Principal component analysis- and tensor decomposition-based unsupervised feature extraction to select more suitable differentially methylated cytosines: Optimization of standard deviation versus state-of-the-art methods

Y-H. Taguchi[a,*], Turki Turki[b,2]

[a]Department of Physics, Chuo University, 1-13-27, Kasuga, Bunkyo-ku, Tokyo, 112-8551, Japan
[b]Department of Computer Science, King Abdulaziz University, Jeddah, 21589, Saudi Arabia

ARTICLE INFO

Keywords:
Tensor Decomposition
Unsupervised Learning
DNA Methylation
Methylation Profiles
Applications in Epigenetics

ABSTRACT

In contrast to RNA-seq analysis, which has various standard methods, no standard methods for identifying differentially methylated cytosines (DMCs) exist. To identify DMCs, we tested principal component analysis and tensor decomposition-based unsupervised feature extraction with optimized standard deviation, which has been shown to be effective for differentially expressed gene (DEG) identification. The proposed method outperformed certain conventional methods, including those that assume beta-binomial distribution for methylation as the proposed method does not require this, especially when applied to methylation profiles measured using high throughput sequencing. DMCs identified by the proposed method also significantly overlapped with various functional sites, including known differentially methylated regions, enhancers, and DNase I hypersensitive sites. The proposed method was applied to data sets retrieved from The Cancer Genome Atlas to identify DMCs using American Joint Committee on Cancer staging system edition labels. This suggests that the proposed method is a promising standard method for identifying DMCs.

1. Introduction

Identification of aberrant methylation (Chen, Lin and Fann, 2016) is important as it plays a critical role in various biological processes. In contrast to the identification of differentially expressed genes (DEGs), effective methods for that of aberrant methylation remain unknown, since it is not directly related to functional units such as genes. There are two approaches to identify aberrant methylation: the identification of differentially methylated cytosines (DMCs) or differentially methylated regions (DMRs). Although these two factors may differ from each other, even if individual cytosines within a region are not differentially methylated and the summation of methylation over the region is differential, it can be considered a DMR. In contrast, even if a region is not a DMR, individual cytosines within the region can be DMCs. This inconsistency provides two ways to identify DMRs. One is a direct method, used for identifying DMRs without identifying individual DMCs. Another is to identify a DMR as an integration of the identified DMCs. Currently, there is no definite consensus on which is the most effective. The identification of DMCs is generally more difficult than that of DMRs due to the huge number of possible methylated CpG sites. For example, there are approximately 28.3 million CpG sites that are predicted to be easily methylated in humans (Luo, Lu and Xie, 2014), which is much larger than the number of human DMRs, which ranges from a few hundred (Skinner, Nilsson, Sadler-Riggleman, Beck, Maamar and McCarrey, 2019) to 20,000 (Shnorhavorian, Schwartz, Stansfeld, Sadler-Riggleman, Beck and Skinner, 2017). As a result, the identification of DMCs is much harder than that of DMRs. To overcome this difficulty, since the microarray era, various computational methods have been proposed to identify DMCs. For example, ChAMP (Tian, Morris, Webster, Yang, Beck, Feber and Teschendorff, 2017) and COHCAP (Warden, Lee, Tompkins, Li, Wang, Riggs, Yu, Jove and Yuan, 2013), developed by illumine, are some of the oldest methods that have been developed to identify DMCs in a specific methylation microarray architecture. For HTS data sets, DMRcate (Peters, Buckley, Chen, Smyth, Goodnow and Clark, 2021), DSS (Park and Wu, 2016), and metilene (Jühling, Kretzmer, Bernhart, Otto, Stadler and Hoffmann, 2015) are some of the most popular methods. However, conventional methods, including these, have common problems that need to be addressed.

1. Conventional methods assume the unrationalized and complicated properties associated with the distribution of methylation.
2. In addition, conventional methods require extended processing times (For example, when we analyzed the GSE77965 data set below, the proposed method required 60 sec for completion whereas ChAMP and COHCAP took 208 sec and 160 sec, respectively. Thus, the conventional methods are two to three times slower) .

To address these problems, we applied principal component analysis (PCA)- and tensor decomposition (TD)-based unsupervised feature extraction (FE) with optimized standard deviation (SD) to identify DMCs. This approach was

*Corresponding author
tag@granular.com (Y-H. Taguchi); tturki@kau.edu.sa (T. Turki)
https://researchmap.jp/Yh_Taguchi/ (Y-H. Taguchi)
ORCID(s): 0000-0003-0867-8986 (Y-H. Taguchi)

recently successfully applied to identify differentially expressed genes (Taguchi and Turki, 2022). We found that the proposed method can outperform modern methods, especially when applied to HTS data.

2. Materials and Methods

2.1. Methylation profiles

We used the following four methylation profiles as representatives. We considered profiles previously used for benchmarking or evaluation of previously proposed tools to avoid employing erroneous profiles. In other words, they were previously confirmed to include several DMCs under effective analysis.

2.1.1. GSE77965

This methylation profile was previously used in a benchmark study (BLUEPRINT consortium, 2016) and is available in GEO. From this data, we retrieved six colorectal cancer samples (GSM2062665 to GSM2062670) and six adjacent normal colon tissue samples (GSM2062671 to GSM2062676), and methylation levels were measured using the Illumina HumanMethylation450 BeadChip. Series matrix GSE77965_series_matrix.txt.gz was used.

2.1.2. GSE42308

This methylation profile was used for the evaluation by COHCAP and is available in GEO. It is composed of three HCT116 cell lines and three HCT116 mutants, also measured using the Illumina HumanMethylation450 BeadChip. Series matrix GSE42308_series_matrix.txt.gz was used.

