HMM-Fisher User Manual

HMM-Fisher website: https://github.com/xxy39/HMM-Fisher

Shuying Sun (ssun5211@yahoo.com)

Xiaoqing Yu (yuxq1120@gmail.com)

August 5, 2015

Contents

1 Overview and Installation

1.1 Overview

The *HMM-Fisher* [1] program identifies differentially methylated (DM) CG sites (DMCs) and regions (DMRs) from either whole genome or targeted bisulfite sequencing (BS) data. This approach first uses a hidden Markov chain to model the methylation signals to infer the methylation state as Not methylated (N), Partly methylated (P), and Fully methylated (F) for each individual sample. Then the Fisher Exact test is used to identify differentially methylated CG sites. Third, identified DM CG sites are summarized into regions based on their status and distance. This program takes aligned BS data in multiple samples and outputs identified DM CG sites and regions.

We will demonstrate the application of HMM-Fisher using a publicly available bisulfite-treated methylation sequencing dataset [2] on chromosome 1 in section 2. This dataset contains eight breast cancer cell lines, including four estrogen receptor positive (ER+) and four negative (ER-) samples. For the purpose of illustration, we treat the ER+ as control group and ER- as test group, and we only use the first 20,000 CG sites on chromosome 1 as an example dataset.

1.2 Installation

HMM-Fisher requires a Linux/Unix system, with R installed. To install HMM-Fisher, the user can download the pipeline from https://github.com/xxy39/HMM-Fisher. After unzipping the file, there are one document and two folders.

HMM.Fisher.user.manual.pdf	A copy of the user manual
HMM.Fisher.code	A folder containing all R source code files used for HMM-Fisher.
example.data	A folder containing all example input data as mentioned in this document, an example.script.txt for running HMM-Fisher (see section 3 for detail), and the output files generated from the example.script.txt (see section 4 for detail)

2 Usages

To identify differentially methylated CG sites and regions, users simply need to call the main function HMM.Fisher (). This function identifies DM regions (DMRs) in three steps:

1. Perform quality control based on coverage *using getMeth(). Users can use the parameter "min.percent" to control the proportion of samples that need to have coverage at each CG sites.*
2. Identify DM CG sites using the HMM-Fisher
    a. Estimate the states (N, Not methylated; P, Partly methylated; F, Fully methylated) for all CG sites using HMM
    b. Perform Fisher Exact test to test for group differences
    c. Identify DM CG sites based on p-values and methylation differences between groups
3. Summarize the filtered DM CG sites into DM regions (Hyper or Hypo), based on their DM status, distance between CG sites, and p-values. For the DM CG sites that are not grouped into regions with other DM CG sites, they will be reported as singletons.

HMM.Fisher

Description

Identify DM CG sites and summarize them into DM regions using the methylation level and coverage data.

Usage

HMM.Fisher (total.reads, meth.reads, n1, n2, chromosome, code.dir, output.dir, . . .)

Arguments

General Information

total.reads	$P \times L$ Matrix. Number of reads covering CG site l in sample p. See section 3.1 for more detail.
meth.reads	$P \times L$ Matrix. Number of methylated reads covering CG site l in sample p. See section 3.2 for more detail.
n1	Numeric. Number of test samples.
n2	Numeric. Number of control samples.
chromosome	Character. The chromosome that users want to analyze, e.g., chromosome =1, or chromosome = 2. The HMM-Fisher processes one chromosome at a time.
code.dir	String. The directory of the source code files of HMM-Fisher (e.g.,

	/home/HMM.Fisher /HMM.Fisher.code). Note, there should be no "/" at the very end.
Output.dir	String. The directory for output files (e.g., /home/HMM.Fisher.results). Note, there should be no "/" at the very end. Five files will be generated from this function. See section 4 for more detail.

Quality Control

min.percent	Numeric between 0 and 1 used in quality control. The CG sites should be covered in at least *min.percent* of the control samples AND of the test samples. Otherwise, the CG sites are dropped. Default = 0.8.

