MDPI

Article

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

Y-H. Taguchi [1,*] and Turki Turki [2]

[1] Department of Physics, Chuo University, Tokyo 112-8551, Japan
[2] Department of Computer Science, King Abdulaziz University, Jeddah 21589, Saudi Arabia; tturki@kau.edu.sa
* Correspondence:tag@granular.com; Tel.: +81-3-3817-1791

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 whose effectiveness toward differentially expressed gene (DEG) identification was recently recognized. The proposed methods can outperform certain conventional methods, including those that must assume beta-binomial distribution for methylation that the proposed methods do not have to assume, especially when applied to methylation profiles measured using high throughput sequencing. DMCs identified by the proposed method are also significantly overlapped with various functional sites, including known differentially methylated regions, enhancers, and DNase I hypersensitive sites. This suggests that the proposed method is a promising candidate for standard methods for identifying DMCs.

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

Citation: Taguchi, Y.-H.; Turki, T. Principal component analysis- and tensor decomposition-based unsupervised feature extraction to select more reasonable differentially methylated cytosines: Optimization of standard deviation versus state-of-the-art methods. *Cells* **2022**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Submitted to *Cells* for possible open access publication under the terms and conditions of the Creative Commons Attri- bution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Identification of aberrant methylation [1] is important since it is supposed to play a critical role in various biological processes. In contrast to the identification of differentially expressed genes (DEGs), approaches towards the identification of aberrant methylation remain unclear, since it is not directly related to functional units such as genes. There are two ways to identify aberrant methylation: the identification of differentially methylated cytosines (DMCs) and that of differentially methylated regions (DMRs). These two might differ from each other; even if individual cytosines within a region are not differentially methylated, if 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. Because of this inconsistency, there are technically 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. At the moment, there is no definite consensus on which is better. The identification of DMCs is generally more difficult than that of DMRs because of the huge number of possible methylated CpG sites. For example, CpG sites predicted as being methylated easily are as many as 28.3 million for human beings [2], which is much larger than the number of human DMRs, which range from a few hundred [3] to 20 thousand [4]. As a result, the identification of DMCs is much harder than that of DMRs. To overcome this difficulty, even in the microarray era before the development of high-throughput sequencing (HTS) technology, various computational methods have been proposed to identify DMCs. For example, ChAMP [5] and COHCAP [6] are some of the oldest methods to identify DMCs in

a specific methylation microarray architecture developed by Illumina. Also, for HTS data sets, DMRcate [7], DSS [8], and metilene [9] are some of the popular methods. Conventional methods, including the aforementioned ones, have common problems to be addressed.

1. They do not assume the rationalized and complicated properties for the distribution of the amount of methylation.
2. In addition, they require extended processing time.

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 recently successfully applied to identify differentially expressed genes [10]. We found that the proposed method can outperform modern methods, especially when applied to HTS architecture.

2. Methods

2.1. Methylation profiles

We used the following four methylation profiles as representatives in this study. 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 if properly analyzed.

2.1.1. GSE77965

This is the methylation profile used in the benchmark study [11] and is available in GEO. Among them, we retrieved six colorectal cancer samples (GSM2062665 to GSM2062670) and six adjacent normal colon tissue samples (GSM2062671 to GSM2062676) whose methylation levels were measured by using the Illumina HumanMethylation450 BeadChip.

2.1.2. GSE42308

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

2.1.3. GSE34864

A part of this methylation profile that originally reported [12] was used for benchmarking in another study [13] and is available in GEO. The profiles used are two oocytes (GSM856495 and GSM856495) and two zygote profiles (GSM856501 and GSM856502) measured by HTS.

2.1.4. EH1072

This methylation profile is included in ExperimentHub [14] in Bioconductor and has been used for the evaluation of DMRcate. It has also been measured 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). Since the comparison of these methods with the proposed method can deal with only pair-wise design, DMCs were identified only for the comparisons between Treg and T-cell control taken from lymph node when methods other than the proposed method are employed.

2.2. PCA-based unsupervised FE with optimized standard deviation

A full explanation is provided in our previous paper [10] and our recent book [15]. In brief, 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

$$\sum_i x_{ij} = 0 \quad (1)$$

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

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., distinct between healthy control and patients), we attributed P-values to ith methylation site assuming that the corresponding PC score, $u_{\ell i}$, obeys the Gaussian distribution (null hypothesis) as

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

where $P_{\chi^2}[> x]$ is the cumulative χ^2 distribution that the argument is larger than x and σ_ℓ is the optimized standard deviation optimized as much as possible to obey the null hypothesis. We assumed a Gaussian distribution for u_{ℓ_i} for several reasons: Primarily, because it worked reasonably well in previous studies [15]. Secondarily, because a Gaussian distribution is expected to occur when random numbers are added with one another because of the central limit theorem. Thus, we assumed the null hypothesis that u_{ℓ_i} obeys the Gaussian distribution, and it showed a good performance, 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 the 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 is as follows: To minimize the SD of binned $h_n = h(1 - P_i)$, σ_h,

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

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

with respect to σ_ℓ, where $N(n < n_0)$ is the number of ns associated with $n < n_0$, i.e., not recognized as outliers but recognized as a part of the Gaussian distribution, we selected the σ_ℓ in eq. (3) that resulted in the smallest σ_h. P-values are corrected using the BH criterion [15] 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

A full explanation is provided in our previous paper [10] and our recent book [15]. Since in this study, TD was applied only to EH1072, we assume that the tensor represents the EH1072 dataset. Suppose that $x_{ijk} \in [0, 1]^{N\times 5\times 3}$ represents methylation of i cytosine at jth cell ($j = 1$: Treg cell from fat, $j = 2$:Treg cell from liver, $j = 3$: Treg cell from skin, $j = 4$:Treg cell from lymph nodes, $j = 5$:T-cell from lymph nodes) of kth biological replicate. x_{ijk} is normalized as

$$\sum_i x_{ijk} = 0 \tag{6}$$

$$\sum_i x_{ijk}^2 = N. \tag{7}$$

We applied higher order singular value decomposition, HOSVD [15], to x_{ijk} and obtained TD as

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

where $G(\ell_1\ell_2\ell_3) \in \mathbb{R}^{N\times5\times3}$ is core tensor, and $u_{\ell_1 i} \in \mathbb{R}^{N\times N}$, $u_{\ell_2 j} \in \mathbb{R}^{5\times5}$ and $u_{\ell_3 k} \in \mathbb{R}^{3\times3}$ are singular value matrices and orthogonal matrices, respectively. To select the $u_{\ell_1 i}$ used to select DMCs, we needed to select $u_{\ell_2 j}$ and $u_{\ell_3 k}$ that have the desired properties: dependence upon cell and independence of 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

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

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

2.4. Overlap between DMCs and various functional regions for microarray

The microarray used in this study is the Illumina HumanMethylation450, whose annotation is available as GPL16304 from GEO. Among the various functional regions annotated, we consider known DMRs, enhancer, and DHS. Fisher's exact test was applied to see if DMCs significantly overlap with DMRs, enhancer, 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 in HTS investigated in this study (i.e., GSE34864 and EH2072). t test was applied to P-values obtained by the proposed method to see if they are distinct 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 default.

