Tensor decomposition and principal component analysis-based unsupervised feature extraction outperforms state-of-the-art methods when applied to histone modification profiles

Sanjiban Sekhar Roy,Y-h. Taguchi

Abstract—Identification of histone modification from datasets that contain high throughput sequencing is difficult. Although multiple methods have been developed to identify histone modification, most of these methods are not specific for histone modification, but are general methods that aim to identify protein binding to the genome. In this study, tensor decomposition and principal component analysis based unsupervised feature extraction with optimized standard deviation, which were successfully applied to gene expression and DNA methylation, were used to identify histone modification. The proposed method identified various histone modifications successfully and also outperformed various state-of-the-art methods. The proposed method is expected to be a *de facto* standard method to identify histone modification.

Index Terms—Tensor decomposition, Principal component analysis, Histone modification, Feature selection, Unsupervised learning

1 Introduction

Identification of histone modification [1] from high throughput sequencing (HTS) datasets is important for the following reasons.

- Histone modification contributes to various functional genomic features, including transcription and alteration of chromatin structures.
- In contrast to variable gene expression, histone modification is more stable; thus, it may be used to characterize the state of the genome.
- In contrast to DNA methylation that has only two states, methylated or not, histones are modified in various ways. Thus, histone modification is related to a more detailed functionality of the genome.
- Since histone modification may activate or suppress transcription, combinations of various histone modifications may have more complicated transcriptional roles.

In spite of the importance of histone modification, compared with their counter parts, gene expression and DNA methylation, histone modification has been less focused upon in genomic studies for the following reasons.

- To identify histone modification throughout the genome, detection of antibody binding to the entire genome is required, in contrast to that of gene expression, which is detected even if only transcription start sites are sequenced.
- In contrast to gene expression that is measured even when only exons are considered or DNA methylation that is meaningful even if only promoter regions are considered, specific regions of the genome cannot be identified since there is very little knowledge on the position-specific functionality of histone modification.
- Because of the above two reasons, HTS for histone modification requires more depth, which is both time consuming and expensive.
- Similarly, the number of datasets required to identify histone modification is larger than that required to measure gene expression and DNA methylation, and requires more computational resources for analysis.

Possibly because of the unpopularity of histone modification identification, fewer methods specific to histone modification have been developed. Because histone modification is often measured using chromatin immunoprecipitation sequencing (ChIP-seq) technology to identify protein binding sites on DNA, so-called peak call programs that have been developed to process protein binding to DNA via ChIP-seq have often been used to process HTS datasets used for histone modification. However, since histone modifications are hardly regarded to be distributed over the genome by forming peaks, these peak calling algorithms are not guaranteed to identify histone modification. For example, although most methods employ binomial distributions to identify the amount of histone modification over the genome, we are unsure if this method is suitable. A more flexible method that does not assume a specific distribution of histone modification over the whole genome is required.

To fulfill this requirement, tensor decomposition (TD) and principal component analysis (PCA)-based unsuper-

- *Sanjiban Sekhar Roy is with the School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India*
- *Y-h. Taguchi is with the Department of Physics, Chuo University, Tokyo 112-8551, Japan.*
  Email: tag@granular.com.

Manuscript received April 19, 2005; revised August 26, 2015.

vised feature extraction (FE) with optimized standard deviation (SD) that were previously successfully applied to gene expression [2] and DNA methylation [3] were tested. In the following, "TD and PCA-based unsupervised FE with optimized SD" is denoted as the proposed method. In contrast to two previous studies [2], [3] where PC or singular value vectors (SVV) obey the Gaussian distribution assumed in the null hypothesis after SD optimization, PC and SVV do not follow a Gaussian distribution, but follow a mixed Gaussian distribution instead; nevertheless, empirically, SD optimization allows the correct identification of histone modification to some extent, even when compared to various state-of-the-art methods.

2 Materials and Methods

2.1 Histone modification profiles

The following histone modification profiles have been used to evaluate the performance of the proposed method. All profiles were retrieved from the Gene Expression Omnibus (GEO) database [4].

