Supplemental file for the manuscript entitled "*Comparing five statistical methods of differential methylation identification using bisulfite sequencing data*" (Revised on January 24, 2016)

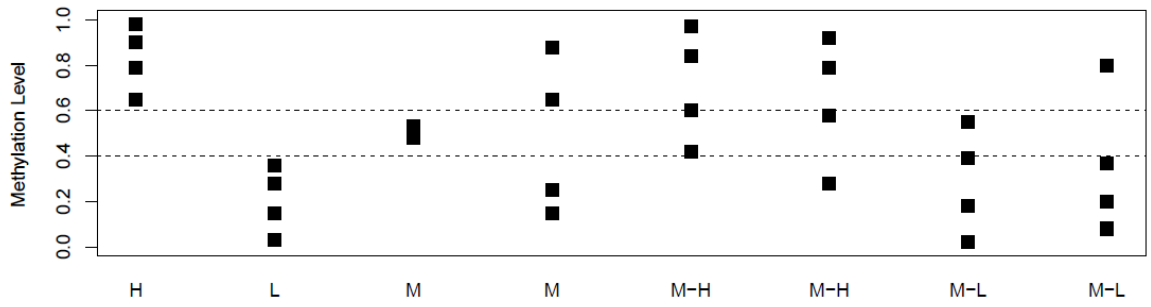
1 Simulating dataset with known DMRs

In order to preserve the natural changes in methylation patterns across CG sites and the variation patterns among samples, all DMRs are chosen based on the methylation and variation status of the "control" group. In detail, the simulation process includes four steps:

1. Categorize CG sites into five classes based on their methylation statuses
2. Group CG sites into regions bases on their classes
    a. Summarize CG sites into four types of regions
    b. Refine and merge regions
    c. Define the patterns of refined regions
3. Select DMRs from the defined regions
4. Simulate methylation levels for DMRs and background CG sites (CG sites not in DMRs) in cancer group

Categorize CG sites

CG sites are categorized into five methylation classes based on their methylation levels and heterogeneity statuses in control group.



Supplemental Figure 1. Examples of the five methylation classes based on methylation levels and variations within control group. H (high methylation): the methylation levels of all four control samples are ≥ 0.6, such that the between sample variation is relatively small; L (low methylation): the methylation levels of all four control samples are ≤ 0.4, such that the between sample variation is relatively small; M (median methylation): the mean of four control samples is within the range of (0.4, 0.6); M-H (median-high methylation): the mean is ≥ 0.6 but the methylation level of the four sample spans larger range compared to the class H; M-L (median-low methylation): the mean is ≤ 0.4 but the methylation level of the four sample spans larger range compared to the class L.

Group CG sites into regions

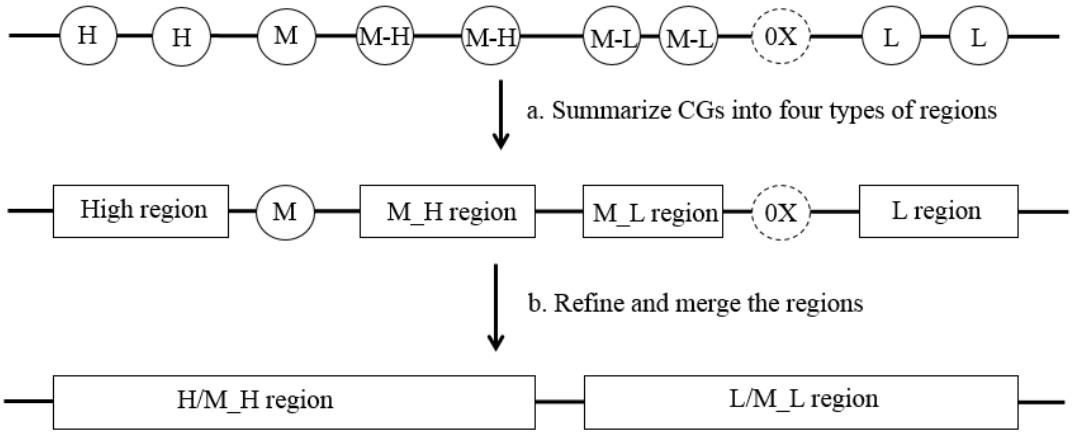
Based on the methylation classes, consecutive CG sites of the same class are grouped together, generating four types of regions: H regions, L regions, M-H regions, and M-L regions (Supplemental Figure 2, state a). The first two types have small variations, while the last two types have larger variations. To group any two consecutive CG sites in the same region, these two CG sites have to meet the following criteria:

(1) They belong to the same methylation class;
(2) Their physical distance is ≤ 100 bp;
(3) There are ≤ 3 CG sites without coverage between them.

Later, the defined regions are further refined and merged together with M CG sites allowed in the regions (Supplemental Figure 2, step b). In particular, two H regions and/or M-H regions are allowed to merge, if the following criteria are satisfied:

(1) Their distance is ≤ 100 bp;
(2) There are ≤ 3 M CG sites between two regions and if one region is a singleton, only 1 M CG is allowed in-between;
(3) There are ≤ 3 CG sites without coverage between them.

By doing this, H/M-H regions are defined as the regions that include mainly H and/or M-H CG sites, and a few M CG sites. Similarly, L/M-L regions are defined as the regions that include mainly L and/or M-L CG sites, and a few M CG sites. This approach ensures that all CG sites within one region have the same or similar methylation status. In addition, only a low frequency of M CG sites is allowed, such that each region could have a clear methylation pattern.



Supplemental Figure 2. Examples of grouping CG sites into regions based on their methylation classes. Step a, summarize CG sites into four types of region; Step b, refine and merge the regions. H, CG sites or regions that have high methylation levels in control group; L, CG sites or regions that have low methylation levels in control group; M-H, CG sites or regions that have large mean and large variation in control group; M-L, CG sites or regions that have small mean and large variation in control group.

Then, the pattern of each region is defined based on the distribution of CG sites with different methylation classes within that region. For instance, for an H/M-H region, if more than 80% of the CG sites are of H class, this region is defined as an "H region"; otherwise, it is defined as an "M-H region". Similarly, for an L/M-L region, if more than 80% of the CG sites are of L class, this region is defined as an "L region"; otherwise, it is defined as an "M-L region". This setting enables us to differentiate the regions with various levels of variation. In particular, H and L regions include CG sites with small variation across the four samples (H and L CG sites); on the other hand, M-H and M-L regions include more CG sites with large variation across the four samples (M-H and M-L CG sites). This step generates 2549 regions, including 817 H, 393 L, 800 M-H, and 449 M-L regions. The lengths of the regions range from 1 bp to 755 bp, with a median of 17 bp.

