# methylDMV: SIMULTANEOUS DETECTION OF DIFFERENTIAL DNA METHYLATION AND VARIABILITY WITH CONFOUNDER ADJUSTMENT

PEI FEN KUAN*, JUNYAN SONG and SHUYAO HE

*Department of Applied Mathematics and Statistics,*
*Stony Brook University,*
*Stony Brook, NY 11794, USA*
**E-mail: peifen.kuan@stonybrook.edu*
*http://www.stonybrook.edu/commcms/ams2/*

DNA methylation has emerged as promising epigenetic markers for disease diagnosis. Both the differential mean (DM) and differential variability (DV) in methylation have been shown to contribute to transcriptional aberration and disease pathogenesis. The presence of confounding factors in large scale EWAS may affect the methylation values and hamper accurate marker discovery. In this paper, we propose a flexible framework called `methylDMV` which allows for confounding factors adjustment and enables simultaneous characterization and identification of CpGs exhibiting DM only, DV only and both DM and DV. The proposed framework also allows for prioritization and selection of candidate features to be included in the prediction algorithm. We illustrate the utility of `methylDMV` in several TCGA datasets. An R package `methylDMV` implementing our proposed method is available at `http://www.ams.sunysb.edu/~pfkuan/softwares.html#methylDMV`.

*Keywords*: DNA methylation; Differential variability; Feature selection; Elastic net.

## 1. Introduction

DNA methylation is an important hallmark of genomic imprinting, transcriptional regulation, X-inactivation and chromosomal stability.[1] The most common DNA methylation process in human involves the addition of a methyl group to the 5-carbon of the cytosine ring. In human, this modification mostly occurs at a CpG site in which a cytosine nucleotide is followed by a guanine nucleotide. Aberrant patterns of DNA methylation have been shown to be a critical mechanism in the development and progression of various diseases, in particular cancer.[2] DNA methylation is one of the most widely studied epigenetics event and has been profiled extensively in large consortiums including the Cancer Genome Altas (TCGA), NIH Roadmap and the Encyclopedia of DNA Elements (ENCODE) projects. These efforts provide research opportunities for secondary analyses of the large datasets to further understand the biology of the disease.

Most of the work in DNA methylation have been focused on identifying DNA methylation markers that exhibit differential average or mean methylation (DM).[3,4] These epigenetic markers have been shown to be promising biomarkers in designing platform for disease diagnosis.[5] Over the last few years, there has been an increasing interest in identifying DNA methylation markers that exhibit differential variability in various diseases, including cancer[6–8] and obesity.[9] These epigenetic variabilities can be attributed to increased plasticity arising from changing environment including varying oxygen tension[10] and is associated with the risk of morphological and neoplastic transformation.[11] These studies opened up new avenues to the

study of DNA methylation, which indicated that simultanenous investigation of both differential mean and variability may delineate the complex patterns of epigenetic regulation in pathophysiology and development of diseases.

One of the most widely used DNA methylation platforms is the Illumina Infinium HumanMethylation450 BeadChip which profiles more than 450,000 CpGs genome wide. The latest phase of the Illumina methylation array is the MethylationEPIC BeadChip which covers approximately 850,000 methylation sites including CpG islands, enhancers and regulatory regions identified from the ENCODE project. The methylation value for each CpG is represented as a *beta* ($\beta$) value, which is the ratio of methylated probe intensities to the total probe intensities, where $0 \leq \beta \leq 1$; $\beta = 0$ and $\beta = 1$ indicate that the CpG is fully unmethylated and methylated, respectively.

An important aspect of differential methylation analysis is to identify CpGs which exhibit differential mean or variance in large scale hypothesis testing. Statistical tests for detecting CpGs which exhibit differential mean methylation include t-tests, non-parametric Wilcoxon rank sum test or limma[12] based on linear models and empirical Bayes approach. On the other hand, several algorithms have been proposed in recent years to identify CpGs which exhibit differential variability in large scale hypothesis testing. For instance, Teschendorff et al. (2012)[8] proposed a regularized version of the Bartlett's test, Ahn et al. (2013)[13] used a score test from generalized regression model, Phipson et al. (2014)[14] proposed a modification of Levene's test, Wahl et al. (2014)[15] introduced a generalized additive models for location, scale and shape (GAMLSS) framework and Kuan (2014)[16] proposed a general linear model with propensity score method for detecting CpGs with differential variability.

CpGs which exhibit differential mean methylation have been utilized in classification algorithm to define methylation signatures for disease subtypes.[17,18] As the methylation arrays encompass $> 450,0000$ CpGs, a common approach in training the classification algorithm is to pre-select features ranked highly by the univariate differential mean methylation as candidate CpGs in the classification algorithm to improve the stability of the algorithm. Motivated by the biological insights of differential variability in methylation, Teschendorff et al. (2012)[8] proposed a method which selected differential variable CpGs using Bartlett's test for inclusion in the prediction algorithm.

Large scale differential methylation analysis requires proper adjustment for confounders to reduce the biases associated with the identified methylation markers. For instance, age[19,20] and cigarette smoking[21,22] have been shown to be associated with DNA methylation; thus in studies to identify methylation markers for cancer or other disease phenotypes, appropriate adjustment for these factors is necessary. In the analysis of differential mean methylation, this can be achieved via a regression framework where confounders are included as covariates in the model. However, in the analysis of differential variability, potential biases due to confounding variables are usually ignored.[8,14]

This paper aims to develop a unified framework to address the limitation of existing work: (1) incorporates adjustment for confounding variables that potentially affect methylation levels, and allows for simultaneous detection of differential mean (DM) and differential variability (DV) in methylation analysis, (2) systematic selection of CpGs which exhibit differential mean

and/or differential variability in the prediction algorithm to improve prediction accuracy and biological interpretation. In Section 2, we describe our proposed approach. This is followed by simulation studies and real data applications in Sections 3 and 4, respectively. The paper concludes with a discussion in Section 5.

