WIFA-Seq : Identifcation of cancer-driver genes in focal genomic alterations from whole genome sequencing data

version 1.0

Ho Jang

Data Mining and Computational Biology Lab (DMCB)

Gwangju Institute of Science and Technology (GIST)

February 26, 2018

Contents

1 Installation

1.1 Requirements

This software runs in Linux system. R program should be installed on your system. Also for running MATLAB binary code, MATLAB Runtime installation is required.

1.2 Installation details

Before running provided programs, following R packages should be installed.

- snowfall R package for using multi processing
- data.table R package
- Rsamtools R package

Also following c codes in './c_libraries' directory should be compiled.

- Shell>R CMD SHLIB d150520008_aggregate.c
- Shell>R CMD SHLIB d170703126_make_fixed_size_bins

Next, install MATLAB Runtime (mathworks.com/products/compiler/mcr.html) for running MATLAB linux binary code in './matlab_libraries'.

2 Workflow

2.1 Quantification of copy number aberration

This is the shell command for quantifying copy number aberrations.

Shell>Rscript quantify_copy_number_aberration.r \
<BAM FILE LIST>\
<DIR OUTPUT>\
<BIN SIZE>\
<NUM PROCESSES>

- **BAM FILE LIST** The file path of tab-delimited text file with following three columns with column names. Three column names should be included.
  - **id** Unique identifier for individual sample
  - **tumor.path** Absolute file path of tumor bam file
  - **normal.path** Absolute file path of normal control bam file

id	tumor.path	normal.path
S0001	/WES/GBM/1.bam	/WES/GBM/4.bam
S0008	/WES/GBM/5.bam	/WES/GBM/7.bam
S0010	/WES/GBM/6.bam	/WES/GBM/9.bam
S0013	/WES/GBM/12.bam	/WES/GBM/17.bam

Figure 1: An example of bam file list

- **DIR OUTPUT** The directory path where the output files will be stored.
- **BIN SIZE** The genomic window size (base-pair).
- **NUM PROCESSES** The number of clusters to process concurrently.

2.2 Identify focal aberrations

This is the shell command for detecting focal copy number aberrations.

```
Shell>Rscript identify_focal_aberration.r \
<SAMPLE FILE LIST>\
<DIR INPUT>\
<DIR OUTPUT>\
<MATLAB RUNTIME>\
<MAX LEVEL>\
<COARSE LEVEL>\
<WAVELET THRESHOLD>\
<BIN THRESHOLD 1>\
<BIN THRESHOLD 2>\
<BIN THRESHOLD 3>\
<BIN SIZE>\
<NUM PROCESSES>
```

- **SAMPLE FILE LIST** List of unique sample identifiers
- **DIR INPUT** The path of input file directory
- **DIR OUTPUT** The path of output file directory
- **MATLAB RUNTIME** The path of MATLAB Runtime directory
- **MAX LEVEL** The wavelet level for all bins of whole genome. See the main article for more detail.
- **COARSE LEVEL** The coarse level for focal copy number alteration identification. See the main article for more detail.
- **WAVELET THRESHOLD** Wavelet transform threshold. Default value is 2.0. You can raise the threshold if you want to remove more noise.
- **BIN THRESHOLD 1** Threshold (base-pair unit) for filtering abnormally short genomic regions in the log2 ratio copy number quantification step. All abnormal region whole length is less than the threshold will be filtered. The default value is 1000 bp.
- **BIN THRESHOLD 2** Threshold (base-pair unit) for filtering abnormally short genomic regions in the log2 ratio copy number quantification step. Candidate genomic regions that are likely to be false focal aberration due to various sequencing bias will be ignored if the length of the genomic region is less than this threshold. The default value is 3000 bp.
- **BIN THRESHOLD 3** Threshold (base-pair unit) for filtering abnormally short genomic regions in the $y_H IGH$ generation step. All abnormal region whole length is less than the threshold will be filtered. The default value is 2000 bp.
- **BIN SIZE** The genomic window size (base-pair).
- **NUM PROCESSES** The number of clusters to process concurrently.

2.3 Identify recurrent aberrations

This is the shell command for identifying recurrent copy number aberrations.

Shell>Rscript identify_recurrent_aberrations.r \
<SAMPLE FILE LIST>\
<DIR INPUT>\
<DIR OUTPUT>\
<MATLAB RUNTIME>\
<NUM PERMUTATIONS>\
<THRES P-VALUE>\
<NUM PROCESSES>

- **SAMPLE FILE LIST** List of unique sample identifiers.
- **DIR INPUT** The path of input file directory.
- **DIR OUTPUT** The path of output file directory.
- **MATLAB RUNTIME** The path of MATLAB Runtime directory.
- **NUM PERMUTATIONS** Total number of permutations for statistical significance testing.
- **THRES P-VALUE** $P-$value for statistically significant genomic regions.
- **NUM PROCESSES** The number of clusters to process concurrently.

Shell>Rscript annotate_using_ensembl_genes.r \
<GENE>\
<DIR INPUT>\
<DIR OUTPUT>

- **GENE** The path of gene annotations
- **DIR INPUT** The path of input file directory
- **DIR OUTPUT** The path of output file directory

2.4 Example

This is the example commands for identifying recurrent copy number aberrations. Sample information file is located in '/home/foo/sampleinfo.txt' and gene information file is located in '/home/foo/GRCh37.75.gtf.genes.txt'. We provide Ensembl gene annotation files for the reference genome GRCh37 and GRCh38. For the wavelet transform, MATLAB Runtime is installed in
'/home/foo/MATLAB/MATLAB_Runtime/v901'.

Copy number aberration quantification results will be generated in '/home/foo/1.copy_number_quantification', focal aberration results will be in '/home/foo/2.focal_aberrations', and recurrent aberration results will be in '/home/foo/3.recurrent_aberrations'.

In this example, only default argument values were used.

```
Shell>Rscript quantify_copy_number_aberration.r \
/home/foo/sampleinfo.txt \
/home/foo/1.copy_number_quantification \
100 \
20
```

```
Shell>Rscript identify_focal_aberration.r \
/home/foo/sampleinfo.txt \
/home/foo/1.copy_number_quantification \
/home/foo/2.focal_aberrations \
/home/foo/MATLAB/MATLAB_Runtime/v901 \
25 \
10 \
2.0 \
1000 \
3000 \
2000 \
100 \
5
```

```
Shell>Rscript identify_recurrent_aberrations.r \
/home/foo/sampleinfo.txt \
/home/foo/2.focal_aberrations \
/home/foo/3.recurrent_aberrations \
1000 \
0.1 \
5
```

```
Shell>Rscript annotate_using_ensembl_genes.r \
/home/foo/GRCh37.75.gtf.genes.txt \
/home/foo/3.recurrent_aberrations \
/home/foo/4.annotation
```