Vol. 00 no. 00 2015
Pages **???**

D^{3}M: Detection of differential distributions of methylation patterns

Yusuke Matsui [1*], Masahiro Mizuta [2], Satoru Miyano [3] and Teppei Shimamura [1*]

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

ABSTRACT

Motivation: DNA methylation is an important epigenetic modification related to a variety of diseases including cancers. One of the key issues of methylation analysis is to detect the differential methylation sites between case and control groups. Previous approaches describe data with simple summary statistics and kernel functions, 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 D^{3}M, 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 outperforms previous methods. Glioblastoma multiforme and lower grade glioma data from The Cancer Genome Atlas and show that our method supports recent biological advances and suggests new insights.

Availability: R implemented code is freely available from `https://cran.r-project.org/web/packages/D3M/` `https://github.com/cran/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 carbon 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) islands. The methylation of promoter regions, in particular, silences cancer suppressor genes.

One of the key issues for methylation analysis is to detect differential methylation site, *i.e.,* significant difference in methylation patterns between case and 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, *limma* (Smyth, *et al.,* 2005), *minfi* (Aryee, *et al.,* 2014), *edgeR* (Robinson, *et al.,* 2010), *DESeq* (Anders, *et al.,* 2010) and *DiffVar* (Phipson, *et al.* 2014) detect the differential methylation sites by testing for significant differences in mean and variance. Other nonparametric approaches exist, such as the Mann-Whitney-Wilcox test (MWW), based on rank statistics, and the Kolmogorov-Smirnov test (KS) or kernel-based approaches, such as M^{3}D (Mayo *et al.,* 2014) with 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, such as the following. The *limma*, *minfi*, *edgeR*, *DESeq*, and *DiffVar* methods are inappropriate when underlying distributions are complicated by being skewed, heavy-tailed, and multimodal. In particular, since cancer cells include heterogeneities, measurements of methylation 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, 2014a). 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; Billard and Diday, 2006; Noirhomme-Fraiture and Diday, 2008). 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.

2 METHODS

Our method is aimed at a distribution-based comparison of methylation patterns 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

*to whom correspondence should be addressed

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 distribution shape. This fact leads to visualization of results using a Q-Q plot to interpret the detected distribution difference including hypo- or hyper-methylation status.

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 control group (*e.g.,* normal subjects) at a CpG site s_i. We represent the data as distribution values by

$$\begin{aligned} F_i(x) &= \Pr\{X(s_i) \leq x\}; x \in [0,1]. \\ G_i(y) &= \Pr\{Y(s_i) \leq y\}; y \in [0,1]. \end{aligned} \tag{1}$$

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;

$$\begin{aligned} \hat{F}_i(x) &:= \frac{1}{n}\sum_{j=1} \mathbf{1}(x_j(s_i) \leq x) \\ \hat{G}_i(y) &:= \frac{1}{m}\sum_{j=1} \mathbf{1}(y_j(s_i) \leq y) \end{aligned} \tag{2}$$

where

$$\mathbf{1}(a \leq b) = \begin{cases} 1 & \text{if} \quad a \leq b \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

2.2 Dissimilarity measure for distributions

The Wasserstein metric is defined by

$$d^q(F_i, G_i) := \int_0^1 |F_i^{-1}(u) - G_i^{-1}(u)|^q du \tag{4}$$

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, 2014a):

$$\begin{aligned} d^2(F_i, G_i) &= \int_0^1 |F_i^{-1}(u) - G_i^{-1}(u)|^2 du \\ &= (\mu_i - \mu_i')^2 + (\sigma_i - \sigma_i')^2 + 2\sigma_i\sigma_i'(1 - \rho_{i,i'}) \end{aligned} \tag{5}$$

where μ_i and σ_i^2 (respectively, μ_i' and $\sigma_i'^2$) are mean and variance of $F_i(x)$ (respectively, $G_i(y)$), and ρ_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

$$d^q(\hat{F}_i, \hat{G}_i) = \int_0^1 |\hat{F}_i^{-1}(u) - \hat{G}_i^{-1}(u)|^q du. \tag{6}$$

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

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.

$$\begin{aligned} &\text{Null hypothesis:} \quad F_i = G_i \\ &\text{Alternative hypothesis:} \quad F_i \neq G_i \end{aligned} \tag{7}$$

We use resampling to construct a null distribution. From the null hypothesis (**??**), we 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 $d_i^* = d(\hat{F}_i^*, \hat{G}_i^*)$ according to (**??**).

