**BIOINFORMATICS**

D³M: Detection of differential distributions of methylation levels

Yusuke Matsui¹, Masahiro Mizuta², Satoru Miyano³ and Teppei Shimamura ¹

¹Nagoya University Graduate School of Medicine, 466-8550, Nagoya, Japan.
²Information Initiative Center, Hokkaido University, 060-0811, Sapporo, Japan.
³Institute of Medical Science, The University of Tokyo, 108-8639, Japan.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

**ABSTRACT**

Motivation: DNA methylation is an important epigenetic modification related to a variety of diseases including cancers. We focus on the methylation data from Illumina’s Infinium HumanMethylation450 BeadChip. One of the key issues for methylation analysis is to detect the differential methylation sites between case and control groups. Previous approaches describe data with simple summary statistics and kernel function, and then use statistical tests to determine the difference. However, a summary statistics-based approach cannot capture complicated underlying structure, and a kernel functions-based approach lacks interpretability of results.

Results: We propose a novel method D³M, for detection of differential distribution of methylation, based on distribution-valued data. Our method can detect high-order moments, such as shapes of underlying distributions in methylation profiles, based on the Wasserstein metric. We test the significance of the difference between case and control groups and provide an interpretable summary of the results. The simulation results show that the proposed method achieves promising accuracy and shows favorable results compared with previous methods. Glioblastoma multiforme and lower grade glioma data from The Cancer Genome Atlas show that our method supports recent biological advances and suggests new insights.

Availability: R implemented code is freely available from https://github.com/ymatts/D3M/
https://cran.r-project.org/package=D3M.
Contact: ymatsui@med.nagoya-u.ac.jp

1 INTRODUCTION

DNA methylation is an epigenetic chemical alternation in which a methyl group is attached to a 5-carbon of a cytosine (C) base. It is closely related to gene expression, silencing, and genomic imprinting, including oncogenesis. Typically, methylation is explained as occurring in cytosine-phosphate-guanine (CpG) sites. The methylation of promoter regions, in particular, silences cancer suppressor genes (Baylin, 2005; Kulis and Esteller, 2010).

We focus on the methylation data from Illumina’s Infinium HumanMethylation450 BeadChip. One of the key issues for methylation analysis is to detect differentially methylated sites, i.e., a significant difference in methylation levels between the case and the control groups at a site. When comparing groups, we often summarize (or aggregate) data in summary statistics, such as mean and variance, and then investigate the difference between the groups. For example, IMA (Wang, et al., 2012) detects the differentially methylated sites by T-test, Empirical Bayes (EB) method or by Mann–Whitney–Wilcox (MWW) test. DiffVar (Phipson, et al., 2014) detects the sites by testing significant difference of variance. Other nonparametric approaches exist, such as Kolmogorov–Smirnov test (KS) or kernel-based approaches, such as maximum mean discrepancy (MMD) (Gretton, et al., 2012). In particular, since KS and MMD consider the underlying distribution structure, they are better suited for use with complicated distributions than methods based on summary statistics.

These approaches are effective in detecting typical differential methylation sites, but are insufficient from some perspectives; underlying distributions are complicated by being skewed, heavy-tailed, and multimodal. In particular, since cancers cells include heterogeneities, measurements of the methylation levels potentially include complex distribution shapes. This observation indicates that we need to consider the underlying structure. The disadvantage of KS and MMD is infeasible interpretability of results because they measure the maximum and kernel distances of distributions, respectively, which are difficult to interpret corresponding to the actual difference of underlying distributions.

We develop a method to detect differential methylation sites with distribution-valued data (Irpino and Verde, 2014). Distribution-valued data are an example of symbolic data analysis (Diday, 1989). This framework can treat complex data such as functional (Ramsey and Silverman 2005), tree (Wang and Marron, 2007), set, interval, and histogram values (Bock and Diday, 2000). The proposed method describes case and control groups using distribution values. We measure the differences between distributions using the Wasserstein metric. We detect the differential methylation sites using a statistical test of significant differences of distribution functions.

