogenicSV/GRCh38"

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

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

ClinVar pathogenic SV annotations

Aim:

ClinVar gives access to the relationships asserted between human variants and observed health status.

Method:

Pathogenic SV are selected based on the following criteria:

- "pathogenic" or "pathogenic/likely pathogenic" clinical significance (CLNSIG)
- "criteria_provided", "_multiple_submitters" or "reviewed_by_expert_panel" SV review status (CLNREVSTAT)
- "Deletion" or "Duplication" SV type (CLNVC)
- ≥ 50 bp in size.

Source files:

The latest update of the ClinVar files are available for free download at:

Genome build GRCh37:

https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/clinvar_20201212.vcf.gz

Genome build GRCh38:

https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/clinvar_20201212.vcf.gz

Dosage sensitive genes/regions annotation (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.

A haploinsufficiency (HI) score of 3 suggests the gene/region to be dosage sensitive for a loss, associated with clinical phenotype.

A triplosensitivity (TS) score of 3 suggests the gene/region to be dosage sensitive for a gain, associated with clinical phenotype.

Method:

Genes and regions with a score of 3 are selected.

Source files:

The latest update of the ClinGen files are available for free download at:

Genome build GRCh37:

ftp://ftp.clinicalgenome.org/ClinGen_gene_curation_list_GRCh37.tsv

ftp://ftp.clinicalgenome.org/ClinGen_region_curation_list_GRCh37.tsv

Genome build GRCh38:

ftp://ftp.clinicalgenome.org/ClinGen_gene_curation_list_GRCh38.tsv

ftp://ftp.clinicalgenome.org/ClinGen_region_curation_list_GRCh38.tsv

dbVarNR pathogenic SV 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. Nevertheless, associated phenotypes are not provided.

Source files:
The latest update of the pathogenic dbVarNR files are available for free download at:

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

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

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

OMIM morbid genes

Aim:
The complete deletion of a morbid gene would be classified as pathogenic.

Method:
The “morbidGenes” are selected and only added to the pathogenic loss SV dataset in AnnotSV.

Source files:
The latest update of the OMIM morbid gene should have been done during the OMIM Gene-based annotation (see section "3.c Gene-based annotation").

Genome build GRCh37 and GRCh38:
To update the known pathogenic loss SV with the morbid gene, run the following commands:

```
cd $ANNOTSV/share/AnnotSV/Annotations_Human/FtIncludedInSV/PathogenicSV/GRCh37
cp $ANNOTSV/share/AnnotSV/Annotations_Human/Gene-based/OMIM/*_morbid.tsv.gz .
cd $ANNOTSV/share/AnnotSV/Annotations_Human/FtIncludedInSV/PathogenicSV/GRCh38
cp $ANNOTSV/share/AnnotSV/Annotations_Human/Gene-based/OMIM/*_morbid.tsv.gz .
```

The OMIM morbid gene coordinates are computed through their gene names and the AnnotSV gene annotations.

e) Known pathogenic SNV/indel annotations

AnnotSV searches for pathogenic SNV/indel from ClinVar completely overlapped with the SV to annotate.

Aim:
Pathogenic variants indicate the region is critical to protein function.

Method:
Pathogenic SNV/indel with all the following requirements are selected:

- "pathogenic" or "pathogenic/likely pathogenic" clinical significance (CLNSIG)
- "criteria_provided", "_multiple_submitters" or "reviewed_by_expert_panel" SV review status (CLNREVSTAT)
- < 50 bp in size.

Annotation columns:
Add 2 annotation columns: "P_snvindel_nb" and "P_snvindel_phen".

Updating the data source (if needed):

- Remove all the files in the following directories:
  "$ANNOTSV/share/AnnotSV/Annotations_Human/FtIncludedInSV/PathogenicSNVindel/GRCh37"
  "$ANNOTSV/share/AnnotSV/Annotations_Human/FtIncludedInSV/PathogenicSNVindel/GRCh38"
- Download and place the files of the different sources in the following directories:
  "$ANNOTSV/share/AnnotSV/Annotations_Human/FtIncludedInSV/PathogenicSNVindel/GRCh37"
  "$ANNOTSV/share/AnnotSV/Annotations_Human/FtIncludedInSV/PathogenicSNVindel/GRCh38"

  The latest update of the ClinVar files are available for free download at:
  Genome build GRCh37:
  https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/clinvar_20201212.vcf.gz

  Genome build GRCh38:
  https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/clinvar_20201212.vcf.gz

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

f) Known benign genes or genomic regions annotation

AnnotSV searches for benign genomic regions completely overlapping the SV to annotate.

Aim:
According to the ACMG technical standards (Riggs et al., 2020), a SV completely contained within an established benign CNV region would be classified as benign (if sharing the same SV type).

Annotation columns:
Add 8 annotation columns: "B_gain_source", "B_gain_coord", "B_loss_source", "B_loss_coord", "B_ins_source", "B_ins_coord", "B_inv_source" and "B_inv_coord".

Benign dataset creation:
For each SV type (Loss, Gain, Ins, Inv), different sources of benign genes or genomic regions have been merged in AnnotSV:

- gnomAD
- ClinVar
- ClinGen
- DGV Gold Standard
- DDD
- 1000 genomes
- Ira M. Hall's lab

Updating the data source (if needed):

- Remove all the files in the following directories:
  "$ANNOTSV/share/AnnotSV/Annotations_Human/SVincludedInFt/BenignSV/GRCh37"
  "$ANNOTSV/share/AnnotSV/Annotations_Human/SVincludedInFt/BenignSV/GRCh38"
- Download and place the files of the different sources in the following directories:
  "$ANNOTSV/share/AnnotSV/Annotations_Human/SVincludedInFt/BenignSV/GRCh37"
  "$ANNOTSV/share/AnnotSV/Annotations_Human/SVincludedInFt/BenignSV/GRCh38"

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

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

gnomAD benign SV 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., 2020). 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.

