AnnotSV Manual

Version 1.1
AnnotSV is a program for annotating structural variations from the human genome.

http://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)

BED: Browser Extensible Data

bp: base pair

CDS: CoDing Sequence

CNV: Copy Number Variation

DDD: Deciphering Developmental Disorders

DECIPHER: DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources

DEL: Deletion

DGV: Database of Genomic Variants

DNA: DesoxyriboNucleic Acid

DUP: Duplication

ENCODE: Encyclopedia of DNA Elements

ExAC: Exome Aggregation Consortium

GRCh37: Genome Reference Consortium Human Build 37

GRCh38: Genome Reference Consortium Human Build 38

HI: Haploinsufficiency

hom: homozygous

htz: heterozygous

ID: Identifier

indel: Insertion/deletion

LoF: Loss of Function

misZ = Z score indicating gene intolerance to missense variation

NAHR: Non-Allelic Homologous Recombination

NM: RefSeq identifiers

OMIM: Online Mendelian Inheritance in Man

pLI = score computed by the ExAc consortium to indicate gene intolerance to a loss of function variation

SNV : Single Nucleotide Variation

SV: Structural Variations

synZ = Z score indicating gene intolerance to synonymous variation

TAD: Topologically Associating Domains

Tcl: Tool Command Language

Tx: transcript

VCF: Variant Call Format

TABLE OF CONTENTS

1. INTRODUCTION

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

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

AnnotSV takes as an input file a classical bed or VCF file describing the SV coordinates. The output file contains the overlaps of the SV with relevant genomic features where the genes refer to NCBI RefSeq genes. In addition to the gene annotations, we provide numerous additional relevant annotations (OMIM, DGV frequencies, compound heterozygosity …).

2. INSTALLATION/REQUIREMENTS/UPDATE

2.1. Tcl (Required)

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

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

2.2. AnnotSV source code (Required)

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

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

The installation requires simply to set the following environment variable:
- $ANNOTSV : “AnnotSV installation directory”

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

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

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

AnnotSV installation directory:

By default the AnnotSV installation directory looks like this:

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

2.3. bedtools (Required)

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

2.4. Annotation sources (Provided)

AnnotSV requires different data sources for the annotation of SV. **In order to provide a ready to start installation of AnnotSV, each annotation source listed below (that do not require a commercial license) is already provided with the AnnotSV sources**. The aim and update of each of these sources are explained below.

Annotation can be performed using either the GRCh37 or GRCh38 build version of the human genome (user defined, see USAGE/OPTIONS), but depending on the availability of some data sources there might be some limitations.

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

a) Gene annotations

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

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

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

Updating the data source (if needed):

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

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

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

b) Promoters annotations

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

Annotation columns:
Adds 1 annotation column: “promoters”

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

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

c) DGV Gold Standard annotations

Aim:

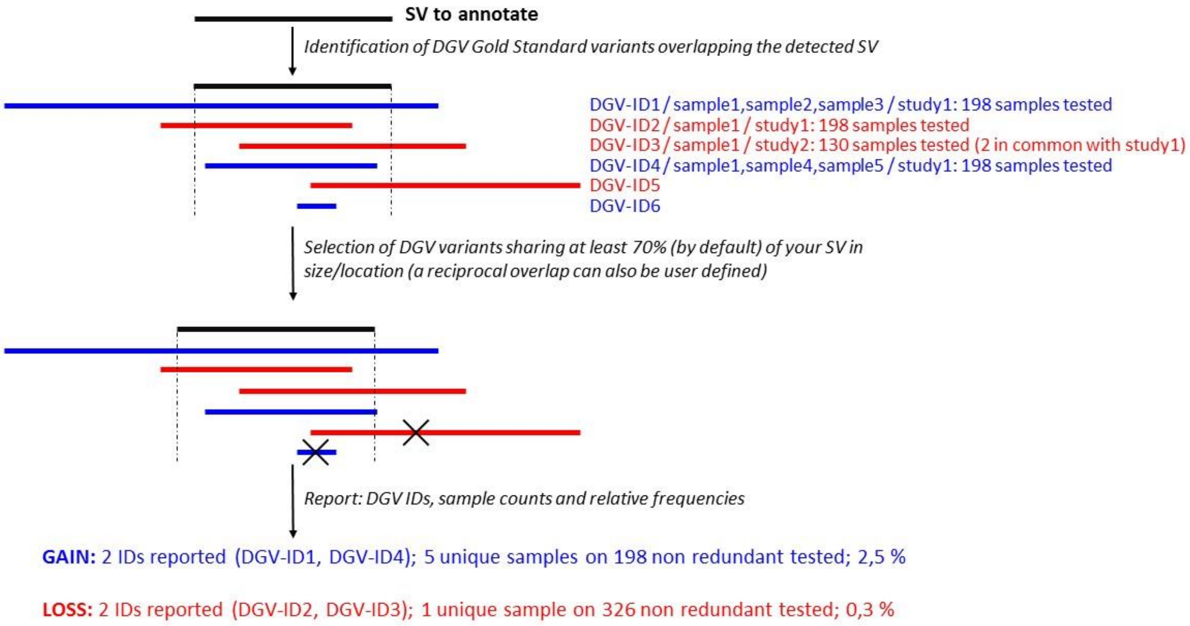
The Database of Genomic Variants (DGV, (MacDonald, et al., 2014)) provides SV defined as DNA elements with a size >50 bp. The content of DGV is only representing SV identified in healthy control samples from large cohorts published and integrated by the DGV team. The annotations will give information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

Annotation columns:

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

Method:

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


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

Warning:

- *Exceptional overestimation of the relative frequencies:*

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

- *Gain/Loss:*

A SV call in DGV can be relative to a specific reference sample, a pool of reference samples or relative to the reference assembly. Since different reference samples may have been used in different studies, what is called as a gain in one study may actually be called a loss in another.

Updating the data source (if needed):

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

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

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

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

d) DECIPHER GENE ANNOTATIONS

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

Annotation columns:
Adds 5 annotation columns: "DDD_status", "DDD_mode", "DDD_consequence", "DDD_disease", "DDD_pmids".

Updating the data source (if needed):

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

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

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

e) DECIPHER FREQUENCY ANNOTATIONS

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

Method:

By default, a DDD CNV is reported if an overlap of ≥70% is found with a SV to annotate. Nevertheless, the user can modify the default behaviour by either use a different percentage or a reciprocal overlap (see SVfromDBoverlap" and "SVtoAnnOverlap" options).

Annotation columns:

Adds 5 annotation columns: "DDD_SV", "DDD_DUP_n_samples_with_SV", "DDD_DUP_Frequency", "DDD_DEL_n_samples_with_SV", "DDD_DEL_Frequency".

Updating the data source (if needed):

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

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

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

f) 1000 GENOMES 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 also the SV. These annotations give additional information on the SV allele frequencies from the 1000 genomes database overlapped by a SV to annotate.

Method:

By default, a 1000g SV is reported if an overlap of ≥70% is found with a SV to annotate. Nevertheless, the user can modify the default behaviour by either use a different percentage or a reciprocal overlap (see "SVfromDBoverlap" and "SVtoAnnOverlap" options).

Annotation columns:

Adds 3 annotation columns: "1000g_event", "1000g_AF" and "1000g_max_AF".