*to whom correspondence should be addressed
2 METHODS
Our method is aimed at a distribution-based comparison of methylation levels in two groups, through site-by-site resolution. We construct distribution functions representing the two groups at each site. Next, we compare the groups using a dissimilarity measure and test statistical significance through site-by-site resampling. We adopt an $L_2$-Wasserstein metric (Ruesehen, 2011) as a dissimilarity measure, a distribution function-based measure of statistical distance. The advantage of this distance is the interpretability of results because the distance can be decomposed into three components, i.e., mean, variance, and shape. This fact leads to visualization of results using a Q-Q plot to interpret the detected distribution difference including hypo- or hypermethylation status. In the following equation, we assume that input data is a beta value that is the ratio between the methylated probe intensity and the overall intensity (Du, et al., 2010). The definition of $i$-th site beta value in Illumina methylation assay is as follows:

$$\text{Beta}_i = \frac{\max(z_i,\text{methy},0) + \alpha}{\max(z_i,\text{unmethy},0) + \max(z_i,\text{methy},0) + \alpha} \quad (1)$$

where $z_i$ is the intensity measured by $i$-th methylated and unmethylated probes, respectively and $\alpha$ is constant.

2.1 Construction of objects
$X(s_i)$ and $Y(s_i)$ ($i = 1, 2, \ldots, S$) represent the beta values in a case group (e.g., cancer subjects) and a control group (e.g., normal subjects) at a CpG site $s_i$. We represent the data as distribution values by

$$F_i(x) = \Pr\{X(s_i) \leq x\}; x \in [0, 1].$$

$$G_i(y) = \Pr\{Y(s_i) \leq y\}; y \in [0, 1]. \quad (2)$$

In practice, let the beta value observations be $x_j(s_i); j = 1, 2, \ldots, n$ and $y_j(s_i); j = 1, 2, \ldots, m$ following $F_i(x)$ and $G_i(y)$, respectively, where $n$ and $m$ are the respective numbers of observations at $s_i$. From the data, we construct the empirical distribution functions;

$$\hat{F}_i(x) := \frac{1}{n} \sum_{j=1}^{n} 1(x_j(s_i) \leq x)$$

$$\hat{G}_i(y) := \frac{1}{m} \sum_{j=1}^{m} 1(y_j(s_i) \leq y) \quad (3)$$

where

$$1(a \leq b) = \begin{cases} 1 & \text{if } a \leq b \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

2.2 Dissimilarity measure for distributions
The Wasserstein metric is defined by the following:

$$d^q(F_i, G_i) := \int_0^1 |F_i^{-1}(u) - G_i^{-1}(u)|^q du \quad (5)$$

where $1 \leq q \leq 2$ and $F_i^{-1}(x)$ and $G_i^{-1}(y)$ indicate quantile functions.

In particular, in the case of $q = 2$, the metric can be decomposed into three components that describe the distribution characteristics, i.e., mean, variance, and shape (Irpino and Verde, 2014):

$$d^2(F_i, G_i) = \int_0^1 |F_i^{-1}(u) - G_i^{-1}(u)|^2 du$$

$$= (\mu_i - \mu_i')^2 + (\sigma_i - \sigma_i')^2 + 2\sigma_i\sigma_i'(1 - \rho_{i,i'}) \quad (6)$$

where $\mu_i$ and $\sigma_i^2$ (respectively, $\mu_i'$ and $\sigma_i'^2$) are mean and variance of $F_i(x)$ (respectively, $G_i(y)$), and $\rho_i$ is the correlation index of the points in the Q-Q plot of $F_i$ and $G_i$.

