knotAnnotSV Manual

Version 1.1.5

https://github.com/mobidic/knotAnnotSV

Please feel free to contact me for any suggestions or bug reports
email: t-guignard@chu-montpellier.fr

TABLE OF CONTENTS

1. INTRODUCTION

knotAnnotSV is a program designed to create customizable html or xlsm files from an AnnotSV output file. This will allow to easily filter and explore each SV. Annotations can be directly displayed as a spreadsheet format or in the web browser or by mouse hovering (tooltip).

By default, the output generation is configured to optimize the analysis by clinicians and biologists. However, thanks to a configuration file, the user can:

- Customize the order, the choice, the name and the description of the annotation columns
- Choose to display the annotation directly or in a tooltip

Moreover, several functionalities (search, highlight, links to public databases, color coded information…) are available to help the user.

2. REQUIREMENTS / INSTALLATION

a) Requirements

The knotAnnotSV program is written in the Perl language. Modern Unix systems have this scripting language already installed. knotAnnotSV requires the following 2 Perl libraries (available via CPAN):

- YAML::XS
- Sort::Key::Natural
- Excel::Writer::XSLX (for spreadsheet output only)

b) Installation

Please use git to download the most recent development tree.

The sources can be cloned to any directory:

```
cd /path/to/install/
git clone https://github.com/mobidic/knotAnnotSV.git
```

3. INPUT

a) Input files

knotAnnotSV takes an AnnotSV output file (‘*file*’.annotated.tsv) and a configuration file as inputs.

Configuration file:

knotAnnotSV uses an indented yaml file (e.g. config_cyto.yaml) to configure the AnnotSV output file. A default configuration file is already distributed with the installation:

config_AnnotSV.yaml

These configuration file is used by knotAnnotSV to:

- Define the POSITION (column ordering) of each field you want to display
- Define in COMMENTLIST which associated fields to display in tooltips (by mouse hovering)
- Define in RENAME which column name to display in the output
- Define in HEADERTIPS some information about the column you want to display in the header tooltips (by mouse hovering)

- Inactivate fields that you don't care about either with a starting '#' or 'POSITION: 0' or by deleting the corresponding line

NOTE: use space instead of tab for indentation, tabs are not allowed

b) Arguments

knotAnnotSV takes several arguments as input including options that are detailed in "USAGE / OPTIONS" section. The different arguments can be passed to the program using the command line.

4. OUTPUT

a) Output formats

a. Web – HTML

Giving an AnnotSV tab-separated values file, an html file is created that can be displayed in a web browser.

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 to continue the analysis.

Windows Browser Compatibility:
Firefox 84
Chrome 87
Safari 12.1+
Edge 87
IE 11

b. Spreadsheet – XLSM

Giving an AnnotSV tab-separated values file, an xlsm file is created that can be displayed in a Spreadsheet software.

Spreadsheet Software Compatibility:
Excel 2010+
WPS
OpenOffice
OnlyOffice
Numbers

b) Output file path(s) and name(s)

Two options (--outDir and --outPrefix) can be used to specify the output directory and/or file name.
By default, the html file is written in the current directory and named by replacing the ".tsv" extension of the AnnotSV file by ".html" or ".xlsm".

c) Compact and expanded visualization modes

AnnotSV context:
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 (i.e. “split”). This is possible thanks to the 2 types of lines provided by AnnotSV:

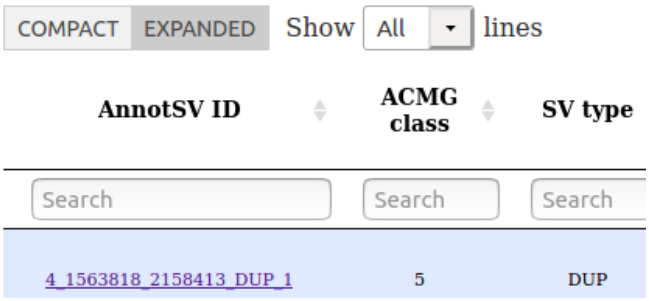
- An annotation on the **“full”** length of the SV. Every SV are reported, even those not covering any gene. This type of annotation gives an estimate of the SV itself.

