Supplementary Material

DM-BLD: Differential Methylation detection using a hierarchical Bayesian model exploiting local dependency

Xiao Wang[1], Jinghua Gu[1], Leena Hilakivi-Clarke[2], Robert Clarke[2] and Jianhua Xuan[1,*]

Table of Contents

S1. Correlation among neighboring CpG sites

We studied the dependency among neighboring CpG sites using three data sets: cell line data, our in-house human data, and TCGA breast cancer data. The cell line data set consists of two cell lines (LCC1 and LCC9), with three biological replicates in each cell line; the in-house human data set consists of samples from two phenotypes, with 6 samples and 5 samples from each group, respectively; the TCGA breast cancer data set consists of 61 estrogen receptor (ER)-positive breast cancer samples divided into two groups, with 41 samples and 20 samples from each group. All of the three data sets were profiled by Illumina 450K, which measured 485,512 CpG sites covering 21,227 genes. The median of the number of CpG sites in each gene is 15, and the median distance between two consecutive CpG sites is about 300 bps. We used correlation coefficient as the metric, and compared the correlation of CpG sites in the following three scenarios: 1) randomly selected CpG sites within 1,000 bases; 2) randomly selected CpG site and its 50 closest neighboring sites; 3) randomly selected CpG sites across the genome. Fig. S1 shows the correlation of CpG sites in the three scenarios using the three data sets.

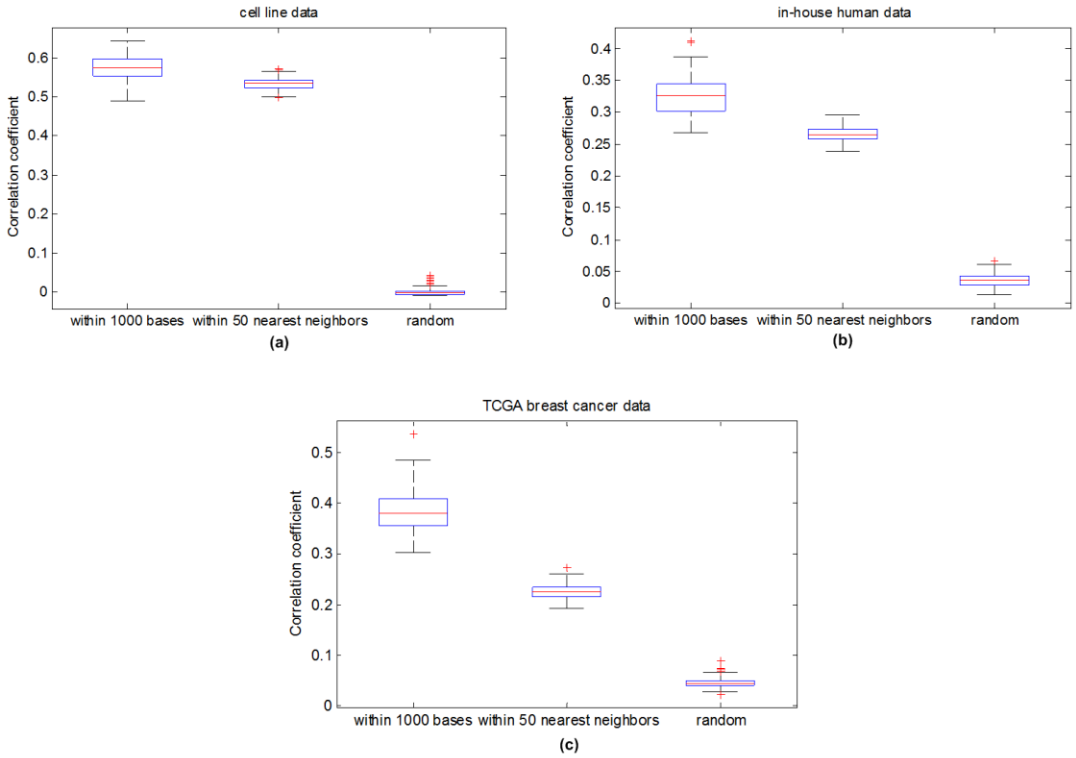


Fig. S1. Correlation of CpG sites calculated from: (a) cell line data; (b) in-house human data; (c) TCGA breast cancer data

S2. Supplemental to Methods

According to Bayes' rule, the joint posterior distribution of the variables and parameters in the DM-BLD model is given by

$$\begin{aligned} &p\left(\boldsymbol{\theta}, \mathbf{d}, \mu_0, \tau_e, \rho, \tau \mid \mathbf{y}\right) \\ &\sim p\left(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{d}, \mu_0, \tau_e, \rho, \tau\right) \times p\left(\boldsymbol{\theta} \mid \rho, \tau\right) \times p\left(\rho\right) \times p\left(\tau\right) \times p\left(\mathbf{d}\right) \times p\left(\tau_e\right) \times p\left(\mu_0\right) \end{aligned} \tag{S1}$$

The conditional posterior distributions of the parameters/variables can be derived as follows.

$$\begin{aligned} &p\left(\theta_i \mid \mathbf{y}_i, \theta_{\partial i}, \mu_0, d_i, \rho, \tau, \tau_e\right) \sim p\left(\mathbf{y}_i \mid \theta_i, \mu_0, d_i, \tau_e\right) \times p\left(\theta_i \mid \theta_{\partial i}, \rho, \tau\right) \\ &\sim \mathrm{N}\left(\frac{\tau\rho\sum_{k=1}^{M_n} w_{k,i}\theta_k + \tau_e\left(\sum_{j=1}^{J_1} y_{i,j} + \sum_{j=1}^{J_2}\left(y_{i,j} - d_i\mu_0\right)\right)}{\tau\left(\rho\sum_{k=1}^{M_n} w_{k,i} + 1 - \rho\right) + J\tau_e}, \frac{1}{\tau(\rho\sum_{k=1}^{M_n} w_{k,i} + 1 - \rho) + J\tau_e} \right) \end{aligned} \tag{S2}$$

$$\begin{aligned} &p\left(\tau \mid \boldsymbol{\theta}, \rho\right) \sim p\left(\boldsymbol{\theta} \mid \tau, \rho\right) p\left(\tau\right) \\ &\sim \mathrm{Gamma}\left(\alpha + \frac{M_n}{2}, \beta + \frac{\sum_{i=1}^{M_n}\left(\left(\rho\sum_{k=1}^{M_n} w_{k,i} + 1 - \rho\right)\left(\theta_i - \frac{\rho\sum_{k=1}^{M_n} w_{k,i}\theta_k}{\rho\sum_{k=1}^{M_n} w_{k,i} + 1 - \rho}\right)^2\right)}{2} \right) \end{aligned} \tag{S3}$$

$$p\left(\rho \mid \boldsymbol{\theta}, \tau\right) \sim p\left(\boldsymbol{\theta} \mid \tau, \rho\right) p\left(\rho\right) \tag{S4}$$

$$p(\tau_e \mid \mathbf{y}, \mu_0, \mathbf{d}, \boldsymbol{\theta}) \sim p(\mathbf{y} \mid \mu_0, \mathbf{d}, \boldsymbol{\theta}, \tau_e)\, p(\tau_e)$$
$$\sim \text{Gamma}\left(\alpha + \frac{J \times M_n}{2}, \beta + \frac{\sum_{i=1}^{M_n}\left(\sum_{j=1}^{J_1}\left(y_{i,j} - \theta_i\right)^2 + \sum_{j=1}^{J_2}\left(y_{i,j} - \theta_i - d_i\mu_0\right)^2 \right)}{2} \right) \quad \text{(S5)}$$

$$p(d_i \mid \mathbf{y}_i, \mu_0, \theta_i, \tau_e, d_{\partial i}) \sim p(\mathbf{y}_i \mid \theta_i, d_i, \mu_0, \tau_e) \times p(d_i \mid d_{\partial i})$$
$$\sim \prod_{j=1}^{J_2} N\left(y_{i,j} \middle| d_i\mu_0 + \theta_i, \frac{1}{\tau_e} \right) \times p(d_i \mid d_{\partial i}) \quad \text{(S6)}$$

$$p(\mu_0 \mid \mathbf{y}, \mathbf{d}, \boldsymbol{\theta}, \tau_e) \sim p(\mathbf{y} \mid \boldsymbol{\theta}, \mu_0, \tau_e, \mathbf{d})\, p(\mu_0)$$
$$\sim \text{N}\left(\frac{\tau_e \sum_{i=1}^{M_n}\sum_{j=1}^{J_2}\left(d_i \times \left(y_{i,j} - \theta_i\right)\right)}{\tau_0 + \tau_e \times J_2 \sum_{i=1}^{M_n} d_i}, \frac{1}{\tau_0 + \tau_e \times J_2 \sum_{i=1}^{M_n} d_i} \right) \quad \text{(S7)}$$

