Identifying structural variants with SVDetect

INSERM - NGS Bioinformatics Course, Roscoff - January 17th 2013

Bruno Zeitouni – contact: svdetect@curie.fr

Most approaches for predicting structural variants (SVs) require you to have paired-end or mate-pair reads. They use the distribution of distances separating these reads to find outliers and also look at pairs with incorrect orientations/orders or mapped on different chromosomes.

Good discussion of some of the issues of predicting structural variation:

Computational methods for discovering structural variation with next-generation sequencing
Paul Medvedev et al, Nature Methods 6, S13 - S20 (2009)

Genome structural variation discovery and genotyping. Alkan C et al, Nat Rev Genet. (2011)

Get SVDetect

Choice 1: Command-line version

Navigate to the SVDetect sourceforge project

http://sourceforge.net/projects/svdetect

More information:

- SVDetect website : http://svdetect.sourceforge.net

- SVDetect paper:

SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data
Zeitouni B et al, Bioinformatics 2010 26: 1895-1896

Try to download the code yourself and install it in your Linux OS based computer.
SVDetect requires a few Perl modules to be installed and has third-party tool dependencies.

Choice 2: Galaxy version

- Go to the Galaxy toolshed website: http://toolshed.g2.bx.psu.edu

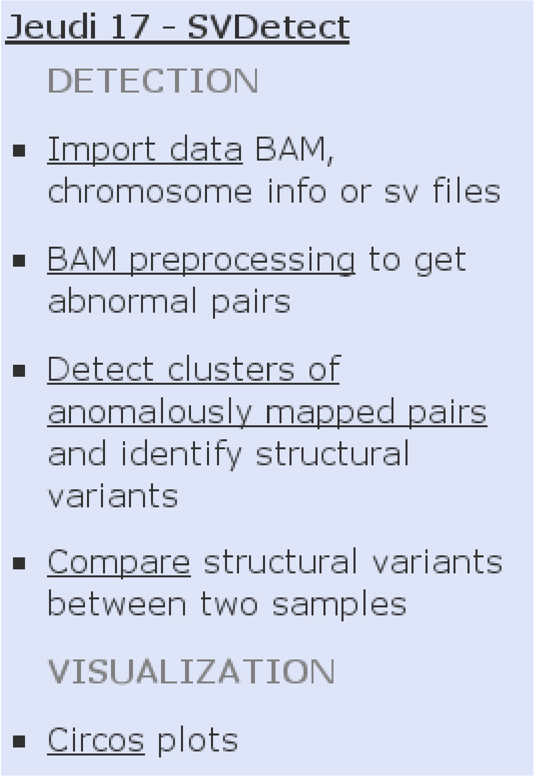
- Search for “svdetect”

- On the SVDetect software page, download the installation archive and follow the instructions to install it on your own Galaxy server

For this workshop, SVDetect has been already set up in the Galaxy Roscoff website:

- Go to the Galaxy server : http://galaxy.sb-roscoff.fr

- Click on the “Jeudi 17 SVDetect” link to see the different tool functions at the left panel (tools) :


Jeudi 17 - SVDetect
DETECTION
Import data BAM, chromosome info or sv files
BAM preprocessing to get abnormal pairs
Detect clusters of anomalously mapped pairs and identify structural variants
Compare structural variants between two samples
VISUALIZATION
Circos plots

Each of the functions takes one or more input files and returns results. They have to be runned successively:

Main programs:

1- Import the data in the adapted file format
2- Preprocess the mapping data before running the variant detection
3- Run the SVDetect process to get clusters of pairs and identify structural variants

Optional programs:

4- Compare SVs between two samples
5- Visualization of SVs through a Circos plot or on the UCSC genome browser

Import data

This is Illumina GAII mate-paired data (having a larger insert size than paired-end data, here 3kb) from whole genome re-sequencing of a Neuroblastoma cell line. The non-tumoral cell line has been also sequenced as the control dataset of the experiment.

The mate-pair sequencing FASTQ files were aligned to the Human hg18 genome with the BWA mapping software using the `bwa aln` and the `bwa sampe` commands with default parameters. Binary alignment SAM files (BAM) for each sample was provided by BWA.