- 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.

Display modes in knotAnnotSV:

Web-html flavour:
Two display modes are available in the web browser by clickable buttons:

- “**Compact**”: only the “**full**” AnnotSV lines are displayed
- “**Expanded**”: the “**full**” and “**split**” AnnotSV lines are displayed

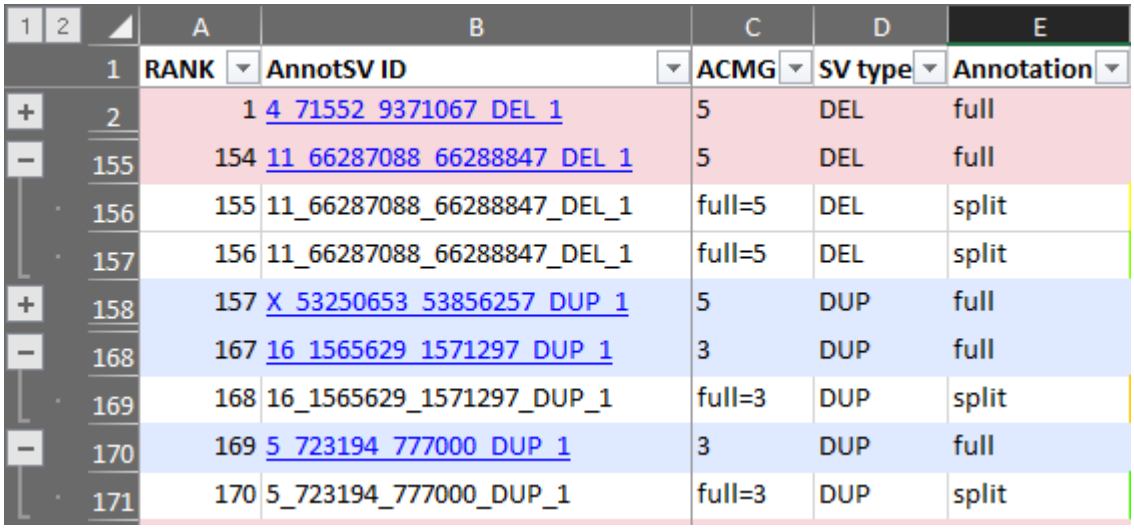


- **Single SV focus:** In the compact mode, a double-click on a « full » line focuses on the relatives « split » lines. Double-click on the same « full » line to come back to the COMPACT mode. Warning: double-click in the first column cells is disabled for this function.

Spreadsheet-xlsm flavour:

Two display modes are available in the spreadsheet grouped into 2 levels. There are clickable buttons on the **top left corner** of the sheet:

- " **1** ": only the "**full**" AnnotSV lines are displayed (similar to ***Compact***)
- " **2** ": All the "**split**" and "**full**" AnnotSV lines are displayed (similar to ***Expanded***)