Let $D_i^* = (d_{i,1}^*, d_{i,2}^*, \ldots, d_{i,B_{all}}^*)$ be all possible distances for the permutation process. Then p-value is

$$P_{all}(d_i) = \frac{\sum_{b=1}^{B_{all}} \mathbf{1}(d_{i,b}^* \geq d_i)}{B_{all}}. \tag{8}$$

Approximation of (**??**) uses the subset of $D_i^*, \tilde{d}_{i,1}^*, \tilde{d}_{i,2}^*, \ldots, \tilde{d}_{i,B}^*$ where $B \leq B_{all}$:

$$P_{sub}(d_i) = \frac{\sum_{b=1}^{B} \mathbf{1}(\tilde{d}_{i,b}^* \geq d_i)}{B}. \tag{9}$$

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,

$$P(d_i) = \begin{cases} \frac{1}{B}\sum_{j=1} \mathbf{1}(\tilde{d}_{i,b}^* \geq d_i) & \text{for} \quad d_i < d_i^{(\min)} \\ \exp(-\lambda_i(d_i - d_i^{(\min)})) & \text{for} \quad d_i \geq d_i^{(\min)} \end{cases} \tag{10}$$

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

2.4 Graphical representation of results

Since the method for detection of methylation, which is based on distance, cannot distinguish the "direction" of the hyper- or hypo-methylation. 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 (Fig **??**). We can see the hyper-methylation with the most significant 1,000 sites (blue lines in Fig **??**).

3 SIMULATION

3.1 Simulation setting

We evaluated the proposed method with simulated datasets. 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 $H_0 : F_i = G_i \leftrightarrow H_1 : F_i \neq G_i$ under significance levels 5% and 1%, and we compare the results to those of the other methods, *i.e.,* DiffVar, MMD, KS, MWW and Welch test (Welch). We used R packages for this simulation, `MissMethyl` (Phipson, *et al.* 2014), `kernlab` (Karatzoglou, *et al.,* 2004), and `base`. The setting of `MissMethyl` is default and that of `kmmd` in `kernlab` with resampling number ($ntimes$) is set by 10,000. Since the distribution distance is decomposed into mean, variance, and shape in (**??**), we conduct seven cases of H_1 (Table 1). Figure 1 shows seven differential methylation cases with `beanplot` (Kampstra, 2008) in which the distribution density functions are described as upper and lower for control and case groups, respectively. The vertical black solid line indicates the distribution mean. Here, we define shape differences of the distributions as the number of modes, *i.e.,* unimodal and bimodal distributions are regarded as different.

We describe the outline of the simulation as follows. We generate the data using two types of distribution. The control and case groups are represented by normal and normal mixture distributions, respectively. In each case, there are 300 samples; 160 and 140 for case and control group, respectively. The details of simulation models are shown in supplemental file S1. First, we evaluate type I errors in case 1 using 5,000 datasets. Next, we evaluate the power in cases 2-8 using 5,000 datasets for each group.