Identifying DM CG Sites

iterations	Numeric. Number of iterations when running HMM-Fisher. Default = 60.
meanDiff.cut	Numeric between 0 and 1. Minimum mean difference of methylation levels between the two groups to call a DM CG site. Default = 0.3.
dist.combine	Numeric. Two consecutive CG sites with distance ≤ *dist.combine* are combined in the Fisher Exact test step. Default = 100 bp.
p.threshold	Numeric between 0 and 1. CG sites with p-value ≤ *p.threshold* can be identified as DM. Default = 0.05.

Summarizing DM regions

max.distance	Numeric. The maximum distance between any two DM CG sites within a DM region. Default = 100 bp.
max.empty.CG	Numeric. The maximum number of CG sites that fail the quality control between any two DM CG sites within a DM region. Default = 3.
max.EM	Numeric. When combining two consecutive DM regions, the maximum number of EM CG sites between these two DM regions. These EM CG sites can be 1) identified as EM by HMM-Fisher but with relatively low p-value (controlled by *max.p*); or 2) identified as DM by HMM-Fisher but with small meanDiff (< *meanDiff.cut*). Default = 1. Note: if either region is a singleton, only 1 EM CG is allowed.
max.p	Numeric between 0 and 1. The maximum p-value for the EM included in the combined DM region. Default = 0.1.
singleton	Logical. Report the singletons or not in summarizing region step? If TRUE (default), the singletons will be reported in the *DMRs.txt*.

3 Input Files and Example Data

HMM-Fisher takes the number of total reads and number of methylated reads as input. Current version of HMM-Fisher takes multiple samples in control and test groups. For the best performance, we recommend at least 4 samples in each of the two groups. Instead of analyzing all CG sites that are sequencing, HMM-Fisher conducts the analysis only for CG sites that pass the quality control based on coverage. To ensure more accurate results, we also recommend filtering out the CG sites with low coverage.

HMM-Fisher processes one chromosome at a time. To analyze multiple chromosomes, we recommend that users prepare separate input files for each chromosome, and run HMM-DM for each chromosome separately.

3.1 total.reads

The *total.reads* file contains the number of reads covering each CG site for all samples. This file contains 1+*n1*+*n2* columns: position for each CG, the number of reads for samples in group1 (e.g., test group), the number of reads for samples in group2 (e.g., control group). Please pay attention to the order of the groups, which is associated with the definition of DM status (see 4.1). The ***totoal.reads.txt*** provided in example.data directory includes 20,000 CG sites on chromosome 1 for 4 test samples and 4 control samples. A sample of this file is shown below.

Box1. mC.matrix input file

pos	test_1	test_2	test_3	test_4	control_1	control_2	control_3	control_4
497	194	90	126	199	177	171	44	138
525	196	92	128	199	176	172	43	139
542	186	89	110	187	143	121	37	136

3.2 meth.reads

The *meth.reads* file contains number of methylated reads covering each CG site for all samples. This file contains 1+*n1*+*n2* columns: position for each CG, the number of reads for samples in group1 (e.g., test group), the number of reads for samples in group2 (e.g., control group).). NOTE that the positions and order of samples should be the same as the ones listed in the above *total.reads* file. The ***meth.reads.txt*** provided in example.data directory includes 20,000 CG sites on chromosome 1 for 4 test samples and 4 control samples. A sample of this file is shown below.

Box2. cov.matrix input file

pos	test_1	test_2	test_3	test_4	control_1	control_2	control_3	control_4
497	175	39	172	88	103	132	195	118
525	171	43	189	88	167	132	191	126
542	135	37	182	83	114	135	177	100

3.3 UNIX command

An example script of running HMM-Fisher is shown in ***example.script.txt*** under the example.data folder. Default settings are used for this example script. Users may change the parameters based on their own needs following the instruction in section 2. Once the input files and parameters are ready, run the following UNIX command to identify the DM CG sites and regions:

```
R CMD BATCH example.script.txt
```

A brief description of this example code is provided below:

Input: This input dataset contain 20,000 CG sites for 8 breast cancer cell lines, including 4 ER+ (BT474, MCF7, ZR751, and T47D), and 4 ER- (BT20, MCF10A, MDAMB231, and MDAMB468) samples. We treat the ER+ as control group and ER- as test group.

