bioRxiv preprint doi: https://doi.org/10.1101/023879; this version posted December 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

BIOINFORMATICS

Vol. 00 no. 00 2015
Pages 1–8

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

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.

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 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 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 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 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 the 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 cancer 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

© Oxford University Press 2015.

bioRxiv preprint doi: https://doi.org/10.1101/023879; this version posted December 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 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. 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:

$$Beta_i = \frac{max(z_{i,methy}, 0) + \alpha}{max(z_{i,unmethy}, 0) + max(z_{i,methy}, 0) + \alpha} \quad (1)$$

where $z_{i,methy}$ and $z_{i,unmethy}$ are the intensity measured by i-th methylated and unmethylated probes, respectively and α 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

$$\begin{aligned} F_i(x) &= \Pr\{X(s_i) \le x\}; x \in [0, 1]. \\ G_i(y) &= \Pr\{Y(s_i) \le y\}; y \in [0, 1]. \end{aligned} \quad (2)$$

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;

$$\begin{aligned} \hat{F}_i(x) &:= \frac{1}{n}\sum_{j=1} \mathbf{1}(x_j(s_i) \le x) \\ \hat{G}_i(y) &:= \frac{1}{m}\sum_{j=1} \mathbf{1}(y_j(s_i) \le y) \end{aligned} \quad (3)$$

where

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

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 $1 \le q \le 2$ 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):

$$\begin{aligned} d^2(F_i, G_i) &= \int_0^1 |F_i^{-1}(u) - G_i^{-1}(u)|^2 du \\ &= (\mu_i - \mu_i')^2 + (\sigma_i - \sigma_i')^2 + 2\sigma_i\sigma_i'(1 - \rho_{i,i'}) \end{aligned} \quad (6)$$

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

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

Technically, we use quantiles to compute the approximation of the (7) 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 (7). 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.

$$\begin{array}{ll} \text{Null hypothesis:} & F_i = G_i \\ \text{Alternative hypothesis:} & F_i \ne G_i \end{array} \quad (8)$$

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

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

Approximation of (9) uses the subset of D_i^*, $\tilde{d}_{i,1}^*, \tilde{d}_{i,2}^*, \ldots, \tilde{d}_{i,B}^*$ where $B \le B_{all}$:

$$P_{sub}(d_i) = \frac{\sum_{b=1}^{B} \mathbf{1}(\tilde{d}_{i,b}^* \ge d_i)}{B}. \quad (10)$$

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.

bioRxiv preprint doi: https://doi.org/10.1101/023879; this version posted December 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

We use an exponential distribution to estimate the distribution tail as follows,

$$P(d_i) = \begin{cases} \frac{1}{B}\sum_{j=1} \mathbf{1}(\tilde{d}^*_{i,b} \geq d_i) & \text{for} \quad d_i < d_i^{(\min)} \\ \exp(-\lambda_i(d_i - d_i^{(\min)})) & \text{for} \quad d_i \geq d_i^{(\min)} \end{cases} \tag{11}$$

where λ_i is a scale parameter and $d_i^{(\min)}$ is a threshold that we set to 99th 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

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 $H_0 : F_i = G_i \leftrightarrow H_1 : F_i \neq G_i$ 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 D³M, MMD, DiffVar, KS, EB, and MWW are close to the significance levels, which indicates that they effectively control type I errors.