2.1.1 GSE24850

This study contained eleven H3K9me3 ChIP-seq mouse nucleus accumbens experiments comprising six controls, three saline-treated samples, and two cocaine samples [5]. Among them, five controls and five treated samples were employed to have an equal number of treated and control samples. Ten bed files that were provided as a Supplementary file in GEO were downloaded. One control samples with a GEO identification (ID) of GSM612984 was discarded and the other ten samples were used.

2.1.2 GSE159075

This study contained various histone modification ChIP-seq experiments using human umbilical vein endothelial cell lines [6]. Among them three H3K4me3 ChIP-seq, three H3K27me3 ChIP-seq, and three H3K27ac ChIP-seq, as well as one input ChIP-seq profile were employed. The corresponding 10 bigwig files were downloaded from the GEO Supplementary file.

2.1.3 GSE74055

This study contained various histone modification ChIP-seq experiments using mouse E14 or DKO cell lines [7]. Among them, 16 H3K4me1 ChIP-seq profiles and the corresponding 16 input profiles (in total 32 profiles) were downloaded in the Bigwig format provided in the GEO Supplementary file.

2.1.4 GSE124690

This study comprised various histone modification CUT&Tag experiments using human H1 or K562 cell lines [8]. Among them, six bulk H3K4me1 profiles (GSM3536499_H1_K4me1_Rep1.bed.gz, GSM3536499_H1_K4me1_Rep2.bed.gz, GSM3536516_K562_K4me1_Rep1.bed.gz, GSM3536516_K562_K4me1_Rep2.bed.gz, GSM3680223_K562_H3K4me1_Abcam_8895.bed.gz, GSM3680224_K562_H3K4me1_ActMot_39113.bed.gz), four bulk H3K3me3 profiles (GSM3536501_H1_K4me3_Rep1.bed.gz, GSM3536501_H1_K4me3_Rep2.bed.gz, GSM3536518_K562_K4me3_Rep1.bed.gz, GSM3536518_K562_K4me3_Rep2.bed.gz), and four bulk H3K27ac profiles (GSM3536497_H1_K27ac_Rep1.bed.gz, GSM3536497_H1_K27ac_Rep2.bed.gz, GSM3536514_K562_K27ac_Rep1.bed.gz, GSM3536514_K562_K27ac_Rep2.bed.gz) were downloaded from the GEO Supplementary file.

2.1.5 GSE188173

This study contained nine ChIP-seq H3K27ac profiles (with one control and one treated with SPT) using patient-derived xenografts of human castration-resistant prostate cancer (18 profiles in total). The corresponding 18 bigwig files were downloaded from the GEO Supplementary file.

2.1.6 GSE159022

This study comprised four H3K4me3 ChIP-seq profiles, four H3K27me3 ChIP-seq profiles, and four H4K16ac ChIP-seq profiles using mouse progenitor cells (two wild type (WT) and two neurofibromin knockouts) [9]. Among them, four H3K27me3 profiles were used and four bigwig files were downloaded from the GEO Supplementary file.

2.1.7 GSE168971

This study contained H3K27ac and H3K9ac ChIP-seq profiles taken from various experimental conditions [10]. Among them, six H3K9ac profiles using C3H-WT mouse liver and the two corresponding inputs were used. The corresponding eight bigwig files were downloaded from the GEO Supplementary file.

2.1.8 GSE159411

This study comprised various ChIP-seq profiles [11]. Among them, four H3K36me3 ChIP-seq profiles (two hiPSC cardiomyocyte and two WT hiPSC) were used. The four corresponding bigwig files were downloaded from the GEO Supplementary file.

2.1.9 GSE181596

This study consisted of four H3K27me3 ChIP-seq profiles (two controls and two treatments) and four H3K4me3 ChIP-seq profiles (two controls and two treatments) in addition to two input profiles that used cells as odontoblasts (treatment was siRNA: si-IKBz) [12]. Among them, four H3K27me3 ChIP-seq profiles were downloaded from the GEO Supplementary file.

2.2 Histone modification profile pre-processing

To apply the proposed method to histone modification profiles, individual histone modification profiles must share region index i. Therefore, the amount of histone modification is averaged within shared regions that is generated by dividing the whole human or mouse genome into equal lengths, L, intervals in each chromosome. $L = 1000$ is used for all histone profiles other than H3K9me3, for which $L = 25000$ is used.