Method:

Putatively benign variants from gnomAD with all the following requirements are selected:

- at least one population allele frequency > 1% **OR** at least 5 homozygous individuals
- "DUP" or "DEL" SV type

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_v2.1_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 it by AnnotSV.

ClinVar benign SV annotations

Aim:

ClinVar gives access to the relationships asserted between human variants and observed health status.

Method:

Benign SV with all the following requirements are selected:

- "benign" or "benign/likely benign" clinical significance (CLNSIG)
- "criteria_provided", "_multiple_submitters" or "reviewed_by_expert_panel" SV review status (CLNREVSTAT)
- "Deletion" or "Duplication" SV type (CLNVC)
- ≥ 50 bp in size.

Source files:

The latest update of the ClinVar files are available for free download at:

Genome build GRCh37:

https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/clinvar_20201212.vcf.gz

Genome build GRCh38:

https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/clinvar_20201212.vcf.gz

Not dosage sensitive genes/regions annotation (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.
An haploinsufficiency (HI) score of 40 suggests the gene/region is not dosage sensitive for a loss.
A triplosensitivity (TS) score of 40 suggests the gene/region is not dosage sensitive for a gain.

Method:

Genes and regions with a score of 40 are selected.

Source files:

The latest update of the ClinGen files are available for free download at:

Genome build GRCh37:

ftp://ftp.clinicalgenome.org/ClinGen_gene_curation_list_GRCh37.tsv
ftp://ftp.clinicalgenome.org/ClinGen_region_curation_list_GRCh37.tsv

Genome build GRCh38:

ftp://ftp.clinicalgenome.org/ClinGen_gene_curation_list_GRCh38.tsv
ftp://ftp.clinicalgenome.org/ClinGen_region_curation_list_GRCh38.tsv

DGV benign SV 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.

Method:

Putatively benign variants from DGV with all the following requirements are selected:

- Allele frequency > 1%

- ≥ 500 individuals tested
- “Loss” or “Gain” SV type

Loss allele frequency is computed as ‘observedlosses’ / (2 x ‘samplesize’).
Gain allele frequency is computed as ‘observedgains’ / (2 x ‘samplesize’).

Source files:
The latest update of the DGV files are available for free download at http://dgv.tcag.ca/dgv/app/downloads.

Genome build GRCh37:

GRCh37_hg19_variants_2020-02-25.txt (see "DGV Variants" section)

Genome build GRCh38:

GRCh38_hg38_variants_2020-02-25.txt (see "DGV Variants" section)

DDD benign SV annotations

Aim:
AnnotSV takes advantage of the common copy-number variants and their frequencies, as used and displayed in DECIPHER.

Method:
Putatively benign variants from DDD with all the following requirements are selected:

- Allele frequency > 1%
- ≥ 500 individuals tested
- “Deletion” or “Duplication” SV type

Source files:
The latest update of the “**population_cnv.txt.gz**” DECIPHER files is available for free download at:

Genome build GRCh37:

https://decipher.sanger.ac.uk/files/downloads/population_cnv.txt.gz

Genome build GRCh38:

The dataset is not yet available from the DDD team.
However, the GRCh37 DDD SV dataset has been lifted over to GRCh38 with the UCSC Lift Genome Annotation tool and is provided as it is by AnnotSV.

1000 genomes benign SV 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. Most of the 1000 genomes data is already included in the gnomAD dataset.

Method:
Putatively benign variants from 1000 genomes with all the following requirements are selected:

- at least one population allele frequency > 1%
- "Gain” or “Loss” SV type

Source files:
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.

Ira M. Hall's lab benign SV 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., 2020). They publicly released SV frequency annotations to guide SV analysis and interpretation in the era of WGS.

Method:
Putatively benign variants from Ira M. Hall's lab with all the following requirements are selected:

- Allele frequency (AF) > 1%
- "DUP" or "DEL" SV type

Data sources:
Supplementary files 1 and 2 from (Abel et al., 2020) are available for free download at:
https://www.biorxiv.org/content/10.1101/508515v1.supplementary-material
Download, uncompress and keep the following files:
Genome build GRCh37:
B37.callset.public.bedpe.gz
Genome build GRCh38:
B38.callset.public.bedpe.gz

g) 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:
Add 2 annotation columns: "GC_content_left", "GC_content_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:
Add 4 annotation columns: “Repeat_coord_left”, “Repeat_type_left”, “Repeat_coord_right” and “Repeat_type_right”.

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.

Segmental duplication annotations

Aim:

Segmental duplications are large duplications of >1Kb of non-RepeatMasked sequence and >= 90% identity normally present in the human genome. They are associated with the non-allelic homologous recombination mechanisms (NAHR). Homologous recombination is thought to be a classical mechanism for promoting either genetic diversity or genomic disease. Moreover, these regions might also cause issues for read-depth SV detection methods. Reads located in a segmental duplication can perfectly map onto two or more genomic positions and lead to a coverage overestimation at these positions.
The SV breakpoints overlap with segmental duplications can therefore give a clue to explain the SV mechanism, but also a clue to filter out false positives in case of read-depth SV detection methods.