Updating the data source (if needed):

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

http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/supporting/GRCh38_positions/ALL.wgs.mergedSV.v8.20130502.svs.genotypes.GRCh38.vcf.gz

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

g) GC CONTENT ANNOTATIONS

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

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

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

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

- Remove all the files in the "$ANNOTSV/Annotations/GCcontent/GRCh37" and/or "$ANNOTSV/Annotations/GCcontent/GRCh38" directories.
- Download and place the human reference genome FASTA file in the "$ANNOTSV/Annotations/GCcontent/GRCh37" and/or "$ANNOTSV/Annotations/GCcontent/GRCh38" directories. The latest update of this file is available for free download at:
  Genome build GRCh37:
  http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/chromFa.tar.gz

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

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

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

h) REPEATED SEQUENCES ANNOTATIONS

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

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

Annotation columns:
Adds 2 annotation columns: “Repeats_coord” and “Repeats_type”

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

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

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

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

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

i) TAD 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 nonrandom organization, topologically associating domains (TAD) emerge as a fundamental structural unit able to separate domains and define boundaries. Disruption of these structures especially by SV can result in gene misexpression (Lupianez, et al., 2016).

AnnotSV reports the TAD boundaries in case there is an overlapped of at least 70% with the SV (user defined, see the "FeaturesOverlap" option in USAGE/OPTIONS).

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

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

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

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

j) 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).

Annotation columns:
Add 3 annotation columns: “Mim Number”, “Phenotypes”, “Inheritance”.

Update:

- Remove all the files in the “$ANNOTSV/Annotations/OMIM” directory.
- Download and place the “**genemap2.txt**” OMIM file in the “$ANNOTSV/Annotations/OMIM” directory. The latest update of this file is available for download following a registration and review process (https://omim.org/downloads/). It is a tab-delimited file containing OMIM's synopsis of the Human gene map including additional information such as genomic coordinates and inheritance.

k) GENE INTOLERANCE ANNOTATIONS

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:

synZ = synonymous Z score
misZ = missense Z score
Positive Z scores indicate gene intolerance to variation.

pLI = score computed by the ExAC consortium
pLI indicates the probability that a gene is intolerant to a loss of function mutation (Nonsense, splice acceptor and splice donor variants caused by SNV). ExAC consider pLI >= 0.9 as an extremely LoF intolerant set of genes.

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

Annotation columns:
Adds 3 annotation columns: “synZ”, “misZ” and “pLI”.

Updating the data source (if needed):

- Remove all the files in the “$ANNOTSV/Annotations/GeneIntolerance” directory.
- Download and place the “**fordist_cleaned_nonpsych_z_pli_rec_null_data.txt**” ExAC file in the “$ANNOTSV/Annotations/GeneIntolerance” directory. The latest update of this file is available for free download at:
  Genome build GRCh37:
  ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3.1/functional_gene_constraint/

Genome build GRCh38:
The dataset is not yet available.

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

I) Haploinsufficiency annotations

Aim:
Haploinsufficiency, wherein a single functional copy of a gene is insufficient to maintain normal function, is a major cause of dominant disease. As detailed in DECIPHER, over 17,000 protein coding genes have been scored according to their predicted probability of exhibiting haploinsufficiency:
- High ranks (e.g. 0-10%) indicate a gene is more likely to exhibit haploinsufficiency
- Low ranks (e.g. 90-100%) indicate a gene is more likely to NOT exhibit haploinsufficiency.

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

Annotation columns:
Add 1 annotation column: “HI_percent”.

Update:

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

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

3. INPUT

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

3.1.SV input file (Required)

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

- VCF is a text file 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.

WARNING:

By default, AnnotSV extracts and reports only some informations from the VCF input file:
- 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) can be reported by setting the "-SVinputInfo" option to 1.

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

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

Additional fields from the BED file are optional and can be reported in the AnnotSV output file (user defined). It can be used to store quality, read depth or other metrics produced by the SV caller.

WARNING:

By default, AnnotSV does not report the additional fields from the BED input file.

All fields can be reported by setting the "-SVinputInfo" option to 1.

3.2.SNV/indel input files - for DELETION filtering (Optional)

AnnotSV can take VCF file(s) with SNV/indel as input to the command line.