2.7. metilene

Metilene was performed with the default options.

2.8. Binomial test for the overlap between two genomic regions

Consider 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} is at least partially overlapped with g_2. Similarly, the region with length ℓ'_{g_2} in g_2 is 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 that 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.

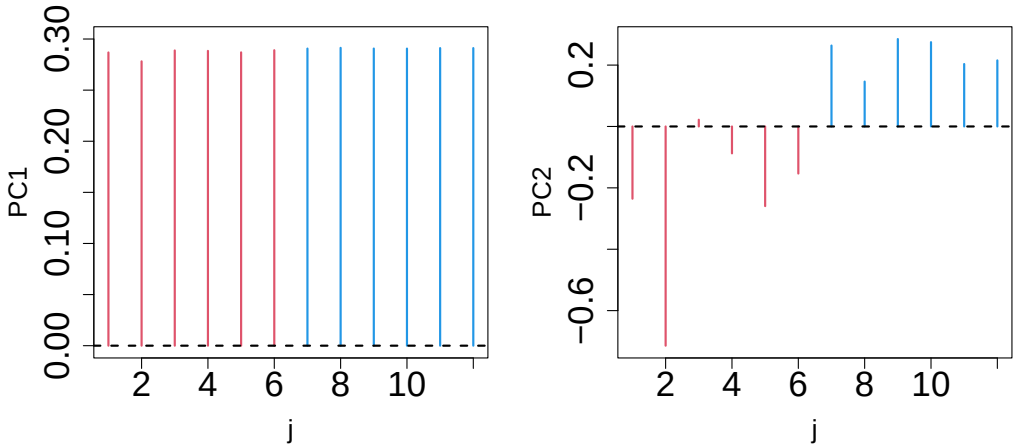

Figure 1. The first and second PC scores (v_{1j} and v_{2j}) attributed to the jth sample for GSE77965. Red: tumors, blue: normal tissues.

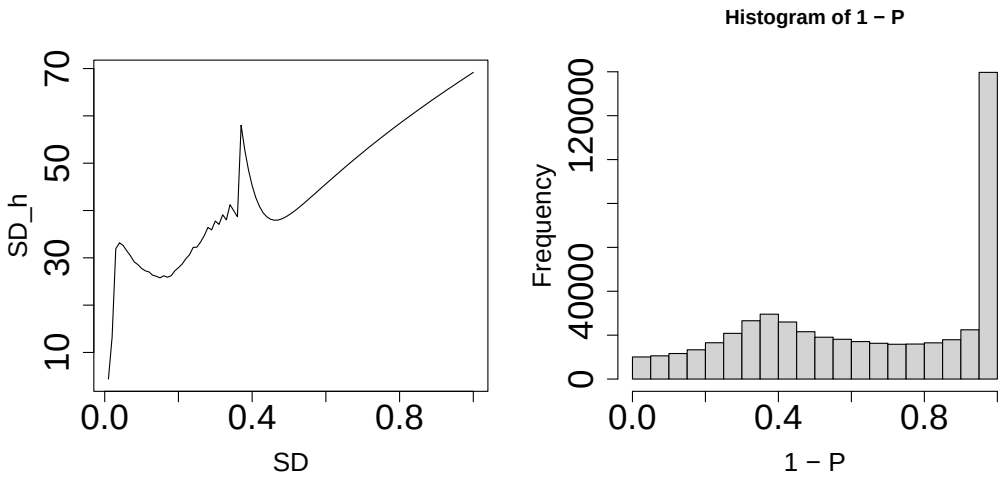

Figure 2. Left: Dependence of standard deviation of h_n on σ_2 for GSE77965. Right: Histogram of $1 - P_i$ for the optimized σ_2.

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. 1, $P = 5.68 \times 10^{-3}$ computed by t test). Then we tried to optimize σ_2 so that u_{2i} obeyed 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. 2). Indeed, the histogram of $1 - P_i$ was close to the ideal case when σ_2 takes the optimized value (Fig. 2). P_is were corrected by 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 information attached to the microarray (i.e., annotations attached to microarray probes). Since they were highly coincident with functional sites (Table 1), our selection of DMCs was valid.

The same process was applied to GSE42308. We again found that the second PC was associated with two classes (Fig. 3, $P = 9.21 \times 10^{-7}$ computed by t test). We also tried to optimize σ_2 so that u_{2i} obeyed a Gaussian distribution as much as possible. We found that

Table 1. Fisher's exact test between identified DMCs and various annotations for GSE77965. DMR: Differentially methylated region, DHS: DNase I hypersensitive site

	adjusted P_i			adjusted P_i	
DMR	> 0.01	< 0.01	Enhancer	> 0.01	< 0.01
NO	378862	69378	NO	321561	61457
YES	19050	18287	YES	76351	26208
odds ratio	5.24		odds ratio	1.80	
P-value	0		*P*-value	0	
	adjusted P_i				
DHS	> 0.01	< 0.01			
NO	352206	73455			
YES	45706	14210			
Odds ratio	1.49				
P-value	8.53×10^{-305}				

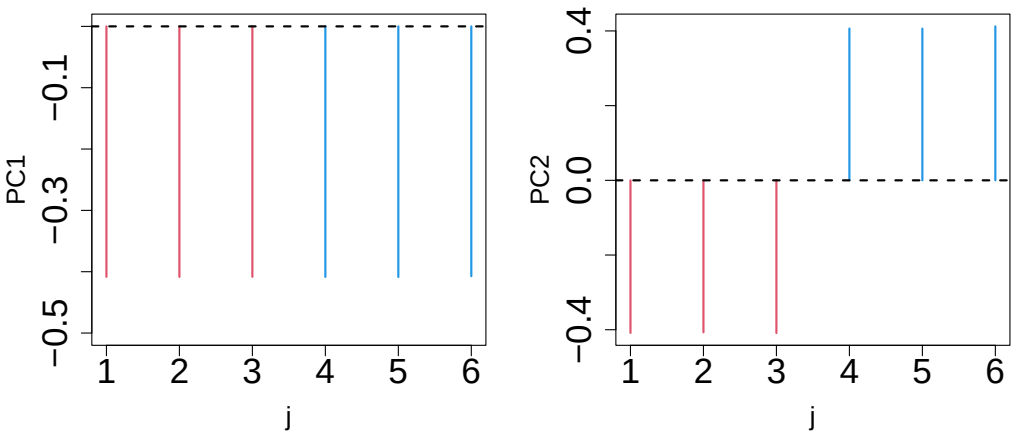

Figure 3. The first and second PC scores (v_{1j} and v_{2j}) attributed to the jth sample for GSE42308. Red: HCT116 cell lines, blue: HCT116 mutants.