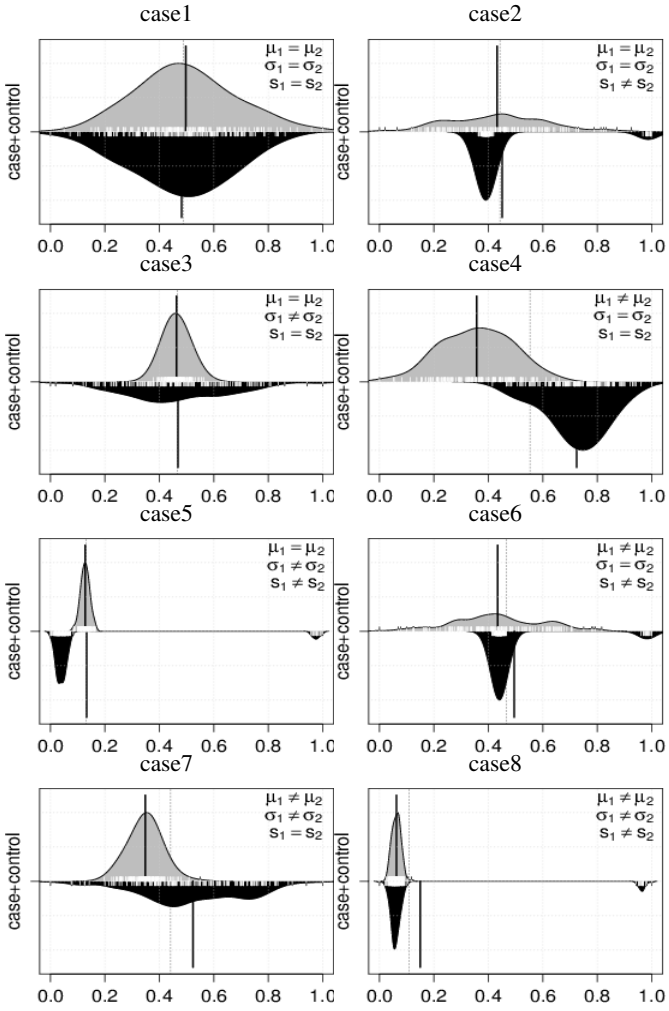

Fig. 1. The beanplot of eight cases

3.2 Simulation results

The results are shown in Table 2. In the first case, it is shown that error rates of D^{3}M, DiffVar, KS, Welch, and WMM are close to the significance levels, which indicates that they effectively control type I errors. In contrast, MMD cannot control type I error at both of the levels of 5% and 1%, *i.e.,* the significance level actually fails.

Furthermore, we investigate the power with cases 2-8. KS detects most of the cases with low variance, with case 8 being an exception. However, KS cannot recognize the difference when the majorities of the two groups overlap with each other (Figure 1, case 8). DiffVar shows high power and low variance for cases where the variances differ. However, DiffVar might capture the other distribution features for the cases with equal variances, leading to uninterpretable results. In this simulation, Welch can appropriately distinguish only the mean difference. MMD succeeds in identifying shape differences in cases 2, 5, and 6. However, it decreases the accuracy in cases 3, 4, and 7, in which the mean and variance differ, and it cannot detect case 8. WMM can detect case 4, 5, and 7, but cannot detect cases in which the means differ under non-normality. D^{3}M outperforms all these other methods and achieves promising accuracy in all cases.

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 patterns in the LGG and GBM groups, and then specify the differential methylation sites. Detection of differential methylation patterns is a clue for revealing epigenetic mechanisms of development from LGG to GBM. We focus on mean, variance, and shape differences using Welch, DiffVar, and D^{3}M and compare the results.

Here we briefly describe the datasets and preprocessing as follows. All the samples are hybridized to Illuminas 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. First, we remove CpG sites on the X and Y

Table 1. Simulation models of eight cases

	$F_i = F_2$	$F_1 \neq F_2$						
	case1	case2	case3	case4	case5	case6	case7	case8
$\mu_1 = \mu_2$	T	T	T	F	T	F	F	F
$\sigma_1 = \sigma_2$	T	T	F	T	F	T	F	F
$s_1 = s_2$	T	F	T	T	F	F	T	F