<!> To use SVDetect, make sure you are using a tool handling paired-read mapping and providing the output results in the SAM format. Ex: BWA, BFAST, Bowtie2

SVDetect requires two types of input files:

- a BAM file, the alignment files of pairs on the genome
- a LEN file, listing chromosome lengths in the following tabulated- txt format:

```
1     chr5       180857966
2     chr11      134452384
```

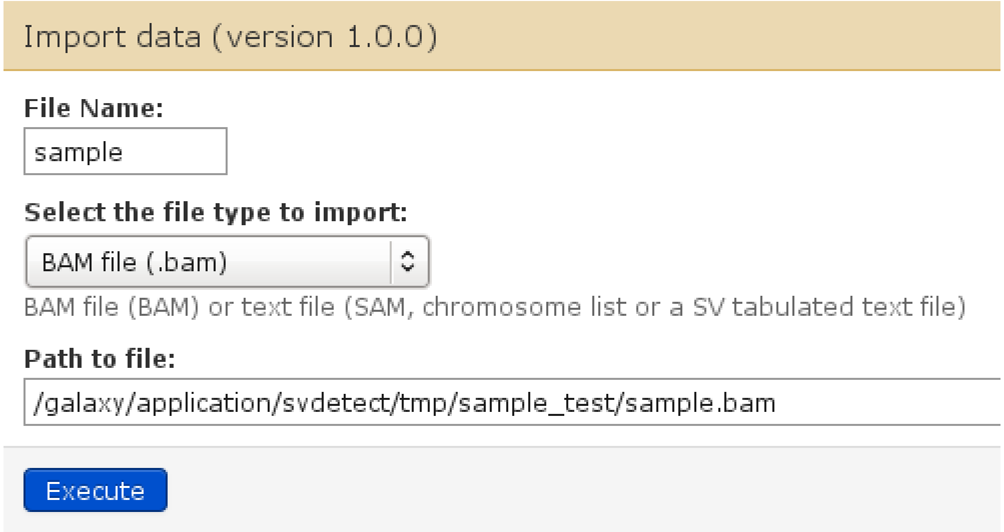
Example:

File name	**Description**	**Sample**
sample.bam	BAM file format, Illumina mate pair	Neuroblastoma cell line
reference.bam	BAM file format, Illumina mate pair	Normal cell line
hs18_chr5_chr11.len	Chromosome file	hg18 *homo sapiens* genome chr5 and chr11 only*

* Only the chromosomes 5 and 11 are used here given the time of the training

The first step to do is to import these data into the Galaxy interface:

- Click on the “Import data BAM, chromosome and sv files”
- To import the sample dataset, fill the corresponding form as followed and “Execute”:



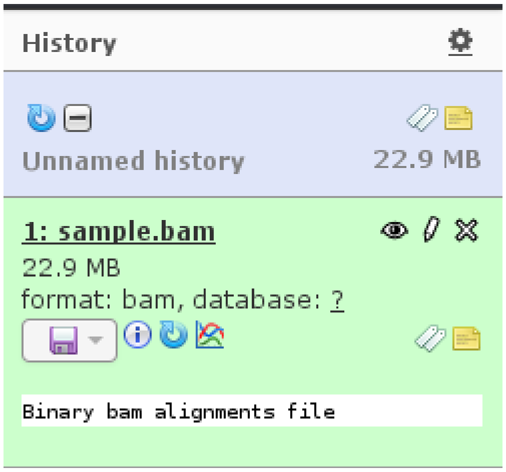
Import data (version 1.0.0)

File Name:
sample

Select the file type to import:
BAM file (.bam)
BAM file (BAM) or text file (SAM, chromosome list or a SV tabulated text file)

Path to file:
/galaxy/application/svdetect/tmp/sample_test/sample.bam

Note: as BAM files are usually big in size, these files have to be present in a location with a open access to Galaxy. This avoids a long uploading process with the “upload file” command of Galaxy. Only a path to the file location has to be provided and the importation is then immediate.



Unnamed history 22.9 MB

1: sample.bam
22.9 MB
format: bam, database: ?

Binary bam alignments file

The sample.bam file is then loaded in the history (right panel).

- Do the same with the reference.bam file by providing the path : /galaxy/application/svdetect/tmp/sample_test/reference.bam

- Load the chromosome file by fill the corresponding fields in the same way:

File Name: hs18_chr5_chr11
File Type: Chromosome info file (.len)
File path: /galaxy/application/svdetect/tmp/sample_test/hs18_chr5_chr11.len