Method:

The Segmental Duplications coordinates are reported.

Annotation columns:

Add 2 annotation columns: "SegDup_left" and "SegDup_right".

Updating the data source (if needed):

AnnotSV needs a UCSC SegDup BED file.

- Remove all the files in the "$ANNOTSV/share/AnnotSV/Annotations_Human/BreakpointsAnnotations/SegDup/GRCh37" and/or "$ANNOTSV/share/AnnotSV/Annotations_Human/BreakpointsAnnotations/SegDup/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, "All Tracks" group, "Segmental Dups" track, "genomicSuperDups" table and "genome" region. Select output format as BED. Choose the following output filename: SegDup.bed. Then, click the get output button.

- Download and place the BED file in the "$ANNOTSV/share/AnnotSV/Annotations_Human/BreakpointsAnnotations/SegDup/GRCh37" and/or "$ANNOTSV/share/AnnotSV/Annotations_Human/BreakpointsAnnotations/SegDup/GRCh38" directories.

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

ENCODE blacklist annotations

Aim:

The human ENCODE blacklist is a comprehensive set of regions that have anomalous, unstructured, or high signal in next-generation sequencing experiments independent of cell line or experiment. The removal of the ENCODE blacklist is an essential quality measure when analyzing functional genomics data.

If you use the blacklist, please cite:
Amemiya, H.M., Kundaje, A. & Boyle, A.P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci Rep 9, 9354 (2019). https://doi.org/10.1038/s41598-019-45839-z

Method:
The ENCODE Blacklist regions and their characteristics are reported.

Annotation columns:
Add 4 annotation columns: "ENCODE_blacklist_left", "ENCODE_blacklist_characteristics_left", "ENCODE_blacklist_right" and "ENCODE_blacklist_characteristics_right".

Updating the data source (if needed):
AnnotSV needs a ENCODE blacklist BED file.

- Remove all the files in the "$ANNOTSV/share/AnnotSV/Annotations_Human/BreakpointsAnnotations/ENCODEblacklist/GRCh37" and/or "$ANNOTSV/share/AnnotSV/Annotations_Human/BreakpointsAnnotations/ENCODEblacklist/GRCh38" directories.
- A current version of the blacklists for hg19 and hg38 ("hg*-blacklist.v2.bed.gz") are available in the "lists" folder of: https://github.com/Boyle-Lab/Blacklist/
- Download, uncompress and place the BED file, renamed "ENCODEblacklist.bed", in the "$ANNOTSV/share/AnnotSV/Annotations_Human/BreakpointsAnnotations/ENCODEblacklist/GRCh37" and/or "$ANNOTSV/share/AnnotSV/Annotations_Human/BreakpointsAnnotations/ENCODEblacklist/GRCh38" directories.

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

GAP annotations

Aim:
Depending on the genome build, several regions of the genome are not yet available. Therefore, they can be misinterpreted due to bad alignment in case of NGS data or badly called in array analysis and then generating false positives calls.

Annotation columns:
Add 2 annotation column: "Gap_left" and "Gap_right".

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

- Remove all the files in the

"$ANNOTSV/share/AnnotSV/Annotations_Human/BreakpointsAnnotations/Gap/GRCh37" and/or "$ANNOTSV/share/AnnotSV/Annotations_Human/BreakpointsAnnotations/Gap/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, "All Tracks" group, "Gap" track, "Gap" table and "genome" region. Select output format as BED. Choose the following output filename: Gap.bed. Then, click the get output button.

- Download and place the BED file in the "$ANNOTSV/share/AnnotSV/Annotations_Human/BreakpointsAnnotations/Gap/GRCh37" and/or "$ANNOTSV/share/AnnotSV/Annotations_Human/BreakpointsAnnotations/Gap/GRCh38" directories.

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

h) 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 (Lupiáñez et al., 2016).

Method:
A TAD boundary is reported if i) the SV overlaps at least 100% 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:
Add 2 annotation columns: "TAD_coordinate", "ENCODE_experiment".
They contain 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.

i) COSMIC annotations (not distributed)

Aim:
COSMIC (Tate et al., 2019), the Catalogue Of Somatic Mutations In Cancer, is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer.

WARNING:
COSMIC data cannot be redistributed. Thus, COSMIC annotation cannot be supplied as part of the AnnotSV sources. Users are required to register in order to download COSMIC data files. More information can be found on their licensing page.

Method:
A COSMIC CNV is reported if the SV overlaps 100% of this feature (user defined, see the "overlap" option in USAGE/OPTIONS).

Annotation columns:
Add 2 annotation columns “Cosmic_ID” and “Cosmic_mut_typ”.

Installing the data source:
AnnotSV needs the “CosmicCompleteCNA.tsv.gz” (2 genome versions available) file from https://cancer.sanger.ac.uk/cosmic/download

- Put the “CosmicCompleteCNA.tsv.gz” file in the corresponding directory:
  “$ANNOTSV/share/AnnotSV/Annotations_Human/FtIncludedInSV/COSMIC/GRCh37/
  or
  “$ANNOTSV/share/AnnotSV/Annotations_Human/FtIncludedInSV/COSMIC/GRCh38/

These files will be reprocessed and then removed the first time AnnotSV is executed.