2.1.3. GSE34864

A part of this methylation profile that was originally reported (Smith, Chan, Mikkelsen, Gu, Gnirke, Regev and Meissner, 2012) was used for benchmarking in another study (Feng, Conneely and Wu, 2014) and is available in GEO. We used two oocyte (GSM856495 and GSM856495) and two zygote profiles (GSM856501 and GSM856502) measured by HTS. Four files with names that start with "GSM" and end in "txt.gz" were used.

2.1.4. EH1072

This methylation profile was included in ExperimentHub (Morgan and Shepherd, 2022) in the Bioconductor and has been used for the evaluation of DMRcate. It has also been analyzed by HTS. It is composed of three Treg cell profiles taken from fat, liver, skin, and lymph nodes and T-cell control taken from lymph nodes (15 samples). For the comparison of conventional methods with the proposed method, since these profiles can only be compared pair-wise, only DMCs between Treg and T-cell control taken from lymph nodes were compared when methods other than the proposed method were employed.

2.1.5. TCGA

Since not all the datasets in TCGA are associated with "ajcc staging system edition", which was decided to employ in the below. we included the following two TCGA data sets: breast and colon cancer, which are well associated with "ajcc staging system edition". After selecting files, we downloaded the Manifest file for batch retrieval using gdc-client. Sample and clinical files were also downloaded for the downstream analyses.

To retrieve desired cancer data sets, we used the following parameters in the GDC data portal.

- Data category: dna methylation
- Data Type : Methylation Beta Value
- Platform : illumina human methylation 450

In addition to those defined above, in order to retrieve the specified cancer data set, we used the following parameters for two data sets.

Breast

- Primary Site: breast
- project : TCGA-BRCA

this results in 895 files.

Colon

- Primary Site: colon
- project : TCGA-COAD

this results in 346 files.

2.2. PCA-based unsupervised FE with optimized standard deviation

A full explanation is provided in our previous paper (Taguchi and Turki, 2022) and our recent book (Taguchi, 2020). In brief, the purpose of the method is to select a limited number of variables associated with the classification; in this study, the classifications were clinical, medical, or biological information, e.g., healthy controls and patients. In this specific study, the included variables were the amounts of methylation of individual cytosines within the genome. Since it is unlikely that all cytosine methylation biologically contributes to the difference between classifications, it was expected that a limited number of cytosines would be selected to represent the distinction between classes. This was achieved by finding principal components (PCs) that coincided with the classification and selecting cytosines whose contribution to the corresponding PCs was attributed to individual cytosines. Evaluation of the contribution was performed by relating P-values to individual cytosines assuming Gaussian distribution; cytosines with small enough P-values were selected as they should have greater contributions than chance alone. See Algorithm S1.

Suppose that $x_{ij} \in [0,1]^{N \times M}$ represents the relative methylation ratio of the ith methylation site of the jth sample (we assume $N > M$). x_{ij} is normalized as

$$\sum_i x_{ij} = 0 \tag{1}$$

$$\sum_i x_{ij}^2 = N. \quad (2)$$

PCA was applied to x_{ij} so that PC score $u_{\ell i} \in \mathbb{R}^{N\times M}$ is attributed to the ith CpG site and PC loading $v_{\ell j} \in \mathbb{R}^{M\times M}$ is attributed to the jth sample. After identifying PC loading of interest (e.g., distinction between healthy control and patients), we attributed P-values to ith methylation site assuming that the corresponding PC score, $u_{\ell i}$, meets the Gaussian distribution (null hypothesis) as

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{\ell i}}{\sigma_\ell}\right)^2\right] \quad (3)$$

where $P_{\chi^2}[> x]$ is the cumulative χ^2 distribution that the argument is larger than x and σ_ℓ is the standard deviation optimized as much as possible to obey the null hypothesis (see below). We assumed a Gaussian distribution for u_{ℓ_i} for several reasons: Firstly, it worked reasonably well in previous studies (Taguchi, 2020). Secondly, a Gaussian distribution is expected to occur when random numbers are added because of the central limit theorem. Thus, we assumed the null hypothesis that u_{ℓ_i} obeys the Gaussian distribution, and it performs well, as demonstrated below. When the histogram of $1 - P_i$ obeys the null hypothesis, it should be flat, excluding a sharp peak located near $1-P_i \sim 1$, which represents is to be selected. The optimization of σ_ℓ is performed with minimization of the standard deviation of $h_n(1 - P_i) \in \mathbb{R}^{10^4}$ for microarray, $\mathbb{R}^{10^3}$ for HTS, $n < n_0$ where h_n represents nth histogram and n_0 is the largest n that satisfies the adjusted $P_i(\in h_n)$ that is larger than P_0, a threshold adjusted P-value.

The optimization of σ_ℓ was as follows: To minimize the SD of binned $h_n = h(1 - P_i)$, σ_h,

$$\sigma_h = \sqrt{\frac{\sum_{n<n_0}\left(h_n - \langle h_n\rangle\right)^2}{N(n<n_0)}} \quad (4)$$

$$\langle h_n\rangle = \frac{\sum_{n<n_0} h_n}{N(n<n_0)} \quad (5)$$

with respect to σ_ℓ, where $N(n < n_0)$ is the number of ns associated with $n < n_0$, i.e., they are not recognized as outliers but are recognized as a part of the Gaussian distribution; thus, we selected the σ_ℓ in eq. (3) that resulted in the smallest σ_h. P-values are corrected using the BH criterion (Taguchi, 2020) and ith associated with adjusted P-values less than P_0, which is typically 0.01, are selected as DMCs.