## 2. Methods

### 2.1. *A framework for simultaneous detection of differential mean (DM) and differential variability (DV)*

Without loss of generality, we describe our proposed framework for detecting differential mean and differential variability between two conditions or groups (e.g., tumor versus normal). A common distribution to model the *beta* values from Illumina methylation arrays is the beta distribution.[23] Since the variance of a beta distribution is a function of the mean, the $\beta$ values exhibit significant heteroscedasticity.[24] To overcome the heteroscedasticity issue, we consider a variance stabilizing transformation via the logit function to the $\beta$ values, i.e., $logit(\beta) = \log[\beta/(1-\beta)]$. Let $x_{ij}$ denote the logit transformed methylation value for sample $i$ and CpG $j$. We first define a deviation measure $r_{ij} = |x_{ij} - \text{wt.med}_i(x_{ij})|$ where $\text{wt.med}_i(x_{ij})$ is the weighted median of CpG $j$ with weights $w_i = 1/2n_{g_i}$, $g_i = 0$ if sample $i$ is a control and $g_i = 1$ if sample $i$ is a case, and $n_0$ and $n_1$ are the respective sample sizes.

We recast the model for simultaneous detection of differential mean and differential variable CpGs using a logistic regression model. Let $y_i$ denote the group membership of sample $i$, where $y_i = 0$ if the sample is a control/normal and $y_i = 1$ if the sample is a case/tumor. $y_i$ is assumed to follow a binomial distribution with $P(y_i = 1) = \pi_i$ and $\log[\pi_i/(1-\pi_i)] = \theta_i$. We consider the four competing models for each CpG:

Model 1: $\theta_i = \beta_0 + \sum_{k=1}^{K} \gamma_k Z_{ik}$ (no DM or DV)
Model 2: $\theta_i = \beta_0 + \beta_m x_{ij} + \sum_{k=1}^{K} \gamma_k Z_{ik}$ (DM only)
Model 3: $\theta_i = \beta_0 + \beta_v r_{ij} + \sum_{k=1}^{K} \gamma_k Z_{ik}$ (DV only)
Model 4: $\theta_i = \beta_0 + \beta_m x_{ij} + \beta_v r_{ij} + \sum_{k=1}^{K} \gamma_k Z_{ik}$ (both DM and DV)

In all models, $\mathbf{Z}_k = (Z_{ik})'$ corresponds to confounding variable $k$, for instance age, smoking status or alcohol consumption. Model 1 is the baseline model which adjusts for confounding variables and assumes that the phenotype is not associated with differential mean (DM) or differential variability (DV). Model 2 (Model 3) assumes that the phenotype is associated with DM (DV) after adjusting for confounders, whereas Model 4 assumes that the phenotype is associated with both DM and DV for a CpG. To identify CpGs which exhibit DM, one can compare Model 1 to Model 2 using likelihood ratio tests or score tests.[25] On the other hand, Model 3 can be compared to Model 1 to obtain p-values associated with DV for each CpG. The comparison of Model 4 and Model 1 identifies CpGs which exhibit either DM or DV. The vector of p-values from each analysis are adjusted via the false discovery rate (FDR)[26] to account for multiple testings. In addition to large scale hypothesis testing framework to identify DM and DV CpGs, another advantage of our proposed model is that it allows for automatic classification of the CpGs into the four classes (1) no DM or DV, (2) DM only, (3) DV only and (4) both DM and DV. This is carried out via a Bayesian Information Criterion

(BIC) to rank the four models for each CpG, i.e., the CpG is categorized into the class with the smallest BIC score.

### 2.2. *Candidate feature selection for prediction modeling*

The BIC used for model ranking within each CpG can also be utilized to aid candidate feature selection to improve the stability of the prediction algorithm. The proposed framework provides flexibility to the user for including top ranking features in constructing prediction model. For instance, if the user is interested in a prediction model using CpGs which exhibit the largest discriminative power in terms of both DV and DM after adjustment for confounding variables, then the subset of CpGs which show the lowest BIC scores for Model 4 are selected as candidate features. On the other hand, if the user is interested in a prediction model using only DM CpGs , then the candidate features correspond to the CpGs which identify Model 2 as the best model using BIC scores.

The selected candidate features are used in the prediction algorithm for constructing classification rule discriminating case from control. In this paper, we consider the elastic net algorithm.[27] The objective function of elastic net consists of a loss function + penalty:

$$\min_{\beta} ||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda \left\{\alpha||\beta||_1 + (1-\alpha)||\beta||^2\right\}$$

where $||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$ and $||\beta||^2 = \sum_{j=1}^{p} \beta_j^2$. The parameters $\lambda$ and $\alpha$ are tuned via cross-validation. Other types of machine learning prediction algorithm can also be used on the selected candidate features, for instance the random forest[28] which is a non-parametric ensemble approach based on a large number of classification trees trained on bootstrap samples.

An R package `methylDMV` implementing our proposed method for testing DM and DV, as well as CpGs ranking by BIC and candidate feature selection is available at `http://www.ams.sunysb.edu/~pfkuan/softwares.html#methylDMV`.