4. VERSIONS OF THE ANNOTATION SOURCES

Annotations source	Version
...Gene annotations	
Gene annotations (RefSeq)	2020-08-17
Gene annotations (ENSEMBL)	2020-10-24
…Regulatory Elements annotations	

Promoter data (RefSeq)	2020-08-17
Promoter data (ENSEMBL)	2020-10-24
EnhancerAtlas 2.0	2019-06-11
GeneHancer	Downloaded by the user
...Gene-based annotations	
DDD disease	2020-12-03
OMIM	2020-11-07
ACMG	ACMG SF v2.0
Gene intolerance (gnomAD)	V2.1.1
Gene intolerance (ExAC)	2016-08-23
Haploinsufficiency (DDD)	2020-07-13
Haploinsufficiency and triplosensitivity (ClinGen)	2020-12-18
Exomiser	2020-08-20 (v2007)
NCBI gene ID	2020-12-18
…Annotations with known pathogenic genes or genomic regions	
ClinVar	2020-12-12
ClinGen	2020-12-18
OMIM	2020-11-07
dbVar	2020-12-02
…Annotations with known pathogenic SNV/indel	
ClinVar	2020-12-12
…Annotations with known benign genes or genomic regions	
gnomAD (GRCh37)	2019-03-06 (v2.1)
ClinVar	2020-12-12
ClinGen	2020-12-18
DGV annotations	2020-02-25
DDD annotations (GRCh37)	2019-03-18
1000 genomes annotations (GRCh37)	2017-05-19
1000 genomes annotations (GRCh38)	2017-11-05
Ira M. Hall's lab annotations	2018-12-31
...Annotations with features overlapping the SV	
...Annotations with features overlapped with the SV	
COSMIC 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	2020-07-16
Segmental Duplication annotations	2020-10-08
ENCODE blacklist annotations	2018 (v2)
GAP regions annotations	2020-10-08

5. SV RANKING/CLASSIFICATION

In order to assist the clinical interpretation of SV, AnnotSV provides on top of the annotations a ranking score to assess SV pathogenicity. This score is an adaptation of the work provided by the joint consensus recommendation of ACMG and ClinGen (Riggs et al., 2020). We especially payed attention to scoring as much as possible recessive SV observed in various dataset (NGS, array based…)

Scoring:

- ≥ 0.99 Pathogenic Class 5
- 0.90 to 0.98 points Likely pathogenic Class 4
- 0.89 to −0.89 points Variant of uncertain significance Class 3
- −0.90 to −0.98 points Likely benign Class 2
- ≤ −0.99 Benign Class 1

Method:

The comprehensive and detailed scoring guidelines are available in the AnnotSV_Scoring_Criteria.xlsx file (see Table1 for loss SV and Table2 for gain SV). In each section, only 1 criterion (from the most pathogenic to the least) is assigned.

To explicit which criteria have been used to support the ranking score, decision criteria are reported by default in the output file (in the "ranking decision criteria" column).

Annotation columns:

Add 3 annotation columns: “AnnotSV_ranking_score”, “AnnotSV_ranking_criteria” and “ACMG_class”.

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 4.3 (Jun 2020) 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 "SV_type" 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 the “USAGE / OPTIONS”) section. The different arguments can be passed to the program in three ways (order of priority):

- Using the command line
- Using a "configfile" located in the same directory as your input file
- Using a "configfile directly in the installation directory in $ANNOTSV/etc/AnnotSV/configfile

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.

Two supplementary fields are **highly recommended**:

1. *SVTYPE* - The SV type (DEL, DUP…)
   => The column number of the BED file with the SV type information should be indicated (see the -svtBEDcol option) in order to be able to classify the SV.
2. Samples_ID – The list of the samples ID for which the SV was detected
   => The column number with the Samples_ID information should be indicated (see the -samplesidBEDcol option)

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)

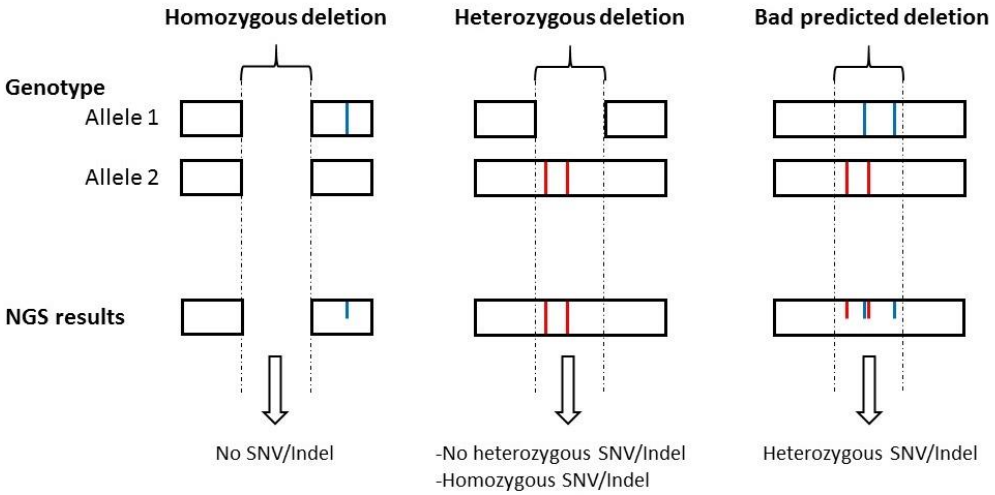
AnnotSV can take VCF file(s) with SNV/indel calls from any sequencing experiment as input to the command line. These annotations report the counts and ratio of homozygous and heterozygous SNV/indel identified from the patients NGS data (user defined samples) and presents in the interval of the **deletion**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

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

Annotation columns:
Add the "Count_hom(sample)", "count_htz(sample)", "Count_htz/allHom(sample)", "Count_htz/total(cohort)" and "Count_total(cohort)" annotation columns.