2.3. TD-based unsupervised FE with optimized standard deviation

The purpose of the method is similar to the previous section, i.e., selecting variables that contribute to the classification. However, the computational method is different as PCA is replaced with TD. See also Algorithm S2. A full explanation is, again, provided in our previous paper (Taguchi and Turki, 2022) and our recent book (Taguchi, 2020). As TD was only applied to EH1072 here, we assumed that the tensor represented the EH1072 dataset.

Suppose $x_{ijk} \in [0,1]^{N\times 5\times 3}$ represents methylation of i cytosine in jth cells ($j = 1$: Treg cells from fat, $j = 2$:Treg cells from liver, $j = 3$: Treg cells from skin, $j = 4$:Treg cells from lymph nodes, $j = 5$:T-cells from lymph nodes) of the kth biological replicate. x_{ijk} is normalized as

$$\sum_i x_{ijk} = 0 \quad (6)$$

$$\sum_i x_{ijk}^2 = N. \quad (7)$$

We applied higher order singular value decomposition, HOSVD (Taguchi, 2020), to x_{ijk} and obtained TD as

$$x_{ijk} = \sum_{\ell_1=1}^{N}\sum_{\ell_2=1}^{5}\sum_{\ell_3=1}^{3} G(\ell_1\ell_2\ell_3)u_{\ell_1 i}u_{\ell_2 j}u_{\ell_3 k} \quad (8)$$

where $G(\ell_1\ell_2\ell_3) \in \mathbb{R}^{N\times 5\times 3}$ is core tensor, and $u_{\ell_1 i} \in \mathbb{R}^{N\times N}$, $u_{\ell_2 j} \in \mathbb{R}^{5\times 5}$ and $u_{\ell_3 k} \in \mathbb{R}^{3\times 3}$ are singular value matrices and orthogonal matrices, respectively. To select the $u_{\ell_1 i}$ used to detect DMCs, we selected $u_{\ell_2 j}$ and $u_{\ell_3 k}$, which had the desired properties of dependence upon cells and independence from replicates. After identifying the $u_{\ell_2 j}$ and $u_{\ell_3 k}$ needed, we checked $G(\ell_1\ell_2\ell_3)$ to see which ℓ_1 shares Gs that have the largest absolute values, with the selected ℓ_2 and ℓ_3. Finally, P-values were attributed to i with the equation

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{\ell_1 i}}{\sigma_{\ell_1}}\right)^2\right] \quad (9)$$

after optimizing σ_{ℓ_1}, as described in the previous subsection. P-values were corrected using the BH criterion and ith was associated with adjusted P-values that were less than P_0, which was typically 0.01, were selected as DMCs.

2.4. Overlap between DMCs and various functional regions for microarray

The microarray used in this study was the Illumina HumanMethylation450, the annotation of which is available as GPL16304 from GEO. Among the various functional regions annotated, we considered known DMRs, enhancers, and DHS. Fisher's exact test was applied to see if DMCs significantly overlapped with DMRs, enhancers, or DHS.

2.5. Identification of coincidence between selected DMC and DHS for HTS

To compare selected DMCs with DHSs, we downloaded http://mirror.vcu.edu/vcu/encode/DNase-seq_Peaks/ wgEncodeDukeDnaseFibrobl.fdr01peaks.hg19.bb, which is the fibroblast DHS in the hg19 genome, to which short reads were mapped from HTS data that were investigated in this study (i.e., GSE34864 and EH2072). t test was applied to P-values obtained by the proposed method to see if there were distinctions between DHSs and non-DHSs in individual chromosomes.

2.6. eDMR

eDMR was performed with the options ACF=F, DMR. methdiff=1, and DMC.methdiff=1, using the normalized third PC score $\frac{u_{3i}}{\sigma_3}$ (for GSE34864) and the normalized second PC score $\frac{u_{2i}}{\sigma_2}$ (for EH1072) as meth.diff where σ_ℓ is the optimized SD. Then, filter.dmr was executed with the options DMR. qvalue =0.01 and mean.meth.diff=1. Other options were set as the default.

2.7. metilene

Metilene was performed with the default options.

2.8. Binomial test for the overlap between two genomic regions

We considered two genomic regions, g_1 and g_2, whose associated sequence lengths are ℓ_{g_1} and ℓ_{g_2}, respectively, within the whole genome g, associated with the sequence length ℓ_g. In g_1, the region with length ℓ'_{g_1} at least partially overlapped with g_2. Similarly, the region with length ℓ'_{g_2} in g_2 at least partially overlapped with g_1. We evaluated the significance of this overlap using the cumulative binomial distribution $P_b(> x, N, p)$, which represents a cumulative probability where the argument is larger than x (which is selected within N) and with an individual probability of the selection p for ℓ'_{g_1} and ℓ'_{g_2} as

$$P_b\left(> \ell'_{g_1}, \ell_{g_1}, \frac{\ell_{g_2}}{\ell_g}\right) \tag{10}$$

and

$$P_b\left(> \ell'_{g_2}, \ell_{g_2}, \frac{\ell_{g_1}}{\ell_g}\right), \tag{11}$$

respectively.

2.9. Enrichment analysis

Enrichment analysis was performed using Enrichr (Kuleshov, Jones, Rouillard, Fernandez, Duan, Wang, Koplev, Jenkins, Jagodnik, Lachmann, McDermott, Monteiro, Gundersen and Ma'ayan, 2016). Gene symbols associated with selected probes were identified in GPL13534, which was downloaded from GEO (GEO ID: GPL13534). The numbers of probes considered were also restricted when COHCAP was applied, since COHCAP identified too many probes as DMCs. P-values computed by COHCAP were used to rank identified probes. Additionally, only top ranked probes were considered so as to have as many gene symbols as those identified by the proposed method. Then, obtained gene symbols were uploaded to Enrichr.