$\sigma_2 = 0.07036153$ could result in the smallest standard deviation of h_n (Fig. 4). In actuality, the histogram of $1 - P_i$ was close to the idealized case when σ_2 takes the optimized value (Fig. 4). P_is were corrected by 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. In order to evaluate if these selected DMCs were reasonable, we compared them with information attached to the microarray (i.e., annotations attached to microarray probes). Since they were highly coincident with functional sites (Table 2), our selection of DMCs was further proved to be reliable. However, the number of false positives obtained was not negligible; thus, further improvement is clearly required.

3.2. HTS

Next, we attempted to validate if the proposed method was effective 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 was associated with significant distinction between two classes. Nevertheless, since we may still be able to identify DMCs with the proposed methods, we employed $\ell = 3$ as that mostly associated with distinction between two classes (Fig. 5). We also tried to optimize σ_3 so that u_{3i} obeyed the Gaussian distribution as much as possible. We found that $\sigma_3 = 0.3499989$ could result in the smallest standard deviation of h_n (Fig. 6). In actuality, the histogram of $1 - P_i$ was close to the idealized case when σ_3 takes the optimized value (Fig.

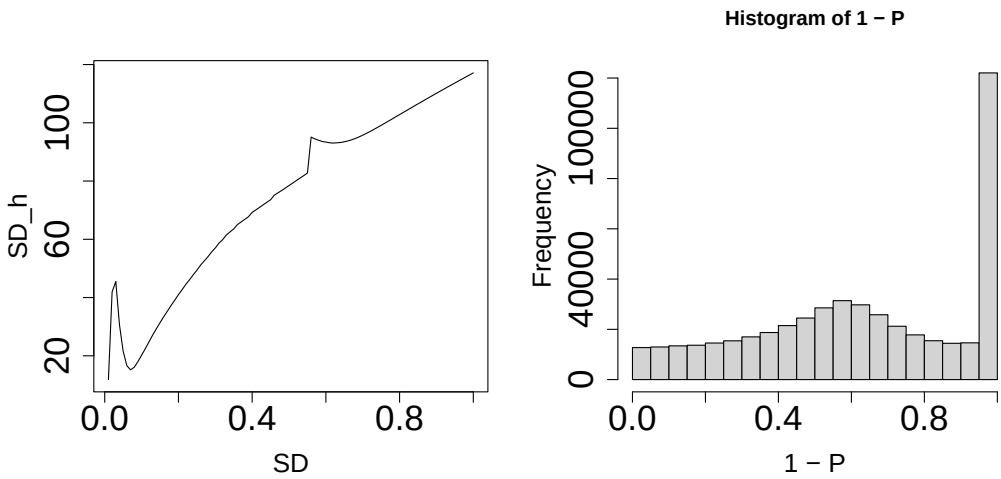

Figure 4. Left: Dependence of standard deviation of h_n upon σ_2 for GSE42308. Right: Histogram of $1 - P_i$ for the optimized σ_2.

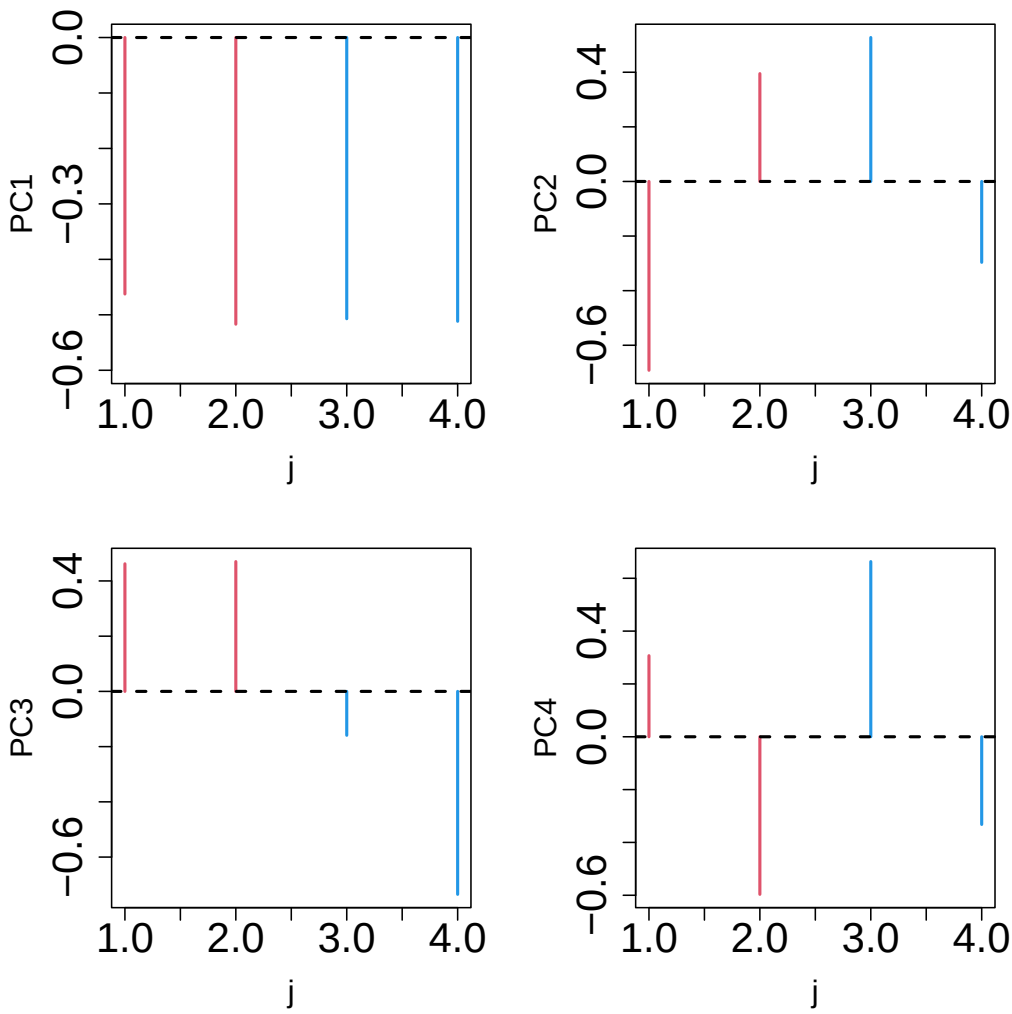

Figure 5. The first to fourth PC scores ($v_{\ell j}, 1 \leq \ell \leq 4$) attributed to the jth sample for GSE34864. Red: oocytes, blue: zygote.