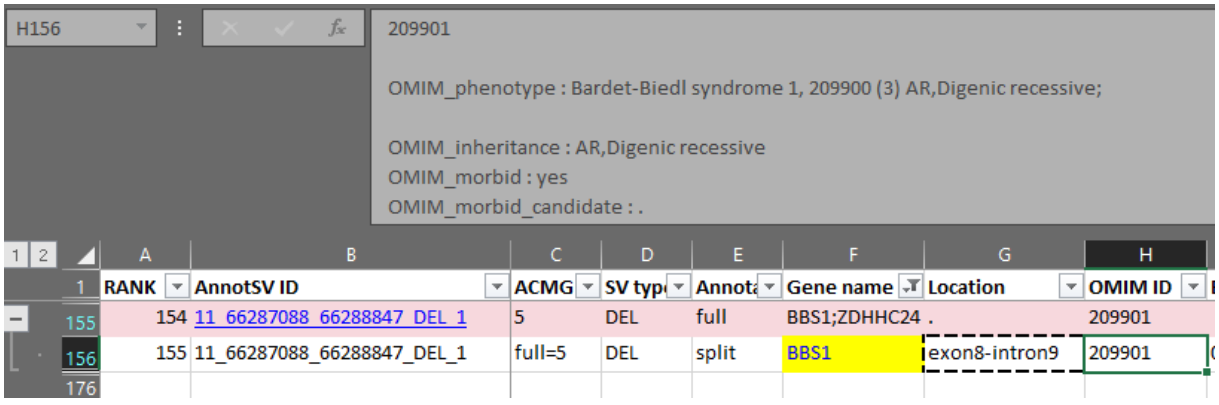


	A	B	C	D	E
1	RANK	AnnotSV ID	ACMG	SV type	Annotation
2	1	4_71552_9371067_DEL_1	5	DEL	full
155	154	11_66287088_66288847_DEL_1	5	DEL	full
156	155	11_66287088_66288847_DEL_1	full=5	DEL	split
157	156	11_66287088_66288847_DEL_1	full=5	DEL	split
158	157	X_53250653_53856257_DUP_1	5	DUP	full
168	167	16_1565629_1571297_DUP_1	3	DUP	full
169	168	16_1565629_1571297_DUP_1	full=3	DUP	split
170	169	5_723194_777000_DUP_1	3	DUP	full
171	170	5_723194_777000_DUP_1	full=3	DUP	split

Moreover, each line is individually expandable or foldable with " **+** " and " **-** " grouping buttons on the left side.

As well, an included vba macro (Excel only) can expand or fold the following lines by double-click on a line.

Complementary annotations are available within each cell. You can extend formula bar to see the whole annotations. (similarly to ***HTML tooltip annotation***). In this example, clicking on the *H156* cell will display all OMIM related annotations.



	A	B	C	D	E	F	G	H
1	RANK	AnnotSV ID	ACMG	SV typ	Annota	Gene name	Location	OMIM ID
155	154	11_66287088_66288847_DEL_1	5	DEL	full	BBS1;ZDHHC24	.	209901
156	155	11_66287088_66288847_DEL_1	full=5	DEL	split	BBS1	exon8-intron9	209901
176								

SV lines color code:
The full lines are highlighted depending on their SV type. Split lines have a white background.

DEL, <CN[0-1]>	DUP, <CN[3-9]>	INS	INV	BND	CPX	Anything else

d) Default order of the annotation lines

By default, the annotation lines are sorted according to these prioritization rules:
FULL Lines:

- ACMG class
  - DEL > DUP > others types
    - Gene Count
      - Exomiser Maximum Score
        - OMIM morbid (or OMIM candidates)

SPLIT Lines:

- Exomiser Score x [sum of weighted parameters]***
  - OMIM morbid > OMIM candidates
    - PhenoGenius Specificity
      - TS with DUP or HI with (DEL or truncated gene)
        - LOEUF bin (applied on split lines only)

*** Exomiser Score x
[Exomiser Score (x50)
+ PhenoGenius Specificity (A=30, B=20, C=10)
+ OMIM morbid (+30) or OMIM candidates (+15)
+ TS and DUP (+20)
+ HI and (DEL or truncated gene) (+20)
+ (10 - LOEUF bin)]

e) Available functions

Top of the HTML page:

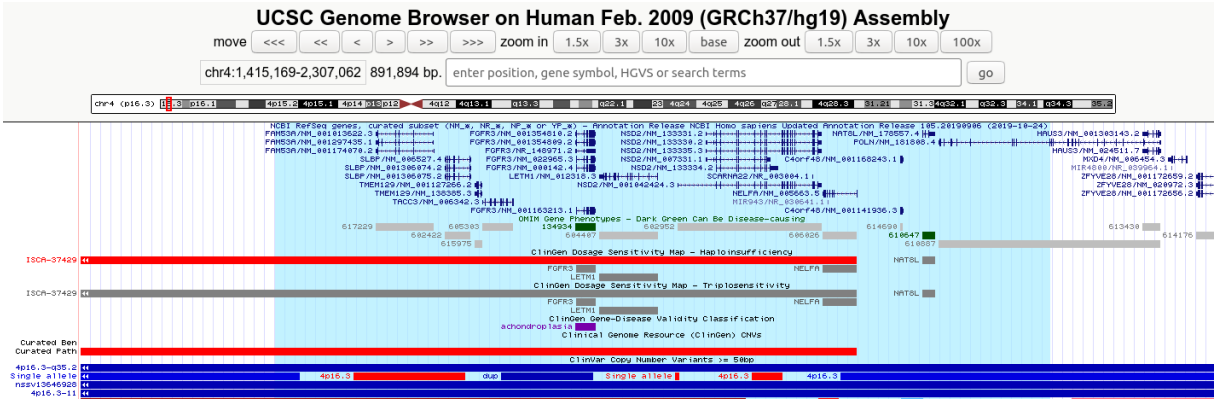
- Click to choose the desired visualization mode. For more explanation, see the « Compact and expanded visualization modes » section.
- Choose the number of lines to be displayed (All or 100) in the upper left corner
- To search for a topic (gene name, phenotype…), just type a word or phrase in the Search box in the upper right corner

Header:

- Display the definition of the column name by mousing hovering (tooltip)
- Click small triangle in each column to sort the data
- Below each column name, a search box is available for:
  - Searching words or extracting matching records
  - Filtering a range of numerical values (with the “>”, “>=”, “<” and “<=” symbols)
    e.g. to select frequencies smaller than 1%, type “<0.01”
  - Filtering out some numerical values (“!= value”)
  - Filtering out some words (“! word”)
- Search is also performed in the tooltips of the OMIM, Pathogenic and Gene Name fields.

Annotation lines:

- The « AnnotSV ID » column is freezed
- Hover the mouse on each annotation to show complementary information (*tooltip for web output only*)
- Double-click on a « full » line to focus on the relatives « split » lines (only for COMPACT mode). Double-click on the same « full » line to come back to the COMPACT mode
- Click on the « AnnotSV ID » to open the SV coordinates in the UCSC Genome browser (the SV region is automatically highlighted in blue and zoomed out by 1.5x)



- Click on the various blue links to access directly to the corresponding public database (OMIM, genecards)

Footer (web only):

- Show the count of filtered lines (as well as the total entries count)
- Navigate between the different pages

NOTE:

- 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.
- To come back to the default annotation lines sorting , the user can click on the header of the first column (AnnotSV ID).

f) "Gene name" box color

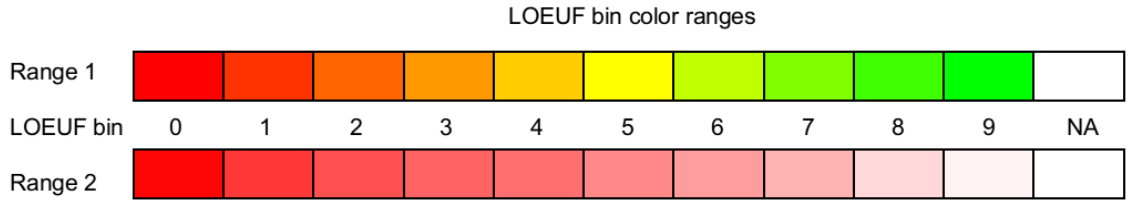
Two different palettes of colors can be used to visualized for each gene its corresponding GnomAD defined LOEUF bin value (user defined, see the USAGE/OPTIONS section), smaller values reflect loss of function intolerance:

- A red-to-green palette (default):

The red color means the lower LOEUF_bin value (0). The green color means the higher LOEUF_bin value (9).

- A sequential palette of red:

Darker colors mean lower LOEUF_bin values (e.g 0).



