# Supplementary file for the manuscript entitled "*HMM-DM: Identifying differentially methylated regions using a hidden Markov model*"

(Revised on Aug 5, 2015)

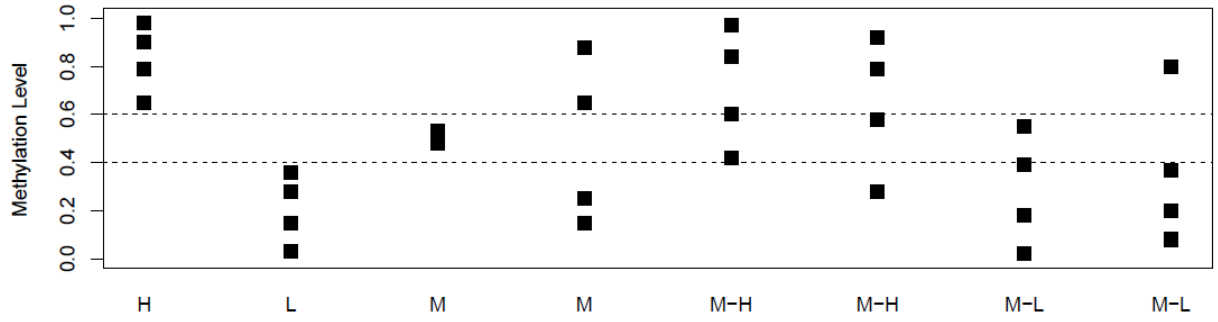
## 1 Simulating dataset with known DMRs

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 status
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 the CG sites**

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




*Supplementary Figure 1. Examples of the five methylation classes obtained 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 samples spans larger range compared to the class H; M-L (median-low methylation): the mean is ≤ 0.4 methylation level of the four samples spans larger rang compared to the class L.*

## Group CG sites into regions

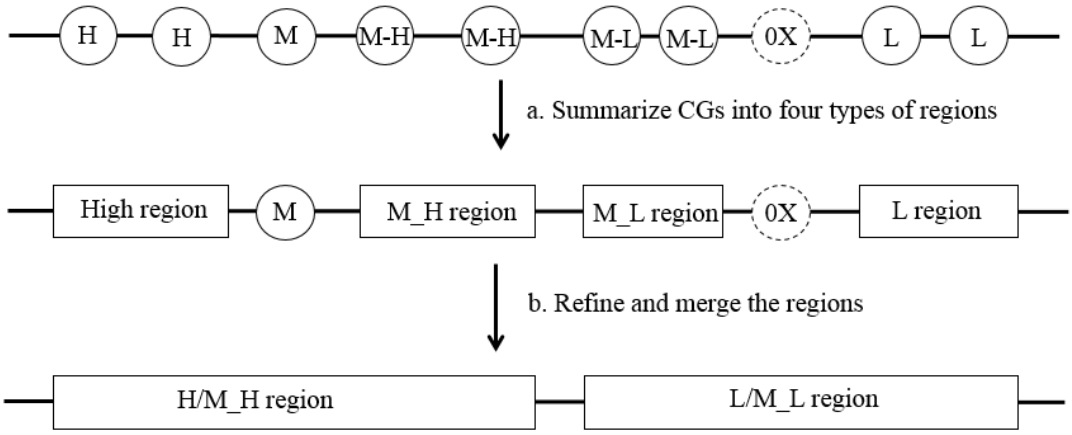
Based on the methylation classes, consecutive CG sites of the same class are grouped together, generating four types of regions: H regions, L regions, M-H regions, and M-L regions (Supplementary Figure 2, state a). The first two types have small variations, while the last two types have larger variations. 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 (Supplementary 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; 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.




*Supplementary 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).

## Select DMRs

From the regions generated above, we then randomly choose 80 DMRs with various methylation status and sizes to insert methylation differences (Supplementary Table 1). For the singletons, we only select the ones without neighboring CG sites in 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.