Table 2. Fisher's exact test between identified DMCs and various annotations for GSE42308. DMR: Differentially methylated region, DHS: DNase I hypersensitive site

	adjusted P_i			adjusted P_i	
DMR	> 0.01	< 0.01	Enhancer	> 0.01	< 0.01
NO	361067	87173	NO	313295	69723
YES	27419	9918	YES	75191	27368
odds ratio	1.50		odds ratio	1.64	
P-value	4.77×10^{-224}		P-value	0	
	adjusted P_i				
DHS	> 0.01	< 0.01			
NO	343294	82367			
YES	45192	14724			
odds ratio	1.36				
P-value	2.37×10^{-188}				

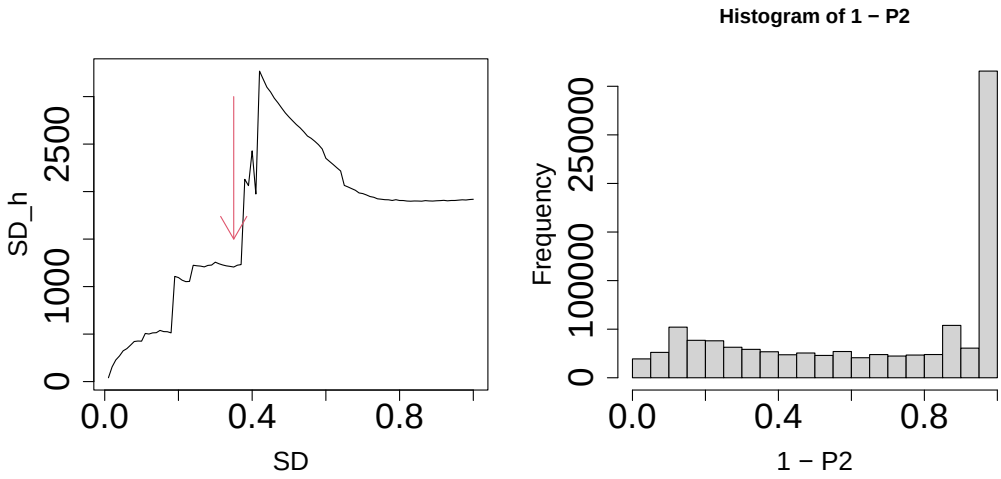

Figure 6. Left: Dependence of standard deviation of h_n upon σ_3 for GSE34864. Vertical red arrow shows the optimized σ_3. Right: Histogram of $1 - P_i$ for the optimized σ_3.

6). In contrast to the cases for microarrays, we lacked annotations for methylation sites. Therefore, we checked the coincidence between DHSs and identified DMCs. We applied a t test to test the null hypothesis that the distribution of P_i was identical between DHS and non-DHS sites in individual chromosomes (Figure 7). 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, EH1072. After applying PCA, we found that v_{2j} was associated with distinction between five cell classes (Fig. 8, $P = 2.22 \times 10^{-3}$ computed by categorical regression). We also tried to optimize σ_2 so that u_{2i} obeyed 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. 9). Indeed, the histogram of $1 - P_i$ was close to the ideal case when σ_3 takes the optimized value (Fig. 9). In contrast to the cases for microarray, we lacked annotations for methylation sites. Therefore, we checked the coincidence between DHSs and identified DMCs. We applied a t test to deny 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 within numerical accuracy (Figure 10). Thus, the proposed method could identify DMCs accurately.

Finally, in order to test if TD also worked as well we applied "TD-based unsupervised FE with optimized SD" to EH1072. We applied HOSVD to x_{ijk}; Fig. 11 shows the first

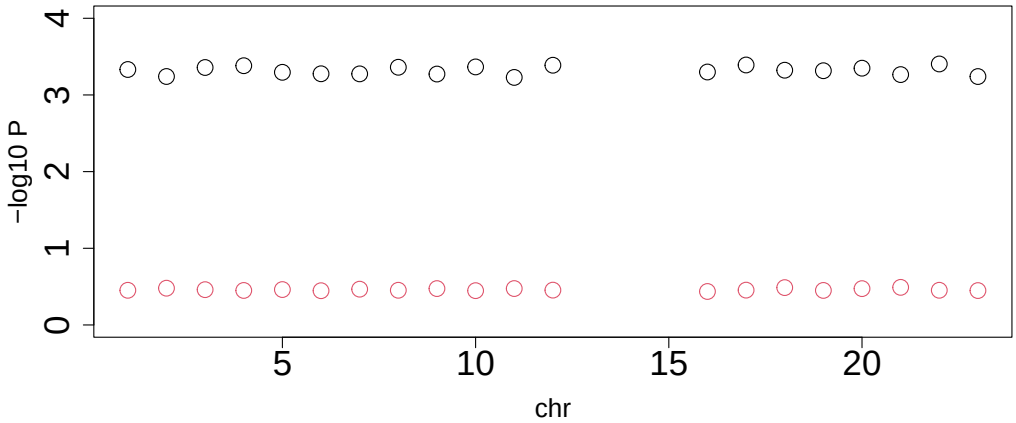

Figure 7. Negative logarithm of mean *P*-values between DHS (black circles) and non-DHS sites (red circles) within each chromosome (horizontal axis) when PCA-based unsupervised FE with optimized SD was applied to GSE34864. *P*-values computed using *t* test between these two were zero. Some chromosomes did not have an associated DHS.