## 3. Simulation studies

We carried out simulation studies to evaluate the effect of confounders on CpG ranking. Specifically, denote $Z_{i1}$ and $Z_{i2}$ as the two confounders, where $Z_{i1} \sim N(0,1)$ and $Z_{i2} \sim$ Bernoulli(0.6) for sample $i$, $i = 1, 2, \ldots, n$. The group indicator $y_i$ was generated from the following model

$$\text{logit}(p_i) = \gamma_0 + \gamma_1 Z_{1i} + \gamma_2 Z_{2i}$$
$$y_i \sim \text{Bernoulli}(p_i)$$

For each CpG $j$ ($j = 1, 2, \ldots, p$), the measurements $x_{ij}$'s were generated from the Gaussian distribution under the assumption that the beta values have been properly transformed (e.g., logit or arcsine transformation), i.e., $x_{ij} \sim N(\mu_{ij}, \sigma_{ij}^2)$ where

(i) $\mu_{ij} = \mu_0 + \alpha_1 Z_{i1} + \alpha_2 Z_{i2}$ and $\sigma_{ij}^2 = \sigma_0^2$ if CpG $j$ is from Model 1 (no DM or DV)
(ii) $\mu_{ij} = \mu_0 + \alpha_g y_i + \alpha_1 Z_{i1} + \alpha_2 Z_{i2}$ and $\sigma_{ij}^2 = \sigma_0^2$ if CpG $j$ is from Model 2 (DM only)
(iii) $\mu_{ij} = \mu_0 + \alpha_1 Z_{i1} + \alpha_2 Z_{i2}$ and $\sigma_{ij}^2 = \sigma_0^2 + \beta_g y_i$ if CpG $j$ is from Model 3 (DV only)
(iv) $\mu_{ij} = \mu_0 + \alpha_g y_i + \alpha_1 Z_{i1} + \alpha_2 Z_{i2}$ and $\sigma_{ij}^2 = \sigma_0^2 + \beta_g y_i$ if CpG $j$ is from Model 4 (both DM and DV)

The proportion of CpGs from Models 1-4 were drawn from a multinomial distribution with $\pi = \left(\pi_1, \frac{1-\pi_1}{3}, \frac{1-\pi_1}{3}, \frac{1-\pi_1}{3}\right)$. We set $\gamma_0 = 1, \gamma_1 = 2, \gamma_2 = -2$ to obtain approximately equal number of cases and controls; and $\alpha_g = 1, \beta_g = 1, \mu_0 = 0, \sigma_0^2 = 1$. We varied $\alpha_1 = \alpha_2 = 0, 0.5, 1, 3, 5$ to reflect the different degrees of confounding in the methylation measurements and $\pi_1 = 0.4, 0.6, 0.8$ for the different mixing proportions of DM and DV CpGs. To evaluate the effect of confounders on the phenotype, i.e., case/control, we also considered the case in which the $y_i$'s were not affected by confounders. Under this scenario, $y_i = 0$ for $i = 1, 2, \ldots, n/2$ and $y_i = 1$ for $i = n/2 + 1, \ldots, n$. For each scenario, the simulation was conducted for $n = 200$ samples and $p = 10000$ CpGs over 100 iterations.

We compared the average accuracy of the BIC ranking procedure in classifying the CpGs into Models 1-4 with (BICadj) and without (BICnoadj) adjustment for confounders. We also included comparison to method which performed tests for DM and DV separately. Two sample t-test and Levene's test were used to identify DM and DV CpGs, respectively. CpG $j$ was classified as DM (DV) if the p-value from t-test (Levene's test) adjusted via the Benjamini-Hochberg procedure[26] $\leq$ FDR. We considered FDR 0.05 and 0.1, and referred to this method as SepTest0.05 and SepTest0.1, respectively.

Figure 1 summarizes the average accuracy for the four methods across the different settings. In scenarios where both the phenotype (case/control status) and methylation measurements were affected by confounders (top row of Figure 1 for $\alpha_1 \neq 0$), the methods which did not adjust for confounders exhibited poor accuracy across different mixing proportions $\pi_1$. For the case where $\alpha_1 = 0$, i.e., methylation measurements were not affected by confounders, the BICadj method showed a slight decrease in accuracy compared to other methods. Bottom row of Figure 1 displays the results for the scenarios where only the methylation measurements were confounded while the phenotype was not affected by confounders. For these cases, the performance of the methods were comparable for $\alpha_1 \leq 1$. The advantages of adjusting for confounders were apparent for $\alpha_1 = 3, 5$, i.e., strong confounding effect in the methylation measurements even in the absence of confounding in case/control status.

## 4. Case studies

### 4.1. *Data preprocessing and normalization*

We illustrated our proposed method, `methylDMV` on three datasets, namely the breast cancer (BRCA), kidney cancer (KIRC) and liver cancer (LIHC) dataset. The breast cancer dataset consisted of 909 samples downloaded from the TCGA data portal and the NCBI gene expression omnibus under accession number GSE67919, whereas the kidney and liver cancer consisted of 475 and 404 samples from the TCGA data portal, respectively. All the samples were profiled using the Illumina Infinium HumanMethylation450 BeadChip.

Preprocessing of the methylation data at the 485,557 CpGs were performed as follows. Probes with detection p-value $> 0.05$ were set to missing and probes with more than 20% missing were filtered. A beta mixture quantile (BMIQ) normalization[29] was applied to the beta values for correction of bias due to the type I and type II probes. Non-specific, cross-hybridized probes,[30,31] probes overlapping with a SNP and probes mapping to repeat regions were filtered. For KIRC and LIHC, we further filtered for CpGs mapping to chromosomes X