These annotations report the counts of homozygous and heterozygous SNV/indel identified from the patients NGS data (user defined samples) and presents in the interval of the SV to annotate.

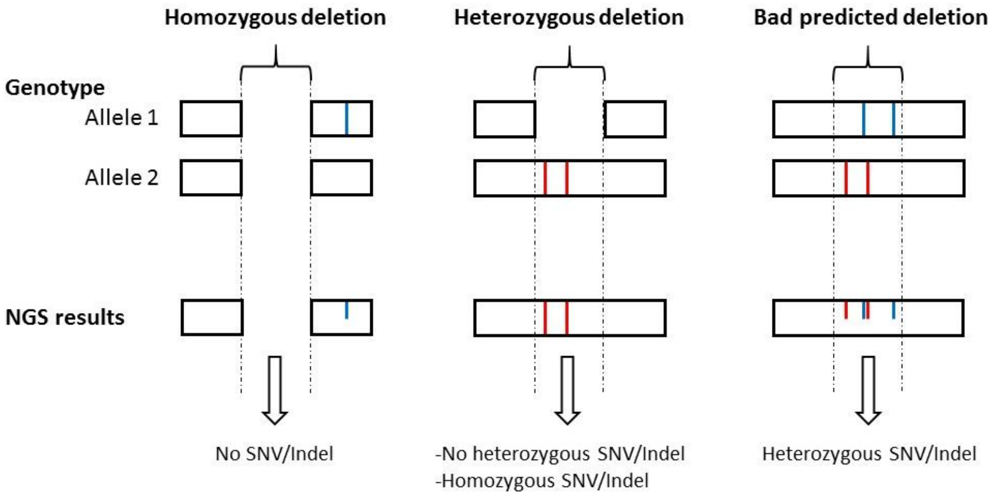
Annotation columns/Usage:

Add the "hom(sample)" and "htz(sample)" annotation columns.

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

Aim:

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



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

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

3.3.Filtered SNV/indel input files - for compound heterozygosity analysis (Optional)

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 presents in the gene overlapped by the SV to annotate, as well in 'healthy' and 'affected' samples (user defined samples). This would be really useful for the user to identify compound heterozygotes with one SNV/indel and one SV.

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

Background:
In recessive genetic disorders, both copies of a certain gene are malfunctioning. This means that the maternally as well as the paternally inherited copy of an autosomal gene harbors a pathogenic mutation. And if the parents are non-consanguineous, compound heterozygosity is the best explanation for a recessive disease.

Aim:
AnnotSV offers an efficient filter to highlight compound heterozygous variants composed of one SV and one SNV/indel in the same gene.
In this way AnnotSV takes in input a VCF file(s) that is already filtered for genotype frequency and effects on protein level. Then, the software extracts from the input VCF file(s) the heterozygous variants (SNV/indel) presents in the gene overlapped by the SV, as well in 'healthy' and 'affected' samples (user defined samples).

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

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

3.4. 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 should be located in the "$ANNOTSV/Annotations" directory.
It is to notice that these files should not contain any of these 2 specific characters "{" and "}" (that would be replaced by "(" and ")").

4. 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.

Output file path and name:
Two options (-outputDir and -outputFile) can be used to specify the output directory and/or file name. 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 "20180312_AnnotSV/mySVinputFile.annotated.tsv".

Output lines:
There are 2 types of lines produced by AnnotSV (*cf* the "AnnotSV type" output column):
- An annotation on the **"full"** length of the SV. Every SV are reported, even those not covering a gene. This type of annotation gives an estimate of the SV event itself.
- An annotation of the SV **"split"** by gene. This type of annotation gives an estimate of the gene composition of the corresponding SV and is meant to analyse the consequences more deeply. Thus, in some cases, when a SV spans over several genes, the output will contain as many annotations lines as covered genes (*cf* example in FAQ). This latter annotation is extremely powerful to shorten the identification of mutation in a implicating a specific gene.