There are two kinds of bed format for histone modification. One is coarse grained histone modification, where

histone modification is averaged within intervals; the other is the genomic loci where histone modification occurs. For the former, coarse grained values are further averaged with the region of length L; for the latter, the regions overlapping the regions of length L are counted.

2.3 PCA-based unsupervised FE with optimized SD

Although one of the proposed methods, PCA-based unsupervised FE with optimized SD, has been fully described in previous studies [2], [3], it is briefly described here.

Suppose histone modification profiles are formatted as a matrix $x_{ij} \in \mathbb{R}^{N\times M}$ that represents histone modification of ith regions in jth samples. Here, we assume that x_{ij} is normalized as

$$\sum_{i=1}^{N} x_{ij} = 0 \quad (1)$$

$$\sum_{i=1}^{N} x_{ij}^2 = N \quad (2)$$

Applying PCA to x_{ij} should obtain

$$x_{ij} = \sum_{\ell=1}^{\min(N,M)} \lambda_\ell u_{\ell i} v_{\ell j} \quad (3)$$

where

$$\sum_i u_{\ell i} u_{\ell' i} = \sum_j v_{\ell j} v_{\ell' j} = \delta_{\ell\ell'} \quad (4)$$

To get $u_{\ell i}$, the eigenvalues and vector of $\sum_j x_{ij} x_{i'j} \in \mathbb{R}^{N\times N}$ must be obtained as

$$\sum_{i'} \left(\sum_j x_{ij} x_{i'j} \right) u_{\ell i'} = \lambda_\ell u_{\ell i} \quad (5)$$

and $v_{\ell j}$ can be obtained as

$$v_{\ell j} = \sum_i u_{\ell i} x_{ij} \quad (6)$$

To select regions of interest, i, first, which $v_{\ell j}$ is associated with the desired property is identified. In this study, the property of interest is that histone modification is independent of samples, that is, $v_{\ell j}$ should be independent of j. After identifying ℓ of interest, an optimal SD of $u_{\ell i}$ is obtained such that $u_{\ell i}$ obeys the Gaussian distribution (null hypothesis) as much as possible.

SD optimization can be done as follows. First, P-values are attributed to the ith region as

$$P_i = P_{\chi^2} \left[> \left(\frac{u_{\ell i}}{\sigma_\ell} \right)^2 \right] \quad (7)$$

where $P_{\chi^2}[> x]$ is the cumulative χ^2 distribution where the argument is larger than x and σ_ℓ is the SD. Then, the histogram h_s of P_i is computed, which is the number of is that satisfy

$$\frac{s}{N_s} \leq 1 - P_i \leq \frac{s+1}{N_s}, 0 \leq s \leq N_s - 1 \quad (8)$$

h_s is expected to have constant values for $s \leq s_0$, with some s_0 and a sharp peak exists at $k_0 < k$. In this case, is included in $h_s, s > s_0$ are selected as associated with significant histone modification. For that, P_is are recomputed with optimized SD and the obtained P_is are corrected using BH criterion [13] and is with adjusted P_is less than 0.01 are selected.

2.4 TD-based unsupervised FE with optimized SD

How PCA is replaced with TD is as follows. Suppose that histone modification is formatted as a tensor $x_{ijk} \in \mathbb{R}^{N\times M\times K}$ in the ith region of the jth sample under the kth condition and TD is obtained as

$$x_{ijk} = \sum_{\ell_1=1}^{N} \sum_{\ell_2=1}^{M} \sum_{\ell_3=1}^{K} G(\ell_1\ell_2\ell_3) u_{\ell_1} \quad (9)$$

where $G \in \mathbb{R}^{N\times M\times K}$ and

$$\sum_i u_{\ell_1 i} u_{\ell_1' i} = \delta_{\ell_1\ell_1'} \quad (10)$$

$$\sum_j u_{\ell_2 j} u_{\ell_2' j} = \delta_{\ell_2\ell_2'} \quad (11)$$

$$\sum_k u_{\ell_3 k} u_{\ell_3' k} = \delta_{\ell_3\ell_3'} \quad (12)$$

with higher order singular value decomposition (HOSVD [13]). Then, after identifying ℓ_2 and ℓ_3 of interest by investigating the dependence of $u_{\ell_2 j}$ and $u_{\ell_3 k}$ on j and k, $u_{\ell_1 i}$ associated with absolutely the largest G is selected with selected ℓ_2 and ℓ_3. The following procedure is the same as above.