- **Count_hom(sample):** Count of homozygous SNV/indel called from the sample and present in the interval of the deletion
- **Count_htz(sample):** Count of heterozygous SNV/indel called from the sample and present in the interval of the deletion
- **Count_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
- **Count_total(cohort):** Total count of SNV/indel called from all the samples of the cohort 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 dependent on sequencing protocols and calling/filtering strategies and cannot be determined as a standard.

Warning:
In the VCF file(s), **the genotype of each variation should be indicated in the FORMAT column 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.

Annotation columns:
Add 1 annotation columns for each sample: **compound_htz(sample).**

Usage:
To add the "**compound_htz(sample)**" 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 column 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 at 100% with the SV will be reported.

By placing the BED file in the "SVincludedInFt" directory, only the features overlapping 100% of the SV 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 the "-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 (see INPUT section) to define there 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 (see INPUT section) and to define there 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 also create another output file: 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

c) "Annotation_mode" 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 by genes. This is possible thanks to the way AnnotSV can present the data. Indeed, there are2 types of lines provided by AnnotSV (*cf* the "Annotation_mode" 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 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 genes covered (*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 Gene-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 annotation modes (full or split), in some cases specific gene annotations are only present while using one of the two modes.

Nomenclature: All the column names begin with an upper case and contain no space character.

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) (deletions have negative values)	X	X	X	X
SV_type	Type of the SV (DEL, DUP, …)	X	X	X	X
Samples_ID	List of the samples ID for which the SV was called	X	X	X	X
REF	Nucleotide sequence in the reference genome (extracted only from a VCF input file)	X	X		X
ALT	Alternate nucleotide sequence (extracted only from a VCF input file)	X	X		X
FORMAT	The FORMAT column from a VCF file	X	X		X
'Sample ID'	The sample ID column from a VCF file	X	X		X
Annotation_mode	Indicate the type of annotation lines generated: - annotation on the SV full length ("full") - annotation on each gene overlapped by the SV ("split")	X	X	X	X

Gene_name	Gene symbol	X	X	X	X
Gene_count	Number of overlapped genes with the SV	X		X	X
Tx[1]	Transcript symbol		X	X	X
Tx_start	Starting position of the transcript		X	X	X
Tx_end	Ending position of the transcript		X	X	X
Overlapped_tx_length	Length of the transcript (bp) overlapping with the SV		X	X	X
Overlapped_CDS_length	Length of the CoDing Sequence (CDS) (bp) overlapped with the SV		X	X	X
Overlapped_CDS_percent	Percent of the CoDing Sequence (CDS) (bp) overlapped with the SV		X	X	X
Frameshift	Indicates if the CDS length is not divisible by three (yes or no)		X	X	X
Exon_count	Number of exons of the transcript		X	X	X
Location	SV location in the gene's Values: txStart, txEnd, exon'i', intron'i' e.g. « txStart-exon3 »		X	X	X
Location2	SV location in the gene's coding regions Values: UTR (no CDS in the gene), 5'UTR (before the CDS start), 3'UTR (after the CDS end), CDS (between the CDS start and the CDS end, can be in an exon or an intron). e.g. « 3'UTR-CDS »		X	X	X
Dist_nearest_SS[2]	Absolute distance to nearest splice site after considering exonic and intronic SV breakpoints		X	X	X
Nearest_SS_type	Nearest splice site type: 5' (donor) or 3' (acceptor)		X	X	X
Intersect_start	Start position of the intersection between the SV and a transcript		X	X	X
Intersect_end	End position of the intersection between the SV and a transcript		X	X	X
RE_gene	Name of the genes regulated by a regulatory element overlapped with the SV to annotate. When available, the regulated gene name is detailed with associated haploinsufficiency (HI), triplosensitivity (TS) and exomiser (EX) scores.	X		X	X
B_gain_source	Origin of the benign Gain genomic regions completely overlapping the SV to annotate	X	X	X	X
B_gain_coord	Coordinates of the benign Gain genomic regions completely overlapping the SV to annotate	X	X	X	X
B_loss_source	Origin of the benign Loss genomic regions completely overlapping the SV to annotate	X	X	X	X
B_loss_coord	Coordinates of the benign Loss genomic regions completely overlapping the SV to annotate	X	X	X	X
B_ins_source	Origin of the benign Ins genomic regions completely overlapping the SV to annotate	X	X	X	X
B_ins_coord	Coordinates of the benign Loss genomic regions completely overlapping the SV to annotate	X	X	X	X
B_inv_source	Origin of the benign Inv genomic regions completely overlapping the SV to annotate	X	X	X	X
B_inv_coord	Coordinates of the benign Inv genomic regions completely overlapping the SV to annotate	X	X	X	X
P_gain_phen	Phenotype of the pathogenic Gain genomic regions completely overlapped with the SV to annotate	X	X	X	X
P_gain_hpo	HPO terms describing the pathogenic Gain genomic regions completely overlapped with the SV to annotate	X	X	X	X
P_gain_source	Origin of the pathogenic Gain genomic regions completely overlapped with the SV to annotate	X	X	X	X