The empirical estimator of the Wasserstein metric is given by

$$d^q(\hat{F}_i, \hat{G}_i) = \int_0^1 |\hat{F}_i^{-1}(u) - \hat{G}_i^{-1}(u)|^q du. \quad (7)$$

Technically, we use quantiles to compute the approximation of the (7) for reducing computational costs. Let $(Q_{1,i}, Q_{2,i}, \ldots, Q_{K,i})$ and $(Q_{1,i}', Q_{2,i}', \ldots, Q_{K,i}')$ be $k$-quantiles of $F_i(x)$ and $G_i(y)$. We calculate $d^2(\hat{F}_i, \hat{G}_i) \approx \sum_{s=1}^{K} (Q_{s,i} - Q_{s,i}')^2$ in the case of $q = 2$, instead of evaluating the integral in (7). Here we simply define $d_i := d(\hat{F}_i, \hat{G}_i)$.

2.3 Detection of differential methylation sites
We use the metric to investigate whether two distributions are significantly different. We pose statistical hypotheses as follows.

Null hypothesis: $F_i = G_i$

Alternative hypothesis: $F_i \neq G_i \quad (8)$

We use resampling to construct a null distribution. From the null hypothesis (8), we jointly permute the observations $(x_1(s_i), x_2(s_i), \ldots, x_n(s_i))$ and $(y_1(s_i), y_2(s_i), \ldots, y_m(s_i))$ to obtain the new distribution functions $\hat{F}^*_{i}(x)$ and $\hat{G}^*_i(y)$. Next, we obtain the new distance $d_i^* = d^2(\hat{F}^*_i, \hat{G}^*_i)$ according to (7).

Let $D_i^* = (d_1^*, d_2^*, \ldots, d_{B_{all}})$ be all possible distances for the permutation process. Then $p$-value is

$$P_{all}(d_i) = \frac{\sum_{k=1}^{B_{all}} 1(d_{k,i}^* \geq d_i)}{B_{all}}. \quad (9)$$

Approximation of (9) uses the subset of $D_i^* = (d_1^*, d_2^*, \ldots, d_{B_{all}})$ where $B \leq B_{all}$:

$$P_{sub}(d_i) = \frac{\sum_{k=1}^{B} 1(d_{k,i}^* \geq d_i)}{B}. \quad (10)$$

In the simulation of section 3 and data analysis in section 4, we set $B = 10000$. The number of permutations $B$ is closely related to the accuracy of the $p$-value. However, resolution of $P_{sub}$ is limited to $1/B$. If we need the very small $p$-values, one solution is to perform a large number of permutations, but it is computationally expensive. A semi-parametric estimation of $p$-value is proposed by Knijnenburg et al. (2009) to obtain more accurate $p$-values.
We use an exponential distribution to estimate the distribution tail as follows,

\[
P(d_i) = \begin{cases} 
\frac{1}{\lambda_i} \sum_{j=1}^{i} 1(d_{i,j} \geq d_i) & \text{for } d_i < d_i^{(\text{min})} \\
\exp(-\lambda_i(d_i - d_i^{(\text{min})})) & \text{for } d_i \geq d_i^{(\text{min})}
\end{cases}
\]

where \(\lambda_i\) is a scale parameter and \(d_i^{(\text{min})}\) is a threshold that we set to 99\% percentile of null distributions. We estimate \(\lambda_i\) using data above the threshold. Technically, we perform the semi-parametric estimation only if \(P_{\text{sub}}(d_i)\) reaches to zero.

### 2.4 Graphical representation of results

The graphical interpretation of the statistical test result is important. One approach is to plot all the distribution (density) functions of candidate sites, but this is infeasible for hundreds of sites. We use a Q-Q plot with two distributions. It enables us to visualize many pairs of distributions at a time, with the directions being easy to interpret. In the actual example shown in section 4, we plotted 1,000 pairs of differentially methylated distributions (Figure 3 (B)). We can see the hyper-methylation with the most significant 1,000 sites (blue lines in Figure 3 (B)).