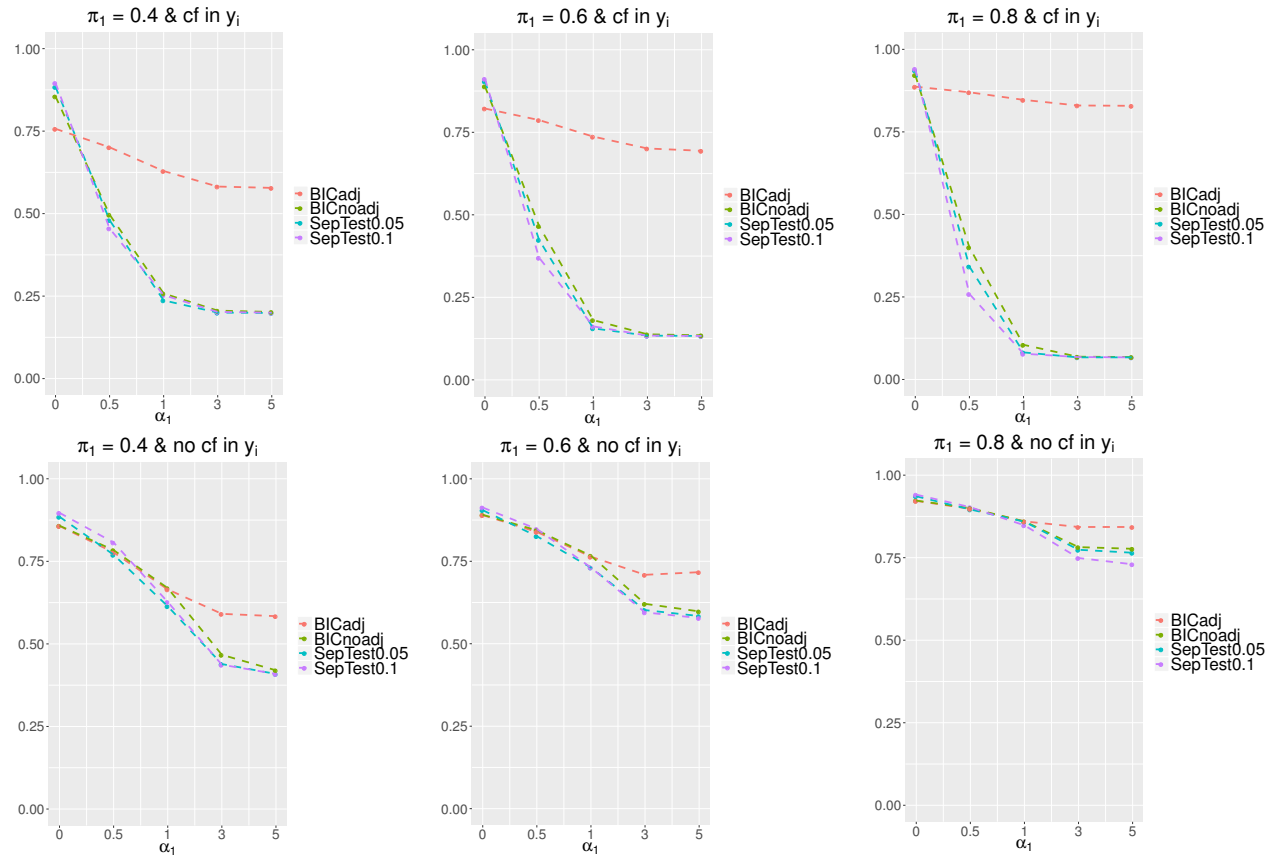


Fig. 1. Average accuracy of CpG classification across $\alpha_1$'s for our proposed BIC ranking with confounding adjustment (BICadj, orange), BIC ranking without confounding adjustment (BICnoadj, green), separate two-sample t-test and Levene's test for DM and DV at FDR 0.05 (SepTest0.05, turquoise) and 0.1 (SepTest0.1, purple). Each panel corresponds to a specific $\pi_1$ value and whether the case control status was affected by confounders (top row: $y_i \sim$ Bernoulli($p_i$), i.e, affected by confounders; bottom row $y_i = 0, i = 1, \ldots n/2$, $y_i = 1, i = n/2 + 1, \ldots n$, i.e, not affected by confounders).

and Y. The normalized datasets consisted of 374,680, 365,896 and 365,658 CpGs for BRCA, KIRC and LIHC, respectively. We performed the following pairwise comparisons:

(i) **KIRC (tumor vs normal)**: Models 1-4 were fitted on $n_0 = 156$ normal (control) and $n_1 = 319$ tumor (case), adjusting for age and race.
(ii) **LIHC (tumor vs normal)**: Models 1-4 were fitted on $n_0 = 47$ normal (control) and $n_1 = 357$ tumor (case), adjusting for age and race.
(iii) **BRCA (tumor vs normal)**: Models 1-4 were fitted on $n_0 = 180$ normal (control) and $n_1 = 729$ tumor (case), adjusting for age and race.
(iv) **BRCA (basal vs luminal A)**: Models 1-4 were fitted on $n_0 = 93$ luminal A (control) and $n_1 = 30$ basal (case), adjusting for age and race.
(v) **BRCA (basal vs luminal B)**: Models 1-4 were fitted on $n_0 = 40$ luminal B (control) and $n_1 = 30$ basal (case), adjusting for age and race.
(vi) **BRCA (luminal B vs luminal A)**: Models 1-4 were fitted on $n_0 = 93$ luminal A (control) and $n_1 = 40$ luminal B (case), adjusting for age and race.