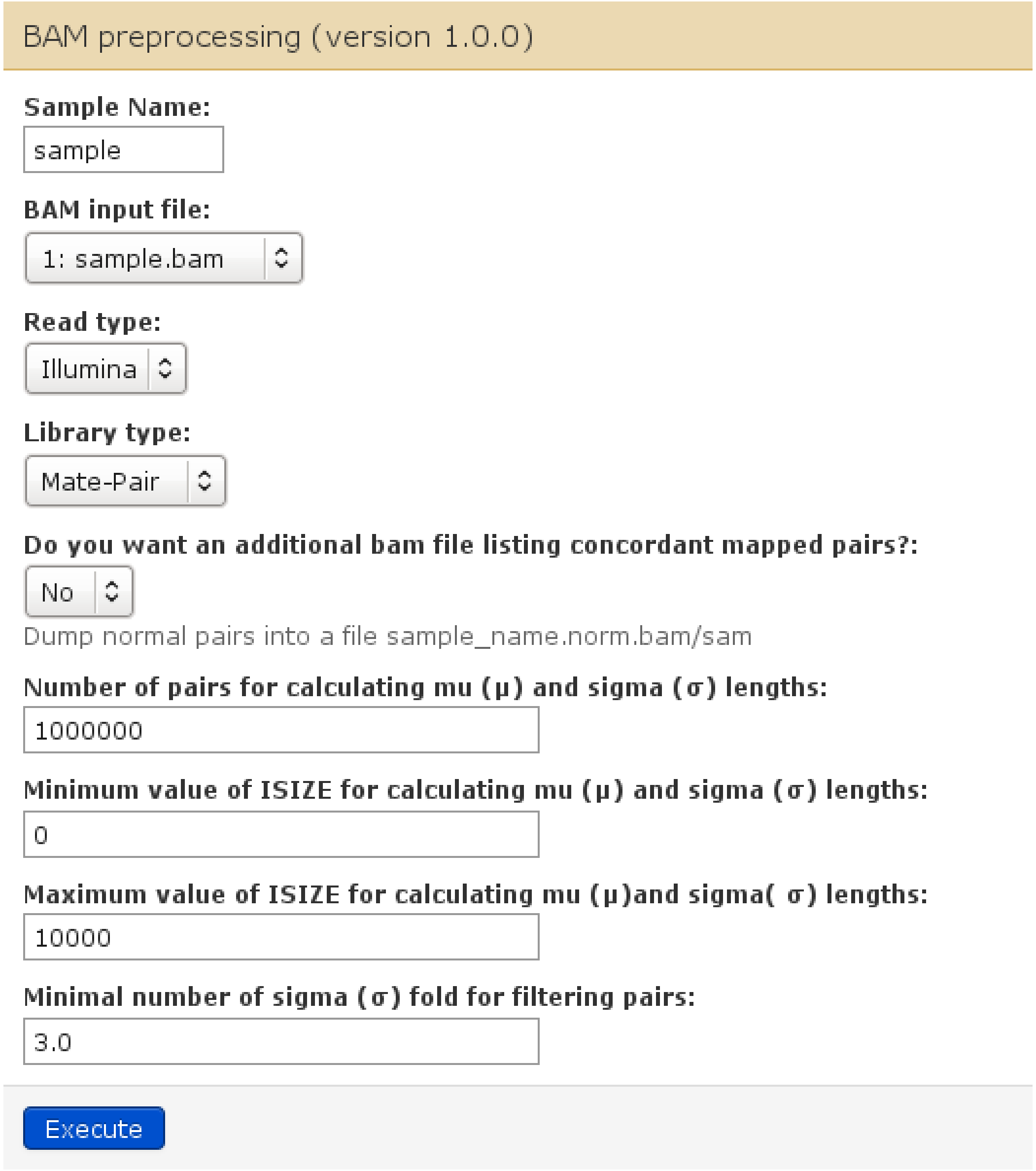
The 2 BAM files and the chromosome files are now in your history.

Preprocess data

The second step is to look at all mapped read pairs and whittle down the list only to abnormal pairs, i.e those that have an unusual insert size (distance between the two reads in a pair), incorrect read order or orientations (strand) or mapped on two different chromosomes.

- Click on the “BAM preprocessing to get abnormal pairs”

- Fill the form as follows to get anomalously mate-pairs reads of the sample dataset:



BAM preprocessing (version 1.0.0)

Sample Name:
sample

BAM input file:
1: sample.bam

Read type:
Illumina

Library type:
Mate-Pair

Do you want an additional bam file listing concordant mapped pairs?:
No
Dump normal pairs into a file sample_name.norm.bam/sam

Number of pairs for calculating mu (μ) and sigma (σ) lengths:
1000000

Minimum value of ISIZE for calculating mu (μ) and sigma (σ) lengths:
0

Maximum value of ISIZE for calculating mu (μ)and sigma(σ) lengths:
10000

Minimal number of sigma (σ) fold for filtering pairs:
3.0

After selecting the corresponding BAM file and the type of reads/paired library, value parameters for filtering abnormal pairs according the insert size have to be set:

  - Number of pairs for operating statistics necessary for the mu and sigma calculation, by default this is set to the first million of pairs in the file,
  - Maximum/Minimum allowed value of the insert size for these calculations,
  - Threshold of the standard deviation to keep pairs out of the range only.