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

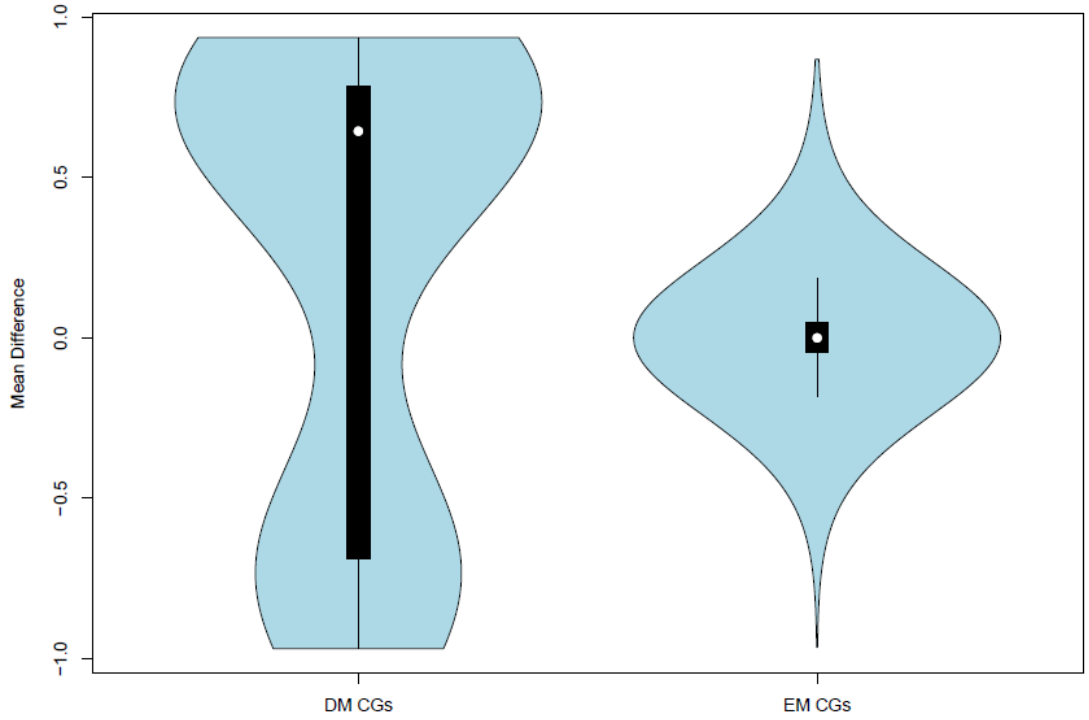
| Size | > 20 | [11, 20] | [3, 10] | [2,1] |
|---|---|---|---|---|
| H DMRs | 2 | 10 | 5 | 5 |
| L DMRs | 5 | 10 | 5 | 5 |
| M-H DMRs | 2 | 5 | 5 | 5 |
| M-L DMRs | 1 | 5 | 5 | 5 |

### 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 obtained by sampling 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. In addition, to ensure a true difference in DMRs with larger variations and/or smaller sizes, we use more stringent uniform distributions for H-M and H-L DMRs and DMRs with $\leq 3$ CG sites (Supplementary Table 2). By doing this, we ensure 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 (Supplementary Figure 3). This step generates 41 hypermethylated DMRs where the test group has higher methylation levels than the control group, and 39 hypomethylated DMRs where the test group has lower methylation levels than the control group.

Supplementary Table 2. Uniform distributions that are used to simulate the test samples in DMRs.

| | > 3 CG sites | ≤ 3 CG sites |
|---|---|---|
| H DMRs | Uniform (0, 0.4) | Uniform (0, 0.2) |
| L DMRs | Uniform (0.6, 1) | Uniform (0.8, 1) |
| M-H DMRs | Uniform (0, 0.3) | Uniform (0, 0.2) |
| M-L DMRs | Uniform (0.7, 1) | Uniform (0.8, 1) |