## 3 SIMULATION

### 3.1 Simulation setting

We evaluated the proposed method with simulated datasets with focus on single cytosine level in the case and control group. Our simulation is intended for the detection of differential methylation sites when there is cancer heterogeneity. Here, the cancer heterogeneity is described by the multiple modes of distributions. We conduct a statistical test for \(H_0 : F_1 = G_1 \leftrightarrow H_1 : F_i \neq G_i\), under significance levels 5\%, and we compare the results to those of the other methods, i.e., DiffVar, MMD, KS, MWW and EB. We used several packages; missMethyl (Phipson, et al., 2014), limma (Ritchie, et al., 2015) written in R and mmd (Gretton et al., 2006) (http://www.kyb.tuebingen.mpg.de/bs/people/arthur/code/mmd.zip) written in MATLAB. The setting of missMethyl is default and mmd with options alpha = 0.5 and MMD METHOD = ‘approxmoments’.

We describe the outline of the simulation as follows. The details are described in supplemental file S1 and R codes are described in S2. The data are generated by using two types of distributions. The control and case groups are represented by normal and normal mixture distributions, respectively. In each case (case 1–case 8), there are 80 samples; 40 samples for case and control groups, respectively. We performed the statistical test with each a method for 50 times in every case (i.e., case 1–case 8) and evaluate averages of a type I error and power. Besides we repeat this process for 100 times to assess the variances of the averages.

### 3.2 Simulation results

The results are shown in Table 2. In the first case, it is shown that error rates of \(D^3M\), MMD, DiffVar, KS, EB, and MWW are close to the significance levels, which indicates that they effectively control type I errors.
DiffVar, KS and D
GBM. We focus on mean, variance, and shape differences using EB, sites. The detection of differential methylation levels is a clue and GBM groups, and then specify the differential methylation
In this analysis, we compare the methylation levels in the LGG
site contains 145 samples and 530 samples in GBM and
which is downloadable from TCGA portal sites. Each CpG
HumanMethylation450K arrays, including 485,577 CpG sites,
follows. All the samples are hybridized to Illumina’s Infinium
4.1 Datasets

and as secondary tumors from LGG (Wiencke et al.,
4.2 Analysis results
Significant differential methylation sites were identified as those having false discovery rate (q-value) (Benjamini, et al.,1995) less than 1%. The Venn diagram (Figure 2 (A)) shows the number of detected probes from the perspective of mean, variance, and shape difference of distributions using EB, DiffVar, KS, and D^3M, and
279,008, 191,050, 297,493, and 255,317 sites are totally detected, respectively. From Figure 2 (A), most of detected sites with the shape difference are overlapped with the sites by the mean and the variance differences, respectively. However, when focusing on top 1000 significant sites (Figure 2 (B)), the overlaps between them become very small. This suggests that the “signal” of differential methylation sites in terms of q-value is quite different from each other. It is important for the further analysis, such as the pathway analysis since we often use the filtered gene set, e.g., using top 1,000 significant sites.

and we show the details in supplemental file in S1 and R codes in S2.

4 ACTUAL EXAMPLE
4.1 Datasets
We apply our method to methylation data of glioblastoma multiforme (GBM) and lower grade glioma (LGG) from The Cancer Genome Atlas (TCGA). GBM is the primary brain tumor that progresses with malignant invasion destroying normal brain tissues (TCGA, 2008), arising through two pathologically distinct routes, de novo and as secondary tumors from LGG (Wiencke et al., 2006). In this analysis, we compare the methylation levels in the LGG and GBM groups, and then specify the differential methylation sites. The detection of differential methylation levels is a clue for revealing epigenetic mechanisms of development from LGG to GBM. We focus on mean, variance, and shape differences using EB, DiffVar, KS and D^3M and compare the results.