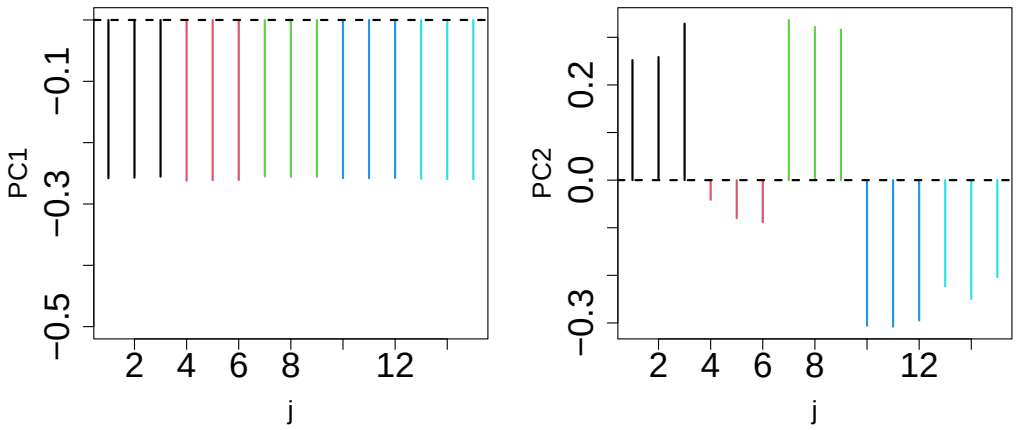

Figure 8. The first and second PC scores (v_{1j} and v_{2j}) attributed to the *j*th sample for EH1072. Black: Treg from fat, red: Treg from liver, green: Treg from skin, blue: Treg from lymph node, cyan: Treg control from lymph node.

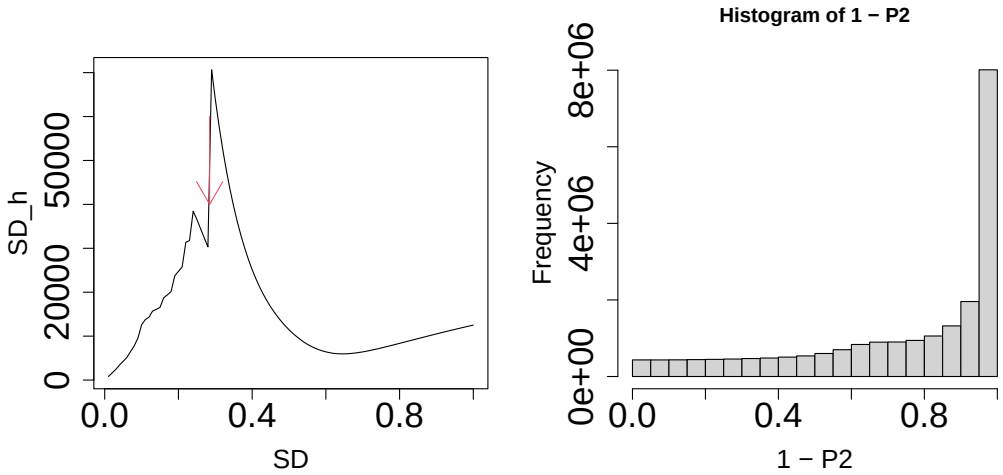

Figure 9. Left: Dependence of standard deviation of h_n on σ_2 for EH1072. Vertical red arrow shows the optimized σ_2. Right: Histogram of $1 - P_i$ for the optimized σ_2.

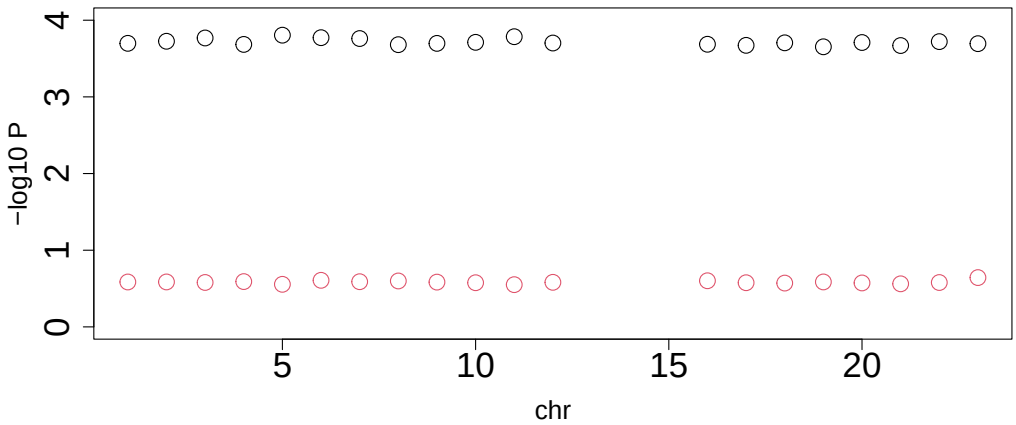

Figure 10. Negative logarithm of mean P-values between DHS (black circles) and non-DHS sites (red circles) within each chromosome (horizontal axis) when PCA-based unsupervised FE with optimized SD was applied to EH2072. P-values computed using t test between these two were zero. Some chromosomes did not have an associated DHS.