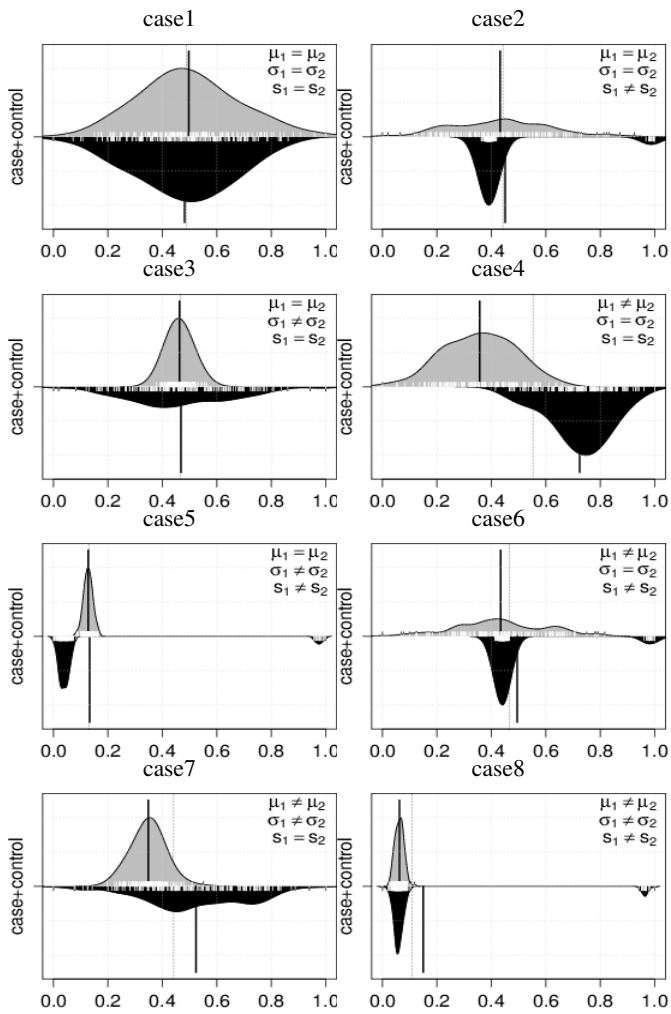

Fig. 1. The beanplot of eight cases

Furthermore, we investigate the power for cases 2 – 8. KS test and MMD show the relatively good performance such as case 2 where the only distribution shape is different. However, they cannot capture the feature of case 8 where the majority of the two groups are overlapped with each other although 15% of minority of distribution exists. Besides, KS testing fails to detect case 6 where 15% of minority distribution is hyper-methylated. DiffVar shows high power for cases where the variances differ, however, it might capture the other distribution features for the cases with equal variances (case 6), leading to uninterpretable results. EB can appropriately distinguish only the mean difference. MWW can detect case 4, and 7, but cannot detect cases in which the means differ under non-normality. The proposed D³M provides preferable results over the eight cases.

On one hand, from this simulation, most of all the distribution features can be captured with D³M, which indicates that it works well for the given situations and it can be applied to various cases flexibly. On the other hand, a small sample situation should also be examined. We conduct the simulation considering such a situation

bioRxiv preprint doi: https://doi.org/10.1101/023879; this version posted December 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Table 1. Simulation models of eight cases

	$F_i = F_2$	$F_1 \neq F_2$						
	case 1	case 2	case 3	case 4	case 5	case 6	case 7	case 8
$\mu_1 = \mu_2$	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. Hypothesis testing in each case (hypotheses are described under the method names). We focus on a probe and generate beta values, 40 samples each in a case and a control group respectively. We apply six methods to this data and evaluate the p-value at the significance level, *i.e.,* 5%. This process is repeated 50 times and returns the averages of type I error (%) and power (%). We represent the power in italics within top three. Furthermore, we also evaluate the standard deviations of the averages; we repeated the process of the evaluating average of type I error and power a 100 times and then obtained the standard deviations of the averages.

