meCytoDiff-User Guide-V1.0

1) Quick Reference

meCytoDiff runs from the command line at linux or unix platform. Meanwhile, gcc (C-compatible compiler) and a kind of mapping tool for bisulfited RNA-Seq (RNA-BisSeq) data are needed to be installed on your computer. meCytoDiff can be downloaded from the meCytoDiff homepage.

(1) Compiling the program

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

```
cd meCytoDiff_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 diff_command which is in the same directory with the folder meCytoDiff_code.

(2) Building an index

meCytoDiff 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 meCytoDiff_code and then the follow commands are executed.

```
cd kallisto_index

../diff_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

meCytoDiff 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 differential analysis at single nucleotide. The file trans_anno records the transcript information and is only used to differential analysis at single nucleotide. The details are as follow.

The input file ${name}_pe.txt contains the following information (1 line per sequence, tab separated):

(1) Seq-id

(2) alignment strand

(3) transcript

(4) start

(5) end

(6) original bisulfite read sequence 1

(7) equivalent transcritome sequence 1 (+2 extra bp)

(8) methylation call string 1

(9) original bisulfite read sequence 2

(10) equivalent transcriptome sequence 2 (+2 extra bp)

(11) methylation call string 2

(12) read 1 conversion

(13) transcriptome conversion

(14) read 1 quality score (Phred33 scale)

(15) read 2 quality score (Phred33 scale)

Note. The detail about the above information may see the user guide of Bismark. The above information can be got by using kinds of mapping tool for RNA-BisSeq data. It is commended to use mapping tool in Episo.

Example:

(1) HWI-D00751:78:C9Y4TANXX:8:1101:1486:1969

(2) -

(3) ENSMUST00000144883

(4) 2623

(5) 2738

(6) AATACAAAAAAATCAAACCATCCTCAAAAC

(7) CAAGTACAGAGGGATCAGGCTATCCTCAGAGC

(8) .h....x.hhh....xh.........x.h.

(9) GAAGGAAGGTAAGGGTTTGGGGATATTGGT

(10) GAAGGAAGGCAAGGGTCTGGGGACACTGGTTG

(11) ....h......x......h.x....

(12) GA

(13) CT

(14) GGGGGGGGGGGGGGGGGGGGGGGGGGGGGG

(15) GGGGGGGGGGGGGGGFFGGGGGFFGGFGGEGG

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

(1) chromosome-location

Example:

(1) chr10-42196932

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 meCytoDiff_code.

(4) Differential analysis of methylation level of isoform

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

```
cd diff_command
```

Next, set parameters in the shell scripts diff_whole.sh and diff.sh:

parameters in diff_whole.sh

```
#parameter setting
samplistA="SRR493366 SRR493367 SRR493368" # the ${name} of ${name}_pe.txt in condition A
samplistB="SRR493369 SRR493370 SRR493371" # the ${name} of ${name}_pe.txt in condition B
idxname="mouse" # the ${idxname} of ${idxname}_transcripts.idx in folder kallisto_index
bs=100 # the number of bootstrap
```

parameters in diff.sh

```
#parameter setting
samplistA="SRR493366 SRR493367 SRR493368" # the ${name} of ${name}_pe.txt in condition A
samplistB="SRR493369 SRR493370 SRR493371" # the ${name} of ${name}_pe.txt in condition B
numA=3 # the number of replicates in condition A and is identical to the number of mapping files in samplistA
numB=3 # the number of replicates in condition B and is identical to the number of mapping files in samplistB
bs=100 # the number of bootstrap
fil=30 # the value of filtering low abundance transcripts
diff=0.05 # the significance level for differential analysis
```

Note. The value of parameter samplistA, samplistB and bs in diff_whole.sh and diff.sh should be identical; fil=30 means that meCytoDiff ignores transcripts where there are less than 30 estimates counts.

Last, execute the shell script diff_whole.sh:

```
./diff_whole.sh
```

(5) Differential analysis of methylation level of single nucleotide on isoform

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

```
cd diff_command
```

Next, set parameters in the shell scripts diff_single.sh and diff.sh:

parameters in diff_single.sh

```
#parameter setting
samplistA="SRR493366 SRR493367 SRR493368" # the ${name} of ${name}_pe.txt in condition A
samplistB="SRR493369 SRR493370 SRR493371" # the ${name} of ${name}_pe.txt in condition B
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 diff.sh

```
#parameter setting
samplistA="SRR493366 SRR493367 SRR493368" # the ${name} of ${name}_pe.txt in condition A
samplistB="SRR493369 SRR493370 SRR493371" # the ${name} of ${name}_pe.txt in condition B
numA=3 # the number of replicates in condition A and is identical to the number of mapping files in samplistA
numB=3 # the number of replicates in condition B and is identical to the number of mapping files in samplistB
bs=100 # the number of bootstrap
fil=30 # the value of filtering low abundance transcripts
diff=0.05 # the significance level for differential analysis
```

Note. The value of parameter samplistA, samplistB and bs in diff_single.sh and diff.sh should be identical; fil=30 means that meCytoDiff ignores transcripts where there are less than 30 estimates counts.

Last, execute the shell script diff_single.sh:

```
./diff_single.sh
```

(6) Results

The results of a meCytoDiff run are placed in the folder diff_results which is in the same directory with the folder meCytoDiff_code. The results of differential analysis are in the file diff_out.tsv and diff_out_single_all.tsv. The diff_out.tsv file contains the information of differential analysis of isoform and should look like this:

```
target_id       estimated_A_m5c A_mean  A_variance      A_pvalue        estimated_B_m5c
B_mean  B_variance      B_pvalue
ENSMUST00000000137.7    0.027253        0.027282        0.000013        0.992002        0.018225
0.017923        0.000015        0.006003
ENSMUST00000000175.5    0.027267        0.027395        0.000004        0.002144        0.036950
0.037161        0.000012        0.999999
ENSMUST00000000349.10   0.010887        0.011308        0.000008        0.009844
0.020787        0.020216        0.000016        0.999598
ENSMUST00000000449.8    0.018622        0.017666        0.000047        0.006523        0.053520
0.053288        0.000195        1.000000
ENSMUST00000000687.7    0.027270        0.027383        0.000006        0.002214        0.033015
0.032962        0.000004        0.989255
```

The diff_out_single_all.tsv file contains the information of differential analysis of single nucleotide and should look like this:

```
site_id  target_id         estimated_A_m5c        A_mean          A_variance       A_pvalue
         estimated_B_m5c         B_mean           B_variance       B_pvalue
chr10-42196932 ENSMUST00000105502.7         0.138462          0.138462          0.000000
         0.000000          0.232877          0.232877          0.000000          1.000000
chr10-52418019 ENSMUST00000023830.15        0.219941          0.219941          0.000000
         0.000000          0.222222          0.222222          0.000000          1.000000
chr10-61428548 ENSMUST00000020288.14        0.227723          0.227723          0.000000
         0.000000          0.234694          0.234694          0.000000          1.000000
chr10-61428549 ENSMUST00000020288.14        0.281553          0.281553          0.000000
         1.000000          0.242424          0.242424          0.000000          0.000000
chr10-61428551 ENSMUST00000020288.14        0.285714          0.285714          0.000000
         1.000000          0.281553          0.281553          0.000000          0.000000
```