Here we briefly describe the datasets and preprocessing as follows. All the samples are hybridized to Illumina’s Infinium HumanMethylation450K arrays, including 485,577 CpG sites, which is downloadable from TCGA portal sites. Each CpG site contains 145 samples and 530 samples in GBM and LGG, respectively. We remove CpG sites on the X and Y chromosomes and SNP control probes (rs1–rs65). We also use HumanMethylation450 v1.2 SNP Update Table to remove SNP related probes by minor allele frequency ≤ 1%. As a result, we get 351,932 probes. Missing values in both groups are inferred using R package pcaMethods (Stacklies, 2007) with pca functions since DiffVar doesn’t accept missing values and the number of the probes including missing values cannot be ignored. We remove the batch effect using the ComBat function in SVA package (Leek, et al., 2012) with default settings. We use batches having more than 2 samples. The details are described in R codes in S3. We finally obtain 141 GBM samples and 530 LGG samples.

Table 1. Simulation models of eight cases

<table>
<thead>
<tr>
<th></th>
<th>$F_1 = F_2$</th>
<th>$F_1 \neq F_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>case 1</td>
<td>case 2</td>
</tr>
<tr>
<td>$\mu_1 = \mu_2$</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>$\sigma_1 = \sigma_2$</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>$s_1 = s_2$</td>
<td>T</td>
<td>F</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Type I</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>case 1</td>
<td>case 2</td>
</tr>
<tr>
<td>D^3M</td>
<td>mean 5.12</td>
<td>93.20</td>
</tr>
<tr>
<td>H_0 : $F_1 = F_2$ , $H_1 : F_1 \neq F_2$</td>
<td>sd 0.07</td>
<td>0.08</td>
</tr>
<tr>
<td>DiffVar</td>
<td>mean 1.80</td>
<td>6.80</td>
</tr>
<tr>
<td>H_0 : $V(F_1) = V(F_2)$ , $H_1 : V(F_1) \neq V(F_2)$</td>
<td>sd 0.04</td>
<td>0.08</td>
</tr>
<tr>
<td>MMD</td>
<td>mean 4.40</td>
<td>100.00</td>
</tr>
<tr>
<td>H_0 : $F_1 = F_2$ , $H_1 : F_1 \neq F_2$</td>
<td>sd 0.07</td>
<td>0.00</td>
</tr>
<tr>
<td>KS</td>
<td>mean 2.56</td>
<td>99.86</td>
</tr>
<tr>
<td>H_0 : $F_1 = F_2$ , $H_1 : F_1 \neq F_2$</td>
<td>sd 0.05</td>
<td>0.01</td>
</tr>
<tr>
<td>EB</td>
<td>mean 4.94</td>
<td>0.66</td>
</tr>
<tr>
<td>H_0 : $E(F_1) = E(F_2)$ , $H_1 : E(F_1) \neq E(F_2)$</td>
<td>sd 0.07</td>
<td>0.02</td>
</tr>
<tr>
<td>MWW</td>
<td>mean 4.68</td>
<td>11.22</td>
</tr>
<tr>
<td>H_0 : $M(F_1) = M(F_2)$ , $H_1 : M(F_1) \neq M(F_2)$</td>
<td>sd 0.07</td>
<td>0.10</td>
</tr>
</tbody>
</table>
Among the detected sites, we investigated sites with the smallest 1,000 \( q \)-value. Heat map and Q–Q plots of the top 1,000 sites are shown in Figures 3 (A) and (B). Comparing heat maps and Q–Q plots, the methylation levels are easy to interpret in the latter. From the Q–Q plot, we could see that the top 1,000 sites tend to be hyper-methylated in LGG (with the reverse in GBM).

Among the top significant 1,000 pairs of distributions of GBM and LGG, we can observe that there are mainly two patterns in terms of the Q–Q plot. Then, we cluster the 1000 curves in the Q–Q plot into two classes; we consider input data as terms of the Q–Q plot. Then, we cluster the 1000 curves in the Q–Q plot, we could see that the top 1,000 sites tend to be hyper-methylated in LGG (with the reverse in GBM).