		Type I	Power						
		case 1	case 2	case 3	case 4	case 5	case 6	case 7	case 8
D^3M	mean	5.12	*93.20*	*95.42*	*84.52*	*98.08*	*84.88*	*99.12*	*99.96*
$H_0 : F_1 = F_2 \quad H_1 : F_1 \neq F_2$	sd	0.07	0.08	0.07	0.11	0.06	0.11	0.03	0.01
DiffVar	mean	1.80	6.80	*93.86*	1.30	*97.62*	11.84	68.26	*86.64*
$H_0 : V(F_1) = V(F_2) \quad H_1 : V(F_1) \neq V(F_2)$	sd	0.04	0.08	0.08	0.04	0.06	0.10	0.14	0.17
MMD	mean	4.40	*100.00*	11.62	9.66	*73.50*	*93.08*	16.36	51.74
$H_0 : F_1 = F_2 \quad H_1 : F_1 \neq F_2$	sd	0.07	0.00	0.10	0.10	0.14	0.08	0.11	0.17
KS	mean	2.56	*99.86*	*32.06*	67.06	73.06	60.22	*89.62*	73.88
$H_0 : F_1 = F_2 \quad H_1 : F_1 \neq F_2$	sd	0.05	0.01	0.15	0.14	0.26	0.16	0.09	0.25
EB	mean	4.94	0.66	5.18	*87.92*	0.10	*78.50*	*87.36*	*96.76*
$H_0 : E(F_1) = E(F_2) \quad H_1 : E(F_1) \neq E(F_2)$	sd	0.07	0.02	0.07	0.10	0.01	0.13	0.11	0.07
MWW	mean	4.68	11.22	5.76	*84.96*	53.36	55.50	83.94	71.36
$H_0 : M(F_1) = M(F_2) \quad H_1 : M(F_1) \neq M(F_2)$	sd	0.07	0.10	0.07	0.12	0.21	0.16	0.12	0.20

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 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. 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* $\leq 1\%$. 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.

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.

bioRxiv preprint doi: https://doi.org/10.1101/023879; this version posted December 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 Q–Q plot into two classes; we consider input data as $Z_i = (Q_{i,1}, Q_{i,2}, \ldots, Q_{i,K}, Q'_{i,1}, Q'_{i,2}, \ldots, Q'_{i,K})(i = 1, 2, \ldots, 1000)$ and define the dissimilarity as $\sqrt{(Z_i - Z_j)^2}\,(i \neq j)$ where $Q_{i,k}$ and $Q'_{i,k}\,(k = 1, 2, \ldots, K)$ is $k-$th quantile in LGG and GBM group, respectively. Then we apply the standard hierarchical clustering (Clusters 1 and 2 in Figure 3 (A)). Typical distribution examples in each cluster are shown in Figure 3 (C). Clusters 1 and 2 contain 626 and 374 sites, respectively.

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 pathways and related genes, ranked with q-values in each cluster.

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 signaling pathway in cluster 1 has been suggested as prompting the cell invasion of GBM (Dominique, *et al.,* 2007) and ERK/MAPK Signaling is reported to be up-regulated in GBM (Liu *et al.* 2013). The enrichments of Caveolar-mediated Endocytosis Signaling and Calcium Signaling are studied in (Dong, *et al.,* 2010;Polisetty *et al.,* 2012). The remaining pathways might be explained elsewhere. Our prediction using D^3M provides a hypothesis that DNA methylation in these pathways might cause the phenotypical difference between 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^3M 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^3M.

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^3M, EB, KS, and DiffVar is shown in Table3.

Table 3. The ranking of the six Wnt gene probes in 351,932 probes. For each method, the upper values are the absolute ranking among 351,932 of the genes, and lower are the their percentages.

	Wnt2	Wnt2b	Wnt3	Wnt4	Wnt7a	Wnt9a
D^3M	342	549	591	829	641	887
	(0.10)	(0.16)	(0.17)	(0.24)	(0.18)	(0.25)
EB	22,541	12,914	3,246	1,811	135	20,747
	(6.40)	(3.67)	(0.92)	(0.51)	(0.04)	(5.90)
KS	4,754	91,240	50,174	41,589	41,066	84,248
	(1.35)	(25.93)	(14.26)	(11.82)	(11.67)	(23.94)
DiffVar	56,950	155,621	147,817	146,033	138,928	242,648
	(16.18)	(44.22)	(42.00)	(41.49)	(39.48)	(68.95)

5 DISCUSSION

Here, we summarize the advantages and disadvantages of D^3M, DiffVar as well as MMD. These methods are designed for detecting differential methylation levels focusing on cancer heterogeneity, 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^3M shows good performance over all eight 8 cases. Specifically, D^3M 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^3M 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^3M in case 7 and case 8.