3. Results

3.1. Microarray

First, we applied the proposed method to methylation profiles measured by microarray architecture. PCA was applied to x_{ij} of GSE77965. Then, we found that v_{2j} was associated with the distinction between two classes (Fig. S1 (a), $P = 5.68\times10^{-3}$ computed by t test). Then, we attempted to optimize σ_2 so that u_{2i} met the Gaussian distribution as much as possible. We found that $\sigma_2 = 0.1506726$ can result in the smallest standard deviation of h_n (Fig. S2 (a)). Indeed, the histogram of $1 - P_i$ was close to the ideal case when σ_2 the optimized value is used (Fig. S2 (a)). P_is were corrected using the BH criterion. We found that 87,665 methylation sites were associated with adjusted P-values less than 0.01 and were regarded as DMCs. To evaluate whether the selected DMCs were reasonable, we compared them with the information attached to the microarray (i.e., annotations attached to microarray probes). The high coincidence with functional sites (Table S1(a)) validated our selection of DMCs.

The same process was applied to GSE42308. We, again, found that the second PC was associated with two classes (Fig. S1(b), $P = 9.21 \times 10^{-7}$ computed by t test). We also attempted to optimize σ_2 so that u_{2i} met a Gaussian distribution as much as possible. We found that $\sigma_2 = 0.07036153$ resulted in the smallest standard deviation of h_n (Fig. S2 (b)). In actuality, the histogram of $1 - P_i$ was close to the idealized case when σ_2 is the optimized value (Fig. S2(b)). P_is were corrected using the BH criterion. We found that 97,091 methylation sites were associated with adjusted P-values less than 0.01 and could be regarded as DMCs. To evaluate if these selected DMCs were reasonable, we compared them with information attached to the microarray (i.e., annotations attached to microarray probes). Their high coincidence with the functional sites (Table S1 (b)) further proved our selection of DMCs to be reliable. However, the number of false positives obtained was not negligible; thus, further improvement is evidently required.

3.2. HTS

Next, we attempted to validate the effectiveness of the proposed method in not only microarrays, but also HTS. To this end, we applied the proposed method to GSE34864. Unfortunately, owing to the small number of samples (4), no PC loading attributed to samples j showed significant distinction between the two classes. Nevertheless, since DMCs appeared to be identifiable using the proposed method, we employed $\ell = 3$ as it was most associated with a distinction between two classes (Fig. S1(c)). We also tried to optimize σ_3 so that u_{3i} met the Gaussian distribution as much as possible. We found that $\sigma_3 = 0.3499989$ resulted in the smallest standard deviation of h_n (Fig. S2 (c)). In actuality, the histogram of $1 - P_i$ was close to the idealized case when σ_3 takes the optimized value (Fig. S2(c)). In contrast to the microarray results, we lacked annotations for methylation sites. Therefore, we checked the coincidence between DHSs and identified DMCs. We applied a t test to assess the null hypothesis that the distribution of P_i was identical between DHS and non-DHS sites in individual chromosomes (Figure S3(a)). Thereafter, we found $P = 0$ for all chromosomes within numerical accuracy. Thus, it was obvious that the proposed method could identify DMCs correctly. However,

the number of false positives obtained was not negligible. Further improvement is, hence, required.

Then, we applied the proposed method to yet another HTS dataset, EH1072. After applying PCA, we found that v_{2j} was associated with distinction between five cell classes (Fig. S1(d), $P = 2.22 \times 10^{-3}$ computed by categorical regression). We also tried to optimize σ_2 so that u_{2i} met the Gaussian distribution as much as possible. We found that $\sigma_2 = 0.2846132$ could result in the smallest standard deviation of h_n (Fig. S2(d)). Indeed, the histogram of $1 - P_i$ was close to the ideal case when σ_3 took the optimized value (Fig. S2(d)). In contrast to the cases for the microarray analysis, we lacked annotations for methylation sites. Therefore, we checked the coincidence between DHSs and identified DMCs. We applied a t test to assess the null hypothesis that the distribution of P_i was identical between the DHS and non-DHS sites in individual chromosomes. We found $P = 0$ for all chromosomes with numerical accuracy (Figure S3(b)). Thus, the proposed method could identify DMCs accurately.

Finally, to test the effectiveness of TD, we applied "TD-based unsupervised FE with optimized SD" to EH1072. We also applied HOSVD to x_{ijk}; Fig. S1(e) shows the first and second $u_{\ell_2 j}$ attributed to jth cells and the first $u_{\ell_3 k}$ attributed to the kth biological replicate. Since u_{2j} represented the dependence upon cells similar to that seen in Fig. S1 (d) and u_{1k} represented the independence of biological replicates, we needed to identify which $u_{\ell_1 i}$ attributed to the ith cytosine site was associated with u_{2j} and u_{1k} (for other $u_{\ell_2 j}$ and $u_{\ell_3 k}$s, see Figs. S4 and S5, which clearly indicated that no other singular value vectors were associated with the desired properties).

Table S2 shows $G(\ell_1, 2, 1)$ which takes the largest absolute value for $\ell_1 = 2$. Thus, although u_{2i} should be used to select DMCs, it is equal to the second PC score attributed to i since HOSVD is mathematically equivalent to SVD applied to unfolded tenor (Taguchi, 2020). Therefore, TD will select the same set of DMCs as PCA. This implies that TD works equally as well in DMC identification.