2.5 Performance evaluation by Enrichr

After regions were selected, Entrez gene IDs included in these regions were listed and were converted to gene symbols using the gene ID converter implemented in DAVID [14], [15]. The obtained list of gene symbols was uploaded to Enrichr [16]. The primary "Epigenomics Roadmap HM ChIP-seq" category was employed to evaluate the performance and the "ENCODE Histone Modifications 2015" category was additionally used for the evaluation if no significant results were obtained in the "Epigenomics Roadmap HM ChIP-seq" category using the proposed method.

2.6 Methods to be compared

The following methods were used in comparison to the proposed method. The criteria to be selected were as follows.

- The method must accept bed or Bigwig file formats as input since the proposed method cannot accept the sam/bam format as input, which is accepted by most popular methods.
- The method must be standalone (not web-based).
- The method must be performed in the Linux platform.
- The method must be implemented as free (open) software.

The dataset tested was GSE24850 (H3K9me3).

2.6.1 MOSAiCS

MOSAiCS [17] was implemented as a bioconductor package [18]. Version 2.32.0 was installed in R [19] and applied to GSE24850.

2.6.2 DFilter

DFilter [20] was implemented as a Linux command-line program. Ver. 1.6 was downloaded from the website https://reggenlab.github.io/DFilter/.

2.6.3 F-Seq2

F-Seq2 [21] was implemented as a Linux command line program. It was downloaded from Github https://github.com/Boyle-Lab/F-Seq2.

2.6.4 HOMER

HOMER [22] was implemented as a Linux command line program. The latest version, HOMER 4.11, which was released on October 24, 2019, was downloaded from the Web site http://homer.ucsd.edu/homer/.

2.6.5 RSEG

RSEG [23] was implemented as a Linux command line program. The latest version, 0.4.9, was downloaded from the web site http://smithlabresearch.org/software/rseg/.

2.7 Gaussian Mixture Analysis

Gaussian mixture analysis was performed using `mclust` package [24] in R [19]. The Mclust function was applied to u_{1i} for GSE24850.

3 Results

A full list of genes and the enrichment analysis is in the Supplementary materials.

3.1 GSE24850

3.1.1 PCA-based unsupervised FE with optimized SD

The proposed method was applied to H3K9me3 modifications in the GSE24850 dataset. PCA was applied to three saline-treated samples and two cocaine samples (five samples in total). The first PC loading, v_{1j}, attributed to samples was independent of j attributed to samples. Then, the corresponding first PC score, u_{1i}, was used for gene selection. SD was optimized and the SD = 0.04264199. The histogram of $1 - P_i$ seemed to obey a combination of two Gaussian distributions and not a single Gaussian distribution (Fig. 1). A small number of regions (1302) were selected among a total of 106,204 regions (i.e., only a few percentages) and the associated 894 Entrez gene IDs were identified. They were further converted to 641 gene symbols, which were uploaded to Enrichr. Table 1 shows the enriched histone modification associated with adjusted P values less than 0.05. All were H3K9me3 modifications, suggesting that the proposed method was very successful. In addition, five out of ten were brain related, which was coincident with the fact that GSE24850 comprised experiments using the nucleus accumbens, strengthening the success of the proposed method.

