Documentation for HMM-DM

Xiaoqing Yu (xxy39@case.edu)

Shuying Sun (ssun5211@yahoo.com)

March 24, 2014

Contents

1 Overview and Installation

1.1 Overview

The *HMM-DM* [2] program identifies differentially methylated (DM) CG sites and regions from whole genome and targeted bisulfite sequencing (BS) data. This approach first uses a Hidden Markov Model to identify differentially methylated CG sites accounting for spatial correlation across CGs and variation across samples, and then summarizes identified DM CG sites 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 demonstrate the use of HMM-DM using a publicly available bisulfite-treated methylation sequencing dataset [1] on chromosome 1 in section 2. This dataset contain eight breast cancer cell lines, including four estrogen receptor positive (ER+: BT474, MCF7, ZR751, and T47D) and four negative (ER-: BT20, MCF10A, MDAMB231, and MDAMB468) 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-DM requires a Linux/Unix system, with R installed. To install HMM-DM, the user can download the pipeline from (https://github.com/xxy39/HMM-DM). After unzipping the file, there are one document and two folders.

HMM.DM.documents.pdf	A copy of the user manual
HMM.DM.code	A folder containing all R source code files used for HMM-DM.
example.data	A folder containing all example input data as mentioned in this document, an example.script.txt for running HMM-DM (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, the users only need to call the main function HMM.DM (). This function identifies DM regions in four steps:

1. Perform quality control based on coverage
2. Identifiying DM CG sites using HMM-DM method
    a. Estimate the differential methylation states (Hyper, hypermethylated; EM, equally methylated; and Hypo, hypomethylated) for all CG sites with HMM
    b. Filter the DM CGs (Hyper or Hypo from step 1) with following criteria
        i. DM CGs with small mean difference are re-classified as EM
        ii. DM CGs with low posterior probability are re-classified as EM
3. Summarize the filtered DM CGs into DM regions, based on their DM states, distance between CGs, and posterior probabilities.

HMM.DM

Description

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

Usage

HMM.DM (total.reads, meth.reads, n.control, n.test, 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.
n.control	Numeric. Number of control samples.
n.test	Numeric. Number of test samples.
chromosome	Character. The chromosome the users want to analyze, e.g., chromosome =1, or chromosome = 2. The HMM-DM processes one chromosome at a time.
code.dir	String. The directory of the source code files of HMM-DM (e.g., /home/HMM.DM /HMM.DM.code). Note, there should be no "/" at the very end.

Output.dir — String. The directory for the output files (e.g., /home/HMM.DM.results). Note, there should be no "/" at the very end. Two matrices will be generated from this function. See section 3.2 for more detail. When analyzing multiple chromosomes, we recommend users specify different *output.dir* for different chromosomes.

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-DM. 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.

post.threshold — Numeric. Filtering based on posterior probability. DM CG sites with posterior probability < *post.threshold* are filtered out. Default = 0.4.

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-DM but with relatively low posterior probability (controlled by *max.post*); or 2) identified as DM by HMM-DM but with small meanDiff (< *meanDiff.cut*). Default = 1. Note: if either region is a singleton, only 1 EM CG is allowed.

max.post — Numeric between 0 and 1. The maximum posterior probability for the EM included in the combined DM region. Default = 0.8.

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-DM takes the number of total reads and number of methylated reads as input. Current version of HMM-DM 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-DM constrain the analysis to the CG sites that pass the quality control based on coverage. To ensure more accurate results, we also encourage the users to first filter out the CG sites with low coverage.

HMM-DM processes one chromosome at a time. To analyze multiple chromosomes, we recommend the users to 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. There are 1+*n.control*+*n.test* columns: position for each CG, the number of reads for samples in group1 (e.g., control group), the number of reads for samples in group2 (e.g., test 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 control samples and 4 test samples. A sample of this file is shown below.