### 4.2. *Feature ranking by BIC scores*

In tumor versus normal comparison within KIRC, LIHC and BRCA datasets, majority of the CpGs were showing either DM or DV or both as shown in Table 1. A large number of CpGs ranked Model 4 (DM and DV) as the best model which indicated that both differential mean and differential variability play important role in distinguishing tumor from normal. In KIRC and BRCA, CpGs showing DM only (Model 2) were enriched in CpG islands, first exons, 200 bp upstream of the transcription start sites (TSS200); whereas CpGs showing DV only (Model 3) were enriched in CpG shores and gene body as shown in Figures 2 and 3. In LIHC, the proportions of DM and DV CpGs mapping to CpG islands were fairly similar, whereas the proportion of DM CpGs mapping to gene body was higher compared to DV CpGs. On the other hand, the subtypes comparison within BRCA identified fewer number of CpGs exhibiting DM or DV. In basal versus luminal A or luminal B comparisons, the proportions of DV CpGs mapping to CpG island and TSS200 were higher than DM CpGs.

Among the lists of DM only CpGs (Model 2) identified by tumor versus normal comparison within KIRC, LIHC and BRCA datasets, 4814 CpGs were in common. On the other hand, there were 1223 and 46885 common CpGs in DV only (Model 3) and both DV and DM (DM&DV) (Model 4) categories, respectively. DAVID (`https://david-d.ncifcrf.gov/home.jsp`) functional annotation enrichment analysis was performed on the genes of mapping to each of the top 1000 common DM only CpGs, DV only CpGs and DM&DV CpGs to identify enriched canonical pathways and biological process ontologies. At FDR $\leq 0.05$, enriched canonical pathways for DM only CpGs include Rho GTPase cycle, Rap1 signaling pathway and NRAGE signals death through JNK; whereas DM&DV CpGs identified olfactory transduction and signaling pathway among the top enriched pathways. On the other hand, DM only CpGs, DV only CpGs and DM&DV CpGs identified processes related to GTPase regulation, regulation of transcription from RNA polymerase II promoter and regulation of ion transmembrane transport, respectively.

Table 1. Number of CpGs identified for each model based on BIC scores for the different datasets and comparisons.

| Data | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|
| KIRC: tumor vs normal | 18685 | 94948 | 44291 | 207972 |
| LIHC: tumor vs normal | 85769 | 52315 | 83296 | 144278 |
| BRCA: tumor vs normal | 33735 | 104575 | 43880 | 192490 |
| BRCA: basal vs luminal A | 201378 | 131085 | 23193 | 19024 |
| BRCA: basal vs luminal B | 198192 | 124764 | 31393 | 20331 |
| BRCA: luminal B vs luminal A | 290963 | 47145 | 31327 | 5245 |

### 4.3. *Elastic net predictive modeling*

The elastic net algorithm[27] was applied to each dataset for constructing a prediction model differentiating case from control. We randomly split the dataset into 80% training and 20%

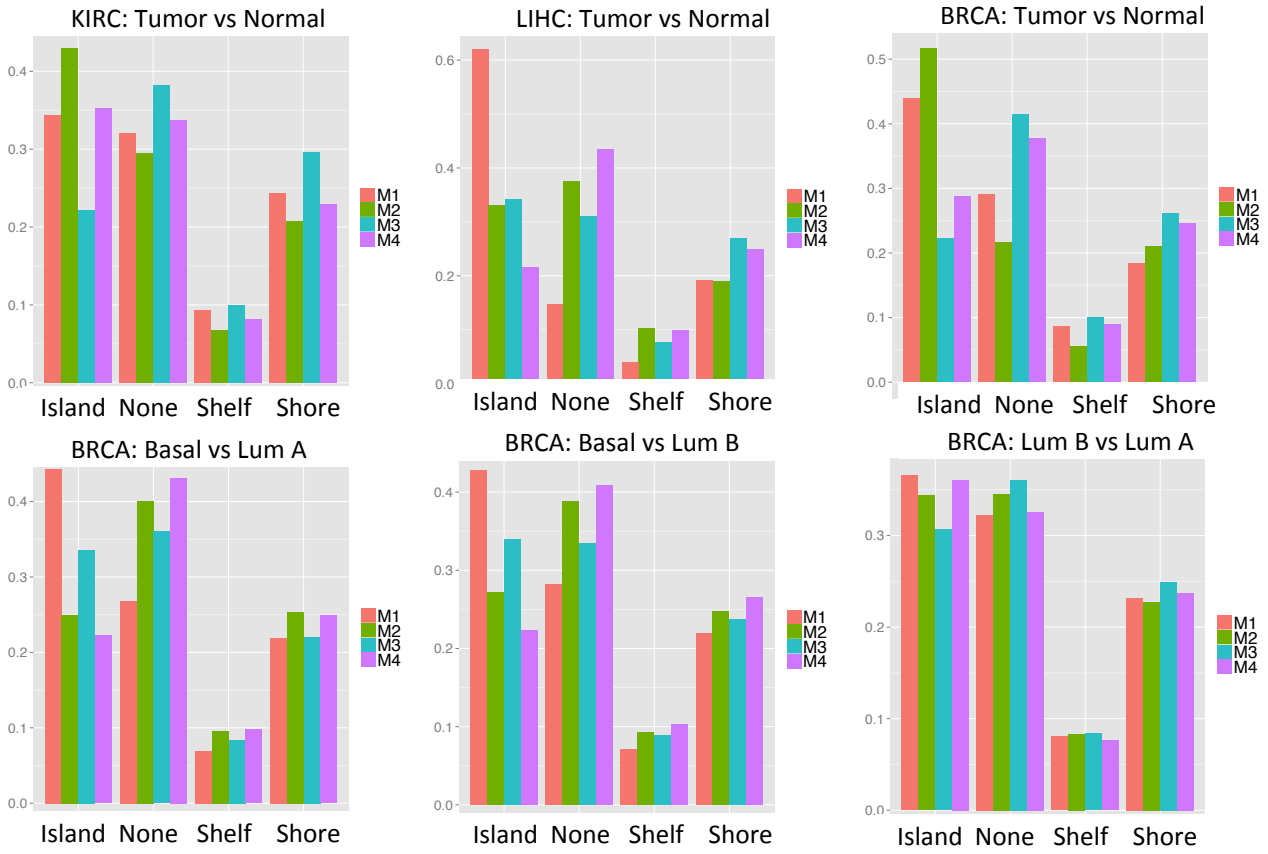


Fig. 2. CpG island, shelf and shore annotation for the proportion of CpGs identified by each model (color code: orange (Model 1), green (Model 2), turquoise (Model 3), purple (Model 4)) for the different datasets and comparisons.