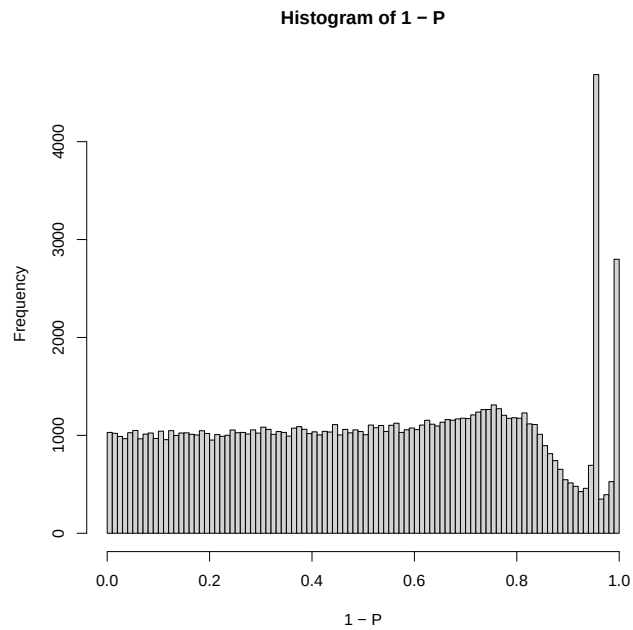

Fig. 1. Histogram of $1 - P_i$ computed using the proposed method for the GSE24850 dataset (H3K9me3)

3.1.2 Comparisons with state-of-the-art methods

Although the proposed method successfully achieved a significant performance, if other state-of-the-art methods achieve a similar performance, the importance of the proposed method drastically decreases. To reject this possibility, various state-of-the-art methods were used to analyze the GSE24850 dataset. No state-of-the-art method had a comparative performance.

3.1.2.1 MOSAiCS: The first state-of-the-art method tested was MOSAiCS. Since MOSAiCS only accepts a pair of ChIP-seq and input datasets (i.e., a control), MOSAiCS was repeatedly applied to five pairs of ChIP-seq data (i.e., three saline and two cocaine samples) and one input profile. MOSAiCS identified 4367, 3648, 2096, 1985, and 5566 peak regions and 1833, 1599, 1136, 1018, and 2223 associated Entrez gene IDs. Finally, 994, 851, 567, 532, and 1184 gene symbols were identified. Next, these five sets of genes were uploaded to Enrichr one by one and how many histone modifications were significantly enriched in the "Epigenomics Roadmap HM ChIP-seq" category of Enrichr was determined. Table 2 shows the performance of MOSAICS, which is clearly inferior to the proposed method, which identifies ten significantly enriched histone modifications, all of which are H3K9me3 modifications. Thus, it was obvious that the proposed method outperformed MOSAiCS.

3.1.2.2 DFilter: The next state-of-the-art method tested was DFilter, which was also applied to the five pairs of data. DFilter identified 25080, 22863, 21371, 23811, and 23369 peak regions and 6286, 5721, 5407, 5987, and 5903 Entrez gene IDs, of which 2621, 2524, 2499, 2631, and 2544 gene symbols were associated with each of the five pairs. These sets of gene symbols were uploaded to Enrichr and no enriched H3K9me3 modifications were observed (Table 2). Thus, DFilter was inferior to the proposed method.

TABLE 1
Enriched histone modification in the "Epigenomics Roadmap HM ChIP-seq" category of Enrichr for the GSE24850 dataset (H3K9me3)

Term	Overlap	P-value	Adjusted P-value
H3K9me3 Brain Mid Frontal Lobe	22/217	2.02×10^{-6}	7.45×10^{-4}
H3K9me3 Brain Inferior Temporal Lobe	18/170	9.31×10^{-6}	1.72×10^{-3}
H3K9me3 CD4 Naive Primary Cells	10/64	3.39×10^{-5}	4.17×10^{-3}
H3K9me3 Brain Anterior Caudate	15/141	4.82×10^{-5}	4.45×10^{-3}
H3K9me3 IMR90	45/814	2.82×10^{-4}	2.08×10^{-2}
H3K9me3 Brain Hippocampus Middle	15/170	3.90×10^{-4}	2.40×10^{-2}
H3K9me3 Stomach Smooth Muscle	17/217	6.56×10^{-4}	3.33×10^{-2}
H3K9me3 Colon Smooth Muscle	10/92	7.24×10^{-4}	3.33×10^{-2}
H3K9me3 CD8 Naive Primary Cells	11/110	8.11×10^{-4}	3.33×10^{-2}
H3K9me3 Brain Cingulate Gyrus	11/115	1.17×10^{-3}	4.33×10^{-2}

TABLE 2
The number of regions/peaks, Entrez gene IDs, and gene symbols selected by various methods and that of the H3K9me3-associated enriched histone modification in the "Epigenomics Roadmap HM ChIP-seq" category of Enrichr with adjusted P-values less than 0.05 for the GSE24850 dataset (H3K9me3).