We also investigate the power of D^3M 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^3M 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), D3M and EB shows the good detection powers.

D^3M can be applied flexibly to differential methylation problems. Simulation results indicate that D^3M 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

bioRxiv preprint doi: https://doi.org/10.1101/023879; this version posted December 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

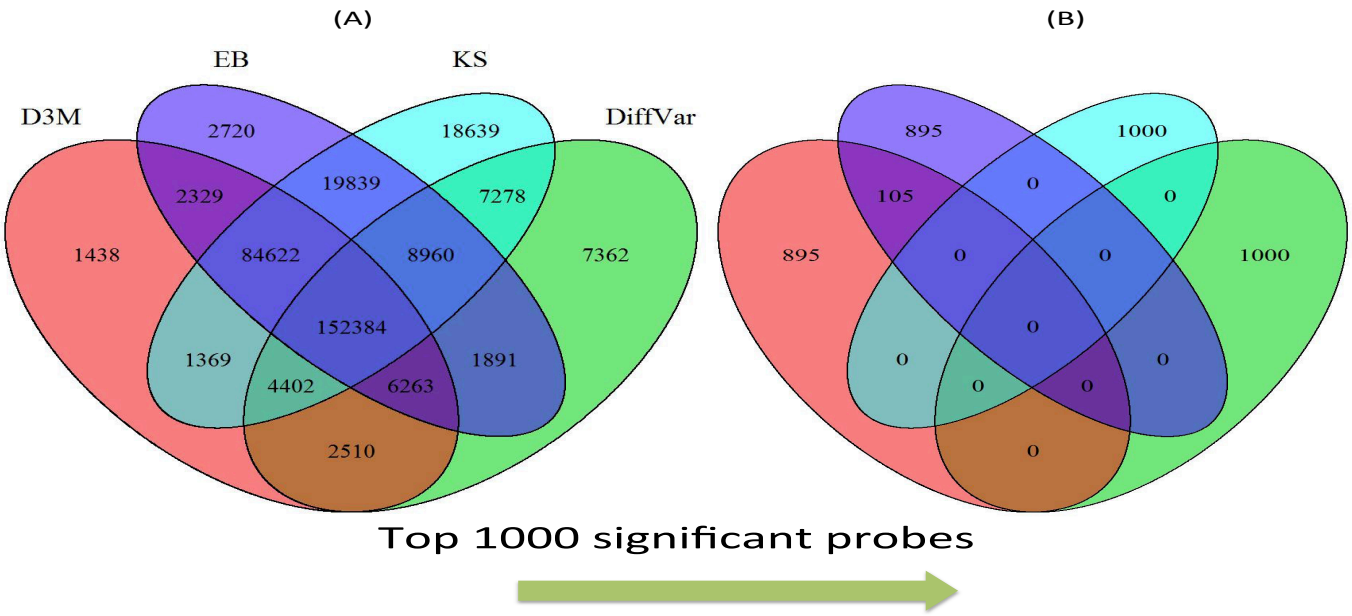

Fig. 2. Venn diagram of detected sites at the significance level 1% (A) and of top significant 1000 sites (B).