P_gain_coord	Coordinates of the pathogenic Gain genomic regions completely overlapped with the SV to annotate	X	X	X	X
P_loss_phen	Phenotype of the pathogenic Loss genomic regions completely overlapped with the SV to annotate	X	X	X	X
P_loss_hpo	HPO terms describing the pathogenic Loss genomic regions completely overlapped with the SV to annotate	X	X	X	X
P_loss_source	Origin of the pathogenic Loss genomic regions completely overlapped with the SV to annotate	X	X	X	X
P_loss_coord	Coordinates of the pathogenic Loss genomic regions completely overlapped with the SV to annotate	X	X	X	X
P_ins_phen	Phenotype of the pathogenic Ins genomic regions completely overlapped with the SV to annotate	X	X	X	X
P_ins_hpo	HPO terms describing the pathogenic Ins genomic regions completely overlapped with the SV to annotate	X	X	X	X
P_ins_source	Origin of the pathogenic Ins genomic regions completely overlapped with the SV to annotate	X	X	X	X
P_ins_coord	Coordinates of the pathogenic Ins genomic regions completely overlapped with the SV to annotate	X	X	X	X
P_snvindel_nb	Number of pathogenic snv/indel from public databases completely overlapped with the SV to annotate	X	X	X	X
P_snvindel_phen	Phenotypes of pathogenic snv/indel from public databases completely overlapped with the SV to annotate	X	X	X	X
TAD_coordinate[3]	Coordinates of the TAD whose boundaries overlapped with the annotated SV (boundaries included in the coordinates)	X		X	X
ENCODE_experiment[3]	ENCODE experiments used to define the TAD	X		X	X
Cosmic_ID	COSMIC identifier	X	X	X	X
Cosmic_mut_typ	Defined as Gain or Loss	X	X	X	X
GC_content_left	GC content around the left SV breakpoint (+/- 100bp)	X		X	X
GC_content_right	GC content around the right SV breakpoint (+/- 100bp)	X		X	X
Repeat_coord_left	Repeats coordinates around the left SV breakpoint (+/- 100bp)	X		X	X
Repeat_type_left	Repeats type around the left SV breakpoint (+/- 100bp) e.g. AluSp, L2b, L1PA2, LTR12C, SVA_D, …	X		X	X
Repeat_coord_right	Repeats coordinates around the right SV breakpoint (+/- 100bp)	X		X	X
Repeat_type_right	Repeats type around the right SV breakpoint (+/- 100bp) e.g. AluSp, L2b, L1PA2, LTR12C, SVA_D, …	X		X	X
Gap_left	Gap regions coordinates around the left SV breakpoint (+/- 100bp)	X		X	X
Gap_right	Gap regions coordinates around the right SV breakpoint (+/- 100bp)	X		X	X
SegDup_left	Segmental Duplication regions coordinates around the left SV breakpoint (+/- 100bp)	X		X	X
SegDup_right	Segmental Duplication regions coordinates around the right SV breakpoint (+/- 100bp)	X		X	X
ENCODE_blacklist_left	ENCODE blacklist regions coordinates around the left SV breakpoint (+/- 100bp)	X		X	X
ENCODE_blacklist_characteristics_left	ENCODE blacklist regions characteristics around the left SV breakpoint (+/- 100bp)	X		X	X
ENCODE_blacklist_right	ENCODE blacklist regions coordinates around the right SV breakpoint (+/- 100bp)	X		X	X
ENCODE_blacklist_characteristics_right	ENCODE blacklist regions characteristics around the right SV breakpoint (+/- 100bp)	X		X	X
ACMG	ACMG genes		X	X	X
HI	ClinGen Haploinsufficiency Score	X	X	X	X

TS	ClinGen Triplosensitivity Score	X	X	X	X
DDD_HI_percent	Haploinsufficiency ranks from DDD	X	X	X	X
DDD_status	DDD status: e.g. confirmed, probable		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_pmid	DDD Pubmed Id		X	X	X
ExAC_synZ	Positive synZ_ExAC (Z score) from ExAC indicate gene intolerance to synonymous variation	X	X	X	X
ExAC_misZ	Positive misZ_ExAC (Z score) from ExAC indicate gene intolerance to missense variation	X	X	X	X
ExAC_delZ	Positive delZ_ExAC (Z score) from ExAC indicate gene intolerance to deletion	X	X	X	X
ExAC_dupZ	Positive dupZ_ExAC (Z score) from ExAC indicate gene intolerance to duplication	X	X	X	X
ExAC_cnvZ	Positive cnvZ_ExAC (Z score) from ExAC indicate gene intolerance to CNV	X	X	X	X
OMIM_ID	OMIM unique six-digit identifier	X	X	X	X
OMIM_phenotype	e.g. Charcot-Marie-Tooth disease		X	X	X
OMIM_inheritance[4]	e.g. AD (= "Autosomal dominant")		X	X	X
OMIM_morbid	Set to “yes” if the SV overlaps an OMIM morbid gene	X	X	X	X
OMIM_morbid_candidate	Set to “yes” if the SV overlaps an OMIM morbid gene candidate	X	X	X	X
LOEUF_bin	Minimal “decile bin of LOEUF” for given transcripts of a gene (lower values indicate more constrained) Values = integer [0-9]	X	X	X	X
GnomAD_pLI	Score computed by gnomAD indicating the probability that a gene is intolerant to a loss of function variation (Nonsense, splice acceptor/donor variants due to SNV/indel).	X	X	X	X
ExAC_pLI	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
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
Count_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
Count_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
Count_htz/allHom(sample) [5]	Ratio for QC filtering: #htz(sample)/#allHom(sample)	X	X		X
Count_htz/total(cohort)	Ratio for QC filtering: #htz(sample)/#total(cohort)	X	X		X
Count_total(cohort)	Total count of SNV/indel called from all the samples of the cohort and present in the interval of the deletion	X	X		X