3.3. Comparisons with other methods

To evaluate the effectiveness of the proposed method, we compared its performance with that of other methods. Although there is a comprehensive review (Chen et al., 2016), we selected COHAMP for microarray architecture and DSS for HTS architecture as representatives from these reviews. As newer ones were not included in this review, we additionally selected ChAMP for microarray architecture and DMRcate and metilene for HTS architecture.

Firstly, we considered the application of the proposed model for microarray architecture. The first method to be compared with the proposed method was COHCAP. First, we applied COHCAP to GSE77965. Table S1 (c) shows the coincidence between the selected and functional regions. When compared with the performance of the proposed method (Table S1(a)), the performance was generally worse. Next, we applied COHCAP to GSE42308. Table S1(d) shows the coincidence between the selected and functional regions. When compared with the performance of the proposed method (Table S1(b)) the performance was slightly better; however, the number of DMRs was much lower than that obtained using the proposed method. In conclusion, the comparisons of GSE77965 and GSE42308 suggest that the overall performance of the proposed method was better than that of COHCAP. To explore whether the superiority of the proposed method is accidental, we also compared it to another method, ChAMP. Table S1(e) shows the coincidence between the selected and functional regions. When compared with the performance of the proposed method (Table S1(a)), although the performance was comparable, the number of regions selected by the proposed method was roughly ten times greater than that selected by ChAMP. Thus, the proposed method can identify more DMRs containing more identified DMCs while retaining a significant overlap with functional regions. Next we applied ChAMP to GSE42308. Table S1(f) shows the coincidence between the selected and functional regions. When compared with the performance of the proposed method (Table S1(b)), the performance and the number of DMRs were comparable. The comparisons using GSE77965 and GSE42308 suggest that the overall performance of the proposed method was better than that of COHCAP, as well as ChAMP. Thus, the superiority of the proposed method, when applied to microarray architecture, over the conventional method was unlikely due to chance. Next, to observe the performance of the proposed model with HTS data, we first tested DSS. For GSE34864, we found that $P = 0$ for the t test denies the probability that the distribution of P-values attributed to DHS and non-DHS sites was identical (Figure S3(c)). For EH1072, since DSS required more than a week for processing, even on using 10 CPU cores, we terminated the computation. We next applied DMRcate to GSE34864 and EH1072. For GSE34864, DMRcate failed to identify any DMCs. For EH1072, we found that $P = 0$ for the t test denies the probability that the distribution of P-values attributed to DHS and non-DHS sites was identical (Figure S3(d)). We also employed metilene, which failed to identify DMCs for either GSE34864 or EH1072. Thus, the overall performance of the proposed method was better than DSS, DMRcate, and metilene. The use of epigenome-wide association studies (EWAS) (Flanagan, 2014) was outside the scope of this model since EWAS requires numerous samples, while the proposed method was effective when the number of samples was less than ten.

Although the effectiveness of the proposed method is questionable since it could not identify DMRs for HTS architecture, we could identify DMRs by integrating identified DMCs. To show this ability, we used eDMR (Li, Garrett-Bakelman, Akalin, Zumbo, Levine, To, Lewis, Brown, D'Andrea, Melnick and Mason, 2013) to identify DMRs by integrating identified DMCs. We found that eDMR identified 6,818 and 12,662 DMRs for GSE34864 and EH1072, respectively. Thus, the proposed method may also have the ability to identify DMRs if a suitable tool is employed. In

contrast, 27 and 12,478 DMRs were identified by metilene for GSE34864 and EH1072, respectively, whilst it failed to identify DMCs for these profiles. Thus, eDMR combined with the proposed method identified more DMRs than metilene.

To confirm the significant overlap of genomic regions selected by eDMR with DHS, we used a binomial distribution for the evaluation (see Methods); we could not employ Fisher's exact test for the HTS, and since eDMR did not attribute P-values to DMRs, we could not employ a t test.

Table S3 shows P-values that are 0 within numerical accuracy. Thus, DMRs selected by eDMR significantly overlapped with DHS.

3.4. Application to TCGA

Above, we have successfully demonstrated the superiority of the proposed method over the conventional methods. In this subsection, we discuss the application of studies associated with biological evaluation using enrichment analyses to strengthen the efficiency of the proposed method. We employed TCGA (see Acknowledgement), which contains versatile cancer data sets. One possible limitation of TCGA is that its DNA methylation data is rarely studied independently; it is usually analyzed in combination with additional data, e.g., gene expression data. Since the purpose of the present study was to demonstrate the effectiveness of the proposed method when applied to *only* DNA methylation alone, integrated analysis with other data is out of its scope. However, to overcome this challenge and to increase the novelty of this method, we have made the following additions. First of all, we used "ajcc staging system edition" (Edge and Compton, 2010) as clinical information. The American Joint Committee on Cancer (AJCC) releases the staging labels of cancer development. Thus, the staging system is a good indicator of the grades of individual samples.

Among the methods compared with the proposed methods, ChAMP and COHCAP can be applied to TCGA, since TCGA employs microarray architecture. Between ChAMP and COHCAP, only COHCAP be used for "ajcc staging system edition", since it has multiple levels, whereas ChAMP is not suitable for analysis of multiple classes, only pairwise comparisons.