With the derived conditional distributions, we develop a Gibbs sampling method for parameter/variable estimation. Samples for the variables ($\boldsymbol{\theta}$, μ_0 and $\mathbf{d}$) and model parameters (τ_e, t and ρ) are drawn iteratively from their conditional distributions. Specifically, samples for variables $\boldsymbol{\theta}$, μ_0 and parameters τ_e, τ are randomly drawn from the corresponding Gaussian or Gamma distribution. For variable $\mathbf{d}$ and parameter ρ, the conditional posterior distributions do not have closed form. Since $\mathbf{d}$ and ρ are finite discrete values, the corresponding conditional probabilities are calculated first, and then new samples of $\mathbf{d}$ and ρ are randomly selected according to the probabilities. Eventually, the Gibbs sampler produces Markov chains of samples of the parameters/variables, from which the

estimates of the parameters/variables can be obtained from their marginal distributions. The estimate of true methylation level γ is obtained through a joint estimation of all variables and model parameters, accounting for the variability (variance) of the samples and the correlation level among neighboring CpG sites. For each CpG site, the estimated methylation change is $\Delta\hat{\gamma}_i = \hat{\gamma}_i^{(2)} - \hat{\gamma}_i^{(1)}$, which is used to determine the methylation score of a gene.

It is worth noting that Gibbs sampling does not guarantee to find the optimal solution to the parameter estimation problem (especially when running for a finite number of iterations). In order to alleviate the potential problem of being trapped in local optima, we have implemented the Gibbs sampling algorithm with an option of multiple runs in addition to one long run of sampling. Specifically, multiple independent runs of Gibbs sampling, with different initializations of the parameters and different random seeds, are implemented in the algorithm. In our DM-BLD software package, we provide an option for multiple independent runs to be used for Gibbs sampling. When using multiple runs, the distributions generated from the runs are checked in the algorithm. In particular, we conduct a fixed number of runs (e.g., five times) and check whether a specific number of different runs (e.g., three times) generate samples from the same distribution. If so, all of the samples from all runs are used for parameter estimation. If not, another set of fixed number of runs will be conducted continually. With the solutions from multiple runs, comparisons of the solutions can indicate whether a global, optimum solution is likely to have been achieved.

S3. Distribution of real methylation data

Using real methylation data, we validated the normality of methylation beta value after the logit transformation using one-sample Kolmogorov-Smirnov test (K-S test). One-sample K-S test is a nonparametric test to compare a sample with a reference probability distribution, which quantifies a distance between the empirical distribution of the sample and the cumulative distribution function of the reference distribution. The null hypothesis is that the samples are drawn from the reference distribution. With the one-sample K-S test as a goodness of fit test for normality of the distribution, samples were first standardized and then compared with a standard normal distribution.

We applied the one-sample K-S test on methylation data acquired from the TCGA breast cancer project. Two groups of samples (41 samples with survival time longer than 5 years, namely the ‘Dead’ group, and 20 samples with survival time less than 5 years, namely the ‘Alive’ group) were tested, respectively. After removing CpG sites with ‘NaN’ value, 390,301 CpG sites out of 485,577 sites were analyzed. As a result, 360,082 (92.26%) sites in the ‘Dead’ group and 378,034 (96.86%) sites in the ‘Alive’ group followed null hypothesis with p-value>0.05, which indicated that they followed normal distribution.

S4. Simulation study

To generate synthetic data to evaluate the performances of the competing methods on the detection of differentially methylated genes, we began by randomly selecting differentially methylated genes and differentially methylated regions (DMRs), and then generated simulation data for the CpG sites (within DMRs or outside DMRs) in multiple scenarios following two different simulation strategies (based on (1) the simulation study used in DMRcate and (2) our Leroux model). In each experiment, we generated a simulated data set for all 450K probes with 20 samples in two conditions/groups, 10 samples for the control group and 10 samples for the case group.

S4.1. Selection of differentially methylated genes and DMRs

Among 21,231 RefSeq genes covered by the Illumina 450K platform, 20,758 genes contained probes/CpG sites in the promoter region. In each simulated data set, 30% of the 20,758 genes were selected as differentially methylated, with 15% hyper-methylated and 15% hypomethylated in the case group, respectively. The other 70% of the 20,758 genes were assigned as non-differentially methylated. For each differentially methylated gene, a promoter-associated neighborhood was randomly assigned as differentially methylated region, while the CpG sites outside the selected regions were assigned as non-differentially methylated. For each CpG site, its neighbors were defined as the CpG sites located within 1000 bps (of both sides) from it. The randomly selected promoter-associated DMRs contained varying number of CpG sites and could be just part of the promoter region of the genes. With the randomly selected DMRs, the methylated values of the CpG sites (within DMRs or outside DMRs) were simulated in the following two different ways: (1) the simulation scheme used in DMRcate; (2) the proposed Leroux model, as described next.

S4.2. Description of the simulation scheme used in DMRcate

We first generated simulation data sets in which the methylation values of CpG sites were generated following the simulation scheme used in DMRcate. The methylation values of CpG sites were generated from different distributions described as follows:

a. CpG sites within DMRs

For each DMR hyper-methylated in the case group, two beta levels were generated by

$$\beta^{(1)} \sim \text{Uniform}(0.01,\ 0.99 - \Delta\beta)\ ;\ \beta^{(2)} \sim \beta^{(1)} + \Delta\beta\ ;$$

For each DMR hypo-methylated in the case group, two beta levels were generated by

$$\beta^{(1)} \sim \text{Uniform}(0.01 + \Delta\beta,\ 0.99)\ ;\ \beta^{(2)} \sim \beta^{(1)} - \Delta\beta\ .$$

$\beta^{(1)}$ ($\beta^{(2)}$) was set as the base methylation level for the control (case) group/condition, and $\Delta\beta$ was a pre-defined true methylation difference between two conditions.

For each DMR, in each condition, the methylation values of all probes in all samples were randomly generated from a beta distribution with its mode equal to the base methylation level. The variance of the data was controlled by a parameter 'K'.

b. CpG sites outside DMRs

The CpG sites outside DMRs were randomly assigned as unmethylated or fully methylated, i.e., half of them were assigned as unmethylated and half of them were assigned as fully methylated. For all unmethylated probes, the methylation values of all samples were randomly sampled from a beta distribution $beta(\mathrm{a}_0, \mathrm{b}_0)$ with its

mode close to 0; for all fully methylated probes, the methylation values of all samples were randomly sampled from another beta distribution $beta(a_1, b_1)$ with its mode close to 1.

Lastly, the beta values of all probes in all samples were adjusted, following the same procedure in DMRcate, to avoid those values very close to 0 or 1.

S4.3. Simulation data generated by the DMRcate scheme

We designed three scenarios to generate simulation data sets following the strategy used in DMRcate, regarding to different proportions of differentially methylated genes and different variances of the beta distributions predefined for data generation. For each scenario, we performed 10 random trails to assess the variance of the performance.