Box3. the input in example data

total.reads	example.total.reads.txt (20,000 × 9, see section 3.1)
meth.reads	example.meth.reads.txt (20,000 × 9, see section 3.2)
n1	4
n2	4
chromrosome	1

Quality control: The data are reduced to CG sites covered in at least 80% of test samples and at least 80% of control samples (*min.percent* = 0.8). After quality control, 5,811 CG sites are left for further analysis.

Identifying DM CG sites: We apply HMM-Fisher to the 5,811 CG sites with 60 *iterations*. Any two consecutive CG sites with distance ≤ 100 bp (*dist.combine* = 100) are combined in Fisher Exact test step. The CG sites with p-value ≤ 0.05 (*p.threshold* = 0.05) and mean methylation difference ≥ 0.3 (*meanDiff.cut* = 0.3).

Summarizing into DMRs: Consecutive DM CG sites are summarized into a DMR if 1) their distance is not larger than100 bp (*max.disance* = 100); 2) between the two CG sites, there are at most 3 CG sites that fail the quality control (*max.empty.CG* = 3). Two DMRs are later merged if 1) they are in the same DM status; 2) At most 1 EM CG site between the two DMRs (*max.EM* = 1) and this CG site has p-value ≤ 0.1 (*max.p* = 0.1).

Output: All results are saved under the output directory defined by parameter *output.dir*

Box4. the output files generated from the example.script.txt

mC.matrix.txt	Methylation levels for the 5,811 CG sites that pass the quality (see section 4.1)
all.CG.txt	DM status for all 5,811 CG sites (see section 4.2)
DM.CG.txt	DM status for the 171 identified DM CG sties (see section 4.3)
joint.prob.ps	Joint probabilities of the likelihood function of 60 (default) iterations, which shows the HMM convergence (see section 4.4)
DMRs.txt	Information for the identified 60 DMRs (see section 4.5)

4 Output Files

4.1 Quality control output: mC.matrix.txt

The first output from HMM-Fisher method contains the methylation ratio for each CG site that passes the quality control. For the sample with 0X coverage (0 in *total.reads*), the methylation ratio is denoted by "NA". The ***mC.matrix.txt*** provided in example.data directory is generated from the example code ***example.script.txt***. It contains 5,811 CG sites that pass the quality control. A sample of this output is shown below in Box 5.

Box5. mC.matrix.txt output file

pos	test_1	test_2	test_3	test_4	control_1	control_2	control_3	control_4
497	0.988701	0.886364	0.886598	0.977778	0.602339	0.956522	0.979899	0.936508
525	0.971591	1.000000	0.964286	0.956522	0.970930	0.949640	0.959799	0.984375
542	0.944056	1.000000	0.978495	0.932584	0.942149	0.992647	0.946524	0.909091

4.2 HMM Fisher raw output: all.CG.txt

This output shows the DM status for ALL CG sites analyzed by HMM-Fisher. An HMM first estimates the methylation states (N, P, F) for each CG. Then in Fisher Exact test step, the consecutive CG sites with distance ≤ *dist.combine* are combined. The output file contains a header line and 22 fields for each CG site. Column *DM.status* indicates the final status for each CG.

1) *DM.stauts* = 1 means "Hyper": CG sites in which the test group has a higher methylation level than the control group (*p.value* ≤ 0.05 and *meanDiff* ≥ 0.3);
2) *DM.stauts* = -1 means "Hypo": CG sites in which the control group has a higher methylation level (*p.value* ≤ 0.05 and *meanDiff* ≤ -0.3);
3) *DM.stauts* = 0 means "EM": CG sites in which the two groups have similar methylation levels.