For our results, for both breast and colon data sets, the second PC attributed to samples, v_{2j}, was significantly associated with "ajcc staging system edition". Then, the PC attributed to sites, u_{2i}, was employed and SD was optimized. As a result, we identified 9,544 and 11,447 probes that were associated with distinctions between multiple classes in "ajcc staging system edition". COHCAP was also applied to the two data sets, and 333,949 and 21,042 probes were selected that were associated with distinctions between multiple classes in "ajcc staging system edition" (when multiple classed were considered, COHCAP utilized ANOVA to select sites). Since the number of probes selected by COHCAP was very large, it is unlikely that all of them were differentially methylated sites. Thus, we restricted the number of probes selected by COHCAP to obtain as many associated gene symbols as those identified by the proposed method (see below). The numbers of gene symbols associated with the probes selected by the proposed method were 4,041 and 2,815 for breast and colon data sets, respectively. Those selected by COHCAP were restricted to 4,326 and 2,974 for breast and colon data sets, respectively, by reducing the numbers of probes to the 6,000 and 5,000 top ranked probes associated with smaller P-values for breast and colon data sets, respectively. We have uploaded the four lists of gene symbols associated with identified methylation sites to Enricher (Kuleshov et al., 2016). The following are some examples of categories associated with enriched terms associated with the identified gene symbols. Figure 1 shows the results for the "Cancer Cell Line Encyclopedia" category of Enricher. The proposed method identified significant cell lines within the top 10 ranked cell lines whereas COHCAP failed (no cell lines associated with significant adjusted P-values less thna 0.05). This is only one example but there could be more cases where the proposed method is successful and COHCAP fails (i.e., the proposed method could identify the top 10 significant cell lines but COHCAP could not; Figs. S6 for "ARCHS4 Cell-lines", S7 for "KEGG 2021 HUMAN", S8 for "GO Biological process 2021", S9 for "Jensen Diseases", S10 for "GO Molecular function 2021", S11 for "GO Cellular Components 2021"), which are summarized in Table 1. Thus, as least for some cases, the proposed method can outperform COHCAP from a biological point of view.

4. Discussion

Although TD- and PCA-based unsupervised FE without optimized SD were developed many years ago and have been successfully applied to a wide range of studies (Taguchi, 2020), the novelty of this study is that we have advanced the method by optimizing the SD so that the histogram of $1 - P_i$ obeys the null hypothesis as much as possible. This advancement is significant since TD- and PCA-based unsupervised FE, without optimized SD, could only identify a very small number of DMCs. Furthermore, it is unlikely that there are no false negatives and the histogram of $1 - P_i$ is far from optimal. The optimization of SDs results in a more ideal histogram of $1 - P_i$ and a sufficient number of DMCs, which competes with state-of-the-art methods, as shown above. In addition, it is rare that a tool can be applicable to both gene expression and DNA methylation measurements by both microarray analysis and HTS without any specific modifications. For example, although Rackham, Dellaportas, Petretto and Bottolo (2015) published a benchmark of DNA methylation tools, no tools specific to gene expression were included. This implies that tools specific to gene expression are not expected to work well for DNA methylation as well. Nevertheless, the proposed method (previously tested for gene expression (Taguchi and Turki, 2022)) was effective for DNA methylation without any specific modifications. Hence, our reporting of this

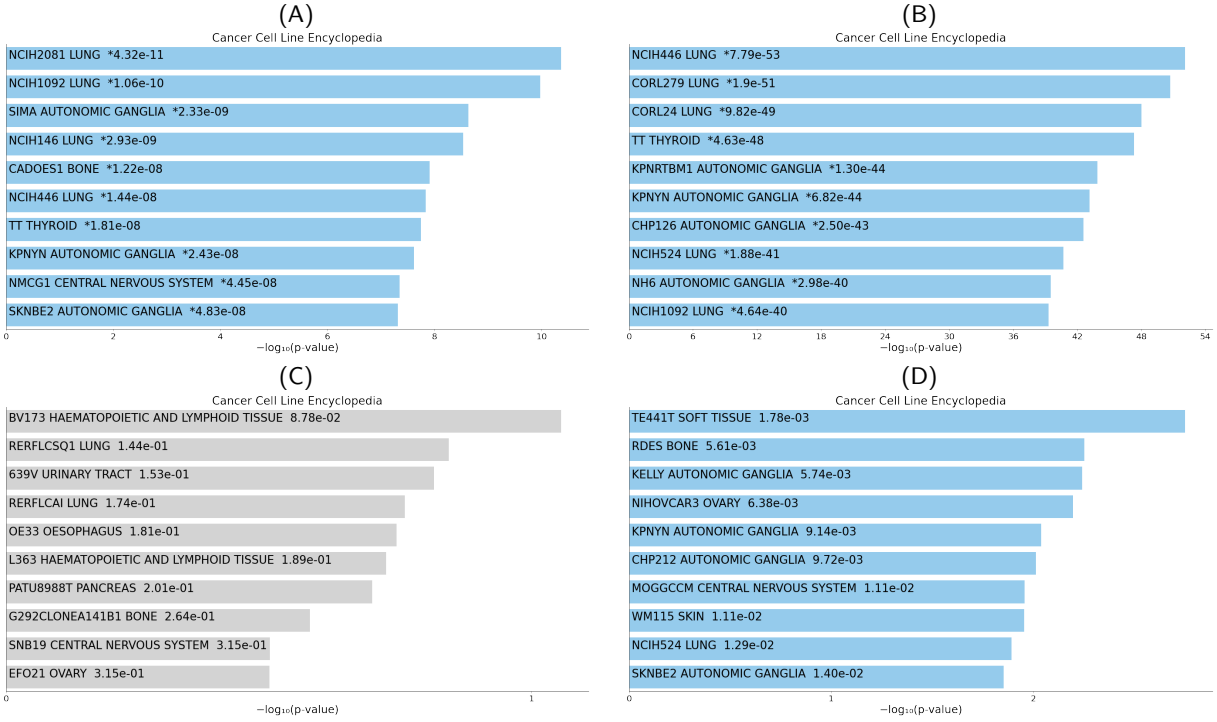

Figure 1: Top 10 ranked cell lines in the "Cancer Cell Line Encyclopedia" category of Enricher for genes selected by the proposed method for (A) breast and (B) colon data sets and those selected by COHCAP for (C) breast and (D) colon data sets, visualized by Appyters (Clarke et al., 2021). Blue colored: raw $P < 0.05$ and asterisked: adjusted $P < 0.05$, respectively.