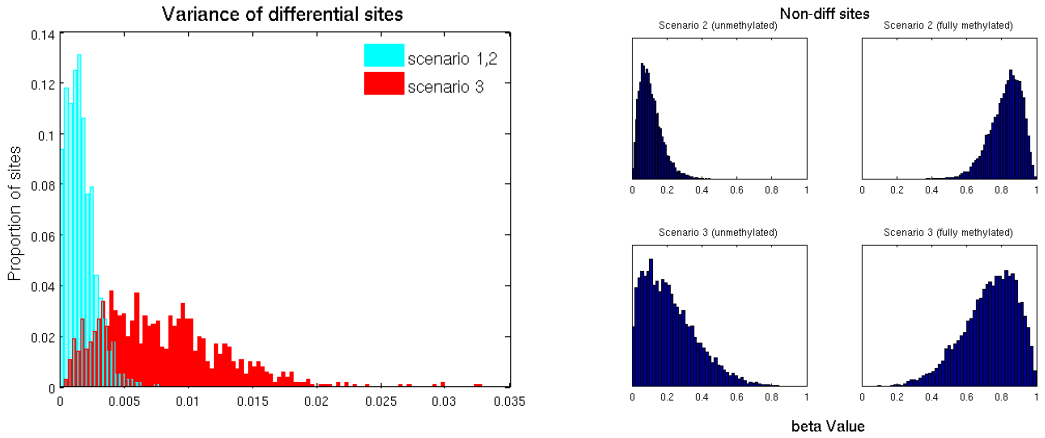
The parameters that affect the beta distributions were set for the three scenarios as listed in Table S1.

Table S1. Parameter settings in three scenarios

	Proportion of DM	$\Delta\beta$	K	a_0	b_0	a_1	b_1
Scenario 1	10%	0.2	100	2.4	20	14	3
Scenario 2	30%	0.2	100	2.4	20	14	3
Scenario 3	30%	0.2	20	1.4	5	5.5	2

The parameter setting in scenario 1 was the same as in the simulation study of DMRcate. In scenario 2, we increased the proportion of differentially methylated genes, and in scenario 3, we increased the variances of the beta distributions to generate simulation data with higher noise, as shown in Fig. S2. Fig. S2(a) shows the variance of true differentially methylated sites among samples in the same

condition. By decreasing K, the variance was higher in scenario 3 than in other two scenarios. Given predefined true difference level $\Delta\beta$, the differential level of the true differentially methylated sites decreased in scenario 3. Fig. S2(b) shows the beta distributions for the non-differentially methylated sites in the two scenarios, which indicates that the variance of non-differentially methylated sites in scenario 3 was higher than scenario 1 and 2. Thus, scenario 3 was more challenging for identifying differentially methylated genes.



(a) (b)

Fig. S2. Simulation data in the three scenarios with at different parameter settings: (a) variance of differentially methylated sites; (b) beta distribution for non-differentially methylated sites.

S4.4. Simulation data generated with the proposed Leroux model

We also generated simulation data sets following the proposed model to mimic the dependency of CpG sites in the neighborhood. In each experiment, the methylation value of CpG sites was simulated by the following steps:

1) Sampling base methylation beta value of all CpG sites

A base methylation beta value of all CpG sites was randomly drawn from the distribution of methylation beta value obtained from a real data set (TCGA breast cancer data set), as shown in Fig. S3.

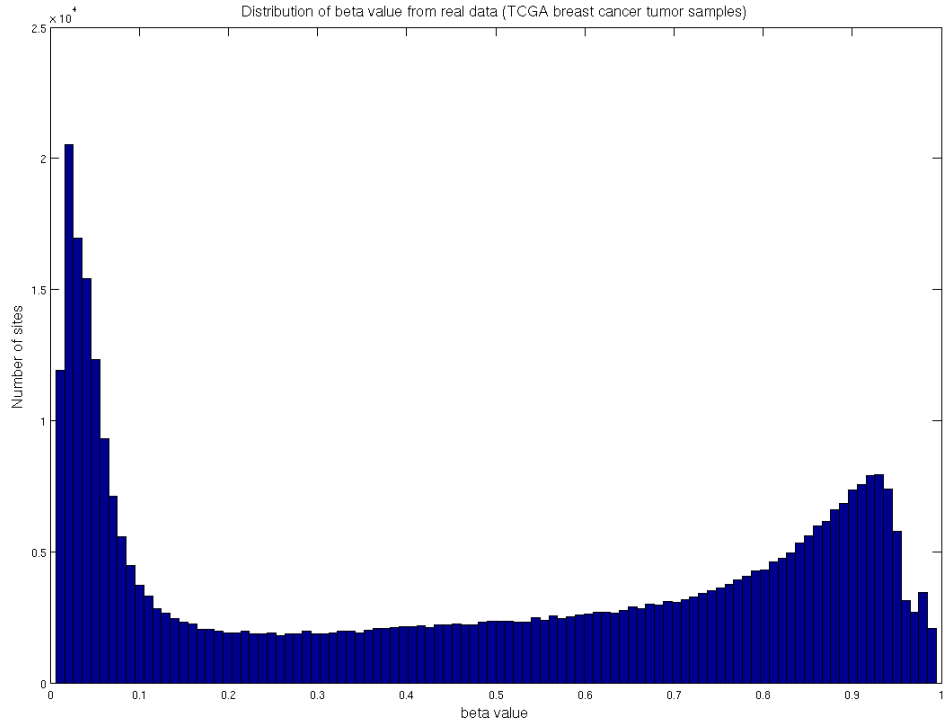


Fig. S3. Distribution of methylation beta value from TCGA breast cancer tumor samples.

2) Sampling true methylation M-value of all CpG sites for each condition

The true methylation M-value (logit transform of beta value) of all CpG sites in the control group was generated using the Leroux model. Specifically, the true methylated M-values were randomly sampled from a multi-variance Gaussian distribution with mean value defined as the logit transform of the base methylation beta value, and variance determined by the neighboring sites. For the non-differentially methylated CpG sites, the true methylation M-value in the case group was the same as in the control group. For CpG sites in the hyper-methylated and hypo-methylated DMRs, the true methylation M-value in the case group were larger or smaller than the control group with difference μ_0 , respectively.

3) Generating methylation M-value for all samples in each condition

In each condition, the methylation M-value for all samples were drawn from normal distribution with mean set as the true methylation M-value, and variance $1/\tau_e$.

Thus, τ_e and μ_0 controlled the differential level of the simulation data set. In the simulation study, we varied τ_e and μ_0 to generate simulation data sets at varying levels of noise and methylation level change. In specific, $\tau_e = 5,\ 2,\ 1$ with $\mu_0 = 0.7$, $\tau = 1$, and $\rho = 0.3$; $\mu = 2,\ 1,\ 0.8$ with $\tau_e = 1, \tau = 1$, and $\rho = 0.3$. Fig. S4 shows the SNR of non-differentially methylated site and differentially methylated sites in the six different scenarios.

