# Quick Guide

PeSV-Fisher is a pipeline for the detection of five general types of structural variants (SVs): deletions, gains, intra- and inter-chromosomal translocations, and inversions, at very reasonable computational costs. The pipeline further provides comprehensive information on co-localization of SVs in the genome, a key aspect for studying biological consequences. The algorithm uses a combination of methods based on paired-reads (PR) and read-depth strategies (RD). PeSV-Fisher has been designed with the aim to facilitate identification of somatic variation, and, as such, it is capable of analysing two or more samples simultaneously, producing a list of non-shared variants between samples, although it can also analyse individual samples.

## 1) Installation:

1. Unpack the source downloaded tarball.

2. Insert in your .bashrc and .bash_profile from HOME directory:

*export PSVFDIR="path of expanded folder, root dir where PeSV-Fisher tool is allocated"*

*export PSVFPYTHON="python3 path"*

*[3. source .bashrc (from your HOME dir). It is for your current bash session. It won't be necessary for your further bash sessions.]*

## 2) Requirements:

The tool has been developed under Linux SO. It has been tested with Fedora 12 and Red Hat (Red Hat 4.4.2-7)

- MySql. Checked : mysql Ver 14.14 Distrib 5.1.39, for redhat-linux-gnu (x86_64)
- Samtools. Checked: samtools-0.1.17
- **Python3.** Checked: python3.1

*[See the requirements section from README for more information]*

## 3) Run:

The tool needs only a configure file. The extension is .cfg, and it must not have any blanks in the middle.

*1. Go to the $PSVFDIR/bin*

*2. ./PeSVFisher configfile.cfg i.e: {$PSVFDIR/config/Template.cfg}*

**3) Configure file .cfg :**

The tool uses a configure file to launch different modules. This file centralizes all the analysis methods and their alternatives. The format is variable=value. You should not introduce blanks in the middle because it can generate errors or bad outputs.

**Definition of the Variables:**

***General variables:***

- ntchr : Int - number of chromosomes to analyze.
- nprocess : Int - number of processes for the paired-reads (PR) strategy.
- ndofcprocess : Int – number of processes for the read depth (RD) strategy.
- basedir : String – path where the PeSV-Fisher is installed.
- sortdir : String - if you don't have the sort command in your path, you should indicate where the tool should go to find it. By default sort.
- samtoolsdir : String - if you don't have the samtools in your path, you should indicate where the tool should go to find it. By default samtools.
- awkdir : String - if you don't have the awk command in your path, you should indicate where the tool should go to find it. By default awk.
- refaprefix : String – prefix (with full path, where the chromosomes of the reference genome are located in your system).
- hg : String – {hg18, hg19}name of the reference genome. By default is hg19.
- datadirbam : Directory where the tool should look for the input data (Bam files). The Bam files should be sorted and indexed. So they should have the sufix .bai in the same path.
- prefix : String – it is used for the non-somatic mode. You should include one or a list of samples ID which you would like to make the analysis separated by ",".
- control : String – It is used for the somatic mode. It indicates that the sample is a control. You can include one or a list of sample IDs separated by ",".
- disease : String – It is used for the somatic mode. It indicates that the sample is a tumor genome. You can include one or a list of sample IDs separated by ",". They should be listed with the same order as the IDs listed in the *control* variable, as they will be used as paired samples (a tumor paired with its normal sample).
- extfile : String – extension file. By default is "tab "
- format : String - pattern margin of files. By default is "\t" , tabular files.
- runALL: Int - Indicates the different modes of execution. See the next section of the Guide.