4.1.Annotation columns available in the output file

The first line of the output file is the column description. In the following table, the annotations available in the AnnotSV output file are described:

Column name	Annotation
SV chrom	Name of the chromosome
SV start	Starting position of the SV in the chromosome
SV end	Ending position of the SV in the chromosome
REF	Nucleotide sequence in the reference genome (extracted only from a VCF input file)
ALT	Alternate nucleotide sequence (extracted only from a VCF input file)

SVTYPE	Type of the SV (extracted only from a VCF input file)
FORMAT	The FORMAT column from a VCF file
Sample ID	The sample ID column from a VCF file
AnnotSV type	Indicate the type of annotation generated: - annotation on the SV full length (“full”) - annotation on each gene overlapped by the SV (“split”)
Gene name	Gene symbol
NM	Transcript symbol[1]
CDS length	Length of the CoDing Sequence (CDS) (bp) overlapping with the SV
tx length	Length of the transcript (bp) overlapping with the SV
location	SV location in the gene (e.g. « txStart-exon1 », « intron3-exon7 »)
intersectStart	Start position of the intersection between the SV and the transcript
intersectEnd	End position of the intersection between the SV and the transcript
promoters	List of the genes whose promoters are overlapped by the SV
DGV_GAIN_IDs	DGV Gold Standard GAIN IDs overlapped with the annotated SV
DGV_GAIN_n_samples_with_SV	Number of individuals with a shared DGV_GAIN_ID
DGV_GAIN_n_samples_tested	Number of individuals tested
DGV_GAIN_Frequency	Relative GAIN Frequency=DGV_GAIN_n_samples_with_SV/DGV_GAIN_n_samples_tested
DGV_LOSS_IDs	DGV Gold Standard LOSS IDs overlapped with the annotated SV
DGV_LOSS_n_samples_with_SV	Number of individuals with a shared DGV_LOSS_ID
DGV_LOSS_n_samples_tested	Number of individuals tested
DGV_LOSS_Frequency	Relative LOSS Frequency=DGV_LOSS_n_samples_with_SV/DGV_LOSS_n_samples_tested
DDD_SV	Deciphering Developmental Disorders (DDD) SV coordinates from the DDD study (data control sets) overlapped with the annotated SV
DDD_DUP_n_samples_with_SV	Number of individuals with a shared DDD_DUP
DDD_DUP_Frequency	DUP Frequency
DDD_DEL_n_samples_with_SV	Number of individuals with a shared DDD_DEL
DDD_DEL_Frequency	DEL Frequency
DDD_status	Deciphering Developmental Disorders (DDD) category e.g. confirmed, probable, possible, …
DDD_mode	Deciphering Developmental Disorders (DDD) allelic requirement e.g. biallelic, hemizygous, …
DDD_consequence	Deciphering Developmental Disorders (DDD) mutation consequence e.g. "loss of function", uncertain, …
DDD_disease	Deciphering Developmental Disorders (DDD) disease name e.g. "OCULOAURICULAR SYNDROME"
DDD_pmids	Deciphering Developmental Disorders (DDD) pmids
1000g_event	1000 genomes event types (e.g. DEL, DUP, ALU, <CN3>...)
1000g_AF	1000 genomes allele frequency
1000g_max_AF	Maximum observed allele frequency across the 1000 genomes populations
synZ	Positive synZ (Z score) indicate gene intolerance to synonymous variation
misZ	Positive misZ (Z score) indicate gene intolerance to missense variation
pLI	Score computed in the ExAc database indicating the probability that a gene is intolerant to a loss of function variation (Nonsense, splice acceptor and donor variants caused by SNV). ExAC consider pLI >= 0.9 as an extremely LoF intolerant set of genes
HI_percent	Haploinsufficiency ranks