Table 2. 5,000 simulations in each case

			Type I	Power						
			case1	case2	case3	case4	case5	case6	case7	case8
D^3M	mean	α=0.05	5.04	99.95	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
		α=0.01	0.95	99.81	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
	sd	α=0.05	0.64	0.17	0.00	0.00	0.00	0.00	0.00	0.00
		α=0.01	0.28	0.56	0.00	0.00	0.00	0.00	0.00	0.00
DiffVar	mean	α=0.05	5.05	74.70	**100.00**	4.10	**100.00**	74.89	**100.00**	**100.00**
		α=0.01	0.99	34.01	**100.00**	0.78	**100.00**	33.88	**100.00**	**100.00**
	sd	α=0.05	0.67	1.25	0.00	0.64	0.00	1.49	0.00	0.00
		α=0.01	0.33	1.67	0.00	0.28	0.00	1.44	0.00	0.00
MMD	mean	α=0.05	0.00	**100.00**	60.88	69.43	**100.00**	**100.00**	89.99	0.00
		α=0.01	0.00	**100.00**	42.49	60.54	**100.00**	**100.00**	85.91	0.00
	sd	α=0.05	0.00	0.00	48.39	46.01	0.00	0.00	30.30	0.00
		α=0.01	0.00	0.00	47.05	48.96	0.00	0.00	35.02	0.00
KS	mean	α=0.05	4.32	**100.00**	99.60	99.84	**100.00**	**100.00**	**100.00**	21.93
		α=0.01	0.80	**100.00**	97.88	99.48	**100.00**	**100.00**	**100.00**	6.60
	sd	α=0.05	0.62	0.00	0.21	0.13	0.00	0.00	0.00	1.23
		α=0.01	0.30	0.00	0.51	0.23	0.00	0.00	0.00	0.85
Welch	mean	α=0.05	4.95	0.85	3.89	99.97	0.00	89.25	99.99	**100.00**
		α=0.01	0.97	0.05	0.69	99.84	0.00	65.04	99.98	**100.00**
	sd	α=0.05	0.64	0.29	0.63	0.05	0.00	1.11	0.03	0.00
		α=0.01	0.27	0.06	0.24	0.13	0.00	1.35	0.05	0.00
WMM	mean	α=0.05	5.02	45.25	7.00	99.96	**100.00**	33.81	99.99	30.87
		α=0.01	0.90	24.17	1.74	99.80	**100.00**	16.73	99.96	12.23
	sd	α=0.05	0.67	1.61	0.82	0.06	0.00	1.36	0.03	1.41
		α=0.01	0.24	1.55	0.39	0.14	0.00	0.98	0.07	0.96

chromosomes and control probes. Missing values in both groups are inferred using R package `pcaMethods`. To distinguish the mean, variance, and shape components, we standardized the values by $(X - E[X]) / \sqrt{V[X]}$ to remove mean and variance effects. Finally, 394,363 sites were used for further analysis.

4.2 Analysis results

Significant differential methylation sites were identified as those having p-values less than 1%. As a result, D^3M, Welch, and DiffVar detected 55,796, 254,334, and 178,395 sites, respectively. Among them, we investigated sites with the smallest 1,000 p-values, including 568, 543, and 513 genes with D^3M, Welch, and DiffVar, respectively. Heat map and Q-Q plots of the top 1,000 sites are shown in Figures 3 and 4. Comparing heat maps and Q-Q plots, the methylation patterns 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).

The Venn diagram shows the number of CpG sites tested for differential methylation using the three methods (Figure 2). The overlaps between D^3M, Welch, and DiffVar are small, indicating that the differential methylation sites based on the shapes include distinct information not relevant to Welch and DiffVar.

Among distributions of the top 1,000 sites, we can observe that there are mainly two distribution types in GBM, and we divide the 1,000 sites into two classes using the distributions in GBM. The clustering procedure is based on the Wasserstein metric (Irpino

and Verde, 2014b). Clusters 1 and 2 contain 713 and 287 sites, respectively. Typical distribution examples in each cluster are shown in Figure 5. Cluster 1 shows two modes for distributions in GBM, whereas cluster 2 shows heavy-tailed distributions in GBM.

Next, we perform enrichment analysis on gene sets in clusters 1 and 2. We used ingenuity pathway analysis (IPA) for 423 and 184 genes in clusters 1 and 2, respectively, and significantly enriched pathways in each cluster using Fisher's exact test. Table **??** shows five pathways and related genes, ranked with p-values in each cluster.

Nearly all the pathways in clusters 1 and 2 have been previously reported as significant pathways in GBM, even though we do not include any information on GBM. The axonal guidance signaling pathway in cluster 1 has been suggested as prompting the cell invasion of GBM (Dominique, *et al.,* 2007). The protein kinase A (PKA) pathway that is dysregulated has been considered to trigger the important steps to cancer genesis (Kiran, *et al.,* 2005), and Prasad, *et al.,* (2003) have indicated that PKA-activated c-AMP inhibits the proliferation and differentiation of GBM. The neuregulin signaling pathway in GBM is investigated by Patricia, *et al.,* (2003), and the effects of death receptor pathway dysregulation is mentioned in Murphy, *et al.,* (2013), Ziegler, *et al.,* (2008), and Krakstad, *et al.,* (2010). In cluster 2, the thioredoxin pathway has been found to play a key role in cancer, including GBM (Powis, *et al.,* 2007; Yacoub, *et al.,* 2010), and Lai, *et al.,* (2014) show that the transcriptional regulatory network in embryonic stem cells is the most significant pathway with genome-wide methylation analysis in GBM. The remaining pathways might be explained elsewhere. Our prediction using D^{3}M provides a hypothesis that DNA methylation in these pathways might cause the phenotypical difference between GBM and LGG.