The ***all.CG.txt*** provided in the example.data directory is generated from the code file ***example.script.txt***. A sample of this output is shown below in Box 6.

Box4. all.CG.txt output file

chr	pos	pos2	p.value	test.s	con.s	test.s2	con.s2	test.mC	con.mC	test.mC2	con.mC2
chr1	497	525	1	1:1:1:1	0.5:1:1:1	1:1:1:1	1:1:1:1	0.99:0.89:0.89:0.98	0.6:0.96:0.98:0.94	0.97:1:0.96:0.96	0.97:0.95:0.96:0.98
chr1	525	542	1	1:1:1:1	1:1:1:1	1:1:1:1	1:1:1:1	0.97:1:0.96:0.96	0.97:0.95:0.96:0.98	0.94:1:0.98:0.93	0.94:0.99:0.95:0.91
chr1	542	NA	1	1:1:1:1	1:1:1:1	<NA>	<NA>	0.94:1:0.98:0.93	0.94:0.99:0.95:0.91	<NA>	<NA>

test.post	con.post	test.post2	con.post2	meanDiff	DM.p	DM.status	index	meanCov.test	meanCov.control
1:1:0.97:1	0.9:1:1:1	1:1:1:1	1:1:1:1	0.0661	EM	EM	1	126.25	158.5
1:1:1:1	1:1:1:1	1:1:1:1	1:1:1:1	0.0069	EM	EM	2	126.75	159.5
1:1:1:1	1:1:1:1	<NA>	<NA>	0.0162	EM	EM	3	113.75	138.5

chr – chromosome number
pos – position for current CG
pos2 – position for the next CG.

p-value – the p-value of Fisher Exact test
test.s – the states of current CG estimated by HMM for test samples, separated by ":". 0, Not methylaited; 0.5, Partly methylated; 1, fully methylated
con.s – the states of current CG estimated by HMM for control samples, separated by ":"
test.s2 – the states of the next CG estimated by HMM for test samples, separated by ":"
con.s2 – the states of the next CG estimated by HMM for control samples, separated by ":"
test.mC – the raw methylation level of current CG in test samples, separated by ":"
con.mC – the raw methylation level of current CG in control samples, separated by ":"
test.mC2 – the raw methylation level of the next CG in test samples, separated by ":"
con.mC2 – the raw methylation level of the next CG in control samples, separated by ":"
test.post – the posterior probabilities of current CG in test samples, separated by ":"
con.post – the posterior probabilities of current CG in control samples, separated by ":"
test.post2 – the posterior probabilities of the next CG in test samples, separated by ":"
con.post2 – the posterior probabilities of the next CG in control samples, separated by ":"
meanDiff – the methylation difference between the two groups = mean(test) – mean (control)
DM.p – the DM status based on p-values. If p-value ≤ *p.threshold* and meanDiff ≥ 0, the CG is identified as Hyper; If p-value ≤ *p.threshold* and meanDiff < 0, the CG is identified as Hypo; otherwise, the CG is identified as EM
DM.status – the FINAL DM status of the CG site considering the mean difference and p-value. If p-value ≤ *p.threshold* and meanDiff ≥ *menDiff.cut*, the CG is identified as Hyper; If p-value ≤ p.threshold and meanDiff < - *menDiff.cut*, the CG is identified as Hypo; otherwise, the CG is identified as EM
index – the index of the CG site in mC.matrix file
meanCov.test – the mean coverage of test group
meanCov.control – the mean coverage of control group
* If the current CG is not combined with the next CG (distance > *dist.threshold*), column pos2, test.s2, con.s2, test.mC2, con.mC2, test.post2, and con.post2 are "NA".

4.3 DM CG output: DM.CG.txt

This output shows the DM CG sites identified by HMM-Fisher. It has the same format as 4.2.