Select DMRs

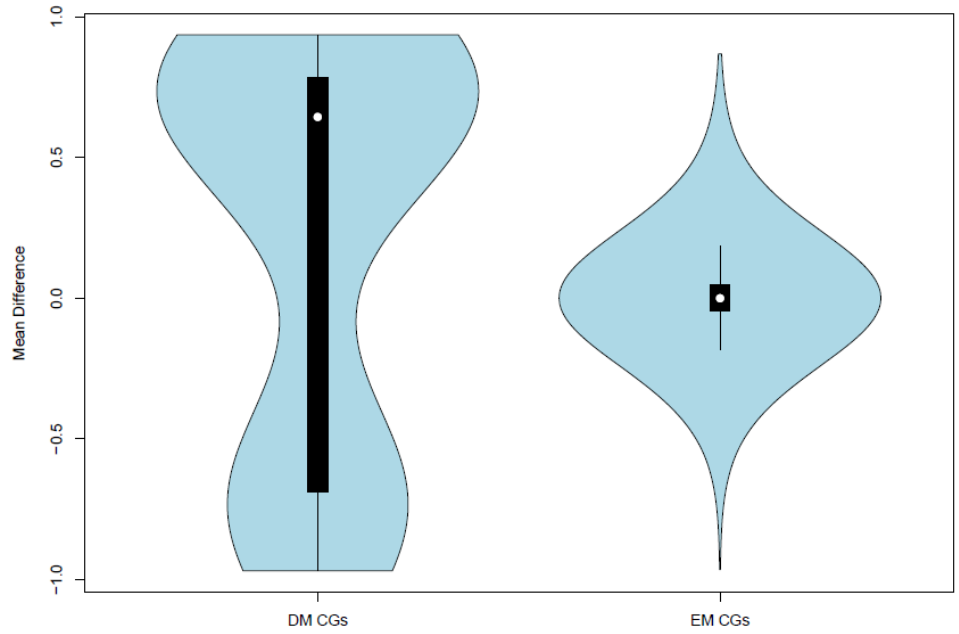
From the regions generated above, we then randomly choose 80 DMRs with various methylation statuses and sizes to insert methylation differences (Supplemental Table 1). For the singletons, we only select the ones without neighboring CG sites within 100 bp. This approach of choosing DMRs based on the methylation status ensures that the natural changes in methylation patterns across CG sites are preserved.

Supplemental Table 1. Number of DMRs with various size and methylation status

Size	> 20	[11, 20]	[3, 10]	[2,1]
H DMRs	2	10	5	5
L DMRs	5	10	5	5
M-H DMRs	2	5	5	5
M-L DMRs	1	5	5	5

Simulate DMRs and background CG sites

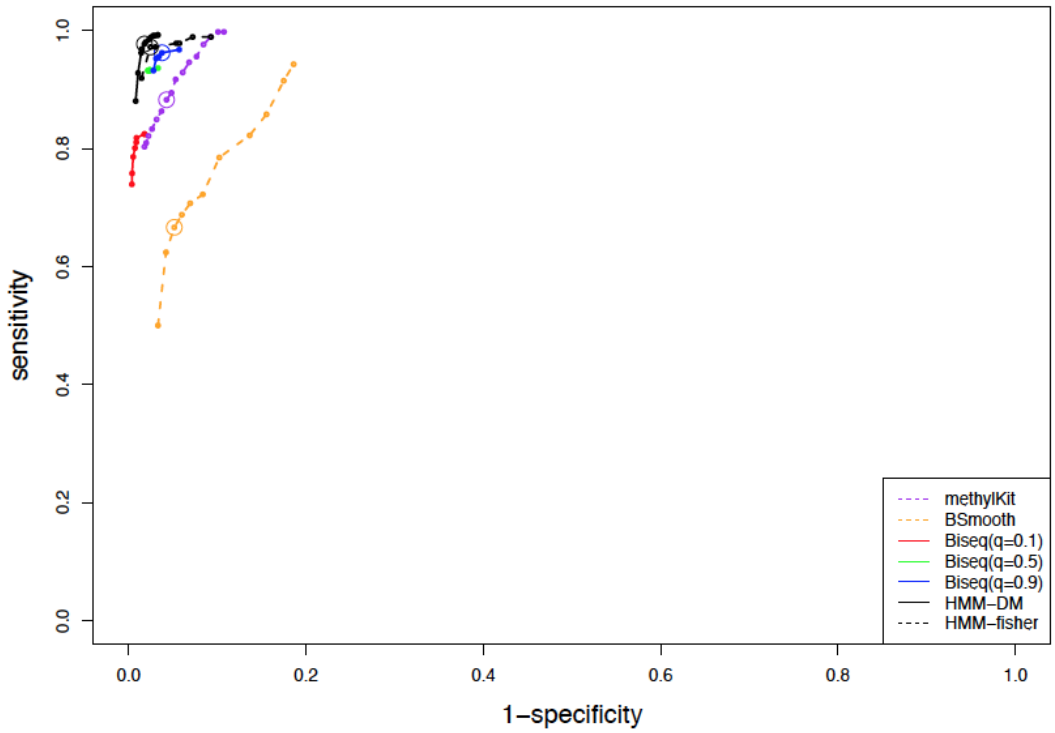
Instead of simulating methylation differences, we simulate methylation levels for the test group in the 80 selected DMRs. For the DMRs that have relatively lower methylation levels in the control group (e.g., L and M-L regions), the test group is sampled from uniform distributions with higher means to generate high methylation levels. For DMRs that have relatively higher methylation levels in the control group (e.g., H and M-H regions), the test group is sampled from uniform distributions with lower means to generate low methylation levels. By doing this, we ensure the actual difference between control and test groups in DM CG sites for all conditions, and this is confirmed by the plot of mean differences between groups for both DM and EM CG sites (Supplemental Figure 3).



