MANUAL VSeq-Toolkit (Version 1.0)

Author: Saira Afzal

VSeq-Toolkit is designed to analyze whole genome or targeted sequencing viral vector gene therapy data. Input format should be FASTQ paired end.

1. Analysis Modes

There are three main analysis modes;

- Mode1 - Contaminant Analysis Mode
- Mode2 - Vector-Vector Fusion Analysis Mode
- Mode3 - Vector-Host Fusion Analysis Mode

All modes can be executed independently or together or Mode2 and Mode3 can be executed at once.

In case of executing all modes together, the contaminant reads detected would be exempted from Mode2 & Mode3 analysis datasets. The final vector-vector fusion reads would be exempted from Mode3 datasets.

In case of executing Mode2 & Mode3 together, the final vector-vector reads would be exempted from Mode3 datasets.

2. Code and Installation

VSeq-Toolkit is available at

Clone the repository from Github;

```
git clone
cd VSeq-Toolkit
```

Note: To check installation was successful see “README.md” file in main toolkit directory.

2. Configuration File & Basic Parameters

The *config.txt* file in the main toolkit directory is used to set the basic parameters in order to execute the VSeq-Toolkit.

There are general parameters that need to be set for executing each mode.

General Parameters

```
#Forward and reverse file of paired end WGS/TES data;
file1
file2

#Path to the output directory for processing and results;
outDir

#Path to the scripts directory within VSeq-Toolkit;
bin

#Quality values for trimming and filtering;
Trimming end until specified or higher quality reached;
qua=20

The minimum read length after trimming;
lenPer=50

#Adapters in the forward and reverse file;
adapter1=
adapter2=

#Path to the third-party tools required for each mode.
#These tools are provided within VSeq-Toolkit ‘thirdPartyTools’
directory.
trimmer
aligner
samtools

#Two modes ‘default’ or ‘sensitive’ can be used for analysis.
#These modes represents sensitivity of analysis.
mode=default
```

The parameters for Mode1

```
#If this mode need to be executed or not (true/false)
contAna=true

#Path to the BWA reference index files
#Before indexing concatenate all references including contaminants,
vector and any respective genome reference together in one file
combinedRef=

#Stringency/Specificity levels of analysis ‘high/moderate/low/null’.
#Recommended medium/high for experimental datasets
stringencyCont=medium
```

```
#Threshold for specifying each read of the pair unique/multiple mapped
within the respective genome
UMthresholdCont=0.95
```

The parameters for Mode2

```
#If this mode need to be executed or not (true/false)
vecVecFusion=true

#Path to the BWA reference index files
#Before indexing concatenate reference together in case of multiple
viral vectors in one file
vecRef=

#Stringency/Specificity levels of analysis ‘high/moderate/low/null’.
#Recommended medium/high depending on the experimental datasets and
vector reference
stringencyVec=low

#Threshold for specifying each read of the pair unique/multiple mapped
within the respective genome
UMthresholdVec=0.95

#Minimum span of each vector region
minMapSpanVec=20

#Maximum unmapped bases between fusion regions
distVecVec=10

#Maximum overlapping bases between fusion regions
#This can be set in accordance with minMapSpanVec parameter
opVecVec=5

#Minimum identity of fusion regions
idenVecVec=95
```

The parameters for Mode3

```
#If this mode need to be executed or not (true/false)
vecGenIS=true

#Path to the BWA reference index files
#Before indexing concatenate reference together in case of multiple
viral vectors in one file
vecRef

#Path to the BWA index files after concatenating reference genome and
vector(s) genomes together in one single file
vecGenRef
```

```
#Stringency/Specificity levels of analysis ‘high/moderate/low/null’.
#Recommended medium/high depending for experimental datasets
stringencyVec=low

#Threshold for specifying each read of the pair unique/multiple mapped
within the respective genome
UMthresholdVec=0.95

#Minimum span of each vector region
minMapSpanVec=20

#Maximum unmapped bases between fusion regions
distVecVec=10

#Maximum overlapping bases between fusion regions
#This can be set in accordance with minMapSpanVec parameter
opVecVec=5

#Minimum identity of fusion regions
idenVecVec=95

#Range for position clustering on genomic fusion/insertion sites
clusterRange=3

#Path to the annotation information table - Refseq
annoTable

#Path to bedtools
bedtools
```

3. Reference and index files

Each module requires specific reference and index files as mentioned in the *“2.Configuration File & Basic Parameters”* section.

In order to combine references together following command can be used;

```
cat reference1 reference2 reference3 > reference.fa
```

To index the reference.fa file use following command;

```
path_to_VSeq-Toolkit/thirdPartyTools/bwa index -a bwtsw reference.fa
```

Important Note:
The string ‘vector’ should be used in the viral vector reference name. E.g., if multiple vector references are analyzed, the reference names can be used as vector1-Name, vector-Name2 etc.

The human or mouse or other reference genome should contain string ‘chr’.

The contaminants should not have ‘vector’ string in the name

4. Executing VSeq-Toolkit

- Create the output directory for running each sample.

- Set the parameters in the *config.txt* file and provide in the config file path of the output directory.

- First export the location of VSeq-toolkit
  export VSeqToolkit=/path_to_VSeq-Toolkit/

- Execute following command on terminal;
  perl path_to_VSeq-Toolkit/scripts/VSeq-TK.pl -c config.txt

- For vector-host fusion module, the annotation refSeq UCSC table is required. It can be downloaded from UCSC https://genome.ucsc.edu/cgi-bin/hgTables by setting;
  group: Genes and Gene Prediction, track: Other RefSeq, table: UCSC RefSeq (refGene), output format: all fields from selected table; output file; plain text

- *The path of the obtained file needs to be provided in the config file as stated in the section “2.Configuration File & Basic Parameters”*

5. Third-party tools

The required third party tools, are packaged within the main directory in “thirdPartyTools” directory *(bwa 0.7.4, samtools 1.3.1, bedtools v2.17.0, skewer 0.1.117)*.

6. Results

The main output result files are;

Mode1
ContAnalysis_DetailsOfReadPairsPerChromsomes.csv – Details of any reference chromosome mapped reads, if reference genome e.g., human or mouse etc are used in analysis
C*ontAnalysis_DetailsOfReadPairsPerContaminatSequence.csv* - Details of any respective contaminant mapped reads
ContAnalysis_DetailsOfReadPairsPerVectorReferenceSequence.csv - Details of any respective vector reference mapped reads

ContAnalysis_FragmentSizeDistribution.csv – Fragment size distribution of mapped reads
ContAnalysis_NumberOfReadPairsPerChromsomes.csv – Stats of any reference chromosome mapped reads, if reference genome e.g., human or mouse etc are used in analysis
ContAnalysis_NumberOfReadPairsPerContaminatORReferenceSequence.csv – Stats of any respective contaminant mapped reads or vector reference mapped reads

Mode2
ISGenomeVector.csv – Clustered main result file
ISGenomeVector.NonUniqueGenome.csv – Clustered main result file with uniquely mapped reference genomic region
ISGenomeVector.UniqueGenome.csv – Clustered main result file with non-uniquely mapped reference genomic region
ISGenomeVector.Unclustered.csv – Unclustered result file

Mode3
**.IntraVector_BreakPoints.csv* – Respective vector–vector breakpoints result file

Other
log – Time log file