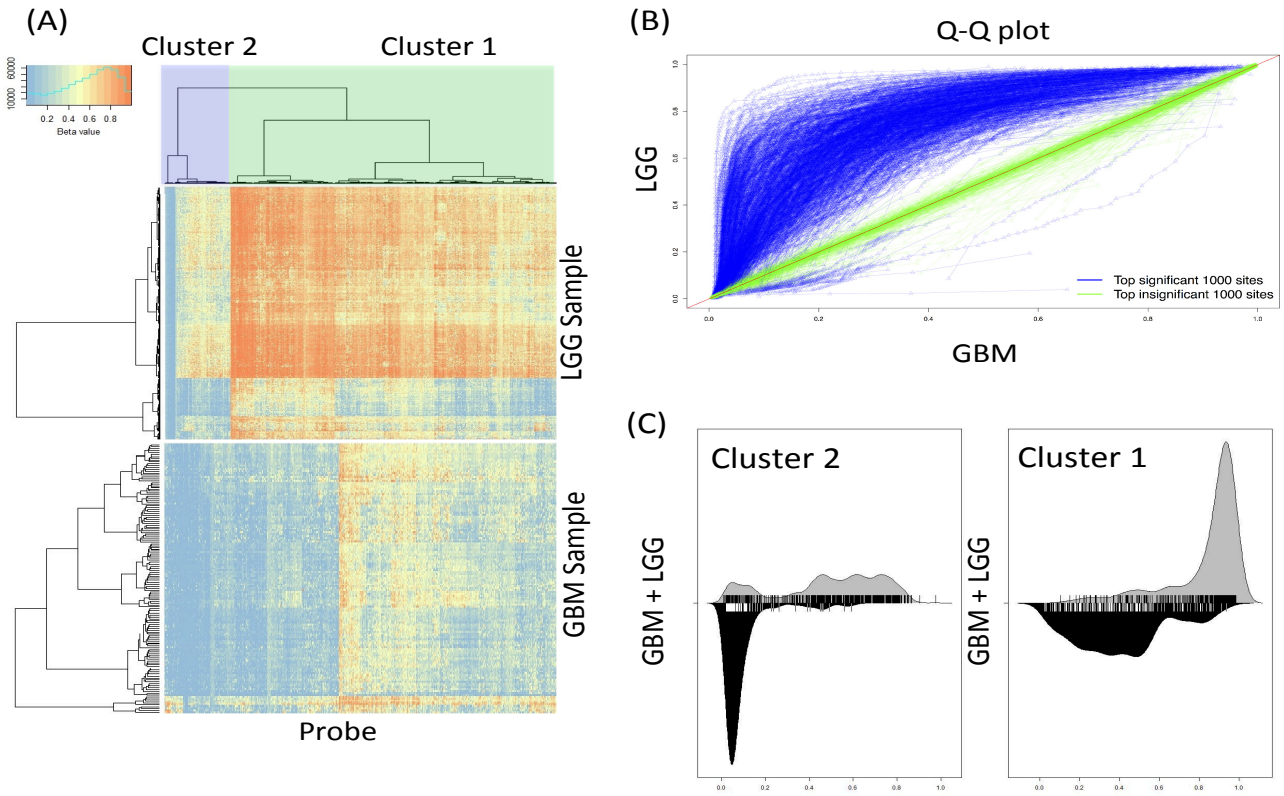

Fig. 3. Distributions of the top 1,000 significant probes. **(A)** The heat map of beta values with 141 LGG and 530 GBM samples at each probe respectively. We jointly clustered quantiles of LGG and GBM probes into the two clusters based on the Wasserstein metric (*i.e.,* we clustered curves in Q-Q plots of the top significant probes into two clusters). **(B)** Q–Q plots of beta values between the LGG and GBM probes (blue lines). D^3M 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. D^3M recognizes strongly heterogeneous distributions and detects such probes as top significant probes.

bioRxiv preprint doi: https://doi.org/10.1101/023879; this version posted December 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Table 4. Pathways detected with the proposed method