We further focus on phosphatase and tensin homolog (PTEN) in neuregulin signaling and protein kinase A signaling pathways, and then compare the ranking based on p-value by D^{3}M with those by other methods. The methylation of PTEN promoter is frequent in LGG and secondary GBM patients, but rare in normal and *de novo* GBM patients (John, *et al.,* 2007). In our result, PTEN belongs to cluster 1, for which the distribution shape for LGG is bimodal, with the majority and minority being hyper- and hypo-methylation, respectively, and the distribution for GBM is unimodal with hypo-methylation. This suggests that demethylation of PTEN in some LGG might trigger transformation from LGG to GBM. PTEN is ranked 922^{nd} out of 394,363 sites (0.23%) with D^{3}M. However, PTEN is not included in the top 1,000 sites with Welch and Differ, being ranked $11{,}424^{th}$ out of 394,363 sites (2.89%) with Welch and $10{,}856^{th}$ out of 394,363 sites (2.75%) with DiffVar.

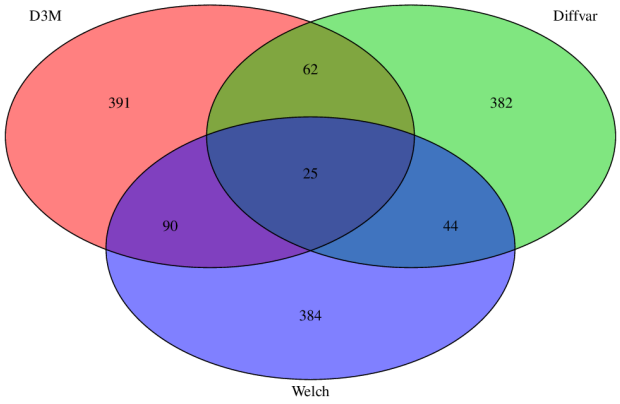

Fig. 2. Venn diagram of genesets with top 1000 sites

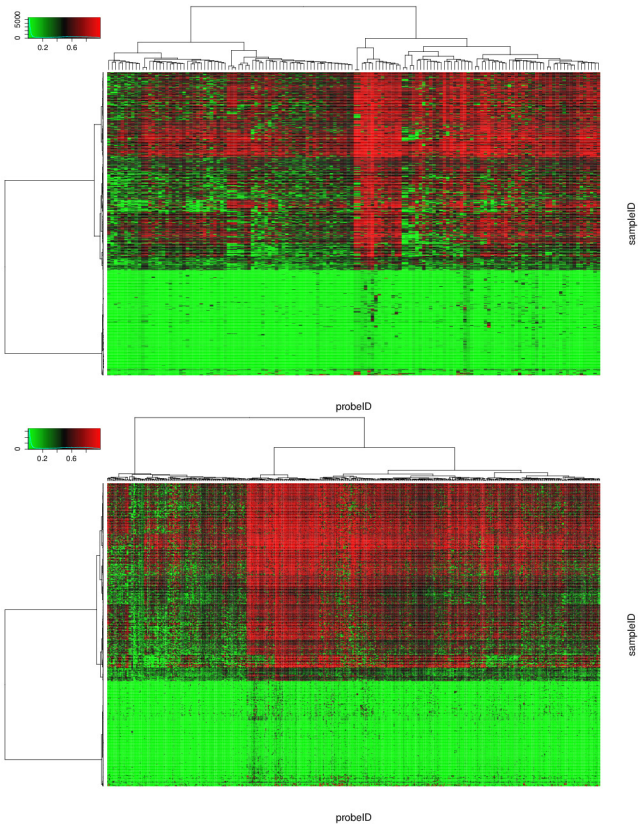

Fig. 3. Heat map of GBM 145 samples (upper) LGG 530 samples (lower) with top 1000 sites