- Execute

  The sample.ab.bam file is then loaded in the history.

- Check the output log file of the command

  Q1: What is the normal insert size of the library?

  Q2: How many abnormal pairs did you get?
  List the different types of these pairs

  Keep in mind the calculated mu and sigma values.

- Modify the different values parameters and observe the statistics changes

- Repeat the procedure with the reference dataset with default parameters
  (keep the first sample.ab.bam in your history)

Get clusters of pairs and Call SVs

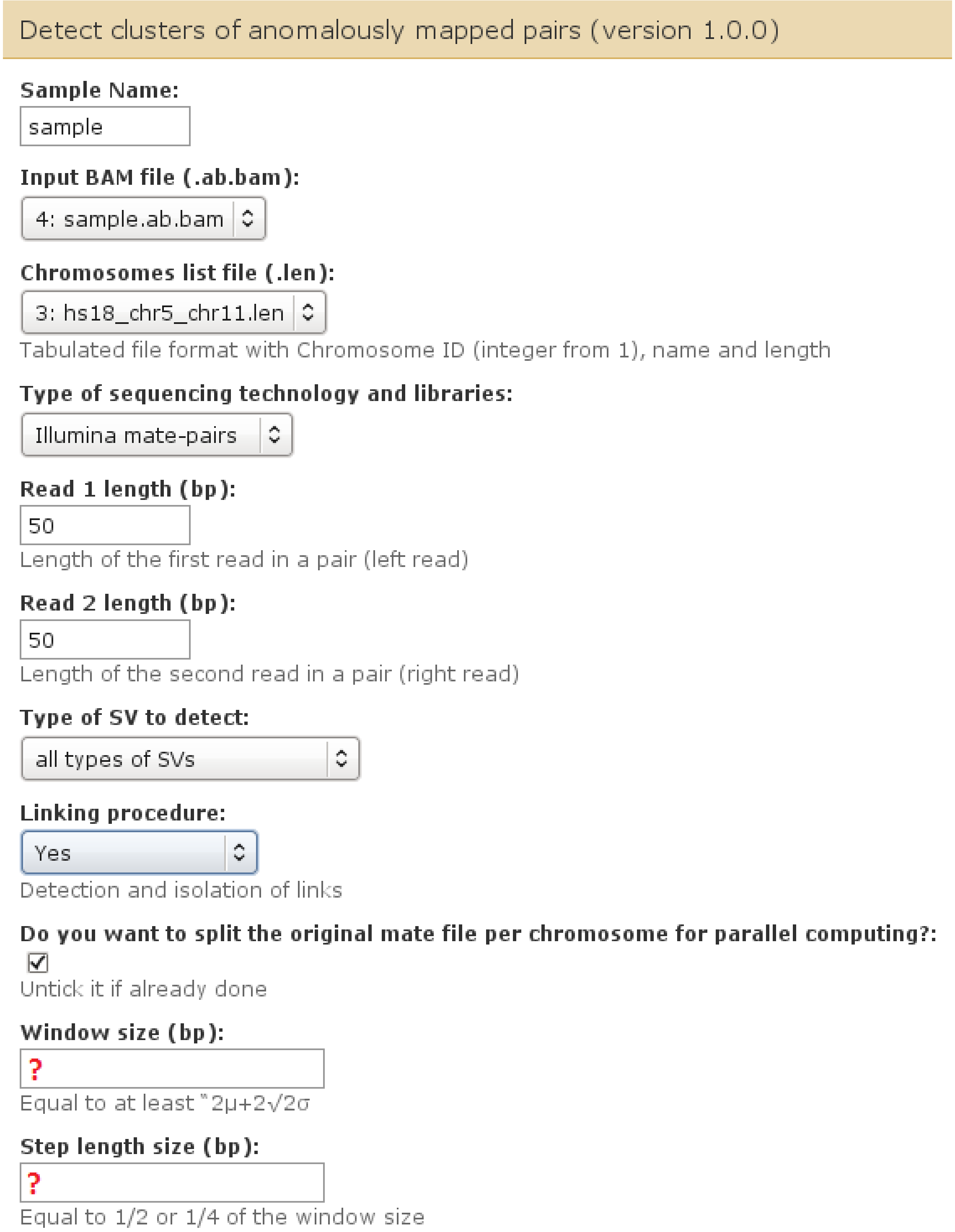
From abnormal mate-pairs of both datasets, we are ready to launch the next step with the linking and filtering functions. This is the main program of SVDetect to isolate clusters of consistent mate-pairs and to call specific SVs from mapped paired-read signatures. After processing this step, the list of significant SVs and the corresponding breakpoints will be available.