The pathways in clusters 1 and 2 include significant pathways in GBM and LGG, which have been previously reported even though we set the variance, but not the mean and the shapes, thus must be included in the analysis.

Next, we perform enrichment analysis on gene sets in clusters 1 and 2. We used ingenuity pathway analysis (IPA) for 385 and 268 genes in clusters 1 and 2, respectively, and significantly enriched pathways in each cluster using Fisher’s exact test. Table 4 shows five examples in each cluster. Clusters 1 and 2 contain 626 and 374 sites, respectively.

The pathways in clusters 1 and 2 include significant pathways in GBM, which have been previously reported even though we do not include any information on GBM. The axonal guidance pathways in cluster 1 has been suggested as prompting the invasion of GBM (Dominique, et al., 2010; Polisetty et al., 2013). The pathways in cluster 2 include significant pathways in GBM and LGG.

We further focus on Wnt in Human Embryonic Stem Cell Pluripotency pathway, and then compare the ranking based on \( p \)-value by \( D^3 \)M with those by other methods. The activation of Wnt family is closely related to cell differentiation of GBM (Rampazzo, et al., 2013). In our analysis, there are 18 Wnt genes on the probes. Among them, six Wnt genes (Wnt2, Wnt2b, Wnt3, Wnt4, Wnt7a, and Wnt 9a) are included in top 1,000 significantly differentially methylated probes using \( D^3 \)M.

We investigate the enrichment of the six Wnt genes in top 1,000 significant probes using Fisher’s exact test and we confirm the enrichment of the genes (the details are described in section S-3-2 in supplemental file S1). The six Wnt genes are included in both the clusters 1 and 2, and the majority of the distribution is hypo-methylated and the minority is hyper-methylated in GBM, vice versa in LGG. This suggests that demethylation of Wnt in some LGG might trigger the activation of Wnt family and prompt transformation from LGG to GBM. The ranking of Wnt genes in \( D^3 \)M, EB, KS, and DiffVar is shown in Table 3.

### Table 3. The ranking of the six Wnt gene probes in 351,932 probes.

<table>
<thead>
<tr>
<th>Wnt2</th>
<th>Wnt2b</th>
<th>Wnt3</th>
<th>Wnt4</th>
<th>Wnt7a</th>
<th>Wnt9a</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D^3 )M</td>
<td>342</td>
<td>549</td>
<td>391</td>
<td>829</td>
<td>164</td>
</tr>
<tr>
<td>( (0.10) )</td>
<td>( (0.16) )</td>
<td>( (0.17) )</td>
<td>( (0.24) )</td>
<td>( (0.18) )</td>
<td>( (0.25) )</td>
</tr>
<tr>
<td>EB</td>
<td>22.541</td>
<td>12.914</td>
<td>3.246</td>
<td>1.811</td>
<td>135</td>
</tr>
<tr>
<td>( (6.40) )</td>
<td>( (3.67) )</td>
<td>( (0.92) )</td>
<td>( (0.51) )</td>
<td>( (0.04) )</td>
<td>( (5.90) )</td>
</tr>
<tr>
<td>KS</td>
<td>4.754</td>
<td>91.240</td>
<td>50.174</td>
<td>41.589</td>
<td>41.066</td>
</tr>
<tr>
<td>( (1.35) )</td>
<td>( (25.93) )</td>
<td>( (14.26) )</td>
<td>( (11.82) )</td>
<td>( (11.67) )</td>
<td>( (23.94) )</td>
</tr>
<tr>
<td>DiffVar</td>
<td>56.950</td>
<td>155.621</td>
<td>147.817</td>
<td>146.033</td>
<td>138.928</td>
</tr>
<tr>
<td>( (16.18) )</td>
<td>( (44.22) )</td>
<td>( (42.00) )</td>
<td>( (41.49) )</td>
<td>( (39.48) )</td>
<td>( (68.95) )</td>
</tr>
</tbody>
</table>

which is caused by epigenetic instability and diversity. Cancer heterogeneity can often be confused with outliers. For example, DiffVar fails to detect simulation case 6 as differential methylation, even though we set the variance, but not the mean and the shapes, to be the same for the two groups. This is because DiffVar deals with minority distributions as outliers and evaluates only those in the majority.