The ***DM.CG.txt*** provided in the example.data directory is generated from the code file ***example.script.txt***. A sample of this output is shown below in Box 7.

Box7. DM.CG.txt output file

chr	pos	pos2	p.value	test.s	con.s	test.s2	con.s2	test.mC	con.mC	test.mC2	con.mC2
chr1	848868	848873	0.0256	0.5:0:0:1	0:0:0:0	1:0.5:0:1	0:0:0:0	0.59:0.2:0.11:0.85	0.08:0.15:0.02:0	0.9:0.47:0.28:0.89	0.18:0.33:0.11:0
chr1	848873	848889	0.0256	1:0.5:0:1	0:0:0:0	1:0:0:0.5	0:0:0:0	0.9:0.47:0.28:0.89	0.18:0.33:0.11:0	0.8:0:0:0.71	0:0.1:0:0

test.post	con.post	test.post2	con.post2	meanDiff	DM.p	DM.status	index	meanCov.test	meanCov.control
0.97:0.93:1:1	1:0.97:1:1	0.97:0.73:1:0.97	1:0.87:1:1	0.3758	hyper	hyper	161	25.5	31
0.97:0.73:1:0.97	1:0.87:1:1	0.9:1:1:0.67	1:1:1:1	0.4785	hyper	hyper	162	24.75	30.25

4.4 DM regions output: DMRs.txt

The identified DM CG sites can be further summarized into DM regions based on the DM status, distance between CG sites, and density of covered CG sites (see Supplemental file for detail). These DM regions are reported in file "*DMRs.txt*". It contains a header line and 11 fields for each DM region. Hyper regions are listed first, followed by Hypo regions. Within each region type, DMRs are ordered based on their positions. A sample of this output (generated from the code file ***example.script.txt***) is shown below in Box 8.

Box8. DMRs.txt output file

chr	start	end	len	DM	num.CG	total.CG	meanCov.test	meanCov.control	meanDiff.mC
chr1	848868	848873	6	hyper	2	2	25.12	30.62	0.4272
chr1	851081	851123	43	hyper	5	5	5.45	8.1	0.6126
chr1	858338	858368	31	hyper	2	2	22.25	12	0.5629
……									
chr1	2243710	2243744	35	hypo	5	5	18.15	18.65	-0.5163
chr1	2373065	2373065	1	hypo	1	1	15	21.75	-0.4574

chr – chromosome number
start – start position for each region
end – end position for each region
len – the length of each region. If "len" is 1, it means it is a DMC (or DM cytosine singleton)
DM – the DM status of this region, "Hyper" or "Hypo"
num.CG – number of DM CG sites within the region
total.CG – number of all CG sites within the region
meanCov.test – mean coverage of the test group
meanCov.control – mean coverage of the control group
meanDiff.mC – the methylation difference between the two groups = mean (test) – mean (control)

4.5 HMM output: joint.prob.ps

The convergence of the model can be checked by plotting the joint probability over iterations for all samples: *joint.prob.ps* in the output directory. Figure 1 shows the joint probabilities of running HMM-Fisher on example data with 60 iterations.