This step requires filling a lot of parameters values that can influence the results and the time of processing. It is important to understand their meaning and to set them properly.

The default parameters may not be adapted to your data features.

- Click on the “Detect clusters of anomalously mapped pairs and identify structural variants” link

- Fill the first part of the form “linking” like below :



Detect clusters of anomalously mapped pairs (version 1.0.0)

Sample Name:
sample

Input BAM file (.ab.bam):
4: sample.ab.bam

Chromosomes list file (.len):
3: hs18_chr5_chr11.len
Tabulated file format with Chromosome ID (integer from 1), name and length

Type of sequencing technology and libraries:
Illumina mate-pairs

Read 1 length (bp):
50
Length of the first read in a pair (left read)

Read 2 length (bp):
50
Length of the second read in a pair (right read)

Type of SV to detect:
all types of SVs

Linking procedure:
Yes
Detection and isolation of links

Do you want to split the original mate file per chromosome for parallel computing?:
☑
Untick it if already done

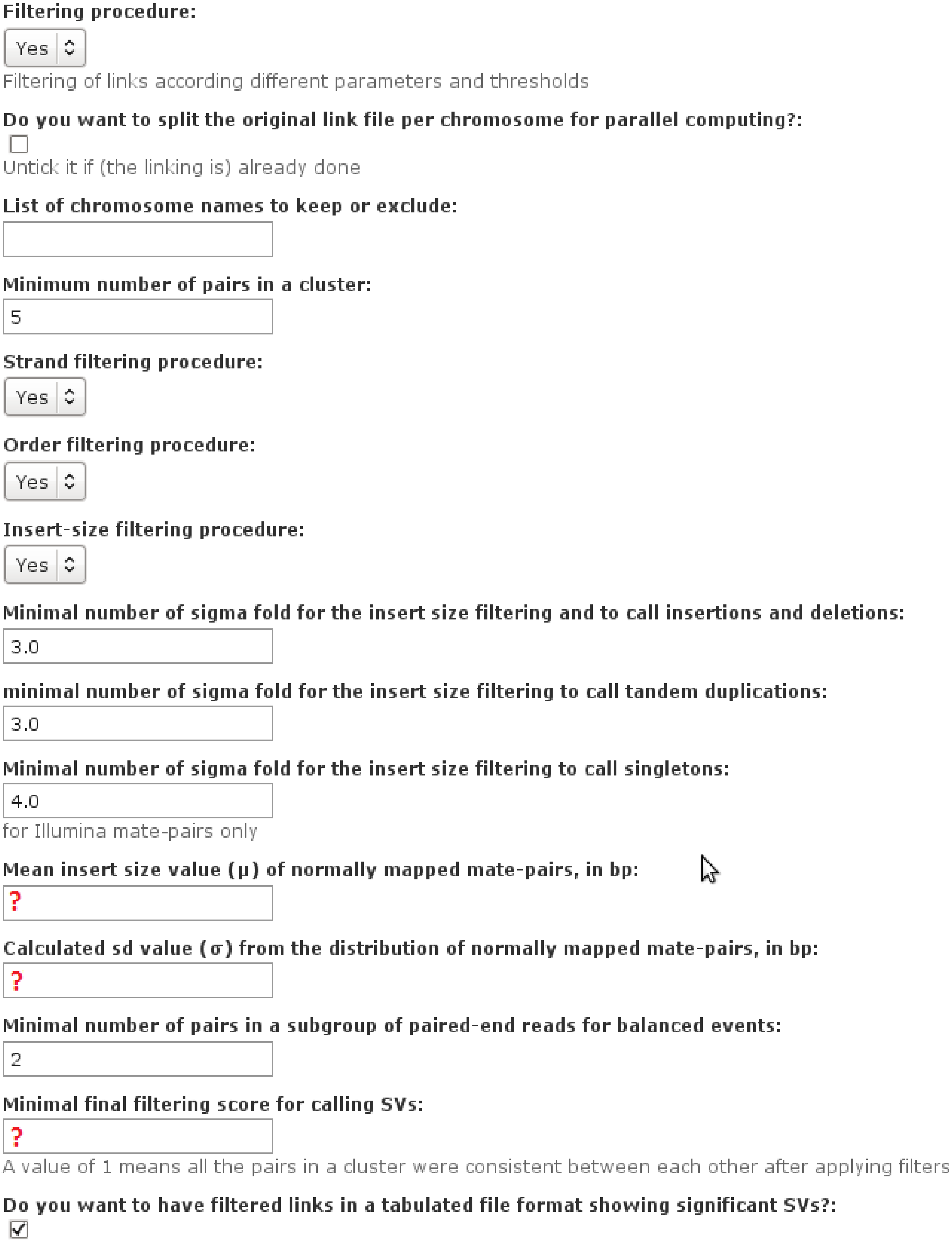
Window size (bp):
?
Equal to at least ˜$2\mu+2\sqrt{2}\sigma$