AnnotSV_ranking_score	SV ranking score following the 2019 joint consensus recommendation of ACMG and ClinGen. Scoring: pathogenic 0.99 or more points, likely pathogenic 0.90 to 0.98 points, variant of uncertain significance 0.89 to −0.89 points, likely benign −0.90 to −0.98 points, benign −0.99 or fewer points.	X		X	X
AnnotSV_ranking_criteria	Decision criteria explaining the AnnotSV ranking score	X		X	X
ACMG_class	SV ranking class 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

[1]*Given one gene, only a single transcript from all transcripts available 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]*AnnotSV calculates the distance to the nearest splice site, upstream and downstream of exonic/intronic SV breakpoints. Then, if several distances were calculated, AnnotSV keeps the smallest one*
[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]*Detailed in the FAQ*
[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 modify the default order of the annotation columns provided by AnnotSV and only select a subset of those . 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 (see INPUT section). There, the user can comment column names with a hash character («#»). An example of configfile is provided in the AnnotSV installation directory.

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 (see INPUT section). To show the options simply type:
$ANNOTSV/bin/AnnotSV -help
or
$ANNOTSV/bin/AnnotSV

OPTIONS:

-annotationsDir: Path of the annotations directory

-bcftools: Path of the bcftools local installation

-bedtools: Path of the bedtools local installation

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

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

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

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

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

-help: More information on the arguments

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

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

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

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

-outputDir: Output path name

-outputFile: Output path and file name

-overlap: Minimum overlap (%) between user features (User BED) and the annotated SV to be reported
Range values: [0-100], default = 100

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

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

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

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

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

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

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

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

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

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

-SVinputInfo: To extract the additional SV input fields and insert the data in the outputfile
Values: 1 (default) or 0

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

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

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

-txFile: Path of a file containing a list of preferred genes transcripts to be used in priority during the annotation (Preferred genes transcripts names should be tab or space separated)

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

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".

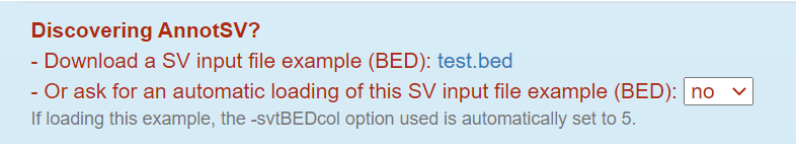
Moreover, an input/output example (the HG00096 individual from the 1000 Genomes project) is available on the AnnotSV website.

11. WEB SERVER

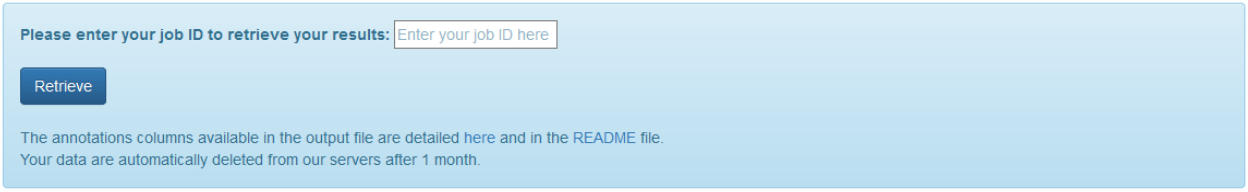
a) AnnotSV annotation and ranking

Annotation and ranking of your SV are freely available online at https://lbgi.fr/AnnotSV/runjob.
User can operate through a web browser, which can be used to select the parameters, run the program, retrieve or visualize/analyze the results.

An SV input file example (BED) is available to easily evaluate AnnotSV online.



A job ID is provided at the time of data submission. It allows user to bookmark and access the results at a later time. The results are available at:
https://lbgi.fr/AnnotSV/retrievejob