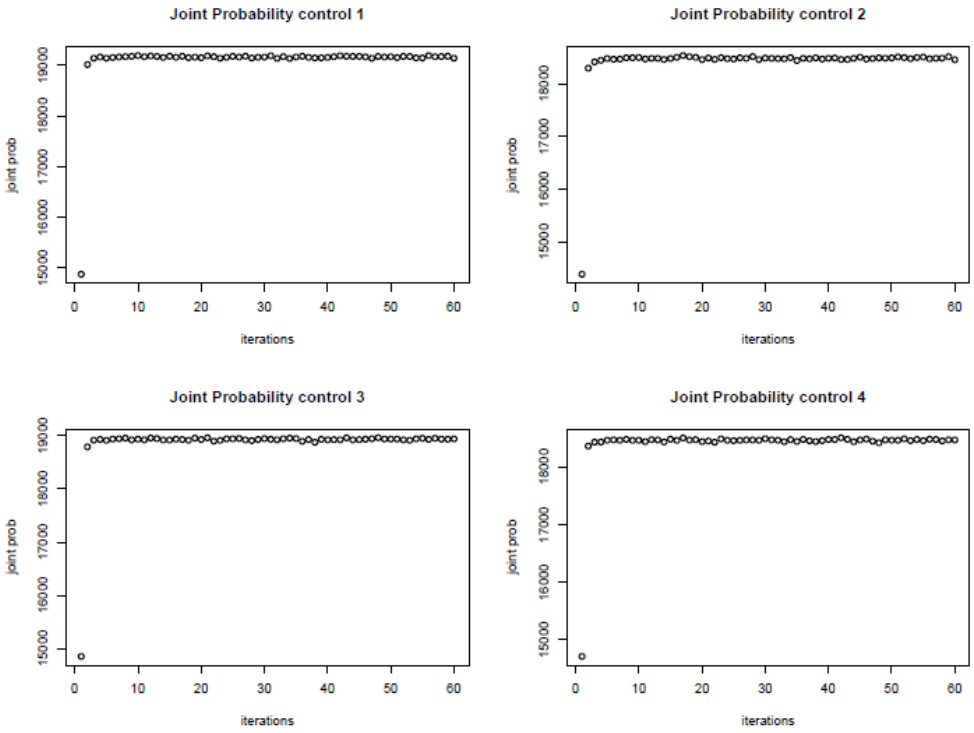

Figure 1. Joint probability of the 4 control samples in example data.

5 Further Analysis

5.1 DMR visualization

We provide an R script *plotDMRs.R* in the *HMM.Fisher.code* directory to plot the identified DMRs.

UNIX command to perform annotation analysis

R CMD BATCH '—args input1 input2 index extend test control header output' HMM.Fisher.code/plotDMRs.R

Arguments

1. ***input1***: The *mC.matrix.txt* output generated by HMM-Fisher program. See section 4.1 for detail.
2. ***input2***: The *DMRs.txt* output generated by HMM-Fisher program. See section 4.5 for detail.
3. ***index***: Vector, which DMR users want to plot in *DMRs.txt* file, e.g., c(13,31:33) means to plot the 13th, and 31th to 33th DMRs in the *DMRs.txt* file.
4. ***extend***: Numeric, how many bp to extend to either side of the region.
5. ***test***: Numeric, number of test samples.
6. ***control***: Numeric, number of control samples.
7. ***header***: Logical, whether ***input2*** file has a header line. T, TRUE; F, FALSE.
8. ***output***: The name for the output .ps file. The file ***example.DMR.plot.ps*** in example.data directory is an output generated from the ***DMRs.txt***. Example of this file is shown in Figure 2.

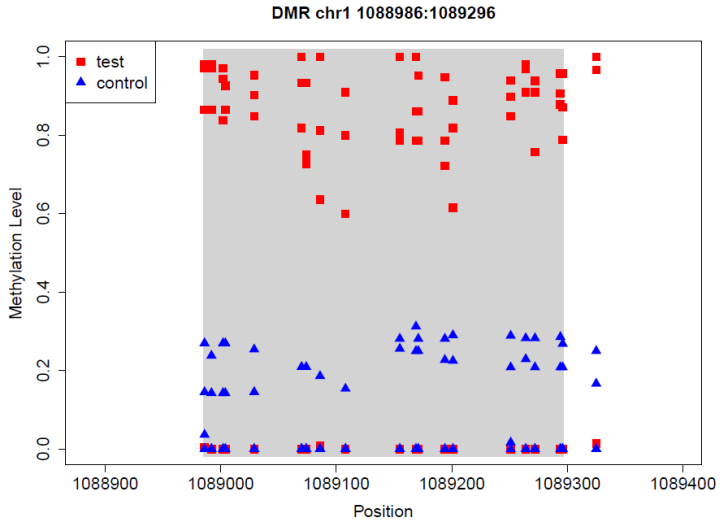

Figure 2. Methylation level of all samples within a detected DMR. The DMR is highlighted in gray.