5 DISCUSSION

Here we summarize the advantages and disadvantageseps of D^{3}M, DiffVar, and MMD, which have all been recently developed. These methods are designed for detecting differential methylation patterns focusing on cancer heterogeneity, which is caused by epigenetic instability and diversity. Cancer heterogeneity can often be confused with outliers. In fact, in our simulations and real data analysis, DiffVar, which is robust to outliers, regards important features of heterogeneity as outliers, and as a result, it fails to detect differential methylation sites. For example, DiffVar detects simulation case 2 as differential methylation, even though we set the mean and variance, but not 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

Table 3. Pathways detected with the proposed method

Cluster	Pathway	$-log(P$-value)	Genes
Cluster 1	Axonal Guidance Signaling	3.96	C9orf3, NFATC4, PLCD1, EFNB2, SEMA6B, GNAO1 SEMA3E, EPHB4, ADAM8, NTN1, TUBA8, ITGA5 ITGA2, EPHA2, NFATC1, MET, EFNA1, PDGFA PRKCZ, BMP7, SEMA5A
	Proten Kinase A Signaling	3.73	NFATC4, PLCD1, PTPN14, CDC14B, PTEN, PDE4A PYGL, NTN1, PTPRN, TGFB2, NFATC1, PDE8A DUSP5, CNGA3, PDE4D, PTPRA, SIRPA, PRKCZ ADCY9
	Neuregulin Signaling	3.53	PTEN, PICK1, ITGA2, NRG3, NRG2 PRKCZ, ITGA5, GRB7
	Death Receptor Signaling	3.40	CFLAR, ACTG1, ACTC1, TNFSF10, PARP14, CASP8 BIRC3, CASP6
	Adipogenesis pathway	3.10	BMPR2, NFATC4, ARNTL, CTBP2, ZNF423, KLF5 RPS6KA1, BMP7, FGFRL1
Cluster 2	Thioredoxin Pathway	3.01	NXN, TXNRD1
	Transcriptional Regulatory Network in Embryonic Stem Cells	2.37	MEIS1, ZFHX3, SET
	Vitamin-C Transport	2.25	NXN, TXNRD1
	Hepatic Fibrosis / Hepatic Stellate Cell Activation	2.24	KLF6, BCL2, COL21A1, TGFB2, COL9A1, COL9A2
	Factors Promoting Cardiogenesis in Vertebrates	2.16	BMP8A, TGFB2, PRKCB, DKK1

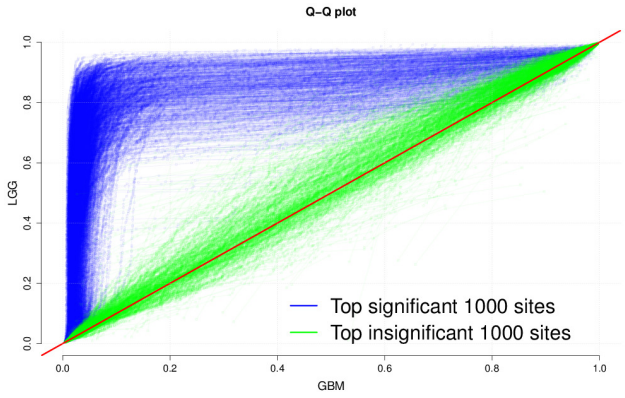

Fig. 4. Q-Q plot of significance and insignificance for top 1,000 sites

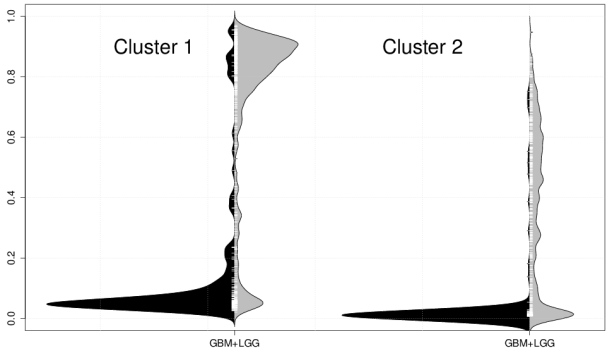

Fig. 5. Distribution instance in clusters 1 and 2

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 is designed to detect higher-order changes, such as shape in methylation profiles based on kernels (Mayo, *et al.,* 2014). However, in our simulation, p-value does not work in the sense of type I error control. M^{3}D based on MMD also cannot derive p-values, substantially just ordering distances over regions. Then, we cannot evaluate error rates probabilistically, which could be a crucial disadvantage when working with actual data.