Step length size (bp):
?
Equal to 1/2 or 1/4 of the window size

Q: What the meaning of the window and the step length sizes?

Complete the expected values showed with “?”

- Keep filling the form with the "filtering" part of the process :



Filtering procedure:

Yes

Filtering of links according different parameters and thresholds

Do you want to split the original link file per chromosome for parallel computing?:

□

Untick it if (the linking is) already done

List of chromosome names to keep or exclude:

Minimum number of pairs in a cluster:

5

Strand filtering procedure:

Yes

Order filtering procedure:

Yes

Insert-size filtering procedure:

Yes

Minimal number of sigma fold for the insert size filtering and to call insertions and deletions:

3.0

minimal number of sigma fold for the insert size filtering to call tandem duplications:

3.0

Minimal number of sigma fold for the insert size filtering to call singletons:

4.0

for Illumina mate-pairs only

Mean insert size value (μ) of normally mapped mate-pairs, in bp:

?

Calculated sd value (σ) from the distribution of normally mapped mate-pairs, in bp:

?

Minimal number of pairs in a subgroup of paired-end reads for balanced events:

2

Minimal final filtering score for calling SVs:

?

A value of 1 means all the pairs in a cluster were consistent between each other after applying filters

Do you want to have filtered links in a tabulated file format showing significant SVs?:

☑

Q: What the meaning of the 3 different thresholds set with sigma?

Complete the expected values showed with “?”

- Leave the last optional fields of outputs as default

- Execute

The process is the most consuming time process. It can take a while. As this dataset is small, the results should be provided quickly after submitting the job to the cluster.

The sample.sv file is then loaded in the history.

- Take a look at the resulting file:

```
chr_type        SV_type BAL_type        chromosome1     start1-end1     average_dist    chromosome2
        start2-end2     nb_pairs        score_strand_filtering score_order_filtering
        score_insert_size_filtering     final_score     breakpoint1_start1-end1
        breakpoint2_start2-end2

INTRA   INSERTION       UNBAL   chr11   165334-165887   1535    chr11   165982-168016   39      100%
        100%    100%    1       165887-169106   164244-165982

INTRA   DELETION        UNBAL   chr11   1912273-1915186         24661   chr11   1937112-1939824
        27      100%    100%    100%    1       1915186-1916045         1936052-1937112

INTRA   INVERSION       UNBAL   chr11   1935007-1936551         24162   chr11   1959072-1960635
        22      100%    100%    -       1       1936551-1938779         1960635-1962844

INTRA   DELETION        UNBAL   chr5    133132181-133134639     4708    chr5    133137034-133139510
        21      100%    100%    100%    1       133134639-133135953     133135738-133137034

                        . . . .
```

Q1: Where are the coordinates ranges of breakpoints?

Q2: Any idea what sorts of mutations produced these three structural variants?

Why?

Q3: Try to schematize the different SVs by their corresponding paired read signatures (coordinates, pairs, breakpoints)

- Repeat exactly the same procedure but with the reference dataset.