Box1. mC.matrix input file

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

3.2 meth.reads

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

Box2. cov.matrix input file

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

3.3 UNIX command

An example script of running HMM-DM is shown in ***example.script.txt*** under the example.data folder. Default settings are used for this example script. The users may change the parameters based on their own data 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
```

All results are saved under the output directory defined by HMM.DM parameter output.dir (see section 2).

A brief description of this example code is provided below:

Input: This RRBS dataset contain 20,000 CG sites from chromosome 1 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 X 9, see section 3.1)
meth.reads	example.meth.reads.txt (20,000 X 9, see section 3.2)
n.control	4
n.test	4
chromrosome	1

Quality control: The data is reduced to CG sites covered in 100% of control and test samples (*min.percent* = 1). After quality control, 5,811 CG sites are left for further analysis.

Identifying DM CG sites: We apply HMM-DM to the 10,000 CG sites with 60 *iterations*. To call a DM CG site, we require 1) this CG is either Hyper or Hypo; 2) its posterior probability is ≥ 04 (*post.threshold* = 0.4); 3) and its 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 at most 100 bp (*max.disance* = 100); 2) between the two CG sites, there are at most 3 CG sites which fail the quality control (*max.empty.CG* = 3). Two DMRs are later merged if 1) they are in the same DM status; 2) there are at most 1 EM CG site between the two DMRs (*max.EM* = 1) and this CG site has posterior probability ≤ 0.8 (*max.post* = 0.8).

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	A quality control output (1867 CG sites, see section 4.1)
all.CG.txt	DM status for all CG sites (1867 CG sites, see section 4.2)
DM.CG.txt	Identified DM CG sties (111 DM CG sites, see section 4.3)
joint.prob.ps	Shows the HMM convergence (see section 4.4)
DMRs.txt	Identified DMRs (32 DMRs, see section 4.5)

4 Output Files

4.1 Quality control output: mC.matrix.txt

The first output from HMM-DM 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 1,867 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	control_1	control_2	control_3	control_4	test_1	test_2	test_3	test_4
497	0.602339	0.956522	0.979899	0.936508	0.988701	0.886364	0.886598	0.977778
525	0.970930	0.949640	0.959799	0.984375	0.971591	1.000000	0.964286	0.956522
542	0.942149	0.992647	0.946524	0.909091	0.944056	1.000000	0.978495	0.932584

4.2 HMM-DM raw output: all.CG.txt

This output from HMM-DM method shows the estimated DM status for each CG site being analyzed. It contains a header line and 12 fields for each CG site. *DM.status* indicates the final status for each CG.

1) *DM.stauts* = 1 means "Hyper": CG sites in which the control group has a higher methylation level than the test group (*mCstatus* =1 and *meanDiff* $\geq$ 0.3);
2) *DM.stauts* = -1 means "Hypo": CG sites in which the test group has a higher methylation level (*mCstatus* = -1 and *meanDiff* $\leq$ -0.3);
3) *DM.stauts* = 0 means "EM": the other 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.

Box6. all.CG.txt output file

chr	pos	Hypo.pos	EM.pos	Hyper.pos	max.p	mCstatus	meanDiff	DM.status	index	meanCov.control	meanCov.test
chr1	497	0	1	0	1	0	-0.0660	0	1	158.5	126.25
chr1	525	0	1	0	1	0	-0.0069	0	2	159.5	126.75
…											
chr1	795361	0.7	0.3	0	0.7	-1	-0.4652	-1	74	70.25	69
chr1	795363	0.8	0.2	0	0.8	-1	-0.5029	-1	75	66.5	67.25

chr – chromosome number
pos – position for each CG
Hypo.pos – posterior probability for Hypo state
EM.pos – posterior probability for EM state
Hyper.pos – posterior probability for Hyper state