Methods	Pair no.	Regions/peaks	Entrez genes	Gene symbols	Enriched histone modification	H3K9me3
Proposed method	—	1302	894	641	10	10
MOSAiCS	1	4367	1833	994	3	1
	2	3648	1599	851	0	0
	3	2096	1136	567	0	0
	4	1985	1018	532	2	2
	5	5556	2223	1184	2	0
DFilter	1	25080	6286	2621	1	0
	2	22863	5721	2524	1	0
	3	21371	5470	2499	1	0
	4	23811	5987	2631	1	0
	5	23369	5902	2544	1	0
F-Seq2	—	—	—	—	—	—
HOMER	—	114727	6771	6747	1	0
RSEG	—	—	—	—	—	—

3.1.2.3 F-Seq2: Although the next state-of-the-art method was F-Seq2, it could not be performed within the available computer memory size. Since the purpose was to demonstrate the superiority of the proposed method over other state-of-the-art methods, the comparison between F-Seq2 and the proposed method was terminated at this stage.

3.1.2.4 HOMER: The next state-of-the-art method is HOMER, which has been used in the ENCODE project and is used for the most recent reports [25]. Since HOMER has the capability to compare five input profiles with five H3K9me3 profiles directly, this was done. HOMER identified 114727 peak regions, 6771 associated Entrez genes, and 6747 gene symbols, which were uploaded to Enrichr. Unfortunately, Enrichr did not identify enriched regions of H3K9me3 modifications (Table 2).

3.1.2.5 RSEG: The last comparison was with the state-of-the-art method RSEG, which resulted in a core dump, although no compile errors were detected. The RSEG demo file also could not be treated without errors; it was too old to be used on the present platform. Actually, although it has been cited many times as one of the popular tools to process ChIP-seq datasets, it has never been really tested in recent publications.

3.2 Histone modification other than H3K9me3

Although the superiority of the proposed method has been demonstrated using H3K9me3 profiles, one of five "core histone marks" proposed by the Roadmap Epigenomics Consortium [26] (H3K4me1/H3K27ac, H3K4me3, H3K36me3, H3K27me3, and H3K9me3), there is a possibility that the proposed method is not effective with other histone modifications. To reject this possibility, the proposed method was tested on various other histone modifications.

3.2.1 GSE159075

Regions, is, with non-zero missing values among three samples were discarded in advance. PCA was applied to three samples with the same histone modification (H3K4me3, H3K27me3, or H3K27ac) and v_{1j} was always associated with the independence of j (biological replicates). The first PC loading, v_{1j}, attributed to samples was independent of j attributed to samples. Next, the corresponding first PC score, u_{1i}, was used for gene selection. SDs were optimized and the SD = 0.19218492 (H3K4me3), 0.8650455 (H3K27me3), and 0.3911769 (H3K27ac). The histogram of $1 - P_i$ seemed to obey a combination of two Gaussian distributions, and not a single Gaussian distribution (Fig. 2). As a result, 34538, 62141, and 61306 regions were selected, and 13692, 5217, and 11604 Entrez gene IDs, and the 13671, 5208, 11590 gene symbols associated with them were identified and uploaded to Enrichr. Table 3 shows the performance evaluation by Enrichr, which clearly works pretty well.

3.2.2 GSE74055

PCA was applied to 16 H3K4me1 modifications divided by the corresponding input (Regions, is, for which input samples with zero were discarded in advance). The u_{1i} could not provide reasonable results; u_{2i} was used for gene

TABLE 3
The number of regions/peaks, Entrez gene IDs, and gene symbols selected by various methods and that of associated enriched histone modification (all and targeted) in the "Epigenomics Roadmap HM ChIP-seq" category of Enrichr with adjusted P-values less than 0.05 for other profiles.*:ENCODE Histone Modifications 2015