Supplemental Figure 3. The mean difference of methylation levels between test and control groups in designed DM and EM CG sites.

Comparing the five methods using simulated data

We apply the five methods to the simulated dataset and the following Supplementary Figure 4 shows the ROC curves for differentially methylated CG sites identified in each method on the full range of [0,1]. In the main text, we use a smaller range to zoom in for a better illustration (Figure 2 of the manuscript).



Supplemental Figure 4. ROC curves for differentially methylated CG sites detected from the five methods on the full range of [0,1].

2 Default settings in real breast cancer dataset

Supplemental Table 2. Results of the five methods with default settings in a real dataset
Shown are the number of DM CG sites, Hypermethylated (Hyper), and hypomethylated (Hypo) CG sites identified in each method. DM CG sites in which ER+ has higher methylation level compared to ER- are defined as hypermethylated, and DM CG sites in which ER+ has lower methylation level are defined as hypomethylated.

Default	**DM**	**Hyper**	**Hypo**
methylKIt (q< 0.01)	6239	3619	2620
Bsmooth (±2)	2540	1906	634
BiSeq	33	33	0
HMM-DM	2326	1789	537
HMM-fisher	1917	1513	404

3 Settings in HMM-DM and HMM-Fisher

As in other methods, we also change parameter settings in HMM-DM and HMM-Fisher. For HMM-DM, key parameters including the prior for transition and emission probabilities are estimated from the data directly. The only two parameters might need to be modified are the size of partition (number of CG sites to break the Markov chain), and the dirichlet prior for transition probabilities. Supplemental Table 3 shows the modified settings and their results in both simulated and real data. All settings show similar results as the default setting. Similarly in HMM-Fisher, there are only two parameters might need to be modified: the standard deviation of the Truncated Normal distribution of emission probabilities for three states, and the dirichlet prior for transition probabilities. All modified settings show similar results as the default setting in both simulated and real data (Supplemental Table 4).

Supplemental Table 3. HMM-DM with different settings in simulated data and real data

Shown are different settings of HMM-DM in simulated and real data. The modified settings include the size of partition, dirichlet distribution of transition prior. The differentially methylated CG sites are defined as the hyper and hyper CG sites with posterior probability > 0.4. For simulated data, the reported results include number of identified DM, the number of true positive DM sites and true positive rate, the number of false positive DM sites and false positive rate, and the number of false negative DM sites. The Set 1 is the default setting of HMM-DM. For real data, the reported results include number of identified DM, DM with mean difference ≥ 0.3, number of Hypermethylated CG sites, and number of Hypomethylated CG sites.