Mim Number	OMIM unique six-digit identifier
Phenotypes	e.g. Charcot-Marie-Tooth disease
Inheritance	e.g. AD (= "Autosomal dominant")[2]
GCcontent_left	GC content around the left SV breakpoint (+/- 100bp)
GCcontent_right	GC content around the right SV breakpoint (+/- 100bp)
Repeats_coord_left	Repeats coordinates around the left SV breakpoint (+/- 100bp)
Repeats_type_left	Repeats type around the left SV breakpoint (+/- 100bp) e.g. AluSp, L2b, L1PA2, LTR12C, SVA_D, …
Repeats_coord_right	Repeats coordinates around the right SV breakpoint (+/- 100bp)
Repeats_type_right	Repeats type around the rignt SV breakpoint (+/- 100bp) e.g. AluSp, L2b, L1PA2, LTR12C, SVA_D, …
TADcoordinates	Coordinates of the TAD whose boundaries overlapped with the annotated SV (boundaries included in the coordinates)
ENCODEexperiments	ENCODE experiments from where the TAD have been defined
compound-htz(sample)	List of heterozygous SNV/indel (reported with "chrom_position") presents in the gene overlapped by the annotated SV
hom(sample)	Number of homozygous variants in the individual "sample" which are presents: - in the SV for the "full" annotation - between intersectStart and intersectEnd for the "split" annotation. Values are extracted from the input VCF file(s)
htz(sample)	Number of heterozygous variants in the individual "sample" which are presents: - in the SV for the "full" annotation - between intersectStart and intersectEnd for the "split" annotation. Values are extracted from the input VCF file(s)

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

[2] *Detailed in the FAQ*

5. USAGE / OPTIONS

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

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

OPTIONS:

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

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

-bedtools: Path of the bedtools local installation

-FeaturesOverlap: Minimum overlap (%) of the features (promoter, TAD…) with the annotated SV to report the features
Range values: [0-100], default = 70

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

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

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

-help: More information on the arguments

-outputDir: Output path name

-outputFile: Output path and file name

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

-SVfromDBoverlap: Minimum overlap (%) of the SV from external databases (DGV, DDD) with the annotated SV to report the features
Range values: [0-100], default = 0

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

-SVtoAnnOverlap: Minimum overlap (%) of the annotated SV with the SV from external databases (DGV, DDD…) to report the features
Range values: [0-100], default = 70

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

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

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

6. Test

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

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

7. FAQ

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

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

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

Q: Can AnnotSV annotate every type of SV?
AnnotSV supports as well VCF or BED format in input.
- BED format doesn't 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.
However, **breakpoints of such features can be annotated by AnnotSV**. In this way, breakpoints positions should be stored in a BED file as follow:

1. chrom - The name of the chromosome (e.g. 3, Y, …) - Preferred without "chr".
2. breakpointStart - The breakpoint position in the chromosome. The first base in a chromosome is numbered 0.
3. breakpointEnd - Equal to "breakpointStart + 1" (The chromEnd base is not included in the display of the feature).

- VCF format supports complex rearrangements with breakends, that can arbitrary be summarized as a set of novel adjacencies, as described in the Variant Call Format Specification VCFv4.3 (Jul 2017).

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 (http://lbgi.fr/AnnotSV/)

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

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

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

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: What is the overlap used by AnnotSV between an annotated SV and features?
By default, a feature (DGV, DDD) is reported if the annotated SV is overlapped at least at 70% by the feature. Nevertheless, the user can choose to use a different percentage or a reciprocal overlap by defining new values of the "SVfromDBoverlap" and "SVtoAnnOverlap" options.

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

Q: How to interpret the presence of my SV in DGV or DDD databases?
The presence of a SV from your sample in DGV or DDD does not necessarily imply a disease causing event. Healthy carriers of pathogenic SV do exist in either databases. When available allele frequency can be helpful

to decide on the status. Nevertheless, DGV is populated with healthy samples whereas DDD is presenting affecting patients.

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

8. REFERENCES

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

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

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

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

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

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

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