Cluster	Pathway	$-log(P$-value)	Genes
Cluster 1	Axonal Guidance Signaling	5.85	SLIT3, ITGB1, SEMA3G, MYL10, WNT3, FES, NRP2, BMP8A, UNC5B WNT2B, PIK3R5, GNAI1, ITGA5, KEL, WNT7A, MAG, ADAM12 NTRK1, PRKAG2, WNT4, BMP7, NTN3, PRKCB
	ERK/MAPK Signaling	4.09	ITGB1, DUSP9, PLA2G4B, PRKAG2, ITGA5, PIK3R5, RAPGEF4, ESR1 PPP2R5A, KSR1, JMJD7-PLA2G4B, PRKCB
	Caveolar-mediated-Endocytosis Signaling	3.72	ITGB1, INS, CD48, ITGA5, ITGB8, JMJD7-PLA2G4B, PRKCB
	Human Embryonic Stem-Cell Pluripotency	3.35	WNT7A, WNT3, BMP8A, SMAD3, NTRK1, WNT2B, PIK3R5, WNT4 BMP7
Cluster 2	TREM1 Signaling	3.55	STAT5A, MPO, CASP1, NOD1, CCL3, ITGAX
	Calcium Signaling	3.28	GRIN2A, TNNT3, ITPR2, CHRNB1, NFATC4, CAMKK2, PPP3CA CAMKK2, PPP3CA, MEF2B

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 $X - \mathrm{E}[X]$ in each group at a site before applying D^3M. This option is provided with R package D^3M; in the function `D3M::d3m`, the logical parameters of `rm.mean` and `rm.var` exists. If we would like to see the variance and shape differences at the same time, we just set the `rm.mean=T` and `rm.var=F`.

The statistical test of D^3M 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^3M 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^3M 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^3M 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^3M will be covered in future work.

6 CONCLUSION

In this study, we proposed a novel method, D^3M, 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^3M 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^3M 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^3M, EB, KS, and DiffVar, respectively; at the FDR $\leq 1\%$, 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

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

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.

Baylin,S.B. (2005) DNA methylation and gene silencing in cancer. *Nat Rev Clin Oncol*, **2**, S4-S11.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc B*, **57**(1), 289-300.

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

Kulis,M. and Esteller,M.(2010) DNA methylation and cancer. *Adv Genet*, 70:27-56.

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

Dong,H. *et al.*(2010) Integrated analysis of mutations, miRNA and mRNA expression in glioblastoma. *BMC Syst Biol*, **4**:163.

Du,P. *et al.* (2010) *BMC Bioinformatics*, 11:587.

Furnari,F.B. *et al.* (2007) Malignant astrocytic glioma: genetics, biology, and paths to treatment. *Genes Dev*, **21**(21), 2683-710.

Gretton,A. *et al.* (2012) A Kernel Two-Sample Test. *J Mach Learn Res*, **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.

Kampstra,P. (2008) Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *J STAT SOFTW*, **28**, Code Snippet 1.

Knijnenburg,T. A. *et al.* (2009) Fewer permutations, more accurate P-values. *Bioinformatics*, **25**, ISMB 2009, i161-i168.

Liu,T. *et al.* (2013) Transcriptional signaling pathways inversely regulated in Alzheimer's disease and glioblastoma multiform. *Sci Rep*, **3**.

Müller,A. (1997). Integral Probability Metrics and Their Generating Classes of Functions. *Adv Appl Prob.*, **29**, 429-443.

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

bioRxiv preprint doi: https://doi.org/10.1101/023879; this version posted December 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Genome Biol, **15**, 465.

Polisetty,R.V. *et al.* (2012) LC-MS/MS analysis of differentially expressed glioblastoma membrane proteome reveals altered calcium signalling and other protein groups of regulatory functions. *Mol Cell Proteomics* **11**(6).

Rampazzo,E. *et al.* (2013) Wnt activation promotes neuronal differentiation of Glioblastoma. *Cell Death Dis*, **4**, 3:500e

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

Ritchie,M.E. *et al.* (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*, **43**(7), e47.

Ritch, P. A. *et al.* (2003) Neuregulin-1 Enhances Motility and Migration of Human Astrocytic Glioma Cells. *J Biol Chem*, **278**(23), 20971 - 20978.

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

Stacklies,W. *et al.* (2007) pcaMethods – a Bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, **23**(9), 1164-1167.

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. *Ann Stat*, **35**(5), 1849-1873.

Wang,D. *et al.* (2012) IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. *Bioinformatics*, **28**(5), 729-730.