GEO ID	Histone modification	Regions/peaks	Entrez genes	Gene symbols	Histone modification All	Histone modification Targeted	Species
GSE159075	H3K4me3	34538	13692	13671	198	54	Human
	H3K27me3	62141	5217	5208	83	56	
	H3K27ac	61306	11604	11590	175	24	
GSE74055	H3K4me1	61329	11890	11858	58*	6*	Mouse
(when PCA replaced with TD)		70187	14220	14187	102*	10*	
GSE124690 (CUT&Tag)	H3K4me1	164466	14893	14866	3*	0*	Human
	H3K4me3	37534	14972	14946	200	54	
	H3K27ac	81249	13086	13061	139	24	
GSE188173	H3K27ac	105438	15579	15548	155	24	Human
GSE159022	H3K27me3	55923	5022	4996	70	56	Mouse
GSE168971	H3K9ac	58490	15460	15452	81*	6*	Mouse
GSE159411	H3k36me3	253326	12282	12270	201	32	Human
GSE181596	H3K27me3	36972	3543	3545	72	56	Human

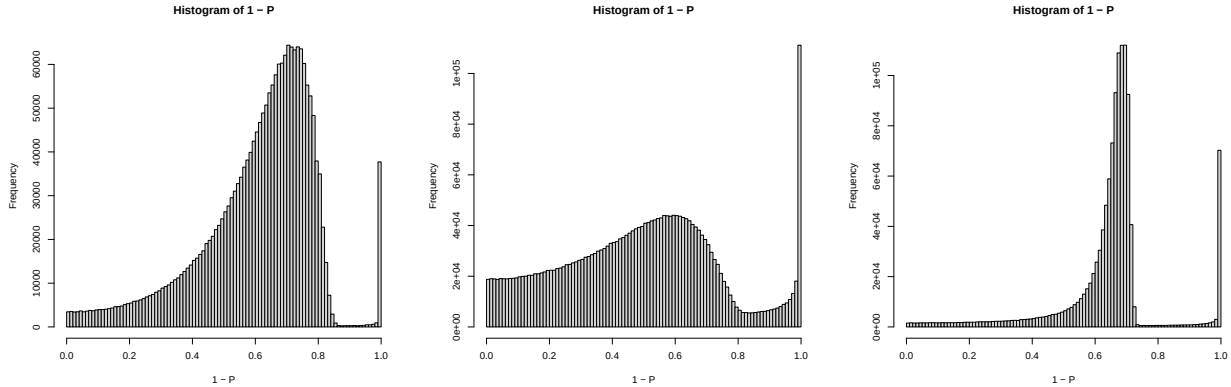

Fig. 2. Histogram of $1 - P_i$ computed by the proposed method using the GSE159075 dataset (from left to right, H3K4me3, H3K27me3, and H3K27ac)

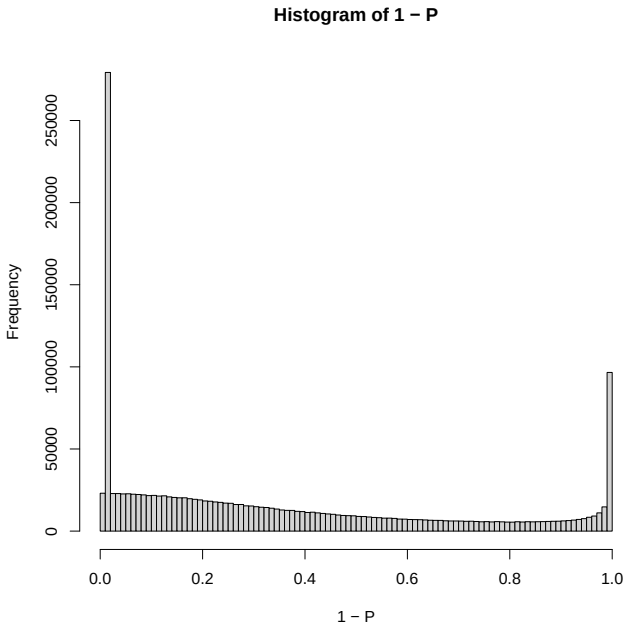

Fig. 3. Histogram of $1 - P_i$ computed by the proposed method using the GSE74055 dataset (H3K4me1)

selection. SDs were optimized and the SD = 0.1873585. Figure 3 shows the histogram of $1 - P_i$ with optimized SD. Next, 61329 regions, 11890 Entrez gene IDs, and 11858 gene symbols were identified, which were uploaded to Enrichr. Table 3 shows the performance evaluation by Enrichr. Its superiority was reduced since no enriched H3K4me1 modifications were identified for the "Epigenomics Roadmap HM ChIP-seq" category. The other category "ENCODE Histone Modifications 2015", which had more enriched histone modifications, needs to be consulted. Regardless, non-zero enriched H3K4me1 profiles could be detected in Enrichr.

