Methyl_ssRSEA.R: User Guide

Ryu Minegishi, Osamu Gotoh, Seiichi Mori
Project for Development of Innovative Research on Cancer Therapeutics
Cancer Precision Medicine Center
Japanese Foundation for Cancer Research

1. Implementation

Methyl-ssRSEA computes a region-set score of the DNA methylome in a sample. Methyl_ssRSEA.R is written in R and executed on the command line (Terminal in Linux/UNIX/OSX or Command Prompt in MS Windows) directly, or a shell script.

The source code of Methyl_ssRSEA.R and the example data are available from the GitHub repository (https://github.com/jfcr-genome).

The directory contains the following materials:

```
-------
[User Guide]
  ./Methyl_ssRSEA_SupplDoc.pdf   # this file

[Source code]
  ./Methyl_ssRSEA.R

[Example data]
  ./sample_EPIC.sh
  ./input_data/
      beta_values_EPIC.csv
      GSM1084801_BCL6_hg19.bed
  ./output/
      ssRSEA_scores_EPIC.csv
  ./probe_coord/
      infinium450k_coord_hg19.csv
      infiniumEPIC_coord_hg19.csv
-------
```

2. Installation

The following packages are required:

- readr (https://CRAN.R-project.org/package=readr)
- dplyr (https://CRAN.R-project.org/package=dplyr)
- tidyr (https://CRAN.R-project.org/package=tidyr)
- tibble (https://CRAN.R-project.org/package=tibble)
- fastmatch (https://CRAN.R-project.org/package=fastmatch)

3. Input files

Input files for Methyl_ssRSEA.R are as below:

(1) Region set

(2) Test data

(3) CpG coordinates

All files are formatted as a simple text with normal Unix line termination (“¥n”).

(1) Region set

A region set is one or more lists of genomic regions to be tested for enrichment. A region set is typically derived from transcription factor binding sites from a ChIP-seq experiment. A region set is made as a tab-delimited matrix of genomic coordinates (chromosome name, start and end position) in the format of Browser Extensible Data (BED).

Required fields are:

a. chr: Chromosome name.

b. start: Start position of the feature in a chromosome. The first nucleotide is numbered 0.

c. end: End position of the feature in a chromosome.

chr	start	end
chr1	713928	714328
chr1	758093	758493
chr1	895835	896235

(2) Test data

A test data is a matrix of β values of CpGs in a Comma-Separated Values (CSV) file with header for sample ID. Row and column indicate CpG probes and samples. When using

Illumina Infinium MethylationEPIC or Methylation450 array, the output file from Illumina GenomeStudio Software or minfi in R can be directly used. In the case of whole-genome bisulfite sequencing (WGBS) data, the CpG-site ID needs to be generated; e.g. it may be generated by connecting chromosome name and genomic position with an underscore line, such as “chr20_61847650”.

Sample

CpG Probes

Probe	200819570057_R03C01	200819570057_R05C01	200819570057_R07C01
cg00000029	0.61083806	0.62336699	0.39735364
cg00000103	0.83058462	0.78256225	0.76610696
cg00000109	0.77658524	0.75820086	0.77539685

Missing value:

A non-numeric character or an empty cell in the place of a β value in the test data matrix is ignored and hence not used for further calculation. When the CpG site does not have sufficient sequencing depth (i.e. depth < 10) in the WGBS data, the β value at the CpG site may not be used for further calculation. Replacing the β value with a non-numerical character, such as “NA”, results in not using the β value in the ssRSEA score calculation. Missing values in the output β-value file from the microarray data, which are typically expressed as “NaN” or empty cell, are also not used for computation.

(3) CpG coordinates

A file of CpG coordinates is to define the genomic position for a CpG probe. The headers of “id”, “chr” and “pos” indicate probe ID (or CpG-site ID for WGBS data), chromosome name, and genomic position, respectively. The probe IDs or CpG-site IDs should be same as those in the test data.

id	chr	pos
cg18478105	chr20	61847650
cg09835024	chrX	24072640
cg14361672	chr9	131463936

4. Output

The output file is made of two rows: the first row is sample ID, and the second row indicates ssRSEA scores.

200819570057_R01C01	200819570057_R05C01	200819570055_R03C01
-268027.2002	-209263.9737	-189503.4809

5. Example run

Execute the following command on the command line.

$ Rscript Methyl_ssRSEA.R <Test data file path> <Region set file path> <CpG coordinates file path> <output file path>

An example script is also provided, as follows:

[sample_EPIC.sh]

```
#!/bin/sh
beta_value_path="input_data/beta_values_EPIC.csv"
region_set_path="input_data/GSM1084801_BCL6_hg19.bed"
probe_coord_path="probe_coord/infiniumEPIC_coord_hg19.csv"
output_path="output/ssRSEA_scores_EPIC.csv"
Rscript Methyl_ssRSEA.R $beta_value_path $region_set_path $probe_coord_path $output_path
```

The files of CpG coordinates for Illumina Infinium MethylationEPIC and HumanMethylation450 are provided in the following directories:

./probe_coord/infiniumEPIC_coord_hg19.csv

./probe_coord/infinium450k_coord_hg19.csv

6. Contact

Dr. Ryu Minegishi (ryu.minegishi@jfcr.or.jp)

Dr. Osamu Gotoh (osamu.gotoh@jfcr.or.jp)

Dr. Seiichi Mori (seiichi.mori@jfcr.or.jp)

Project for Development of Innovative Research on Cancer Therapeutics,

Cancer Precision Medicine Center,

Japanese Foundation for Cancer Research