A. Simulated data

Set	Partition	Transition prior	Identified DM	TP (TP rate)	FP (FPR)	FN
1	200 CG	dirichlet (10,10,10)	1220	922 (99.25%)	298 (3.19%)	7
2	200 CG	dirichlet (1,1, 1)	1150	920 (99.03%)	230 (2.54%)	9
3	200 CG	dirichlet (5,5,5)	1190	922 (99.25%)	268 (2.95%)	7
4	200 CG	dirichlet (20,20,20)	1251	920 (99.03%)	331 (3.65%)	9
5	500 CG	dirichlet (5,5, 5)	1193	922 (99.25%)	271 (2.99%)	7
6	500 CG	dirichlet (10,10,10)	1194	921 (99.14%)	273 (3.01%)	8
7	500 CG	dirichlet (20,20,20)	1192	920 (99.03%)	272 (3.00%)	9
8	500 CG	dirichlet (50,50,50)	1272	918 (98.82%)	354 (3.90%)	11

B. Real data

Set	Partition	Transition prior	Identified DM	DM with meandiff ≥ 0.3	Hyper	Hypo
1	200 CG	dirichlet (10,10,10)	2639	2326	537	1789
2	200 CG	dirichlet (1,1, 1)	2169	1913	377	1536
3	200 CG	dirichlet (5,5,5)	2351	2118	468	1650
4	200 CG	dirichlet (20,20,20)	2942	2563	633	1930
5	500 CG	dirichlet (5,5, 5)	2189	1969	390	1679
6	500 CG	dirichlet (10,10,10)	2252	2033	426	1607
7	500 CG	dirichlet (20,20,20)	2443	2184	495	1689
8	500 CG	dirichlet (50,50,50)	2825	2460	596	1864

Supplemental Table 4. HMM-Fisher with different settings in simulated data and real data

Shown are different settings of HMM-Fisher in simulated and real data. The modified settings include the standard deviation of the Truncated Normal distribution for emission probability of state (N, P, F), dirichlet distribution of transition prior. CG sites with p-value ≤ 0.05 in Fisher's exact test are defined as DM CG sites. For simulated data, the reported results include number of identified DM, the number of true positive DM sites and true positive rate, the number of false positive DM sites and false positive rate, and the number of false negative DM sites. The Set 1 is the default setting of HMM-DM. For real data, the reported results include number of identified DM, DM with mean difference ≥ 0.3, number of Hypermethylated CG sites, and number of Hypomethylated CG sites.

A. Simulated data

Set	**Sd of Emission probability**	**Transition prior**	**Identified DM**	**TP (TP rate)**	**FP (FPR)**	**FN**
1	(0.12, 0.15, 0.13)	dirichlet (1, 1, 1)	1171	903(97.20%)	271 (2.91%)	26
2	(0.15, 0.15, 0.15)	dirichlet (1, 1, 1)	1182	905 (97.43%)	277 (3.04%)	24
3	(0.1, 0.1, 0.1)	dirichlet (1, 1, 1)	1180	906 (97.52%)	274 (3.02%)	23
4	(0.2, 0.2, 0.2)	dirichlet (1, 1, 1)	1180	906 (97.52%)	274 (3.02%)	23
5	(0.12, 0.15, 0.13)	dirichlet (10, 10, 10)	1173	902 (97.09%)	271 (2.99%)	27
6	(0.12, 0.15, 0.13)	dirichlet (50, 50, 50)	1058	896 (96.45%)	162 (1.79%)	33

B. Real data

Set	**Sd of Emission probability**	**Transition prior**	**Identified DM**	**DM with meandiff ≥ 0.3**	**Hyper**	**Hypo**
1	(0.12, 0.15, 0.13)	dirichlet (1, 1, 1)	2704	1917	404	1513
2	(0.15, 0.15, 0.15)	dirichlet (1, 1, 1)	2605	1851	399	1452
3	(0.1, 0.1, 0.1)	dirichlet (1, 1, 1)	2545	1850	394	1456
4	(0.2, 0.2, 0.2)	dirichlet (1, 1, 1)	2495	1707	366	1341
5	(0.12, 0.15, 0.13)	dirichlet (10, 10, 10)	2685	1899	403	1496
6	(0.12, 0.15, 0.13)	dirichlet (50, 50, 50)	2662	1888	396	1492