In general, the significance of an outlier depends on the context of analysis (Aggarwal, 2013). When an outlier arises from measurement error not relevant to signals of interest, we must remove them prior to analysis. In contrast, when an outlier arises from an unusual event including new findings that we seek, we use them for further analysis. In this case, cancer heterogeneity could be regarded as an abnormal event compared with normal cases, and thus must be included in the analysis.

MMD was originally developed as a distribution-free two sample test based on a kernel. Both the Wasserstein metric and MMD are family of integral probability metrics (Müller, 1997) designed to capture not only such basic difference as those between the means but also such higher-order differences as those between the distribution shapes of the groups. From the simulation in section 3.1, we see that MMD shows good detection performance in cases such as case 2 and 6. \( D^3 \)M shows good performance over all eight cases. Specifically, \( D^3 \)M shows excellent performance in cases 7 and 8. Considering that case 8 is most frequently observed in actual data, this result is preferable in applications.

In the study of sample size effect in section S-2-1, using the same simulation model as in section 3.1, the result indicates that the superiority of \( D^3 \)M to other methods are reduced compared to large sample case, although all the methods decrease in power (Table S.3). On the other hand, MMD retains its power in case 2 and case 6, and so does \( D^3 \)M.

We also investigate the power of \( D^3 \)M with other simulation models using beta distributions considering small sample size in section S-2-2 and S-2-3 in supplemental file S1. The result indicates that EB and MWW detect differences well in the small samples (\( n = 12 \), \( n = 20 \)) and so do \( D^3 \)M and MMD in the moderate sample sizes (\( n = 28 \), \( n = 36 \)). In case of unbalanced sample sizes between the case and the control groups (Table S.5), \( D^3 \)M and EB shows the good detection powers.

\( D^3 \)M can be applied flexibly to differential methylation problems. Simulation results indicate that \( D^3 \)M can detect not only shape differences but also summary statistics differences as effectively as EB and DiffVar, i.e., natural results from the decomposition (6). This suggests that if we cannot obtain sufficient power using a
Fig. 2. Venn diagram of detected sites at the significance level 1% (A) and of top significant 1000 sites (B).

Fig. 3. Distributions of the top 1,000 significant probes. (A) The heat map of beta values with 141 LGG and 530 GBM samples at each probe respectively. We jointly clustered quantiles of LGG and GBM probes into the two clusters based on the Wasserstein metric (i.e., we clustered curves in Q-Q plots of the top significant probes into two clusters). (B) Q–Q plots of beta values between the LGG and GBM probes (blue lines). D³M detects the probes with hyper-methylated LGG probes compared with GBM probes. Green lines are the top 1,000 insignificant probes (as negative controls). (C) The instances of distributions with Clusters 1 and 2 shown in (A-1) and (A-2), respectively. The distributions appear to be highly heterogeneous especially in LGG probes. D³M recognizes strongly heterogeneous distributions and detects such probes as top significant probes.
Table 4. Pathways detected with the proposed method