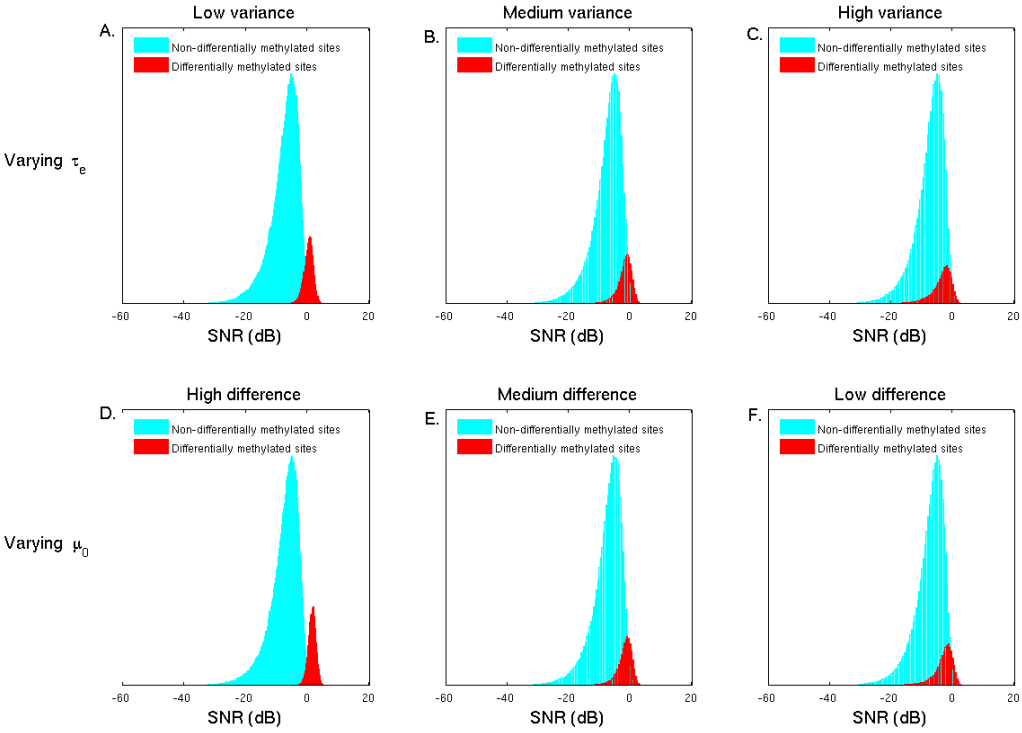


Fig. S4. SNRs of non-differentially methylated and differentially methylated sites at varying noise and methylation change.

S4.5. Implementation of the competing methods

We implemented four existing DMR detection methods, i.e., Bumphunter, DMRcate, Comb-P, and Probe Lasso, and calculated the p-values of the genes from the detected DMRs as follows:

1) Bumphunter: default setting with 100 permutations, where the cutoff was determined from the 100 permutations at default setting. We used the reported p-value of the area as the p-value of the detected DMRs, and assigned a gene's p-value as the minimum p-value of the DMRs associated with (or mapped to) the gene.

2) DMRcate: default settings with pcutoff = 1, lambda = 1000, and C = 2. We used the reported meanpval as the p-value of the detected DMRs, and assigned a gene's p-value as the minimum p-value of the DMRs associated with the gene.

3) Probe Lasso: default setting with adjPval = 1 and DMRpval = 1. We used the reported dmr.Pval as the p-value of the detected DMRs, and assigned a gene's p-value as the minimum P value of the DMRs associated with the gene.

4) Comb-P: p-value from Limma was used as the input, with seed = 0.5. We used the reported z_sidak_p as the p-value of the detected DMRs, and assigned a gene's p-value as the minimum p-value of the overlapped DMRs mapped/linked to the gene.

For the genes with no DMRs detetced, their p-values were set as 1.0.

S4.6. Results on simulation data generated by the DMRcate procedure

In each scenario, we performed 10 random experiments/trials to assess the variance of the performance. Fig. S5 is the boxplot of the AUCpr of the competing methods in 10 random trails for the three scenarios.

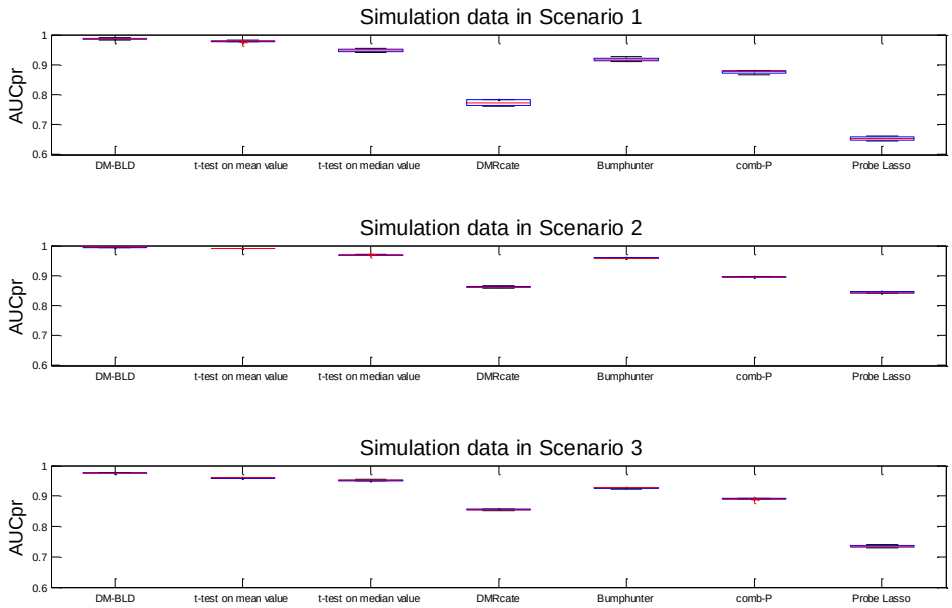


Fig. S5. Performance comparison on simulation data generated following the simulation scheme used in DMRcate.

S4.7 Results on simulation data generated with the Leroux model

Fig. S6 shows the performance on detecting differentially methylated genes at different differential levels between two groups/conditions.

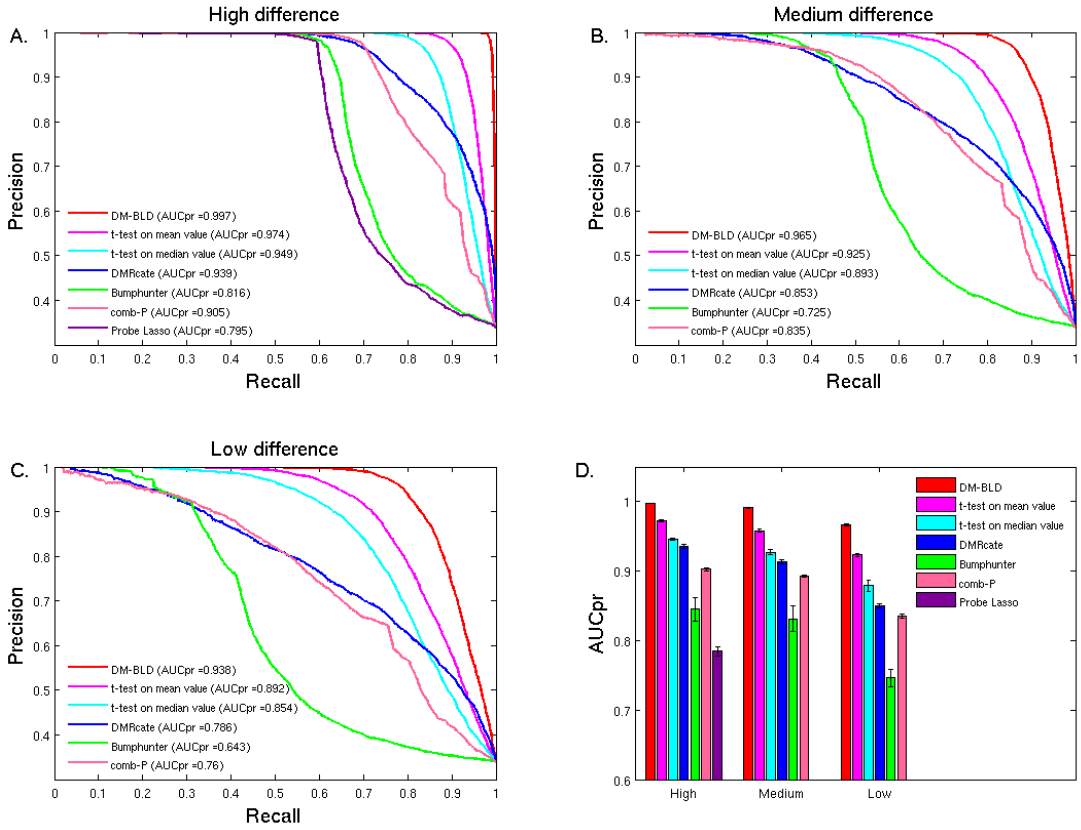
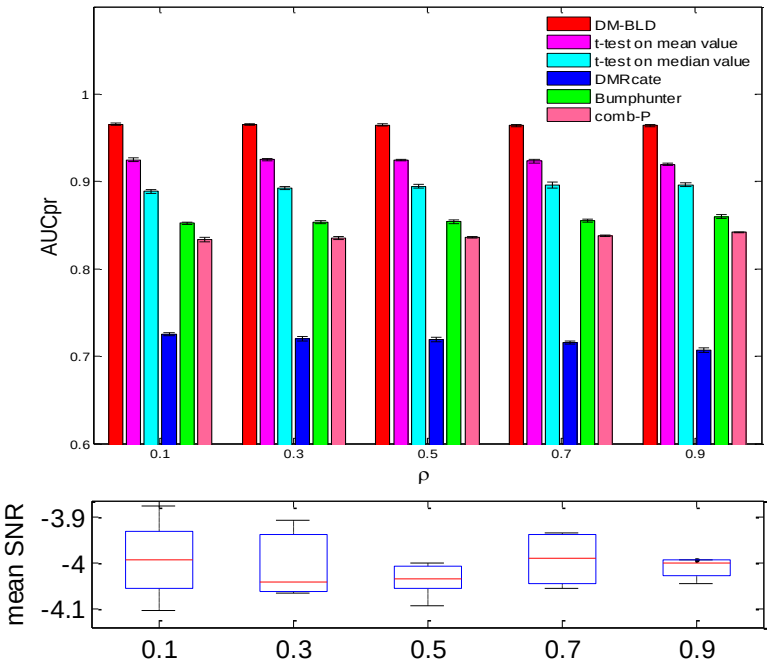