Table 1
Number of significant (i.e., adjusted P-values are less than 0.05) variables (e.g., cell lines, pathways, and diseases e.t.c.) among the top 10 ranked factors in individual categories for the proposed method and COHCAP

	Proposed method		COHCAP		Figs.
Category	Breast	Colon	Breast	Colon	
Cancer Cell Line Encyclopedia	10	10	0	0	1
ARCHS4 Cell-lines	10	10	0	0	S6
KEGG 2021 HUMAN	10	10	2	0	S7
GO Biological process 2021	10	10	1	9	S8
Jensen Diseases	10	10	3	7	S9
GO Molecular function 2021	10	10	6	7	S10
GO Cellular components 2021	9	10	6	0	S11

remarkable finding. Thus, although this advancement may be small, the improvement in outcomes will be large.

As can be seen in Figs. S2 (a), (b), (c) and (d), the SD of h_n becomes zero when the optimized SD is equal to zero. This is inefficient since all is have $P = 0$ and are selected as DMCs. This results in an empty $\{n < n_0\}$ since no is have a larger P_i than the threshold value P_0, no matter how small P_0 is. Thus, we need to find the smallest SD to have the smallest SD of h_n with an excluding SD of zero.

We employed a t test for HTS instead of a Fisher's exact test as, for microarray architecture, the selection was performed among probes with annotations. Thus, we could compare two classifications such as functional probes and non-functional probes versus selected probes and non-selected probes. Nevertheless, for HTS, since there was only a list of DHSs and DMCs, there were no common background sets from which we could define non-selected and non-DHS sites. Thus, we could not perform a Fisher's exact test, and opted to apply a t test to P-values between DHS and non-DHS as this was possible even without a common background set.

In conclusion, in this paper, we applied the proposed method "PCA- and TD-based unsupervised FE with optimized SD" to identify DMCs. Identified DMCs were significantly overlapped with various functional sites. This technique also outperformed some modern methods, especially when applied to HTS architecture. We also identified DMRs by integrating DMCs. The proposed method does not have to assume non-trivial distribution such as the beta-binomial distribution for methylation and, instead, simply assumes that the PC score obeys a Gaussian distribution, which is rationalized in the framework for probabilistic PCA (Tipping

and Bishop, 1999); therefore, the proposed method is more adequate than other current methods that must assume beta-binomial distribution.

CRediT authorship contribution statement

Y-H. Taguchi: Planning the research, Analyses, Valuation of the results, discussions, and outcomes, Wring the manuscript. **Turki Turki:** Valuation of the results, discussions, and outcomes, Wring the manuscript.

Funding

This work was supported by KAKENHI [grant numbers 20H04848, and 20K12067] to YHT.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data availability

All data used in this study are cited in the manuscript and are available by public access. Sample R source code is available at https://github.com/tagtag/PCAUFEOPSD/ (Accessed 5th January 2023).

Conflicts of interest

The authors declare no conflict of interest.

Acknowledgement

The results shown here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

References

BLUEPRINT consortium, 2016. Quantitative comparison of DNA methylation assays for biomarker development and clinical applications. Nature Biotechnology 34, 726–737. URL: https://doi.org/10.1038/nbt.3605, doi:10.1038/nbt.3605.

Chen, D.P., Lin, Y.C., Fann, C.S.J., 2016. Methods for identifying differentially methylated regions for sequence- and array-based data. Briefings in Functional Genomics 15, 485–490. URL: https://doi.org/10.1093/bfgp/elw018, doi:10.1093/bfgp/elw018, arXiv:https://academic.oup.com/bfg/article-pdf/15/6/485/8263147/elw018.pdf.

Clarke, D.J., Jeon, M., Stein, D.J., Moiseyev, N., Kropiwnicki, E., Dai, C., Xie, Z., Wojciechowicz, M.L., Litz, S., Hom, J., Evangelista, J.E., Goldman, L., Zhang, S., Yoon, C., Ahamed, T., Bhuiyan, S., Cheng, M., Karam, J., Jagodnik, K.M., Shu, I., Lachmann, A., Ayling, S., Jenkins, S.L., Ma'ayan, A., 2021. Appyters: Turning jupyter notebooks into data-driven web apps. Patterns 2, 100213. URL: https://www.sciencedirect.com/science/article/pii/S2666389921000234, doi:https://doi.org/10.1016/j.patter.2021.100213.

Edge, S.B., Compton, C.C., 2010. The American Joint Committee on Cancer: the 7th Edition of the AJCC Cancer Staging Manual and the Future of TNM. Annals of Surgical Oncology 17, 1471–1474. doi:10.1245/s10434-010-0985-4.

Feng, H., Conneely, K.N., Wu, H., 2014. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. Nucleic Acids Research 42, e69–e69. URL: https://doi.org/10.1093/nar/gku154, doi:10.1093/nar/gku154, arXiv:https://academic.oup.com/nar/article-pdf/42/8/e69/25364302/gku154.pdf.

Flanagan, J.M., 2014. Epigenome-wide association studies (EWAS): Past, present, and future, in: Methods in Molecular Biology. Springer New York, pp. 51–63. URL: https://doi.org/10.1007/978-1-4939-1804-1_3, doi:10.1007/978-1-4939-1804-1_3.