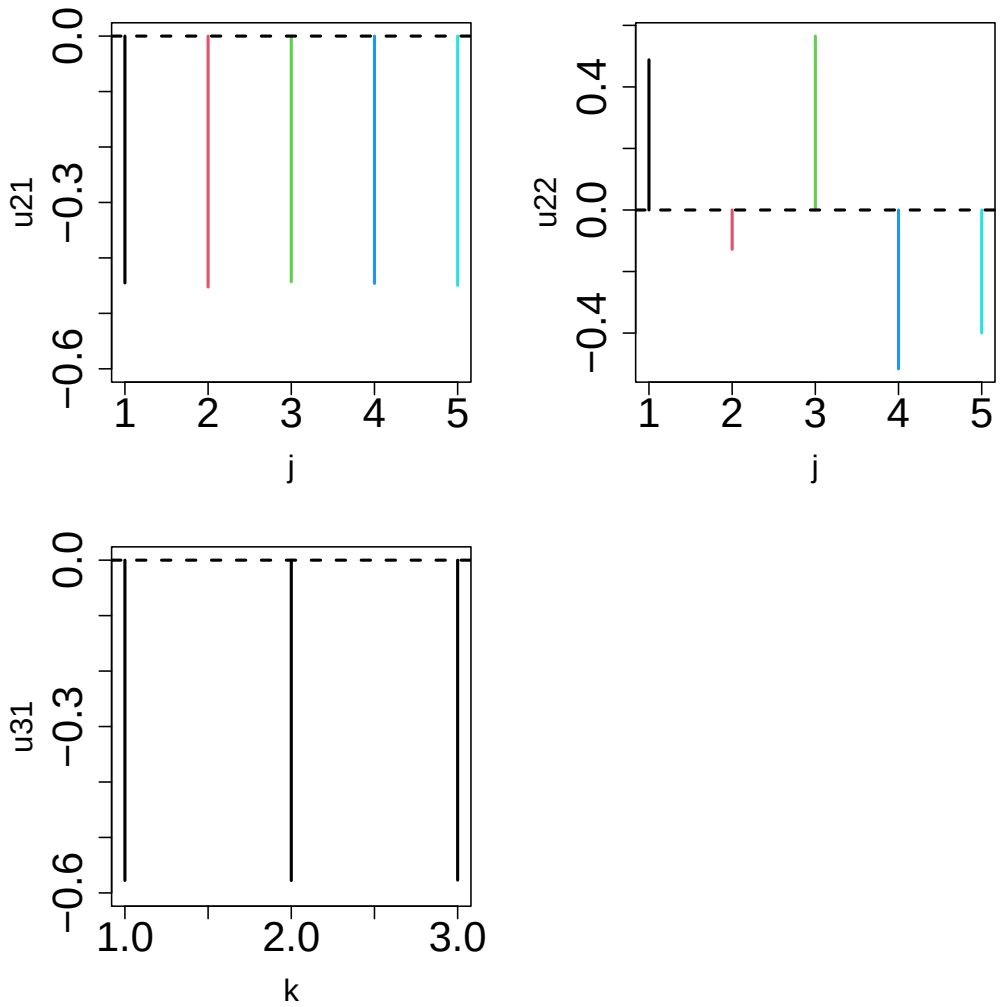

Figure 11. Upper: The first and second $u_{\ell_2 j}$ attributed to jth cells. Black: Treg from fat, red: Treg from liver, green: Treg from skin, blue: Treg from lymph nodes, cyan: Treg control from lymph node. Lower: The first $u_{\ell_3 k}$ attributed to biological replicates.

Table 3. $G(\ell_1, 2, 1)$

ℓ_1	$G(\ell_1,2,1)$	ℓ_1	$G(\ell_1,2,1)$
1	-10	6	44
2	2949	7	7
3	6	8	31
4	5	9	-22
5	-18	10	12

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_{22} represented the dependence upon cells similar to that seen in Fig. 8 and u_{31} represented the independence of biological replicates, we needed to identify which u_{1i} attributed to the ith cytosine site was associated with u_{22} and u_{31} (for other $u_{\ell_2 j}$ and $u_{\ell_3 k}$s, see Figs. S1 and S2, which clearly indicate that no other singular value vectors were associated with the desired properties). Table 3 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 [15]. Therefore, the same set of DMCs will be selected even when using TD instead of PCA. This implies that TD works equally well in DMC identification.

Table 4. Fisher's exact test between identified DMCs and various annotations when COHCAP was applied to GSE77965. DMR: Differentially methylated region, DHS: DNase I hypersensitive site

	adjusted P_i			adjusted P_i	
DMR	> 0.01	< 0.01	Enhancer	> 0.01	< 0.01
NO	447353	887	NO	382429	589
YES	37233	104	YES	102157	402
odds ratio	1.41		odds ratio	2.55	
P-value	1.51×10^{-3}		P-value	1.62×10^{-43}	
	adjusted P_i				
DHS	> 0.01	< 0.01			
NO	424801	860			
YES	59785	131			
odds ratio	1.08				
P-value	4.11×10^{-1}				

Table 5. Fisher's exact test between identified DMCs and various annotations when COHCAP was applied to GSE42308. DMR: Differentially methylated region, DHS: DNase I hypersensitive site

	adjusted P_i			adjusted P_i	
DMR	> 0.01	< 0.01	Enhancer	> 0.01	< 0.01
NO	434162	14078	NO	371646	11372
YES	35215	2122	YES	97731	4828
odds ratio	1.86		odds ratio	1.61	
P-value	5.86×10^{-129}		P-value	8.38×10^{-154}	
	adjusted P_i				
DHS	> 0.01	< 0.01			
NO	412252	13409			
YES	57125	2791			
odds ratio	1.50				
P-value	3.61×10^{-75}				

3.3. Comparisons with other methods

To evaluate the performance of the proposed method, we compared its performance with that of other methods. Although there is a comprehensive review [1], 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, ChAMP for microarray architecture and DMRcate and metilene for HTS architecture were additionally selected.

At first, we considered its application to microarray architecture. The first method to be compared with the proposed method is COHCAP. First, we applied COHCAP to GSE77965. Table 4 shows the coincidence between the selected and functional regions. If compared with the performance by the proposed method (Table 1), the performance was generally worse. Next, we applied COHCAP to GSE42308. Table 5 shows the coincidence between the selected and functional regions. If compared with the performance by the proposed method (Table 2), although the performance was slightly better than the proposed method, the number of DMRs was much less than that obtained using the proposed method. In conclusion, the comparisons using GSE77965 and GSE42308 suggest that the overall performance of the proposed method was better than COHCAP. To explore whether the superiority of the proposed method is accidental, we also tried yet another method, ChAMP. Table 6 shows the coincidence between the selected and functional regions. If compared with the performance by the proposed method (Table 1), although the performance was comparable, the number of selected regions by the proposed method was roughly ten times larger than that by ChAMP. Thus, the proposed method could identify more DMRs containing more identified DMCs while retaining a significant overlap with functional regions. Next we applied ChAMP to GSE42308. Table 7 shows the coincidence between

Table 6. Fisher's exact test between identified DMCs and various annotations when ChAMP was applied to GSE77965. DMR: Differentially methylated region, DHS: DNase I hypersensitive site

	adjusted P_i			adjusted P_i	
DMR	> 0.01	< 0.01	Enhancer	> 0.01	< 0.01
NO	442353	5887	NO	377718	5300
YES	35069	2268	YES	99704	2855
odds ratio	4.86		odds ratio	2.04	
P-value	0		P-value	1.18×10^{-186}	
	adjusted P_i				
DHS	> 0.01	< 0.01			
NO	419579	6082			
YES	57843	2073			
odds ratio	2.47				
P-value	5.50×10^{-231}				

Table 7. Fisher's exact test between identified DMCs and various annotations when ChAMP was applied to GSE42308. DMR: Differentially methylated region, DHS: DNase I hypersensitive site

	adjusted P_i			adjusted P_i	
DMR	> 0.01	< 0.01	Enhancer	> 0.01	< 0.01
NO	374324	73916	NO	324955	58063
YES	28338	8999	YES	77707	24852
odds ratio	1.61		odds ratio	1.79	
P-value	1.86×10^{-283}		P-value	0	
	adjusted P_i				
DHS	> 0.01	< 0.01			
NO	356284	69377			
YES	46378	13538			
odds ratio	1.50				
P-value	1.01×10^{-301}				

the selected and functional regions. If compared with the performance by the proposed method (Table 2), 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 an unlikely occurrence. Next, when considering HTS, we first tested DSS. For GSE34864, we found that $P = 0$ for the t test that denies the probability that the distribution of P-values attributed to DHS and non-DHS sites was identical (Figure 12). For EH1072, since DSS required more than a week of processing, even 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 that denies the probability that the distribution of P-values attributed to DHS and non-DHS sites was identical (Figure 13). 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) [16] was out of scope since EWAS requires numerous samples while the proposed method was effective even when the number of samples was less than ten.