Fig. S6. Performance on the detection of differentially methylated genes at varying levels of difference between two phenotypes. Precision-recall curves at (A) high difference; (B) medium difference; (C) low difference; (d) AUCpr in each scenario with 10 experiments.

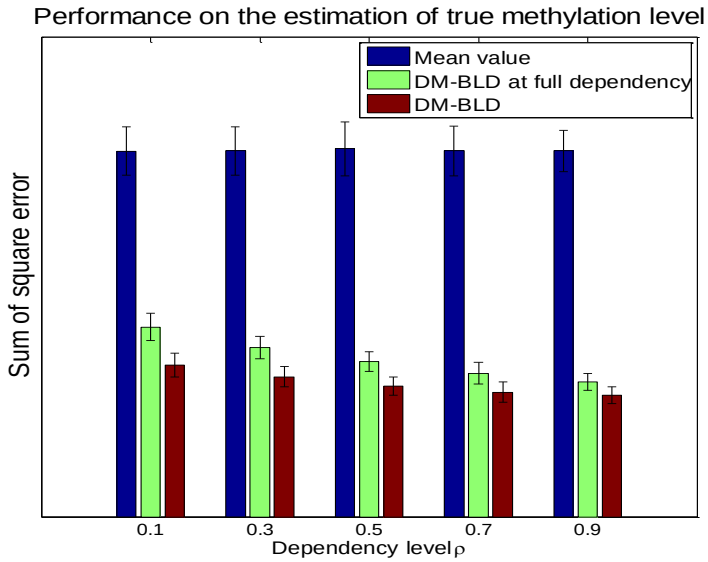
S4.8 Simulation study on varying dependency level

To assess the effectiveness of DM-BLD on various levels of local dependency, we further varied parameter ρ to generate simulation data sets. Specifically, ρ varied from 0.01 to 0.9 with interval 0.2, while $\mu_0 = 1$, $\tau = 1$, $\tau_e = 1$. The higher ρ

is, the higher the local dependency. As shown in Supplementary Fig. S7(a), varying local dependency levels did not directly affect the differential level of CpG sites. However, it impacts on the estimation of the methylation level of CpG sites, as shown in Fig.S7(b). Since the mean value of the samples did not take dependency into account, the performance was similar among all different dependency levels. The performance of DM-BLD increased with increasing dependency levels, since more information can be incorporated from the neighbors. When the dependency level was low, the performance of DM-BLD was much better than that of DM-BLD at full dependency (i.e., where ρ was simply set as 0.999), indicating that the dependency level needed to be correctly estimated. The performance of DM-BLD on the identification of differentially methylated genes consistently outperformed the other methods across different scenarios, as shown in Fig. S7(a).



(a)



(b)

Fig. S7. Performance comparison on varying dependency level ρ. (a) AUCpr for the performance on differentially methylated gene detection; (b) performance on the estimation of true methylation level of the CpG sites.

S5. Experiment on real data for breast cancer recurrence study

S5.1 TCGA ER-positive breast cancer tumor samples

We collected a set of estrogen receptor positive (ER+) breast cancer tumor samples from TCGA to study the molecular mechanism underlying recurrence. The tumor samples were divided into two groups according to the survival time. Specifically, tumor samples from patients who were still alive with a follow-up time longer than 5 years were grouped as 'Alive' (indicating 'late recurrence'), while those from patients dead within 5 years were grouped as 'Dead' (indicating 'early recurrence'). Fig. S8 shows the distribution of the survival time of the patients.

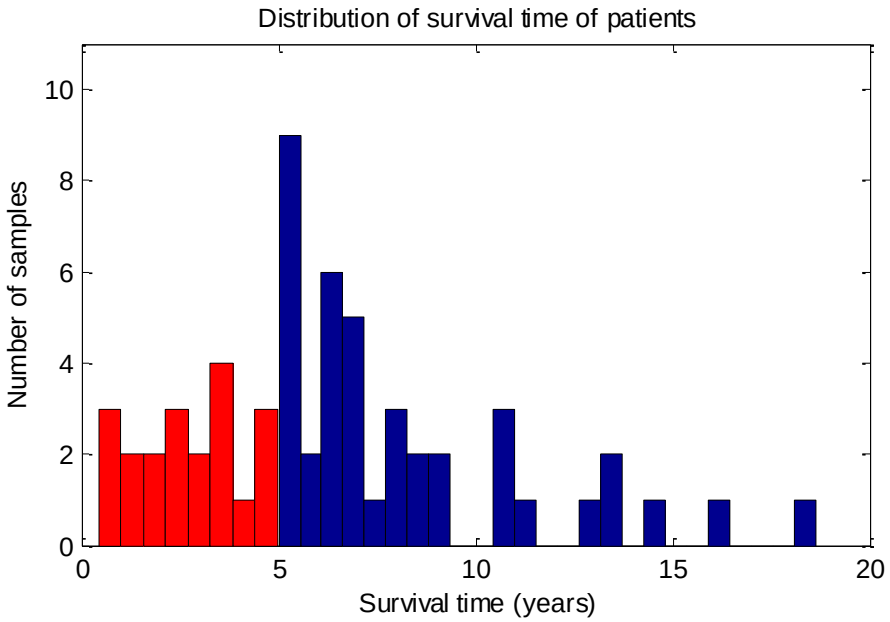


Fig. S8. Histogram of patients' survival time: the 'Dead' group is shown in red; the 'Alive' group is shown in blue

S5.2 Permutation test

To assess the significance of differentially methylated genes, we performed permutation-based statistical tests. Specifically, we randomly permuted both sample labels and CpG site location, and performed DM-BLD over 100 random trials. Such permutation disrupts both the association between samples and phenotypes, and the correlation structure among neighboring CpG sites. We performed two significance tests over the 100 random trails as follows:

- In the first test, the observed (or estimated) differential methylation score of each gene was tested against the ‘global’ null distribution; the ‘global’ null distribution was estimated from the differential methylation scores of all the genes in consideration, as obtained with the 100 random trials. Note that the ‘global’ null distribution was the aggregated distribution calculated from all the genes.