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Pathway</th>
<th>(-\log(P\text{-value}))</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>Axonal Guidance Signaling</td>
<td>5.85</td>
<td>SLIT3, ITGB1, SEMA3G, MYL10, WNT3, FES, NRP2, BMP7A, UNC5B, WNT2B, PIK3R5, GNA11, ITGA5, KEL, WNT7A, MAG, ADAM12, NTRK1, PRKAG2, WNT4, BMP7, NTN3, PRKCB</td>
</tr>
<tr>
<td></td>
<td>ERK/MAPK Signaling</td>
<td>4.09</td>
<td>ITGB1, DUSP9, PLA2G4B, PRKAG2, ITGA5, PIK3R5, RAPGEF4, ESRT, PPP2R5A, KSR1, JMD7-PLA2G4B, PRKCB</td>
</tr>
<tr>
<td></td>
<td>Caveolar-mediated-Endocytosis Signaling</td>
<td>3.72</td>
<td>ITGB1, INS, CD48, ITGA5, ITGB8, JMD7-PLA2G4B, PRKCB</td>
</tr>
<tr>
<td></td>
<td>Human Embryonic Stem-Cell Pluripotency</td>
<td>3.35</td>
<td>WNT7A, WNT3, BMP7A, SMA1D3, NTRK1, WNT2B, PIK3R5, WNT4 BMP7</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>TREM1 Signaling</td>
<td>3.55</td>
<td>STAT5A, MPO, CASP1, NOD1, CCL3, ITGA</td>
</tr>
<tr>
<td></td>
<td>Calcium Signaling</td>
<td>3.28</td>
<td>GRIN2A, TNNT3, ITPR2, CHRN1B1, NFA5TC3, CAMKK2, PPP3CA, CAMKK2, PPP3CA, MEF2B</td>
</tr>
</tbody>
</table>

simple summary statistics approach, we have other options to add shape information. If we would like to see the results of variance and shape differences simultaneously, we remove the mean from the data using \( X - \bar{X} \) in each group at a site before applying \( D^M \). This option is provided with R package \( D^M \); in the function \( D^M \_i3m \), the logical parameters of \( rm\_mean \) and \( rm\_var \) exist. If we would like to see the variance and shape differences at the same time, we just set the \( rm\_mean=F \) and \( rm\_var=F \).

The statistical test of \( D^M \) relies on resampling and requires computational time to calculate \( p\)-values. However, we could reduce the resampling time using a semi-parametric approach (Knijnenburg, et al., 2009).

A current limitation of \( D^M \) is that it deals with univariate distributions. In a case of the study of large sample sizes, we can deal with covariates, such as age and gender, but our method currently does not incorporate them into the model, and the user needs to remove the effects of covariates; for example, using residuals with regression analysis, prior to the analysis. The extension of \( D^M \) to multivariate distribution relies on the estimation of the Wasserstein metric between the empirical multivariate distributions. The approximation of the Wasserstein metric between the multivariate distributions has been studied recently in (Applegate, et al., 2011), and we could derive the null distributions based on that approximation.

\( D^M \) does not support a spatial correlation in the current form. The spatial information will enhance the power of detection for differential methylated sites. This could be accomplished by spatially weighted average of Wasserstein distance over a fixed range of locus. These extensions of \( D^M \) will be covered in future work.

6 CONCLUSION

In this study, we proposed a novel method, \( D^M \), for detecting differential methylation sites based on distribution-valued data. We showed that distribution shape includes interesting information other than that found using mean- and variance-based methods. A simulation study indicated that \( D^M \) is capable to detect various situations.

In the application to the GBM and LGG dataset in the TCGA cohort, we identified 1,000 sites with the smallest \( q\)-values. Most of the sites detected by \( D^M \) show strong heterogeneity and tend to be hyper- and hypo-methylated in LGG and GBM, respectively, as found in previous studies.

Since the GBM and LGG dataset contains a large number of significantly different sites, including 279,008, 191,050, 297,493 and 255,317 sites for \( D^M \), EB, KS, and DiffVar, respectively; at the FDR \( < 1\% \), it is difficult to understand the methylation levels at these sites. In the future, it would be of interest to develop a method that describes the diversity of methylation levels.

REFERENCES

Applegate,D. et al. (2011) Unsupervised Clustering of Multidimensional Distributions Using Earth Mover Distance. Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 636-644, ACM.