Although one might also wonder if the proposed method was useless since it could not identify DMRs for HTS architecture, we could identify DMRs by integrating identified DMCs. To show this capability, we used eDMR [17] that could identify DMRs by integrating identified DMCs. We found that eDMR could identify 6,818 and 12,662 for GSE34864 and EH1072, respectively. Thus, the proposed method may also have the capability to identify

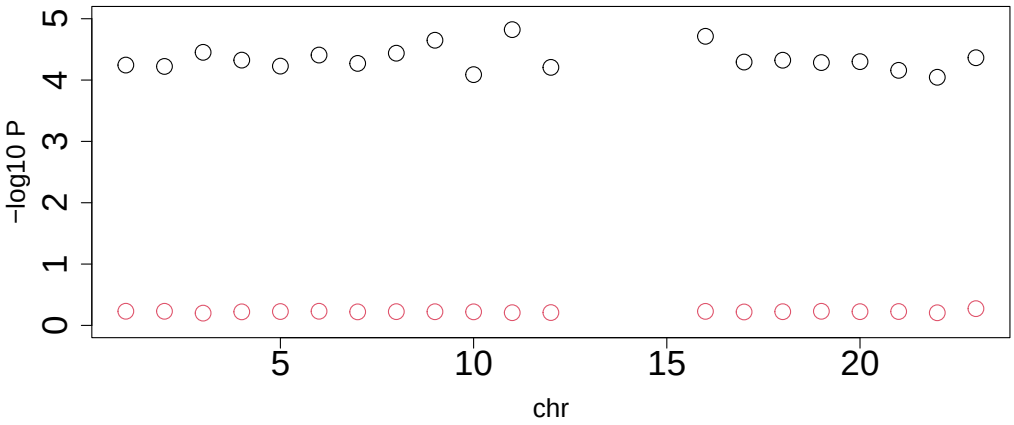

Figure 12. Negative logarithm of mean *P*-values between DHS (black circles) and non-DHS sites (red circles) within each chromosome (horizontal axis) when DSS was applied to GSE34864. *P*-values computed using *t* test between these two were zero. Some chromosomes did not have an associated DHS.

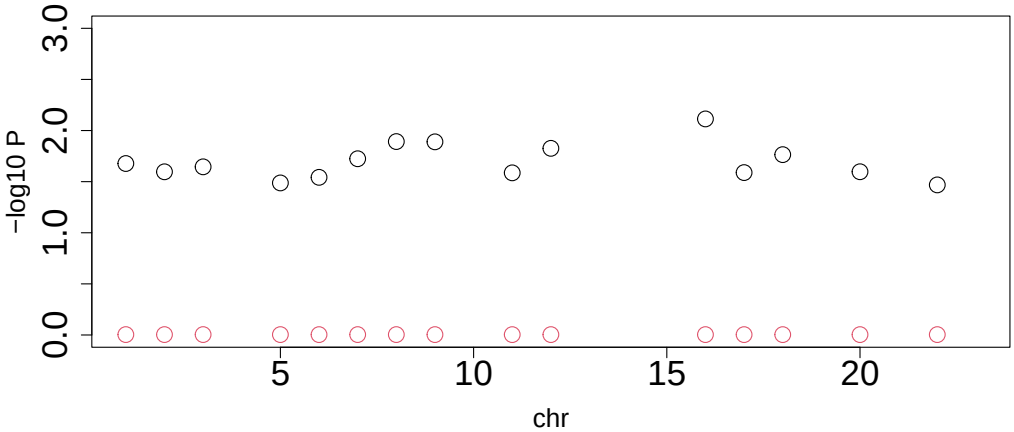

Figure 13. Negative logarithm of mean *P*-values between DHS (black circles) and non-DHS sites (red circles) within each chromosome (horizontal axis) when DMRcate was applied to EH1072. *P*-values computed using *t* test between these two were zero. Some chromosomes did not have an associated DHS.

DMRs if a suitable tool is employed. In contrast, 27 and 12,478 DMRs were identified for GSE34864 and EH1072, respectively, by metilene, which failed to identify DMCs for these profiles. Thus, eDMR combined with the proposed method had the capability to identify more DMRs, at least, than metilene.

In order to confirm that genomic regions selected by eDMR were significantly overlapped 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 8 shows P-values that are 0 within numerical accuracy. Thus, DMRs selected by eDMR were significantly overlapped with DHS.

Table 8. Binomial test for the coincidence between DHS and DMRs, identified by eDMR.

		GSE34864	EH1072
Length of	Human genome (ℓ_g)	3,117,275,501	
	DHS (ℓ_{g_1})	61,495,656	
	Overlap (ℓ'_{g_1})	86,523	53,907
	DMR (ℓ_{g_2})	3,246,111	742,103
	Overlap (ℓ'_{g_2})	418,206	25,541
$P_b\left(> \ell'_{g_1}, \ell_{g_1}, \frac{\ell_{g_2}}{\ell_g}\right)$		0	0
$P_b\left(> \ell'_{g_2}, \ell_{g_2}, \frac{\ell_{g_1}}{\ell_g}\right)$		0	0

4. Discussion

One might think that this study does not have enough novelty to be published as an independent study, since TD- and PCA-based unsupervised FE without optimized SD were invented many years ago and were successfully applied to a wide range of studies [15]; an advancement introduced in this study is only an optimization of 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 identify a very small number of DMCs. It is unlikely that there are no false negatives and the histogram of $1 - P_i$ is far from the ideal one. Upon optimizing SDs, we can now have a more ideal histogram of $1 - P_i$ and a sufficient number of DMCs which can compete 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 measurement by both microarray analysis and HTS without any specific modifications. For example, although Rackham et al [18] 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 [10]) was effective for DNA methylation without any specific modifications. This finding is remarkable enough to be reported. Thus, although the advancement may be small, the improvement in outcomes is large.

As can be seen in Figs. 2, 4, 6, and 9, the SD of h_n takes zero when the optimized SD is equal to zero. This is useless 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 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 Fisher's exact test because, for microarray architecture, the selection was performed among probes with annotations. Thus, we could compare two classifications, i.e., functional probes and non-functional probes versus selected probes and non-selected probes. Nevertheless, for HTS, since we only had a list of DHS and DMC, there were no common background sets from which we could define non-selected and non-DHS sites. Thus, we could not perform Fisher's exact test, and opted

to apply t test to P-values between DHS and non-DHS as this was possible even without a common background set.