- In the second test, the observed (or estimated) differential methylation score of each gene was tested against its corresponding ‘local’ null distribution; the ‘local’ null distribution was estimated from the differential methylated sores of the gene obtained in the 100 random trails. Note that the ‘local’ null distribution was gene-specific, i.e., each gene had its own null distribution.

In the significance test, the null hypothesis was that the observed methylation score was drawn from the null distribution, and the p-value for each gene was calculated by assuming the null distribution was Gaussian-distributed. Benjamini-Hochberg correction (Benjamini and Hochberg, 1995) was used to estimate the FDR-adjusted p-value. Fig. S9 shows the histogram of the adjusted p-values calculated from the ‘global’ null distribution and the ‘local’ null distribution, respectively.

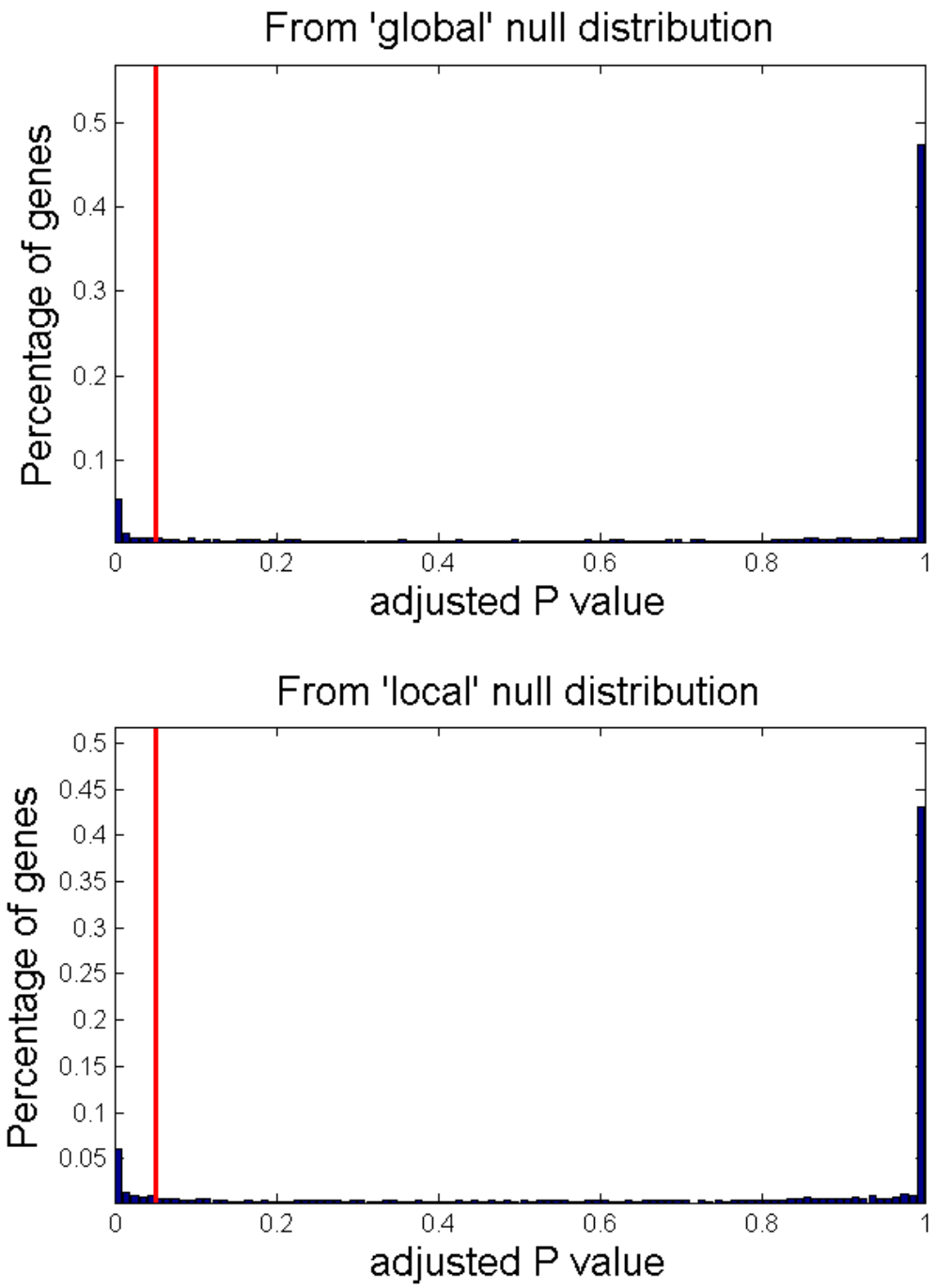


Fig. S9. Histogram of p-values calculated from permutation tests. The red lines presents adjusted p-value = 0.05.

S5.3. Comparison with the competing methods

We also applied Bumphunter (v1.6.0), DMRcate (v1.2.0), Comb-P and Probe Lasso (part of the ChAMP (v1.4.1) package) onto the breast cancer data. We used the same parameter settings as in the simulation studies for the competing methods. Probe Lasso did not report any differentially methylated regions; thus, it was not included in the comparison. Bumphunter reported methylation regions associated with 1,246 genes, where 236 genes were differential with p-value < 0.05. Comb-P reported methylation regions associated with 748 genes, where 721 genes were differential with p-value < 0.05. DMRcate reported 3,347 differentially methylated genes with p-value < 0.05. The Venn diagram of the genes detected by the four methods was shown in Fig. S10. Bumphunter identified much fewer differentially methylated genes as compared with the other methods.

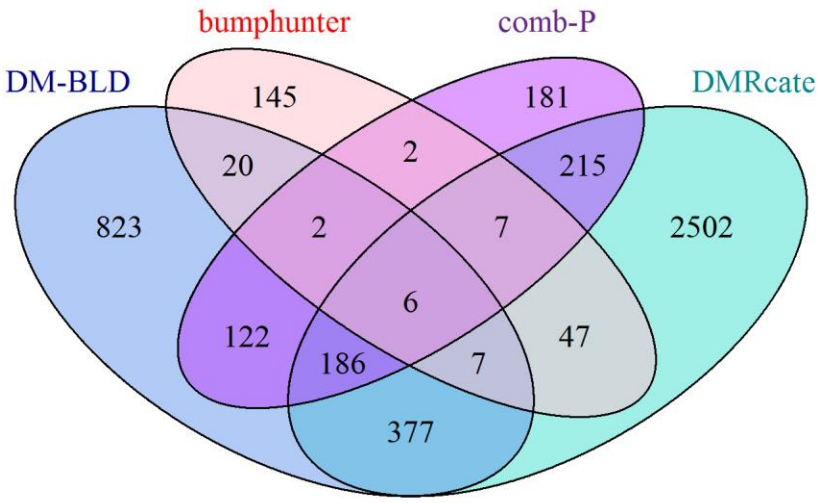
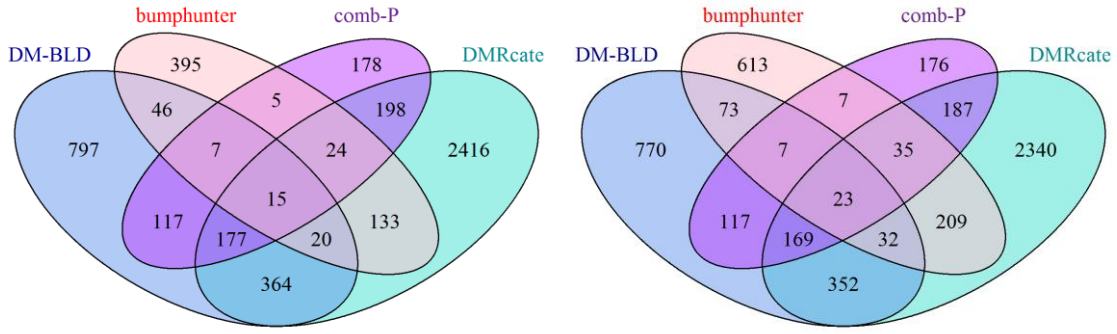


Fig. S10. Venn diagram of the differentially methylated genes detected by the competing methods.