NOTE: Depending on the overlapped gene part, the « Gene name » box is fully colored or not:

- Gene totally overlapped: fully colored box

TMEM129

- Gene partially overlapped: half colored box

5' of the gene is overlapped:

TMEM129

3' of the gene is overlapped:

TMEM129

For the spreadsheet ouput, partial overlap is highlighted with a black dashed line border on the side, in the Location column.

F	G
Gene name	Location
BBS1	exon8-intron9
ZDHHC24	exon5-txEnd
KDM5C	txStart-intron1
RIBC1	txStart-txEnd

g) "RE_gene" processing

Gene functions can be impacted by SV overlap in different ways:

- The SV overlaps partially/totally the gene
- The SV overlaps partially/totally the gene AND a regulatory element of the gene
- The SV overlaps ONLY a regulatory element of the gene

In order to easily distinguish genes where only a regulatory element (and not the gene itself) is overlapped with the SV, the "overlapped regulatory elements" field is split in two sets:

The first set shows genes whose ONLY regulatory elements are overlapped (hence absent from «Gene name» list).

Separated by multiple hyphens (-----), the second set contains genes for which the SV overlaps both the gene and a regulatory element of the gene.

Furthermore, the regulated genes are sorted according to their exomiser score and OMIM morbid features.

5. USAGE / OPTIONS

To run knotAnnotSV, the default command line is the following:

```
cd /path/to/install/knotAnnotSV
```

1. Web-flavour output:

```
perl ./knotAnnotSV.pl --configFile config_AnnotSV.yaml --annotSVfile
file.annotated.tsv
```

The command line can be completed by some options.
To show the options simply type:

```
perl ./knotAnnotSV.pl

--configFile <YAML config file for customizing output>
--annotSVfile <AnnotSV annotated file>
--outDir <output directory (default = current dir)>
--outPrefix <output file prefix (default = "")>
--genomeBuild <Genome Assembly reference (default = hg19)>
--LOEUFcolorRange <Number to define which color to use for LOEUF bin:
1 (red-to-green), 2 (red-shades-only). (default = 1)>
--datatableDir <Local Path to DataTables directory containing css and
js files (default = "", requires web connection)>
```

Or

2. Spreadsheet-flavour output:

```
perl ./knotAnnotSV2XL.pl --configFile config_AnnotSV.yaml --
annotSVfile file.annotated.tsv
```

The command line can be completed by some options.
To show the options simply type:

```
perl ./knotAnnotSV2XL.pl

--configFile <YAML config file for customizing output>
--annotSVfile <AnnotSV annotated file>
--outDir <output directory (default = current dir)>
--outPrefix <output file prefix (default = "")>
```

```
--genomeBuild <Genome Assembly reference (default = hg19)>
--LOEUFcolorRange <Number to define which color to use for LOEUF bin: 1 (red-to-green), 2 (red-shades-only). (default = 1)>
--geneCountThreshold <Maximum number genes (split lines) to output, omim morbid genes or exomiser > 0.7 will be kept anyway (inactive by default, 40 is advised )>
```

6. Test

In order to validate the knotAnnotSV installation and to facilitate its use we have provided with the installation an input/output example in the « example » folder.

1. Change to the repo directory, and run the example

```
cd /path/to/install/knotAnnotSV/
perl ./knotAnnotSV.pl --annotSVfile ./example/example.annotated.tsv --configFile ./config_AnnotSV.yaml --outDir ./example/
```

or

```
perl ./knotAnnotSV2XL.pl --annotSVfile ./example/example.annotated.tsv --configFile ./config_AnnotSV.yaml --outDir ./example/
```

2. Display the html output on a web browser / the xlsm output on a spreadsheet software.

3. Have fun with the exploring!

7. FAQ

Q: What does knotAnnotSV means ?
If pronounced like « knot a knot SV », knotAnnotSV is a word play meaning that this tool creates a link between the SV data to be analysed :o)