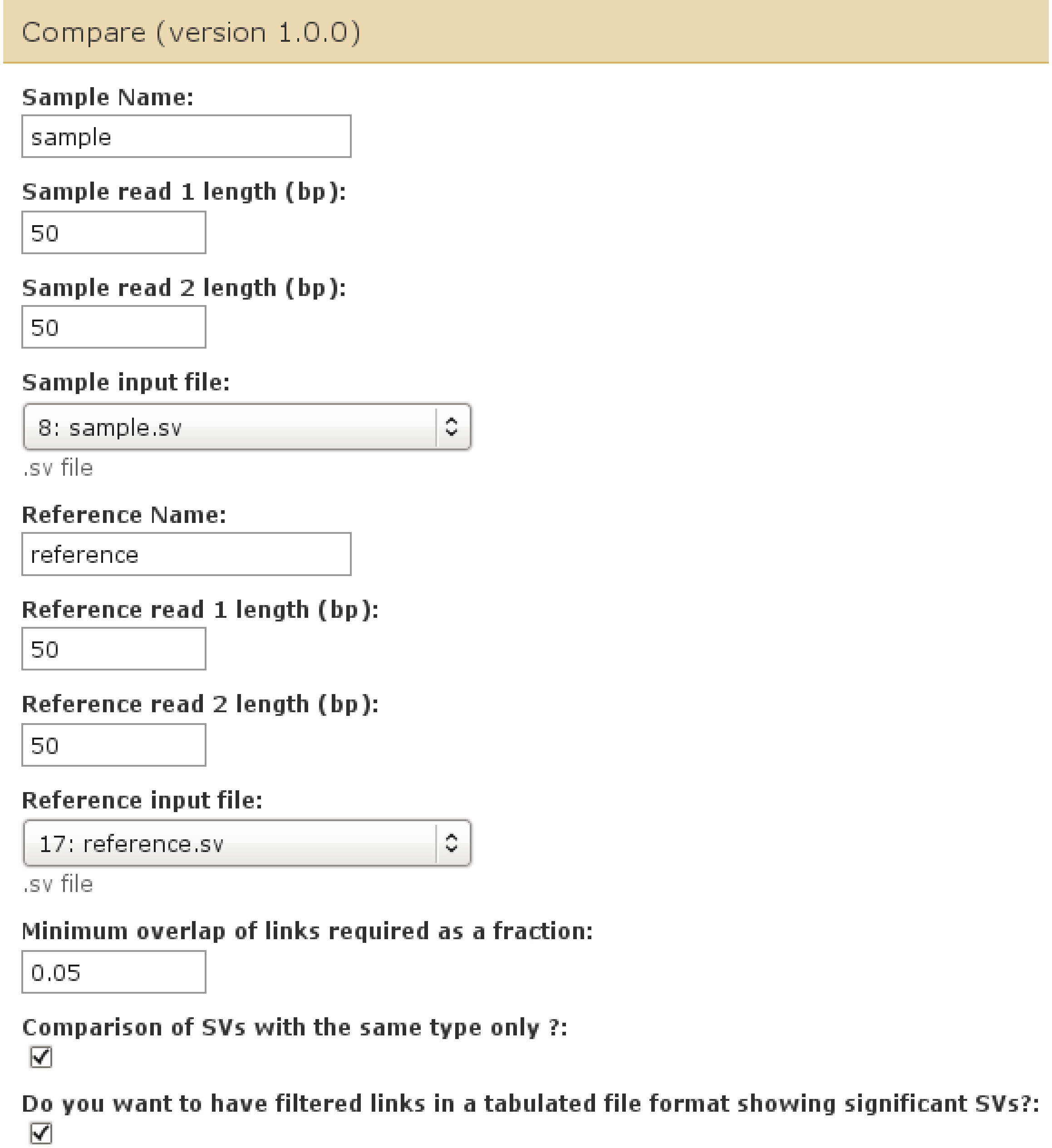
<!> You can reload the same parameters with the ↻ button

Get specific SVs and outputs

A very useful option of SVDetect is to compare two samples and to get both common SVs and sample-specific SVs. In our example, we want to see only the SVs that are specific to the cancer sample Neuroblastoma and to discard those which are constitutively present in the cell line.

- Click on the “Compare structural variants between two samples”

- Fill the form as follows to compare SVs between the sample and the reference dataset:



Compare (version 1.0.0)

Sample Name: sample

Sample read 1 length (bp): 50

Sample read 2 length (bp): 50

Sample input file: 8: sample.sv

.sv file

Reference Name: reference

Reference read 1 length (bp): 50

Reference read 2 length (bp): 50

Reference input file: 17: reference.sv

.sv file

Minimum overlap of links required as a fraction: 0.05

Comparison of SVs with the same type only ?:
☑

Do you want to have filtered links in a tabulated file format showing significant SVs?:
☑

- Leave the last optional fields of output as default (see the next chapter for more details)

- Execute

  After completion, you can find 3 new sv files in your history:

  - common.compared.sv: list of SVs commons to the two datasets
  - sample.compared.sv: list of SVs specific to the neuroblastoma sample
  - reference.compared.sv: list of SVs specific to the cell line

Graphical visualization of SVs

SVDetect has a nice option to visualize SVs graphically by providing output files that can be read either by Circos to produce drawings or by the UCSC Genome browser to see the corresponding mapped pairs on the genome.

This option can be used directly either after the filtering step or after making the comparison between samples.