test set. The parameters $\lambda$ and $\alpha$ were tuned using 10 fold cross-validation on the training set. The random partitioning of data into training and test set was repeated 10 times. We compared the following methods for selecting top 2000 CpGs from the training set to be included as candidate features:

(i) **Set 1**: Logit transformed beta values $x_{ij}$ of the top 2000 CpGs among the CpGs which ranked model 2 as the best model.
(ii) **Set 2**: Absolute deviation measure $r_{ij}$ of the top 2000 CpGs among the CpGs which ranked model 3 as the best model.
(iii) **Set 3**: Both the logit transformed beta values $x_{ij}$ and absolute deviation measure $r_{ij}$ of the top 2000 CpGs among the CpGs which ranked model 4 as the best model.

We evaluated the performance of the prediction algorithm on the test set in terms of area under the receiver operating characteristics curve (AUC), accuracy (Acc)= $\frac{TP+TN}{n_0+n_1}$, sensitivity (Sn)= $\frac{TP}{TP+FN}$, specificity (Sp)= $\frac{TN}{TN+FP}$ and Matthew's correlation coefficient (Mcc)= $\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$, averaged over the 10 iterations. The results are presented in Table 2. The prediction model for predicting tumor from normal in KIRC, LIHC

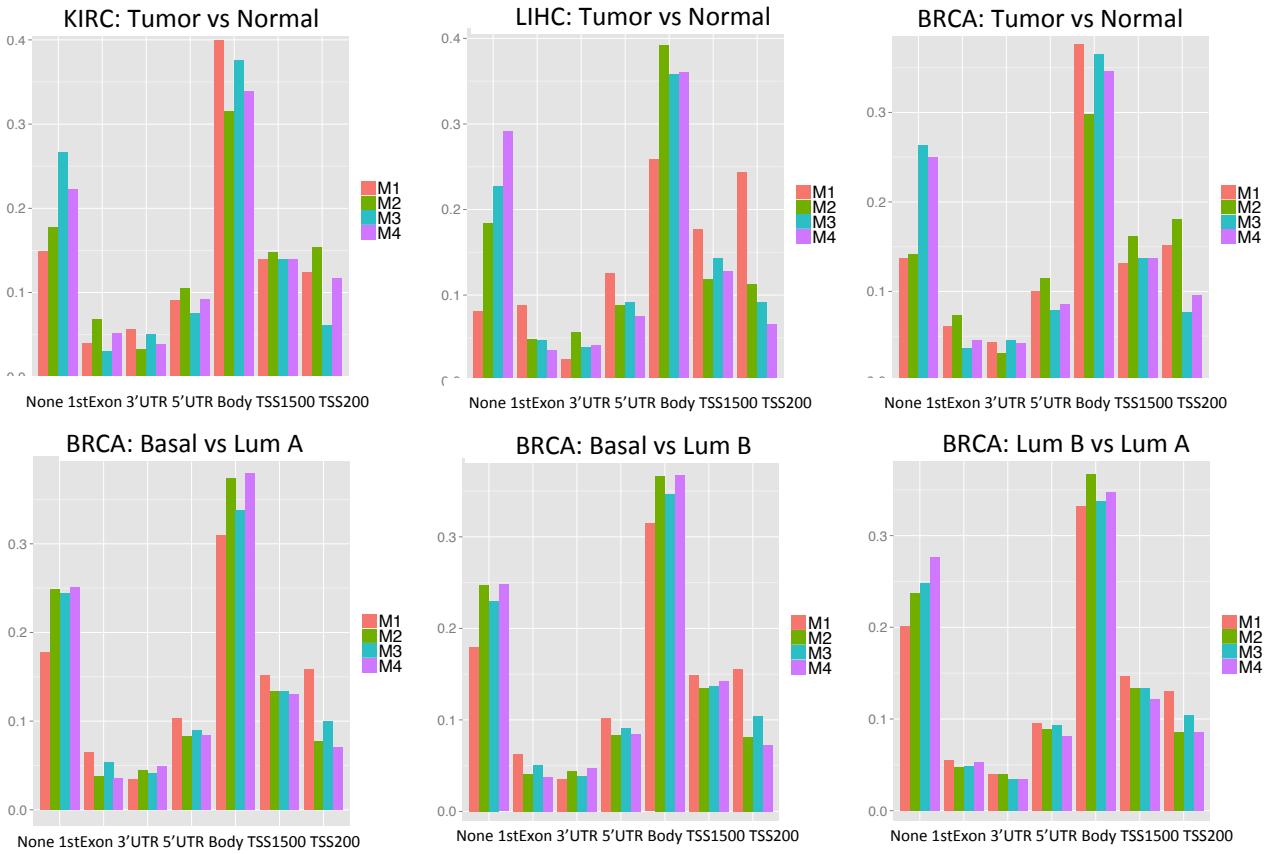


Fig. 3. Gene annotation for the proportion of CpGs identified by each model (color code: orange (Model 1), green (Model 2), turquoise (Model 3), purple (Model 4)) for the different datasets and comparisons.

and BRCA had high accuracy and AUC, and were comparable across the different candidate feature sets. Similar patterns were observed in basal versus luminal A and basal versus luminal B comparisons, indicating that DNA methylation was able to differentiate the more aggressive subtype (basal) from the less aggressive subtypes (luminals A and B) regardless of whether DM or DV CpGs were used. On the other hand, the prediction algorithm for predicting luminal A from luminal B subtypes exhibited lower accuracy compared to the previous comparisons, indicating that it is harder to differentiate these two subtypes based on DNA methylation.

