Episo_Kallisto-User Guide-V1.0

1) Quick Reference

Episo needs a working version of Perl and it is run from the command line. Meanwhile, Bowtie1 and Kallisto (0.44.0 or high version) need to be installed on your computer. First you need to download a transcript annotation file from the Ensembl or NCBI websites. Episo supports the reference trancriptom sequence files in FastA format, allowed file extensions are either .fa or .fasta.

(1) Compiling the program

After downloading Episo_Kallisto_code, you should change directories to where the Episo_Kallisto_code is located and execute the follow commands.

cd Episo_Kallisto_code

chmod u=rwx,g=rx,o=x compile.sh

./ compile.sh

After compiling, all executable files and control files (the expand file name is ctl) are in folder m5c_command which is in the same directory with the folder Episo_Kallisto_code.

(2) Building an index

Episo_Kallisto needs to run the program Kallisto which is developed by Nicolas L Bray and is used to quantify abundances of transcripts from RNA-Seq data. Kallisto requires processing a transcriptome file to create a "transcriptome index". To begin, the folder kallisto_index should be created in the same directory with the folder Episo_Kallisto_code and then the follow commands are executed.

cd kallisto_index

../m5c_command/kallisto index -i ${idxname}_transcripts.idx transcripts.fasta.gz

Note. The file transcripts.fasta.gz can be downloaded from the Ensembl or NCBI websites. If we analyse the transcript from mouse, the ${idxname} is mouse.

(3) Preparing input files

Episo_Kallito needs three input files: ${name}_pe.txt, site_info.txt and trans_anno. The file ${name}_pe.txt records the mapping information of RNA-BisSeq data. The file site_info.txt records the site location information and is only used to estimate m^5c level at single nucleotide. The file trans_anno records the transcript information and is only used to estimate m^5c level at single nucleotide. The details are as follow.

(A) Generating transcriptome indexing for the file ${name}_pe.txt

Usage: bismark_genome_preparation [options] <path_to_trancriptome_folder>

Note. The reference trancriptome sequence file downloaded from the Ensembl or NCBI websites should be put in the <path_to_transcriptome_folder>.

A typical trancriptome indexing could be like this:

bismark_genome_preparation --path_to_bowtie /usr/local/bowtie --verbose /data/transcriptome/

(B) Generating the file ${name}_pe.txt

Usage: bismark-liu [options] <transcriptiome_folder> -1 <mates> -2 <mates>

A typical calling example could be like this:

bismark-liu --path_to_bowtie /usr/local/bowtie --vanilla --sam -n 2 /data/transcriptome/ -1 example_1.fastq -2 example_2.fastq

This will produce three output files:

(a) example_1.fastq_bismark_pe.txt (contains all alignments and methylation call strings)

(b) example_1.fastq_bismark_pe_mul.txt (contains the transcript information, where the alignment belongs)

(c) example_1.fastq_bismark_PE_report.txt (contains alignment and methylation summary)

Note. The options "vanilla" and "sam" are necessary and the bowtie version must be bowtie1. The program compare-paired and the file trans_anno must be in the same directory in which bismark-liu is.

(C) The input file site_info.txt contains the following information (1 line per site, tab separated):

(1) chromosome-location

Example:

(1) chr10-42196932

(D) The input file trans_anno contains the following information (1 line per transcript, tab separated):

(1) chromosome-id

(2) transcript-id

(3) start of the first exon

(4) end of the first exon

(5) start of the second exon

(6) end of the second exon

.

.

.

(2n+1) start of the nth exon

(2n+2) end of the nth exon

Example (mouse):

(1) chr1

(2) ENSMUST00000070533

(3) 3214482

(4) 3216968

(5) 3421702

(6) 3421901

(7) 3670552

(8) 3671498

Please note that the above three input files should be put in the folder inputmapping which should be created in the same directory with the folder Episo_Kallisto_code.

(4) Estimating the methylation level of each isoform

To begin, first change directories to where the command files are located:

cd m5c_command

Next, set parameters in the shell scripts m5c_whole.sh and m5c.sh:

parameters in m5c_whole.sh

```
#parameter setting
samplistA="SRR493366 SRR493367 SRR493368" # the ${name} of ${name}_pe.txt
idxname="mouse" # the ${idxname} of ${idxname}_transcripts.idx in folder kallisto_index
bs=100 # the number of bootstrap
```

parameters in m5c.sh

```
#parameter setting
samplistA="SRR493366 SRR493367 SRR493368" # the ${name} of ${name}_pe.txt
numA=3 # the number of replicates and is identical to the number of mapping files in samplistA
bs=100 # the number of bootstrap
```

```
fil=30 # the value of filtering low abundance transcripts
```

Note. The value of parameter samplistA and bs in m5c_whole.sh and m5c.sh should be identical; fil=30 means that Episo_Kallisto ignores transcripts where there are less than 30 estimates counts.

Last, execute the shell script m5c_whole.sh:

./m5c_whole.sh

(5) Estimating the methylation level of single nucleotide on isoform

To begin, first change directories to where the command files are located:

cd m5c_command

Next, set parameters in the shell scripts m5c_single.sh and m5c.sh:

parameters in m5c_single.sh

```
#parameter setting
samplistA="SRR493366 SRR493367 SRR493368" # the ${name} of ${name}_pe.txt
idxname="mouse" # the ${idxname} of ${idxname}_transcripts.idx in folder kallisto_index
bs=100 # the number of bootstrap
total=100 # the number of sites in the file site_info.txt
lg=115 # the length of bisulfite read in the mapping file ${name}_pe.txt
```

parameters in m5c.sh

```
#parameter setting
samplistA="SRR493366 SRR493367 SRR493368" # the ${name} of ${name}_pe.txt
numA=3 # the number of replicates and is identical to the number of mapping files in samplistA
bs=100 # the number of bootstrap
fil=30 # the value of filtering low abundance transcripts
```

Note. The value of parameter samplistA and bs in m5c_single.sh and m5c.sh should be identical; fil=30 means that Episo_Kallisto ignores transcripts where there are less than 30 estimates counts.

Last, execute the shell script diff_single.sh:

./m5c_single.sh

(6) Results

The results of Episo_Kallisto are placed in the folder m5c_results which is in the same directory with the folder Episo_Kallisto_code. The results of estimated m^5c level are in the file m5c_out and m5c_out_single_all. The m5c_out file contains the information of m^5c level of isoform and should look like this:

target_id	estimated_m5c	mean	variance
ENSMUST00000000137.7	0.027253	0.027282	0.000013

The m5c_out_single_all file contains the information of m^5c level of single nucleotide and should look like this:

site_id	target_id	estimated _m5c	mean	variance
chr10-27048035	CUFF.4589.4	0.164527	0.159813	0.001396