5. Conclusion

In this paper, we applied the proposed method "PCA- and TD-based unsupervised FE with optimized SD" to the identification of DMCs. Identified DMCs were significantly overlapped with various functional sites. This technique also outperformed some modern methods, especially when applied to HTS architecture. We can also identify 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 PC score obeys a Gaussian distribution, which is rationalized in the framework for probabilistic PCA [19]; therefore, the proposed method is more adequate than other modern methods that must assume beta-binomial distribution.

Author Contributions: Y.-H.T. planned the research and performed the analyses. Y.-H.T. and T.T. evaluated the results, discussions, and outcomes and wrote the manuscript. All authors have read and agreed to the submitted version of 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 Statement: 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 28th February 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chen, D.P.; Lin, Y.C.; Fann, C.S.J. Methods for identifying differentially methylated regions for sequence- and array-based data. *Briefings in Functional Genomics* **2016**, *15*, 485–490, [https://academic.oup.com/bfg/article-pdf/15/6/485/8263147/elw018.pdf]. https://doi.org/10.1093/bfgp/elw018.
2. Luo, Y.; Lu, X.; Xie, H. DynamicAluMethylation during Normal Development, Aging, and Tumorigenesis. *BioMed Research International* **2014**, *2014*, 1–12. https://doi.org/10.1155/2014/784706.
3. Skinner, M.K.; Nilsson, E.; Sadler-Riggleman, I.; Beck, D.; Maamar, M.B.; McCarrey, J.R. Transgenerational sperm DNA methylation epimutation developmental origins following ancestral vinclozolin exposure. *Epigenetics* **2019**, *14*, 721–739, [https://doi.org/10.1080/15592294.2019.1614417]. PMID: 31079544, https://doi.org/10.1080/15592294.2019.1614417.
4. Shnorhavorian, M.; Schwartz, S.M.; Stansfeld, B.; Sadler-Riggleman, I.; Beck, D.; Skinner, M.K. Differential DNA Methylation Regions in Adult Human Sperm following Adolescent Chemotherapy: Potential for Epigenetic Inheritance. *PLOS ONE* **2017**, *12*, 1–18. https://doi.org/10.1371/journal.pone.0170085.
5. Tian, Y.; Morris, T.J.; Webster, A.P.; Yang, Z.; Beck, S.; Feber, A.; Teschendorff, A.E. ChAMP: updated methylation analysis pipeline for Illumina BeadChips. *Bioinformatics* **2017**, *33*, 3982–3984, [https://academic.oup.com/bioinformatics/article-pdf/33/24/3982/25168400/btx513.pdf]. https://doi.org/10.1093/bioinformatics/btx513.
6. Warden, C.D.; Lee, H.; Tompkins, J.D.; Li, X.; Wang, C.; Riggs, A.D.; Yu, H.; Jove, R.; Yuan, Y.C. COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis. *Nucleic Acids Research* **2013**, *41*, e117–e117, [https://academic.oup.com/nar/article-pdf/41/11/e117/25338160/gkt242.pdf]. https://doi.org/10.1093/nar/gkt242.
7. Peters, T.; Buckley, M.; Chen, Y.; Smyth, G.; Goodnow, C.; Clark, S. Calling differentially methylated regions from whole genome bisulphite sequencing with DMRcate. *Nucleic Acids Research* **2021**, *49*, e109–e109, [https://academic.oup.com/nar/article-pdf/49/19/e109/41071508/gkab637.pdf]. https://doi.org/10.1093/nar/gkab637.
8. Park, Y.; Wu, H. Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics* **2016**, *32*, 1446–1453, [https://academic.oup.com/bioinformatics/article-pdf/32/10/1446/16920397/btw026.pdf]. https://doi.org/10.1093/bioinformatics/btw026.
9. Jühling, F.; Kretzmer, H.; Bernhart, S.H.; Otto, C.; Stadler, P.F.; Hoffmann, S. metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome Research* **2015**, *26*, 256–262. https://doi.org/10.1101/gr.196394.115.
10. Taguchi, Y.h.; Turki, T. Tensor decomposition- and principal component analysis-based unsupervised feature extraction to select more reasonable differentially expressed genes: Optimization of standard deviation versus state-of-art methods. *bioRxiv* **2022**, [https://www.biorxiv.org/content/early/2022/02/22/2022.02.18.481115.full.pdf]. https://doi.org/10.1101/2022.02.18.481115, https://doi.org/10.1101/2022.02.18.481115.

11. BLUEPRINT consortium. Quantitative comparison of DNA methylation assays for biomarker development and clinical applications. *Nature Biotechnology* **2016**, *34*, 726–737. https://doi.org/10.1038/nbt.3605.
12. Smith, Z.D.; Chan, M.M.; Mikkelsen, T.S.; Gu, H.; Gnirke, A.; Regev, A.; Meissner, A. A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* **2012**, *484*, 339–344. https://doi.org/10.1038/nature10960.
13. Feng, H.; Conneely, K.N.; Wu, H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Research* **2014**, *42*, e69–e69, [https://academic.oup.com/nar/article-pdf/42/8/e69/25364302/gku154.pdf]. https://doi.org/10.1093/nar/gku154.
14. Morgan, M.; Shepherd, L. *ExperimentHub: Client to access ExperimentHub resources*, 2022. R package version 2.2.1.
15. Taguchi, Y.H. *Unsupervised Feature Extraction Applied to Bioinformatics*; Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-22456-1.
16. Flanagan, J.M. Epigenome-Wide Association Studies (EWAS): Past, Present, and Future. In *Methods in Molecular Biology*; Springer New York, 2014; pp. 51–63. https://doi.org/10.1007/978-1-4939-1804-1_3.
17. 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.; et al. An optimized algorithm for detecting and annotating regional differential methylation. *BMC Bioinformatics* **2013**, *14*. https://doi.org/10.1186/1471-2105-14-s5-s10.
18. Rackham, O.J.L.; Dellaportas, P.; Petretto, E.; Bottolo, L. WGBSSuite: simulating whole-genome bisulphite sequencing data and benchmarking differential DNA methylation analysis tools. *Bioinformatics* **2015**, *31*, 2371–2373, [https://academic.oup.com/bioinformatics/pdf/31/14/2371/17086629/btv114.pdf]. https://doi.org/10.1093/bioinformatics/btv114.
19. Tipping, M.E.; Bishop, C.M. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **1999**, *61*, 611–622. https://doi.org/10.1111/1467-9868.00196.