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

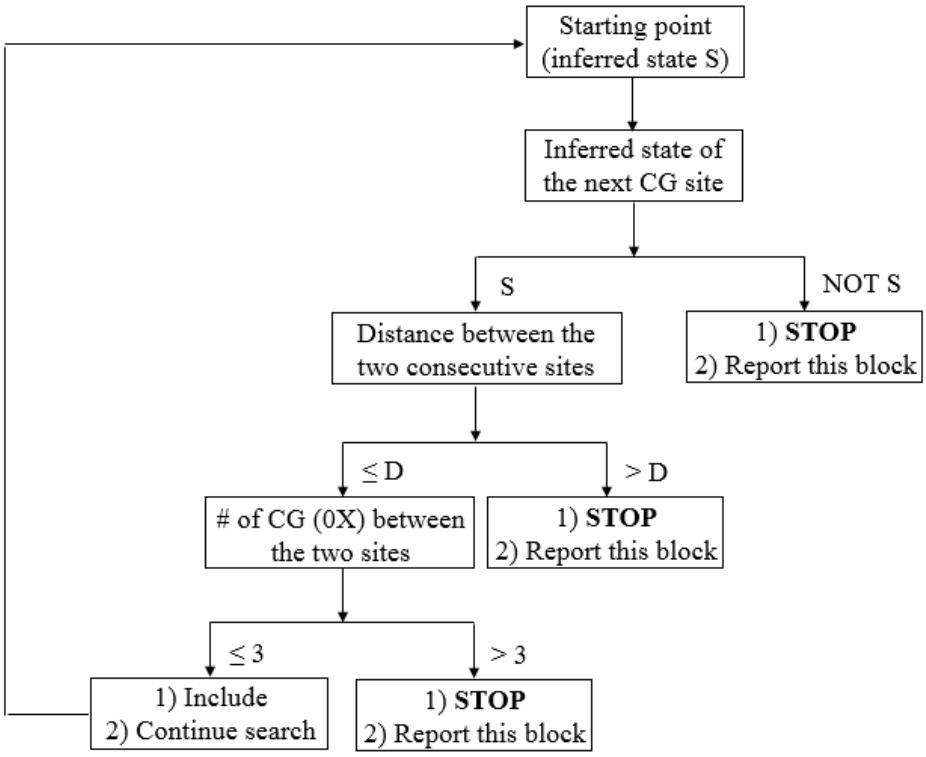
## 2 Summarizing identified DM CG sites into regions

After obtaining the DM CG sites, we develop a searching strategy to group the DM CG sites with same states together, by incorporating the distance between any consecutive CG sites and their sequencing coverage. The strategy contains the following criteria:

(1) Two consecutive CG sites have the same state.
(2) The distance between these two consecutive regions is less than the threshold D bp.
(3) Between two consecutive CG sites, there are at most three CG sites without coverage.

Supplementary figure 4 shows an example of summarizing a Hyper region. Our searching strategy treats the first CG site in the chain that is inferred as "Hyper" as the starting point, then search for the next (neighboring) CG site along the chain. If the neighboring CG site is inferred differently from the starting point, that is, if it is "EM" or "Hypo", then the search stops and the starting point is reported as a Hyper block with a single CG site – a singleton Hyper site. If the neighboring CG site is also inferred as the starting point – "Hyper", then the strategy checks the distance between these two consecutive CG sites. If the distance is larger than the threshold D bp, the search stops and reports the starting point as a singleton Hyper site. If the distance is shorter than the threshold, the strategy continues on to check the number of CG sites without coverage between these two CG sites. If there are more than 3 CG sites without coverage, the search stops and reports the starting point as a singleton Hyper site; otherwise, this neighbor CG site will be included in this Hyper block and the search continues to the next CG site until it reaches

another stopping point. Once the search stops, a Hyper block will be reported and a new search will start from the next "Hyper" CG site along the chain, such that this "Hyper" CG site will be treated as a new starting point. Note that the search results in a Hyper block and each Hyper region could be a singleton or a region with at least two CG sites. The same strategy is applied to identify Hypo regions.




*Supplementary Figure 4. Strategy diagram for identifying Hyper and Hypo regions.*

The identified differential methylation regions are later merged into larger blocks. Between two differential methylation regions, occasionally, there are CG sites that are inferred as EM but with low posterior probabilities, which means the states of these CG sites are estimated with less confidence. Also, some CG sites are estimated as differentially methylated by HMM but later called as EM due to their low mean difference. For cases like this, combining two consecutive differential methylation regions can result in a larger block and capture more information related to methylation patterns in further analysis. Therefore, to combine any two consecutive methylation regions, we use the following four criteria:

1) Two consecutive regions have the same state; that is, they are both Hyper or both Hypo.
2) The distance between these two consecutive regions is less than the threshold D bp.
3) Between the two regions, there is one CG site inferred as EM but with low posterior probability, or there is one CG site inferred as differential methylated but with smaller mean difference.
4) Between the two regions, there are at most three CG sites without coverage.