***Variables related to the paired-read strategy* (PR)*:***

- sv : Int - Indicates the different cluster types that you would like to call in the paired-reads (PR) strategy module. See next section of the Guide. By default 1 to do the complete analysis.
- quantile: frequency of read-pairs that we will assume as concordant, i.e. those falling into the expected insert range. By default 0.99.
- numclones : Int – minimum number of read-pairs within a cluster. By default 2.
- somatic : Int – {0,1} The tool will perform the somatic analysis if the value is set to 1, then the tool will use the paired-samples in control and disease fields. When the value is set to 0, independent analyses will be performed for each sample, in this case the tool use the prefix list value. By default 1 (somatic).
- posid : Int - ID position for the input of PeSV-Fisher core. By default 0

- poschr1 : Int - Position of first chromosome for the input of PeSV-Fisher core. By default 1
- posini : Int - Postion of where the first read start into reference. By default is 2
- posend : Int - Position of where the second start into reference. By default is 3
- poslen : Int - Position of insert size between reads. By default is 4
- poschr2 : Int - Position of the chromosome of the second read. By default is 5
- postrand1: Int – Position of the strand of first read. By default is 6
- postrand2 : Int – Position of the strand of the second read. By default is 7
- poscore1 : Int – Position of the aligment score of the first read. By default is 8
- poscore2 : Int – Position of the aligment score of the second read. By default is 9

***Variables related to the read-depth strategy* (RD)*:***
- lenwin : Int - length of windows for read depth (RD) analysis. By Default 100
- movilwin : Int – number of windows for further read-depth signal smoothing using mobile means analysis. By default is 10.
- agada: Float – input AGADA variable. By default is 0.2. Check GADA doc.
- tgada: Float – input TGADA variable. By default is 4. Check GADA doc.
- mgada : Float – input MGADA variable. By default is 10. Check GADA doc.
- sgada : Float – input SGADA variable. By default is 0.5. Check GADA doc.
- bgada : Float – input BGADA variable. By default is 0. Check GADA doc.

***Variables related to the definition of structural variants, SVs:***
- limitinf : Float – upper limit value of read-depth normalized signal to consider a copy neutral state. By default is 0.4
- limitsup : Float – upper limit value of read-depth normalized signal to consider a copy gain state. By default is 6.0.

***General variables:***
- msg : Int – Messages are available for the logs files . By default is 1
- debug : Int – {0,1}. The value 1 indicates that you want to keep all temporally files. 0 is to delete them. See the next section.
- login : login to enter in the mysql database
- passwd : password to enter in the mysql database
- database : name of your database
- conn : connection string of your database

***runALL values :***

| values | modes |
|---|---|
| 0 | The tool only runs the samtools module. Extract the sam files from bam files. |
| 1 | **By default value.** The tool will run all modules: samtools, paired-reads strategy (PR), read-depth strategy (RD) and definition of SV modules. |
| 11 | The tool will run paired-reads strategy (PR), read-depth strategy (RD) and definition of SV modules. The process needs the sam files allocated in the $PSVFDIR/tmp/splitchr directory by chromosomes. |
| 2 | The tool will run samtools and paired-reads strategy (PR). |
| 3 | The tool will run samtools and read-depth strategy (RD) modules. |
| 4 | The tool will run only the paired-reads strategy (PR). The process needs the sam files allocated in the $PSVFDIR/tmp/splitchr directory by chromosomes. |
| 5 | The tool will run only read-depth strategy (RD). The process needs the sam files allocated in the $PSVFDIR/tmp/splitchr directory by chromosomes. |

***sv values :***

By default the values will be 0 and *it is the recommendable value.*
0 - The clustering process analyze all read-pairs.
1 - The clustering process analyze only ordori-pairs.
2 - The clustering process analyze only ori-pairs.
3 - The clustering process analyze only chrpos-pairs.
4 - The clustering process analyze only chrposori-pairs.
5 - The clustering process analyze only order-pairs.
6 - The clustering process analyze ordori-, ori- and order-pairs.

***Example:***

A complete analysis to obtain somatic variants using a tumour and a normal blood sample of a patient with chronic lymphocytic leukaemia.