For Bumphunter we used the default setting, where the cutoff was determined from the 100 permutations (as described before in S4.5), for our previous analysis. We adjusted the 'cutoff' of Bumphunter, and reran the analysis. As a result, with p-value < 0.05, 645 of 8616 reported genes were identified as differentially methylated with 'cutoff = 0.01'; 999 out of 9687 reported genes were identified as differentially methylated with 'cutoff = 0'. Fig. S11 shows the Venn diagram with results from different implementation of Bumphunter. Fig. S12 shows the percentage of differentially expressed genes in the top differentially methylated genes detected by Bumphunter at different values of 'cutoff', where the genes were ranked by p-value of the genes summarized from the result of Bumphunter. With different settings of 'cutoff', the number of differentially methylated genes identified by Bumphunter varied. However, in terms of consistency with differentially expressed genes, the top ranked genes had similar performance.



(a) (b)

Fig. S11. Venn diagram of the differentially methylated genes from the competing methods. (a) 'cutoff' = 0.01 used in Bumphunter; (b) 'cutoff' = 0 used in Bumphunter.

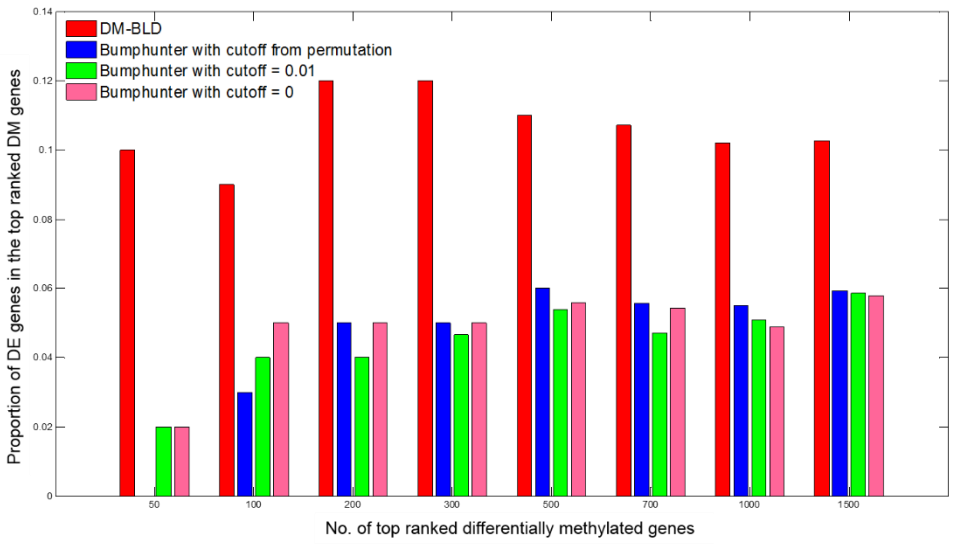


Fig. S12. Proportion of differentially expressed (DE) genes among the top ranked differentially methylated (DM) genes detected by Bumphunter at different 'cutoff' values.

S5.4. Characterization of the common and unique gene sets

We compared the genes identified by our proposed method only to the genes that are also detected by other methods in terms of differential level and number of CpG sites. For Bumphunter, we used the result with default setting, where the 'cutoff' is determined from the permutation test. First, we compared the absolute difference of beta value and the SNR of the CpG sites in the detected DMRs of the two sets of genes, as shown in Fig. S13(a) and (b). From one-tail two-sample K-S test, the absolute difference of beta value and SNR were significantly lower in the unique gene set. We also tested on the number of CpG sites across the whole gene region and the number of CpG sites in DMRs, as shown in Fig. S13(c) and (d). The K-S test showed that the number of sites across the gene region was significantly higher in the common gene set and the number of sites in DMRs was significantly higher in the unique gene set, as shown in Table S2.

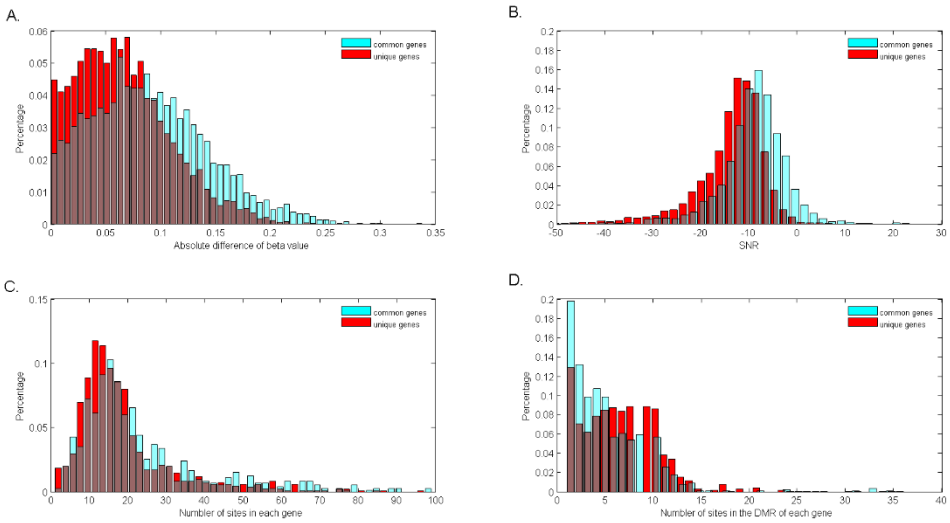


Fig. S13. Common differentially methylated genes v.s. Unique differentially methylated genes detected by DM-BLD. (A) absolute difference of beta value; (B) SNR; (C) number of CpG sites associated with gene; (D) number of CpG sites in DMR.

Table S2. P-value from K-S test

	P-value from K-S test
Absolute difference of beta value (">": larger in common genes than in unique genes)	5.79e-89
SNR (">": larger in common genes than in unique genes)	6.44e-200
Number of sites in each gene (">": larger in common genes than in unique genes)	1.77e-7
Number of sites in DMRs ("<": less in common genes than in unique genes)	2.36e-15

S5.5. Differentially expressed genes detected from RNA-seq data

We analyzed the mRNA expression data of the same set of patients and detected differentially expressed genes. We downloaded the RNA-seq data (Level 1) of all of the 61 samples profiled by Illumina HiSeq 2000 RNA Sequencing Version 2 analysis from the TCGA data portal, and then performed alignment using 'TopHat 2 (TopHat v2.0.12)' (http://ccb.jhu.edu/software/tophat/index.shtml) with UCSC hg19 as the reference sequence. With the isoform structure annotation file (RefSeq genes) downloaded from the UCSC genome browser database (http://genome.ucsc.edu/), we applied the cuffdiff 2 method (cuffdiff 2.2.1; http://cole-trapnell-lab.github.io/cufflinks/) to identify differentially expressed isoforms by analyzing samples from the two groups: the 'Dead' group vs. the 'Alive' group. Differentially expressed genes were defined as genes with differentially expressed isoforms with p-value less than 0.05. As a result, 1101 differentially expressed genes were identified.