## 5. Discussion

The promise and power of DNA methylation for therapeutics and diagnostics have been demonstrated in various diseases including cancer. Advancements in biotechnology enable large scale and population based epigenome-wide profiling of DNA methylation for identifying differential mean (DM) and differential variability (DV) CpGs. In these studies, covariates such as demographic and clinical factors may be confounded with both DNA methylation and disease phenotypes. One way to circumvent this problem is via randomization. However, this approach is not always feasible especially in case control studies. Moreover, in DNA

Table 2. Average AUC, Mcc, Accuracy (Acc), Sensitivity (Sn) and Specificity (Sp) for the different datasets and comparisons.

| Candidate feature | AUC | Mcc | Acc | Sn | Sp |
|---|---|---|---|---|---|
| KIRC: tumor vs normal | | | | | |
| Set 1 | 1.000 | 0.998 | 0.999 | 0.998 | 1.000 |
| Set 2 | 1.000 | 0.991 | 0.996 | 0.994 | 1.000 |
| Set 3 | 1.000 | 0.998 | 0.999 | 0.998 | 1.000 |
| LIHC: tumor vs normal | | | | | |
| Set 1 | 0.996 | 0.933 | 0.986 | 0.992 | 0.940 |
| Set 2 | 0.994 | 0.913 | 0.981 | 0.985 | 0.950 |
| Set 3 | 0.997 | 0.929 | 0.984 | 0.987 | 0.960 |
| BRCA: tumor vs normal | | | | | |
| Set 1 | 1.000 | 0.976 | 0.992 | 0.997 | 0.975 |
| Set 2 | 0.999 | 0.969 | 0.990 | 0.993 | 0.978 |
| Set 3 | 1.000 | 0.976 | 0.992 | 0.997 | 0.972 |
| BRCA: basal vs luminal A | | | | | |
| Set 1 | 0.996 | 0.947 | 0.980 | 0.950 | 0.989 |
| Set 2 | 0.987 | 0.848 | 0.944 | 0.817 | 0.984 |
| Set 3 | 0.995 | 0.947 | 0.980 | 0.950 | 0.989 |
| BRCA: basal vs luminal B | | | | | |
| Set 1 | 0.998 | 0.905 | 0.950 | 0.950 | 0.950 |
| Set 2 | 0.996 | 0.905 | 0.950 | 0.967 | 0.938 |
| Set 3 | 0.998 | 0.889 | 0.943 | 0.950 | 0.938 |
| BRCA: luminal B vs luminal A | | | | | |
| Set 1 | 0.798 | 0.339 | 0.741 | 0.425 | 0.874 |
| Set 2 | 0.720 | 0.287 | 0.722 | 0.413 | 0.853 |
| Set 3 | 0.791 | 0.380 | 0.767 | 0.413 | 0.916 |

methylation studies using whole blood sample, the different cell types have been shown to be confounded with the measured methylation levels.[32] In such cases, confounding factors need to be properly accounted for to avoid biases in DNA methylation biomarker detection. There are several approaches for DM analysis which allow for confounders adjustment,[33] however to the best of our knowledge existing DV analysis approaches are not tailored for confounders adjustments, except for our earlier work[16] which proposed a DV only analysis in the presence of confounders within large scale hypothesis testings framework. This paper extends our earlier work which allows for simultaneous detection of DM and DV in large scale hypothesis testings framework, and at the same time provides a candidate feature selection mechanism

for the prediction algorithm.

We showed that the analysis on KIRC, LIHC and BRCA TCGA datasets identified DM and DV CpGs which mapped to different CpG and gene annotations. For instance, in tumor versus normal comparisons, a larger proportion of DM CpGs mapped to CpG island and TSS200, whereas in basal versus luminal A or B comparisons, a larger proportion of DV CpGs mapped to these regions, suggesting that DM and DV CpGs regulate transcription differently. An R package `methylDMV` implementing this flexible framework is available at `http://www.ams.sunysb.edu/~pfkuan/softwares.html#methylDMV`.

DNA methylation generated from high resolution arrays including Illumina Infinium HumanMethylation450 BeadChip may induce a natural correlation structure among neighboring CpGs. An immediate extension of our current framework is to model the dependence structure and borrow information from nearby CpGs to improve the power of detecting DM and DV CpGs. Two of such approaches are (1) the hidden Markov model and local index of significance method as in Kuan et al. (2012),[34] and (2) the smoothing and bump hunting method as in Jaffe et al (2012),[7] which can possibly be adapted into our current `methylDMV` framework for detecting DM and DV CpGs.

## Acknowledgments

The authors thank the reviewers for their constructive comments and suggestions.