max.p – the maximum posterior probability of the three states

mC.status – the state of this CG (the state with the highest posterior probability). -1, Hypo; 0, EM; 1, Hyper.
meanDiff – the mean difference of methylation level between the two groups (control group – test group)
DM.status – the DM status of the CG site considering the mean difference. For a given CG site, if the mC.status is -1 or 1, while the absolute value of meanDiff is less than the meanDiff.cut parameter provided by the user (default = 0.3), this CG site will be identified as a EM.
index – the index of the CG site in mC.matrix file
meanCov.control – the mean coverage of control group
meanCov.test – the mean coverage of test group

4.3 DM CG output: DM.CG.txt

This output shows the DM CG sites identified by HMM-DM. 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	Hypo.pos	EM.pos	Hyper.pos	max.p	mCstatus	meanDiff	DM.status	index	meanCov.control	meanCov.test
chr1	795361	0.7	0.3	0	0.7	-1	-0.4652	-1	74	70.25	69
chr1	795363	0.8	0.2	0	0.8	-1	-0.5029	-1	75	66.5	67.25
chr1	848868	0.7667	0.2333	0	0.7667	-1	-0.3759	-1	161	31	25.5
chr1	848873	0.8	0.2	0	0.8	-1	-0.4785	-1	162	30.25	24.75

4.4 HMM output: joint.prob.ps

The convergence of the model can be checked by examining the plot the joint probability over iterations, file *joint.prob.ps* in the output directory. Figure 1 shows the joint probabilities of running HMM-DM on example data with 60 iterations.

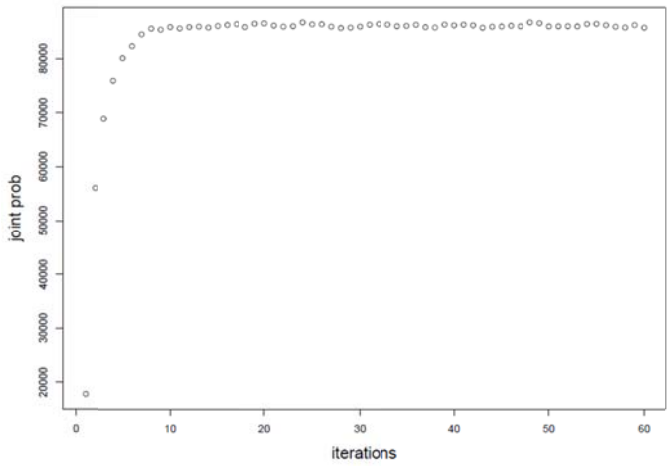

Figure 1. Joint probability of applying HMM-DM to the example data.

4.5 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 regions. Hyper regions are listed first, followed by Hypo regions. Within each region type, the 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.control	meanCov.test	meanDiff.mC	meanPost
chr1	2243626	2243744	119	hyper	9	10	28.44	28.44	0.4675	0.8741
chr1	2260304	2260304	1	hyper	1	1	23.25	23.25	0.547	0.7333
chr1	2373065	2373081	17	hyper	2	2	21.88	21.88	0.5836	0.8334
…										
chr1	1061913	1062006	94	hypo	5	7	8.25	8.25	-0.5109	0.8667
chr1	1062332	1062420	89	hypo	10	10	52.48	52.48	-0.456	0.87

chr – chromosome number
start – start position for each region
end – end position for each region
len – the length of each region
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.control – mean coverage of the control group
meanCov.test – mean coverage of the test group
meanDiff.mC – the methylation difference between the two groups = mean (control) – mean (test)
meanPost – the mean posterior probability of DM CG sites within this region

5 Annotation Analysis

We also provide an R script *annotation.R* in the *HMM.DM.code* directory if the 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 the 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.DM.code/annotation.R

Arguments

1. ***input1***: the *DM.CG.txt* output generated by HMM-DM program. See section 4.3 for detail.
2. ***input2***: the annotation file downloaded from 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. Promoter region for a specific gene is defined as the *distance* bp extended from the start and end of the gene.
4. ***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	858379	hypo	-0.5839	10.25:18.5	SAMD11	NA
chr1	885064	hypo	-0.4008	19.25:9.5	NA	KLHL17;NOC2L

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 level between the two groups (control – test)

meanCov.control – the mean coverage of control group

meanCov.test – the mean coverage of test 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

R CMD BATCH '--args DM.CG.txt refGene.txt 1000 annotation.txt' HMM.DM.code/annotation.R

6 References

1. 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.
2. Yu X, Sun S. **HMM-DM: Identifying differentially methyalted regions using a Hidden Markov model**.