S5.6. Interaction of the identified functional genes in PPI network

To study the interaction of the identified genes, we first mapped the differentially expressed genes to the Protein-Protein interaction (PPI) network from the Human Protein Reference Database (HPRD) (Keshava Prasad, et al., 2009). Fig. S14(A) shows that the major connected network is largely downregulated in the 'Dead' group as compared to that in the 'Alive' group. In the PPI network of differentially expressed genes, there are two modules of interacting genes with differential methylation activity between two groups. The two modules, potentially regulated by DNA methylation, are shown in Fig. S14(B) and highlighted in yellow and blue in Fig. S14(A).

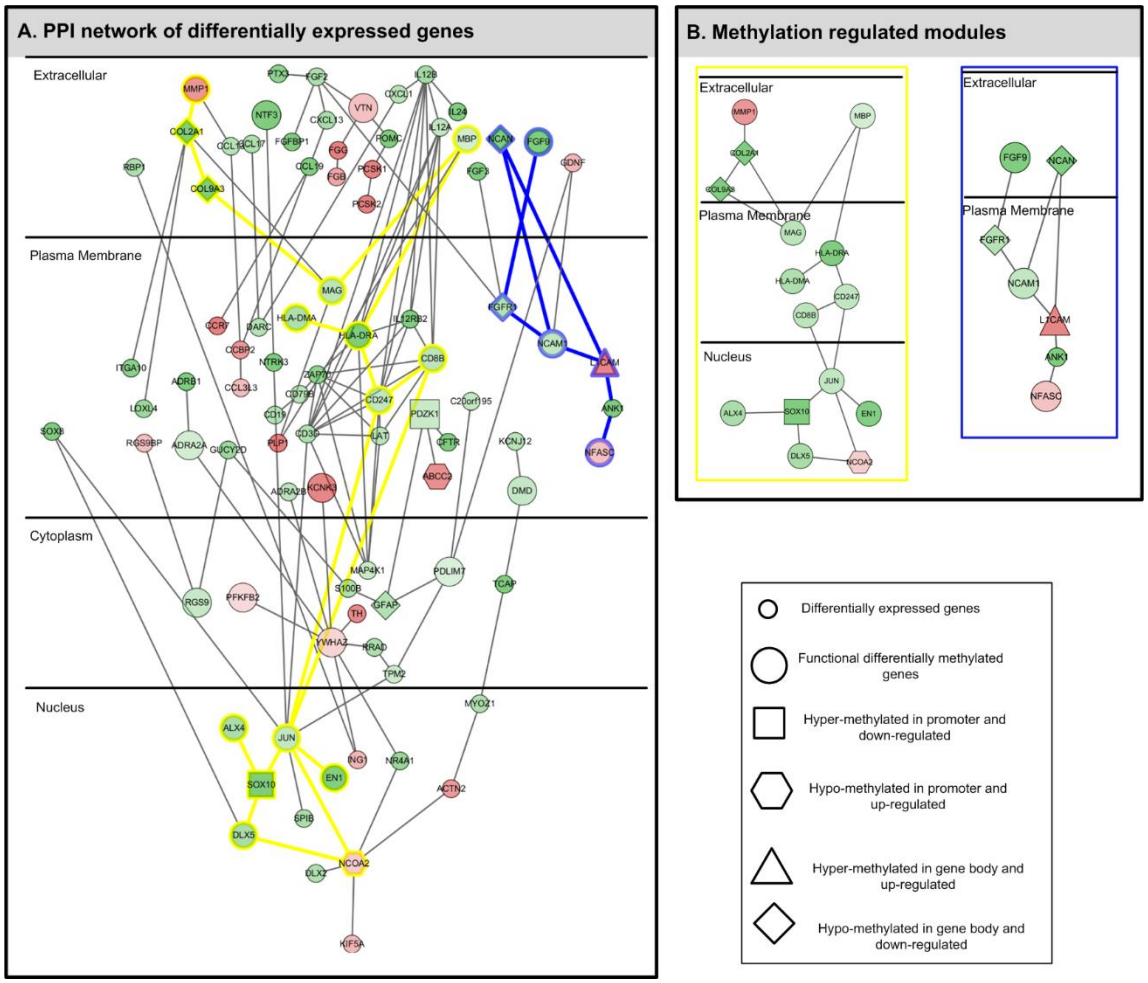


Fig. S14. Network of differentially expressed genes and methylation regulated modules. (a) A PPI network of differentially expressed genes; (b) methylation regulated modules with interacting genes that are differentially expressed and also differentially methylated.

S6. Polycomb target genes detected from ChIP-seq data

As polycomb group (PcG) proteins are essential epigenetic regulators, we identified Polycomb target genes from ChIP-seq data, and then checked the overlap with the identified hyper-methylated genes. Specifically, we first download the ChIP-seq data of EZH2, SUZ12, H3K4me3 and H3K27me3 in embryonic stem cells from ENCODE (https://www.encodeproject.org/). Then, we used MACS (Zhang, et al., 2008) with default setting for peak calling. Finally, we matched the peaks to genes using GREAT (McLean, et al., 2010) with the regulatory region defined as $\pm 2K$ from the transcriptions start sites (TSS). The gene sets identified from the four ChIP-Seq studies were shown in Fig. S15. As a result, 2,589 common genes from the four studies were detected as the Polycomb target genes.

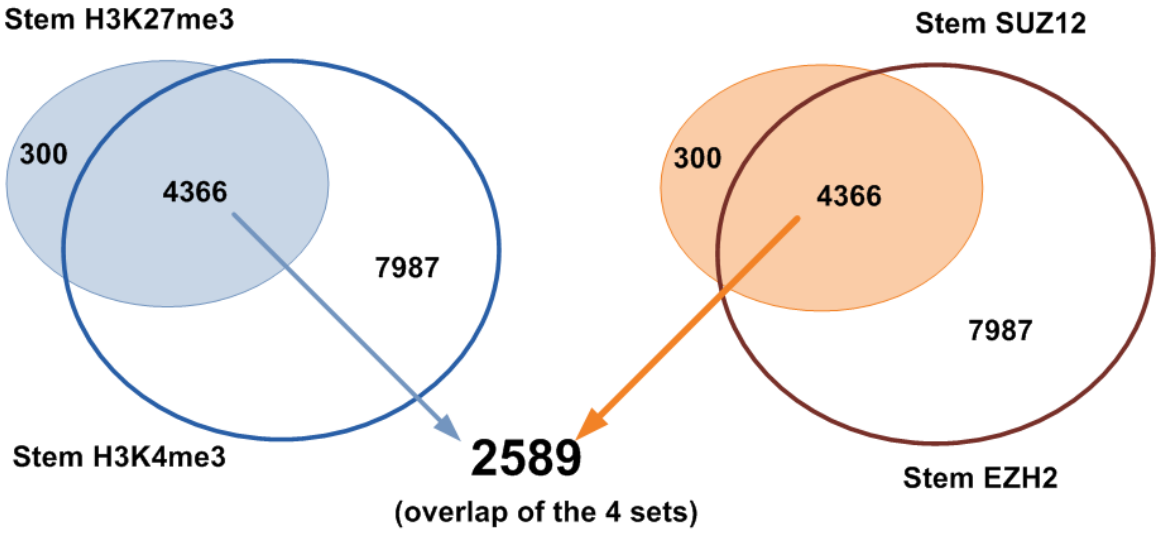


Fig. S15. Number of genes identified from four ChIP-seq studies on stem cell

Reference

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300.

Keshava Prasad, T.S.*, et al.* (2009) Human Protein Reference Database--2009 update, *Nucleic acids research*, **37**, D767-772.

McLean, C.Y.*, et al.* (2010) GREAT improves functional interpretation of cis-regulatory regions, *Nature biotechnology*, **28**, 495-501.

Zhang, Y.*, et al.* (2008) Model-based analysis of ChIP-Seq (MACS), *Genome biology*, **9**, R137.