Jühling, F., Kretzmer, H., Bernhart, S.H., Otto, C., Stadler, P.F., Hoffmann, S., 2015. metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. Genome Research 26, 256–262. URL: https://doi.org/10.1101/gr.196394.115, doi:10.1101/gr.196394.115.

Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., McDermott, M.G., Monteiro, C.D., Gundersen, G.W., Ma'ayan, A., 2016. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Research 44, W90–W97. URL: https://doi.org/10.1093/nar/gkw377, doi:10.1093/nar/gkw377, arXiv:https://academic.oup.com/nar/article-pdf/44/W1/W90/18788036/gkw377.pdf.

Li, S., Garrett-Bakelman, F.E., Akalin, A., Zumbo, P., Levine, R., To, B.L., Lewis, I.D., Brown, A.L., D'Andrea, R.J., Melnick, A., Mason, C.E., 2013. An optimized algorithm for detecting and annotating regional differential methylation. BMC Bioinformatics 14. URL: https://doi.org/10.1186/1471-2105-14-s5-s10, doi:10.1186/1471-2105-14-s5-s10.

Luo, Y., Lu, X., Xie, H., 2014. DynamicAluMethylation during normal development, aging, and tumorigenesis. BioMed Research International 2014, 1–12. URL: https://doi.org/10.1155/2014/784706, doi:10.1155/2014/784706.

Morgan, M., Shepherd, L., 2022. ExperimentHub: Client to access ExperimentHub resources. R package version 2.2.1.

Park, Y., Wu, H., 2016. Differential methylation analysis for BS-seq data under general experimental design. Bioinformatics 32, 1446–1453. URL: https://doi.org/10.1093/bioinformatics/btw026, doi:10.1093/bioinformatics/btw026, arXiv:https://academic.oup.com/bioinformatics/article-pdf/32/10/1446/16920397/btw0

Peters, T., Buckley, M., Chen, Y., Smyth, G., Goodnow, C., Clark, S., 2021. Calling differentially methylated regions from whole genome bisulphite sequencing with DMRcate. Nucleic Acids Research 49, e109–e109. URL: https://doi.org/10.1093/nar/gkab637, doi:10.1093/nar/gkab637, arXiv:https://academic.oup.com/nar/article-pdf/49/19/e109/41071508/gkab637.pdf.

Rackham, O.J.L., Dellaportas, P., Petretto, E., Bottolo, L., 2015. WGBSSuite: simulating whole-genome bisulphite sequencing data and benchmarking differential DNA methylation analysis tools. Bioinformatics 31, 2371–2373. URL: https://doi.org/10.1093/bioinformatics/btv114, doi:10.1093/bioinformatics/btv114, arXiv:https://academic.oup.com/bioinformatics/article-pdf/31/14/2371/17086629/btv1

Shnorhavorian, M., Schwartz, S.M., Stansfeld, B., Sadler-Riggleman, I., Beck, D., Skinner, M.K., 2017. Differential dna methylation regions in adult human sperm following adolescent chemotherapy: Potential for epigenetic inheritance. PLOS ONE 12, 1–18. URL: https://doi.org/10.1371/journal.pone.0170085, doi:10.1371/journal.pone.0170085.

Skinner, M.K., Nilsson, E., Sadler-Riggleman, I., Beck, D., Maamar, M.B., McCarrey, J.R., 2019. Transgenerational sperm dna methylation epimutation developmental origins following ancestral vinclozolin exposure. Epigenetics 14, 721–739. URL: https://doi.org/10.1080/15592294.2019.1614417, doi:10.1080/15592294.2019.1614417, arXiv:https://doi.org/10.1080/15592294.2019.1614417. pMID: 31079544.

Smith, Z.D., Chan, M.M., Mikkelsen, T.S., Gu, H., Gnirke, A., Regev, A., Meissner, A., 2012. A unique regulatory phase of DNA methylation in the early mammalian embryo. Nature 484, 339–344. URL: https://doi.org/10.1038/nature10960, doi:10.1038/nature10960.

Taguchi, Y.H., 2020. Unsupervised Feature Extraction Applied to Bioinformatics. Springer International Publishing. URL: https://doi.org/10.1007/978-3-030-22456-1, doi:10.1007/978-3-030-22456-1.

Taguchi, Y.h., Turki, T., 2022. Adapted tensor decomposition and PCA based unsupervised feature extraction select more biologically reasonable differentially expressed genes than conventional methods. Scientific Reports 12, 17438. doi:10.1038/s41598-022-21474-z.

Tian, Y., Morris, T.J., Webster, A.P., Yang, Z., Beck, S., Feber, A., Teschendorff, A.E., 2017. ChAMP: updated methylation analysis pipeline for Illumina BeadChips. Bioinformatics 33, 3982–3984. URL: https://doi.org/10.1093/bioinformatics/btx513, doi:10.1093/bioinformatics/btx513, arXiv:https://academic.oup.com/bioinformatics/article-pdf/33/24/3982/25168400/btx513.pdf.

Tipping, M.E., Bishop, C.M., 1999. Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61, 611–622. URL: https://doi.org/10.1111/1467-9868.00196, doi:10.1111/1467-9868.00196.

Warden, C.D., Lee, H., Tompkins, J.D., Li, X., Wang, C., Riggs, A.D., Yu, H., Jove, R., Yuan, Y.C., 2013. COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis. Nucleic Acids Research 41, e117–e117. URL: https://doi.org/10.1093/nar/gkt242, doi:10.1093/nar/gkt242, arXiv:https://academic.oup.com/nar/article-pdf/41/11/e117/25338160/gkt242.pdf.