We want to get a visualization of the specific SVs of the cancer sample.

- Click on the "Compare structural variants between two samples"

- Fill the form in the same way than before (use ↻ **)** but activate the "Output file conversion" option to "Yes"

- Tick on the two boxes for the two file format conversion options (Circos and USCS)

- Add the following color-codes according the number of pairs present in the SV clusters:

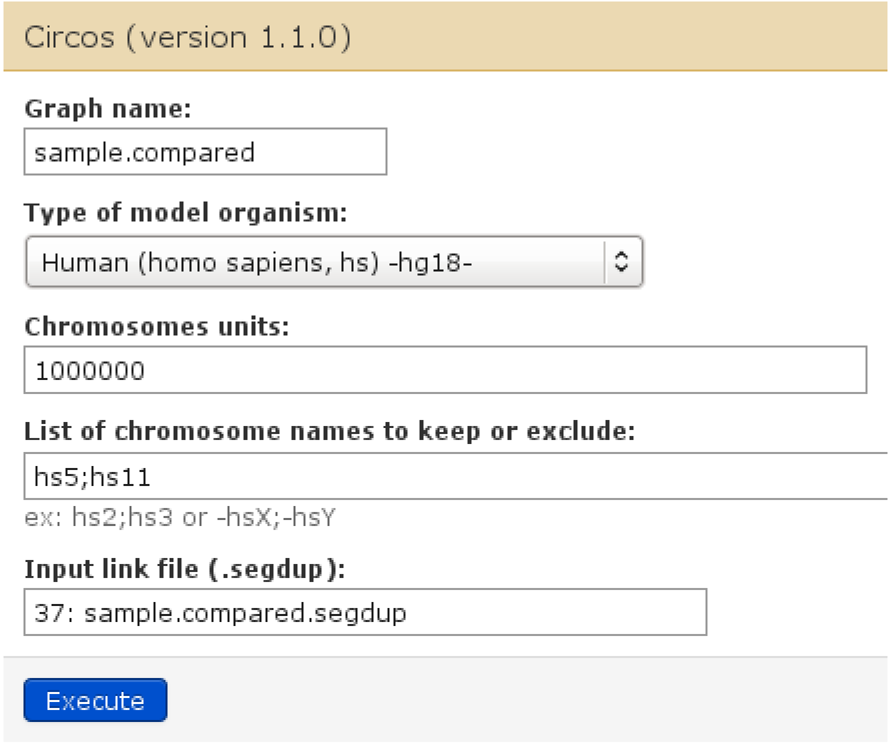
Color-code 1	Color: black	Interval: 1,3
Color-code 2	Color: blue	Interval: 4,6
Color-code 3	Color: green	Interval: 7,10
Color-code 4	Color: red	Interval: 11,1000

- Execute

After completion, you can find two additional datafiles for each set of comparison:

- xxx.compared.segdup: SVs in the Circos compatible format
- xxx.compared.bed: list of reads pairs included in SV clusters compatible with the UCSC browser (BED format)

- Click on the “Circos plots”
- To get the corresponding Circos of the sample-specific SVs, fill the form as follows:



Circos (version 1.1.0)

Graph name: sample.compared

Type of model organism: Human (homo sapiens, hs) -hg18-

Chromosomes units: 1000000

List of chromosome names to keep or exclude: hs5;hs11

ex: hs2;hs3 or -hsX;-hsY

Input link file (.segdup): 37: sample.compared.segdup

<!> Check the special field to plot the chromosomes chr5 and chr11 only.
You can try to leave blank this field and to look at the different resulting plot

- Execute
- Click on the 👁 button of the resulting “sample.compared.png” file to look at the result.
  Zoom in the Circos plots if needed.

Q: Can you connect the SVs in the Circos plot to those present in the SV file?

See the only inter-chromosomal rearrangement detected (“TRANSLOC”) and keep in mind the corresponding genomic coordinates.
Do the same for a deletion viewed both in the SV table and in the Circos plot.

- Click on the “sample.compared.bed” link in your history and save the BED file on your desktop by using the download 💾 button

- Go to the UCSC genome browser website : http://genome.ucsc.edu/

- Then Click on the “Genome Browser” link and select “Human” and “Mar. 2006 (NCBI36/hg18)”

- Click on the “add custom track” button and upload the bed file with “Parcourir” and looking for the file previously saved

- Click on the “go to genome browser” button

- In the “position, gene symbol or search terms” field, put the coordinates of the SVs that you have detected and visualize the correspond pairs on the human genome

- Repeat the search for different types of SVs and identify the possible impact of them on the genome