## References

1. V. Rakyan, T. Down, N. Thorne, P. Flicek, E. Kulesha, S. Graf, E. Tomazou, L. Backdahl, N. Johnson, M. Herberth, K. Howe, D. Jackson, M. Miretti, H. Fiegler, J. Marioni, E. Birney, T. Hubbard, N. Carter, S. Tavare and S. Beck, *Genome Research* **18**, 1518 (2008).
2. M. Esteller, *Annual Review Pharmacological Toxicology* **45**, 629 (2005).
3. R. Irizarry, C. Ladd-Acosta, B. Carvalho, H. Wu, S. Brandenburg, J. Jeddeloh, B. Wen and A. Feinberg, *Genome Research* **18**, 780 (2008).
4. P. Wang, Q. Dong, Z. Chong, P. Kuan, Y. Liu, W. Jeck, W. Jiang, G. S. nd T. Tan, J. Andersen, T. Auman, J. Hoskins, A. Misher, C. Moser, S. Yourstone, J. Kim, K. Cibulskis, S. Getz, H. Hunt, S. Thorgerisson, L. Roberts, D. Ye, K. Guan, Y. Xiong, L. Qin and D. Chiang, *Oncogene* **32**, 3091 (2012).
5. K. Conway, S. Edmiston, Z. Khondker, P. Groben, X. Zhou, H. Chu, P. Kuan, H. Hao, C. Carson, M. Berwick, D. Olilla and N. Thomas, *Pigment Cell and Melanoma Research* **24**, 352 (2011).
6. K. Hansen, W. Timp, H. Bravo, S. Sabunciyan, B. Langmead, O. McDonald, B. Wen, H. Wu, Y. Liu, D. Diep, E. Briem, K. Zhang, R. Irizarry and A. Feinberg, *Nature Genetics* **26**, 768 (2011).
7. A. Jaffe, A. Feinberg, R. Irizarry and J. Leek, *Biostatistics* **13**, 166 (2012).
8. A. Teschendorff and M. Widschwendter, *Bioinformatics* **28**, 1487 (2012).
9. X. Xu, S. Su, V. Barnes, C. Miguel, J. Pollock, D. Ownby, H. Shi, H. Zhu, H. Snieder and X. Wang, *Epigenetics* **8**, 522 (2013).
10. A. Feinberg and R. Irizarry, *Proc. Natl Acad. Sci. USA* **107**, 1757 (2010).
11. A. Teschendorff, A. Jones, H. Fiegl, A. Sargent, J. Zhuang, H. Kitchener and M. Widschwendter, *Genome Medicine* **4**, p. DOI: 10.1186/gm323 (2012).
12. G. Smyth, *Statistical Application in Genetics and Molecular Biology* **3**, p. 3 (2004).
13. S. Ahn and T. Wang, *Pacific Symposium of Biocomputing* , 69 (2013).

14. B. Phipson and A. Oshlack, *Genome Biology* **15**, DOI: 10.1186/s13059 (2014).
15. S. Wahl, N. Fenske, S. Zeilinger, K. Suhre, C. Gieger, M. Waldenberger, H. Grallert and M. Schmidt, *BMC Bioinformatics* **15**, DOI: 10.1186/1471 (2014).
16. P. Kuan, *Statistical Applications in Genetics and Molecular Biology* **13**, 645 (2014).
17. O. Stefansson, S. Moran, A. Gomez, S. Sayols, C. Arribas-Jorba, J. Sandoval, H. Hilmarsdottir, E. Olasfdottir, L. Tryggvadottir, J. Jonasson, J. Eyfjord and M. Esteller, *Molecular Oncology* **9**, 555 (2015).
18. J. Zhuang, M. Widschwendter and A. Teschendorff, *BMC Bioinformatics* **13**, DOI: 10.1186/1471 (2012).
19. S. Horvath, *Genome Biology* **14**, p. R115 (2013).
20. M. Jung and G. Pfeifer, *BMC Biology* **13**, doi: 10.1186/s12915 (2015).
21. M. Dogan, B. Shields, C. Cutrona, L. Gao, F. Gibbons, R. Simons, M. Monick, G. Brody, K. Tan, S. Beach and R. Philibert, *BMC Genomics* **15**, DOI: 10.1186/1471 (2014).
22. K. Lee and Z. Pausova, *Frontiers in Genetics* **4**, p. doi: 10.3389/fgene.2013.00132 (2013).
23. A. Houseman, B. Christensen, R. Yeh, C. Marsit, M. Karagas, M. Wrensch, H. Nelson, J. Wiemels, S. Zheng, J. Wiencke and K. Kelsey, *BMC Bioinformatics* **9**, doi:10.1186/1471 (2008).
24. P. Du, X. Zhang, C. Huang, N. Jafari, W. Kibbe, L. Hou and S. Lin, *BMC Bioinformatics* **11** (2010).
25. C. Rao, *Proceedings of the Cambridge Philosophical Society* **44**, 50 (1948).
26. Y. Benjamini and Y. Hochberg, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **57**, 289 (1995).
27. H. Zou and T. Hastie, *Journal of the Royal Statistical Society, Series B* **67**, 301 (2005).
28. L. Breiman, *Journal of Machine Learning* **45**, 5 (2001).
29. A. Teschendorff, F. Marabita, M. Lechner, T. Bartlett, J. Tegner, D. Gomez-Cabrero and S. Beck, *Bioinformatics* **29**, 189 (2013).
30. E. Price, A. Cotton, L. Lam, P. Farre, E. Emberly, C. Brown, W. Robinson and M. Kobor, *Epigenetics and Chromatin* **6** (2013).
31. Y. Chen, M. Lemire, S. Choufani, D. Butcher, D. Grafodatskaya, B. Zanke, S. Gallinger, T. Hudson and R. Weksberg, *Epigenetics* **8**, 203 (2013).
32. A. Houseman, W. Accomando, D. Koestler, B. Christensen, C. Marsit, H. Nelson, J. Wiencke and K. Kelsey, *BMC Bioinformatics* **13**, 189 (2012).
33. M. Ritchie, B. Phipson, D. Wu, Y. Hu, C. Law, W. Shi and G. Smyth, *Nucleic Acids Research* **43**, p. e47 (2015).
34. P. Kuan and D. Chiang, *Biometrics* **68**, 774 (2012).