Example *Template.som.cfg into the $PSVFDIR/config directory* :

```
ntchr=25
nprocess=5
ndofcprocess=3
basedir={absolute dir path of PeSV-Fisher}
sortdir=sort
samtoolsdir=samtools
awkdir=awk
refaprefix={absolute path+prefix of the reference genome, by chromosome}
hg=hg19
datadirbam={absolute path of data directory: The bam and .bai files}
prefix=-
control={sample control name-ID}
disease={sample tumor name-ID}
extfile=tab
format=\t
runALL=1
sv=0
quantile=0.99
numclones=2
somatic=1
posid=0
poschr1=1
posini=2
posend=3
poslen=4
poschr2=5
postrand1=6
postrand2=7
poscore1=8
poscore2=9
lenwin=100
movilwin=10
agada=0.2
tgada=4
mgada=10
sgada=0.5
bgada=0
limitinf=0.4
limitsup=6.0
msg=1
debug=1
login={yourlogin of database}
passwd={your password of database}
database={the name of your database in the  mysql}
conn={host name of server of your database}
```

**Notes:** the blue fields are those you should modify, and the green are the specific variables for this example. As you can see somatic is linked to the control and disease fields, so in this case the prefix value is not important.

**Summary:**

cd $PSVFDIR/bin
./PeSVFisher ../config/Template.som.cfg

*you could redirect the logs to file and use the nohup command if you would like to close the terminal:*
*nohup ./PeSVFisher ../config/Template.Nonsom.cfg >& ../log/template.log&*

copy the results
*nohup ./PeSVFisher ../config/Template.som.cfg >& ../log/template.som.log&*

**Notes:**

- If you would like to run consecutively more than one somatic studies, you can create a list

  control={samplecntrl-ID1},{samplecntrl-ID2}
  disease={sampletumor-ID1}.{sampletumour-ID2}

- If your samples are aligned using the hg18 you must modify the variables :

  refaprefix={absolute path+prefix of the reference genome hg18, by chromosome}
  hg=hg18

**Running the test data. NA19240.bam**

*For requeriments see the section requeriments from README file*
*For Installation see the Installation section*

In this case, you would like to make a fully analysis from one single sample aligned using hg18:

1) Download the test data from gd.crg.eu/tools
   wget http://gd.crg.eu/tools/testdata/NA19240.bam
   wget http://gd.crg.eu/tools/testdata/NA19240.bam.bai

2) Create a directory for the bam file and move it there
   mkdir $PSVFDIR/test/data
   mv NA19240.bam* $PSVFDIR/test/data

3) Create a directory for the reference genome hg18 data set by chromosome and download them
   mkdir $PSVFDIR/test/hg18
   rsync -avzP rsync://hgdownload.cse.ucsc.edu/goldenPath/hg18/chromosome $PSVFDIR/test/hg18

4) Modify the configure file : Template.som.cfg
   cp $PSVFDIR/config/Template.som.cfg $PSVFDIR/config/Template.Nonsom.cfg
   vim $PSVFDIR/config/Template.Nonsom.cfg

*Note : Change {$PSVDIR} for the absolute path of PeSV-Fisher:*

```
ntchr=25
nprocess=5
ndofcprocess=3
basedir={$PSVFDIR}/
sortdir=sort
samtoolsdir=samtools
awkdir=awk
refaprefix={$PSVFDIR}/test/hg18/chr
hg=hg18
datadirbam={$PSVFDIR}/test/data/
prefix=NA19240
control=-
disease=-
extfile=tab
format=\t
runALL=1
sv=0
quantile=0.99
numclones=2
somatic=0
posid=0
poschr1=1
posini=2
posend=3
poslen=4
poschr2=5
postrand1=6
postrand2=7
poscore1=8
poscore2=9
lenwin=100
movilwin=10
agada=0.2
tgada=4
mgada=10
sgada=0.5
bgada=0
limitinf=0.4
limitsup=6.0
msg=1
debug=0
login={yourlogin of database}
passwd={your password of database}
database={the name of your database in the  mysql}
conn={host name of server of your database}
```

5) Run
   cd $PSVFDIR/bin
   nohup ./PeSVFisher ../config/Template.Nonsom.cfg >& ../logs/template.Nonsom.out&

6) See RESULTS file