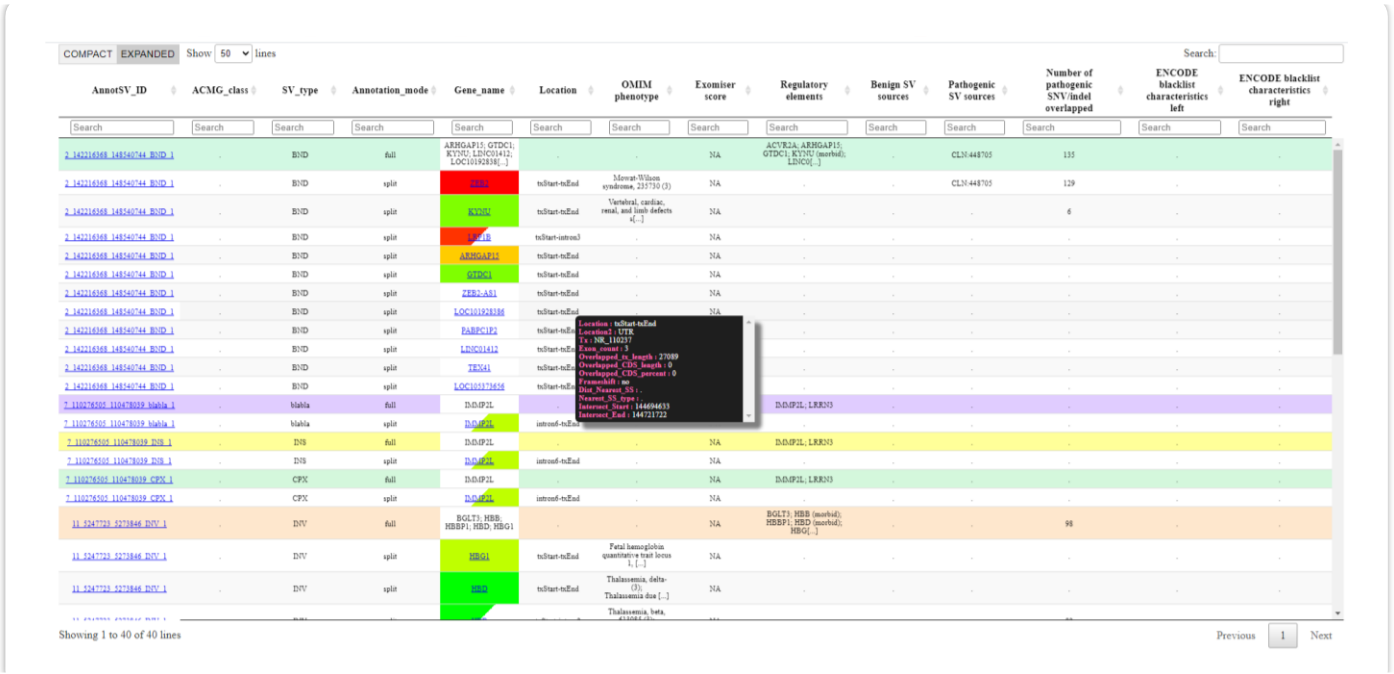


Moreover, this job ID will give access to the status of the job (running or finished).

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

b) Visualization of the annotation data

AnnotSV results visualization are powered by knotAnnotSV. The user can visualize, filter and analyze the annotation data thanks to different user-friendly functions (search/filtering box, tooltip, links to public databases, color coded information...)



The interface is well detailed in the knotAnnotSV README file:
https://github.com/mobidic/knotAnnotSV/blob/master/README.knotAnnotSV_v1.0.pdf

At each stage of the analysis process, all the set-up filters are locally stored. The html file can thus be closed at any time by biologists and then re-opened on the same computer to continue the analysis.

The knotAnnotSV source code is available under the GNU GPL licence and is downloadable on GitHub.

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 (Jun 2020).

- 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' Annotation_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 Annotation_mode = "full") and one based on each gene within the SV (corresponding to the Annotation_mode = "split"). Thus, you will have access to:

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

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

Regarding these "split" lines,

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

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

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

Q: 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 "-annotationMode" 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: Why do I get negative values in the SV_length column?

It is to notice that deletions have negative values. Other SV types have positive values.

Q: What does the candidateGenesFile parameter refer to?

The candidateGenesFile contains the candidate genes of the user. This information is used to filter out the SV annotations that do not overlap a candidate gene (-candidateGenesFiltering yes).

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.

AnnotSV creates a report of unannotated variants (".unannotated.tsv" file).

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

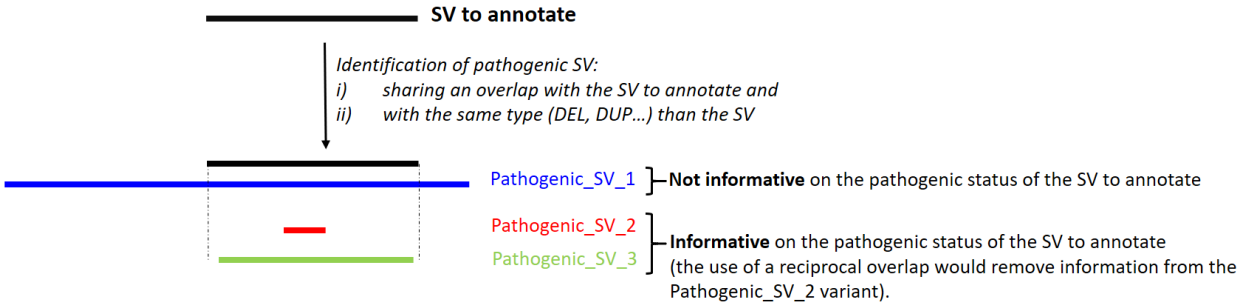
Q: How overlaps (%) are calculated?
AnnotSV provides different types of annotations:

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



=> 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. 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
```

Q: I'm getting the error: « unexpected token "END" at position 0; expecting VALUE » while running Exomiser. How can I fix this problem?

You are facing a memory issue. Please, try increasing RAM/MEM on your compute node.

Q: What is knotAnnotSV?

knotAnnotSV is a freely accessible web interface that allows you to explore your annotated SV dataset in a user-friendly way. This interface is well detailed in the "README.knotAnnotSV_*'version'*.pdf" file available on Github: https://github.com/mobidic/knotAnnotSV/

13. REFERENCES

Abel, H.J., Larson, D.E., Regier, A.A., Chiang, C., Das, I., Kanchi, K.L., Layer, R.M., Neale, B.M., Salerno, W.J., Reeves, C., et al. (2020). Mapping and characterization of structural variation in 17,795 human genomes. Nature.

Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alföldi, J., Francioli, L.C., Khera, A.V., Lowther, C., Gauthier, L.D., Wang, H., et al. (2020). A structural variation reference for medical and population genetics. Nature *581*, 444–451.

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.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993.

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.

Riggs, E.R., Andersen, E.F., Cherry, A.M., Kantarci, S., Kearney, H., Patel, A., Raca, G., Ritter, D.I., South, S.T., Thorland, E.C., et al. (2020). Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). Genetics in Medicine *22*, 245–257.

Smedley, D., Jacobsen, J.O.B., Jäger, 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.

Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., et al. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res *47*, D941–D947.