D^{3}M detects differences of all moments with underlying distributions based on the Wasserstein metric.

Simulation results indicate that D^{3}M can detect not only shape differences but also mean and variance differences, as effectively as Welch and DiffVar. Thus, the proposed method can be applied to differential methylation analysis for general purposes. The limitation of D^{3}M is that it requires sufficient sample size to construct distribution values to some extent. Empirically, because quantiles are used in the calculation of the Wasserstein metric, it requires at least 100 samples. The statistical test 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).

6 CONCLUSION

In this study, we proposed a novel method, D^{3}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^{3}M can detect differential methylation sites in various cases of distributions for which other methods, Welch, DiffVar, KS, MWW, and MMD, failed.

In the application to the GBM and LGG dataset in the TCGA cohort, we identified 1,000 sites with the smallest p-values. Most of the sites detected by D^{3}M show strong heterogeneity and tend

to be hyper- and hypo-methylated in LGG and GBM, respectively, as found in previous studies. Furthermore, mean-, variance-, and shape-based methods mutually detected differential methylation sites, because overlapped sites included up to approximately 20% of each other. Thus, distribution shape differences can provide new insights regarding methylation patterns.

Since the GBM and LGG dataset contains a large number of significantly different sites, including 55,796, 254,334, and 178,395 sites for D^3M, Welch, and DiffVar, respectively, at the 1% significance level, it is difficult to understand the methylation patterns at these sites. In the future, it would be of interest to develop a method that describes the diversity of methylation patterns.

REFERENCES

Aggarwal, C. C. (2013) *Outlier Analysis*, Springer New York.

Anders, S., Huber, W. (2010) expression analysis for sequence count data. *Genome Biology*, **11**:R106.

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., Irizarry, R. A. (2014) Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, **30**, 1363-1369.

Billard, L. and Diday, E. (2006) *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, Wiley Chichester.

Bock, H. H. and Diday, E. (2000) *Analysis of Symbolic Data*, Springer, Berlin Heidelberg.

Diday, E. (1989) Introduction a l'analyse des donnees symboliques. RR-1074,inria-00075485.

Furnari, F. B., Fenton, T., Bachoo, R. M., Mukasa, A., Stommel, J. M., Stegh, A., Hahn, W. C., Ligon, K. L., Louis, D. N., Brennan, C., Chin, L., DePinho, R. A. and Cavenee, W. K. (2007) Malignant astrocytic glioma: genetics, biology, and paths to treatment. *Genes Dev*, **21**(21), 2683-710.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. and Smola, A. (2012) A Kernel Two-Sample Test. *Journal of Machine Learning Research*, **13**, 723-773.

Irpino, A. and Verde, R. (2014a) Basic statistics for distributional symbolic variables: a new metric-based approach, *Adv Data Anal Classif.* **9**, 143-175.

Irpino, A., Verde, R. and De Carvalho, Francisco de A.T. (2014b) Dynamic clustering of histogram data based on adaptive squared Wasserstein distances, *Expert Systems with Applications*, **41**(7),3351-3366.

Kampstra, P. (2008) Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Jornal of Statistical Software*, **28**, Code Snippet 1.

Karatzoglou, A., Smola, A., Hornik, K. and Zeileis, A. (2004) kernlab – An S4 Package for Kernel Methods in R. *Jornal of Statistical Software*, **11**(9).

Knijnenburg, T. A., Wessels, L. F. A. , Reinders, M. J. T. and Shmulevich, I. (2009) Fewer permutations, more accurate P-values. *Bioinfomatics*, **25**, ISMB 2009, i161-i168.

Krakstad, C, Chekenya, M. (2010) Survival signaling and apoptosis resistance in glioblastomas: opportunities for targeted therapeutics. *Mol Cancer.* ,**9**:135.

Lai, R. K., Chen, Y., Guan, X., Nousome, D., Sharma, C., Canoll, P., Barnholtz-Sloan, J. (2014). Genome-Wide Methylation Analyses in Glioblastoma Multiforme. *PLoS ONE*, **9**(2), e89376.