Example command line that generate *example.DMR.plot.ps*

R CMD BATCH '--args mC.matrix.txt DMRs.txt c(26:27,59) 100 4 4 T example.DMR.plot' HMM.Fisher.code/plotDMRs.R

5.2 Annotation

We also provide an R script *annotation.R* in the *HMM.Fisher.code* directory if users want to perform annotation analysis. This R script takes the *DM.CG.txt* output from HMM-DM program and the annotation file downloaded from UCSC table browser as input, and generates the annotation information for each DM CG identified. If users want to use other annotation resources, the *annotation.R* script can be easily revised to fit their need.

UNIX command to perform annotation analysis

R CMD BATCH '--args input1 input2 distance output' HMM.Fisher.code/annotation.R

Arguments

1. ***input1***: The *DM.CG.txt* output generated by HMM-Fisher program. See section 4.3 for detail.
2. ***input2***: The annotation file downloaded from the UCSC table browser for your genome of interest. To download this file, go to http://genome.ucsc.edu/, click "Table Browser" on the right menu. Select your "**genome**" of interest and "**assembly**", which should be consistent with the reference genome you use to align bisulfite sequencing reads. Select *"Genes and Gene prediction tracks"* from the **"group"** drop-down menu, and select *"Refseq Genes"* from the **"track"** drop-down menu. Select "*all fields from selected table*" for the "**output format**". Type in the file name (e.g., refGene.txt) in "**output file**", then click "**get output**" to download the annotation file.
3. ***distance***: The distance of the promoter regions. The promoter region for a specific gene is defined as the *distance* bp extended from the start and end of the gene.
4. ***header1***: Logical, whether ***input1*** file has a header line. T, TRUE; F, FALSE.
5. ***header2***: Logical, whether ***input2*** file has a header line. T, TRUE; F, FALSE.
6. ***output***: The annotation output file. This file contains 7 fields for each CG in *DM.CG.txt*. The ***annotation.txt*** provided in example.data directory is generated from the ***DM.CG.txt***. Example of this file is shown in Box 9.

Box9. Output of *annotation.R*

chr	pos	DM	meanDiff.mC	meanCov	genes	promoters
chr1	795361	hyper	0.4652	69:70.25	FAM41C	NA
.....						
chr1	1061913	hyper	0.3752	13.25:12	NA	LOC254099

chr – chromosome number

pos – position for each CG in *DM.CG.txt*

DM – the DM status of each CG

meanDiff.mC – the mean difference of methylation levels between the two groups (test – control)

meanCov – the mean coverage of test group: the mean coverage of control group

genes – list of genes that contain this CG site in gene bogy regions, separated by ":". Labeled as "NA" if not covered by any gene in gene bogy regions.

promoters – list of genes that contain this CG site in their promoter regions, separated by ":". Labeled as "NA" if not covered by any gene in promoter regions.

Example command line that generates *annotation.txt*

R CMD BATCH '--args DM.CG.txt refGene.txt 1000 T T annotation.txt' HMM.Fisher.code/annotation.R

6 References

[1] Sun S, Yu X: **HMM-Fisher: Identifying differential methylation using a hidden Markov model and Fisher's exact test**. *manuscript submitted* in 2015.

[2] Sun Z, Asmann YW, Kalari KR, Bot B, Eckel-Passow JE, Baker TR, Carr JM, Khrebtukova I, Luo S, Zhang L *et al*: **Integrated analysis of gene expression, CpG island methylation, and gene copy number in breast cancer cells by deep sequencing**. *PLoS One* 2011, **6**(2):e17490.