Mayo, T. R., Schweikert, G. and Sanguinetti, G. (2014) M3D: a kernel-based test for spatially correlated changes in methylation profiles. *Bioinfomatics*, **31**(6), 809-816.

Mucignat-Caretta, C., Cavaggioni, A., Redaelli, M., Malatesta, M., Zancanaro, C., and Caretta, A. (2008) Selective distribution of protein kinase A regulatory subunit RIIα in rodent gliomas. *Neuro-Oncology*, **10**(6), 958-967.

Murphy, Á. C., Weyhenmeyer, B., Schmid, J., Kilbride, S. M., Rehm, M., Huber, H. J., Senft, C., Weissenberger, J., Seifert, V., Dunst, M., Mittelbronn, M., Kögel, D., Prehn, J. H. M. and Murphy, B. M. (2013) Activation of executioner caspases is a predictor of progression-free survival in glioblastoma patients: a systems medicine approach. *Cell Death & Disease*, **4**(5), e629.

Nadella, K. S., and Kirschner, L. S. (2005) Disruption of Protein Kinase A Regulation Causes Immortalization and Dysregulation of D-Type Cyclins. *Cancer Res.*, **65**:10307-10315.

Noirhomme-Fraiture, M and Diday, E (2008) *Symbolic Data Analysis and the SODAS Software*, Wiley Chichester.

Phipson, B. and Oshlack, A. (2014) DiffVar: a new method for detecting differential variability with application to methylation in cancer and aging. *Genome Biology*, **15**, 465.

Powis, G., Kirkpatrick, D. L. (2007) Thioredoxin signaling as a target for cancer therapy. *Curr Opin Pharmacol*, **7**, 392-7.

Prasad, K. N., Cole, W. C., Yan, X. D., Nahreini, P., Kumar, B., Hanson, A., Prasad, J. E. (2003) Defects in cAMP-pathway may initiate carcinogenesis in dividing nerve cells: a review. *Apoptosis*. **8**, 579-586.

Ramsay, J. O. and Silverman, B. W. (2005) *Functional Data Analysis* (2^{nd} edition). Springer-Verlag.

Ritch, P. A., Carroll, S. L., and Sontheimer, H. (2003) Neuregulin-1 Enhances Motility and Migration of Human Astrocytic Glioma Cells. *The Journal Of Biological Chemistry*, **278**(23), 20971 - 20978.

Robinson, M. D., McCarthy, D. J., Smyth, G. K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139-140.

Rueshendorff, L. (2011) Wasserstein metric, *Encyclopedia of Mathematics*.

Smyth, G. K.: Gentleman, R., Carey, V., Dudoit, S., Irizarry, R. and Huber, W. (eds.). (2005) Limma: linear models for microarray data. *In Bioinforma Comput Biol Solut using R Bioconductor*, 397-420, Springer New York.

The Cancer Genome Atlas Research Network. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**(7216), 1061-1068.

Wang, H. and Marron, J. S. (2007) Object Oriented Data Analysis: Sets Of Trees. *The Annals of Statistics*, **35**(5), 1849-1873.

Wiencke, J. K., Zheng, S., Jelluma, N., Tihan, T., Vandenberg, S., Tamgüney, T., Baumber, R., Parsons, R., Lamborn, K. R., Berger, M. S., Wrensch, M. R., Haas-Kogan, D. A. and Stokoe, D. (2003) Methylation of the PTEN promoter defines low-grade gliomas and secondary glioblastoma. *Neuro Oncol*, **9**(3), 271-279.

Yacoub, A., Hamed, H. A., Allegood, J., Mitchell, C., Spiegel, S., Lesniak, M. S., Ogretmen, B., Dash, R., Sarkar, D., Broaddus, W. C., Grant, S., Curiel, D. T., Fisher, P. B. and Dent, P. (2010) PERK-dependent regulation of ceramide synthase 6 and thioredoxin play a key role in mda-7/IL-24-induced killing of primary human glioblastoma multiforme cells. *Cancer Res*, **70**(3), 1120-9.

Ziegler, D. S., Kung, A. L., Kieran, M. W. (2008) Anti-apoptosis mechanisms in malignant gliomas. *J Clin Oncol*, **26**, 493-500.