3.2.3 GSE124690

PCA was applied to six H3K4me1, four H3K4me3, and four H3K27ac ChIP-seq profiles separately and v_{1j} was always associated with the independence of j (biological replicates). However, v_{2j} was used for H3K4me1 modifications as u_{2i} could not provide good results. Next, the corresponding first PC score, u_{1i}, was used for gene selection for H3K4me3 and H3K27ac modifications and the second PC score, u_{2i}, was used for gene selection for H3K4me1 modifications. SDs were optimized and the SD = 0.5468109 (H3K4me1), 0.5600861 (H3K4me3), and 0.4509331 (H3K27ac). The histogram of $1 - P_i$ seemed to obey a combination of two Gaussian distributions and not a single Gaussian distribu-

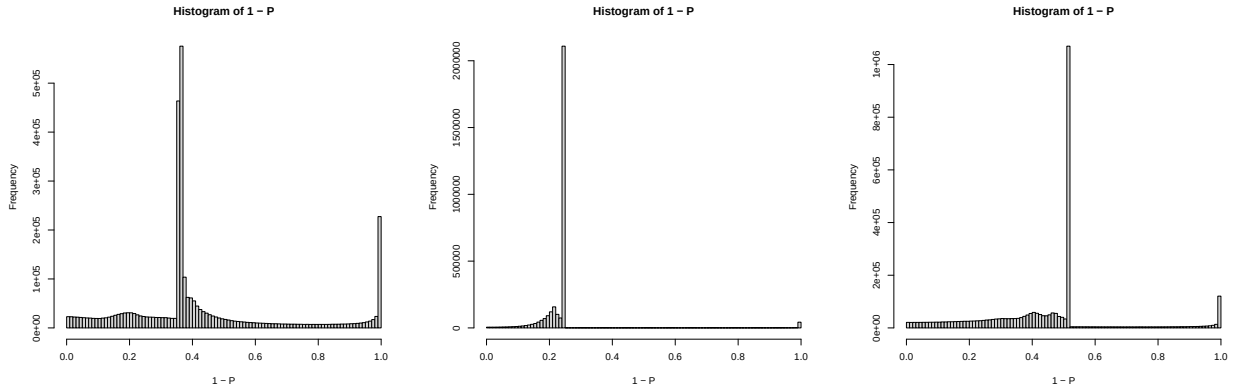

Fig. 4. Histogram of $1-P_i$ computed by the proposed method using the GSE124690 dataset (From left to right, H3K4me1, H3K4me3, and H3K27ac)

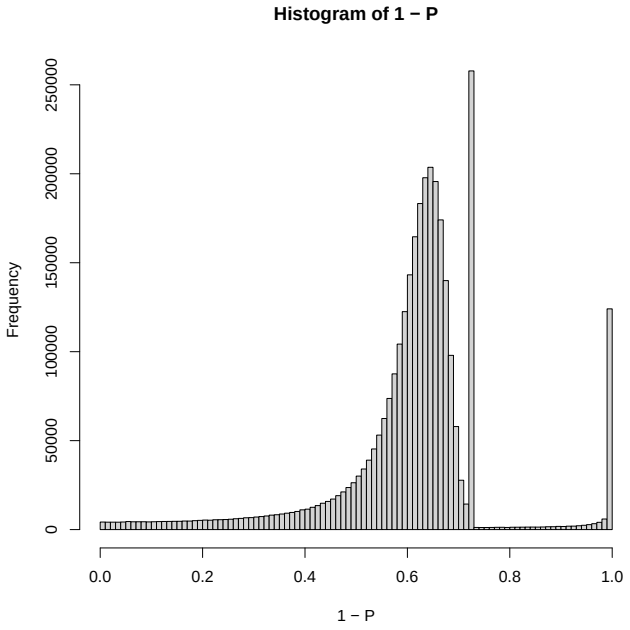

Fig. 5. Histogram of $1-P_i$ computed by the proposed method using the GSE188173 dataset (H3K27ac)

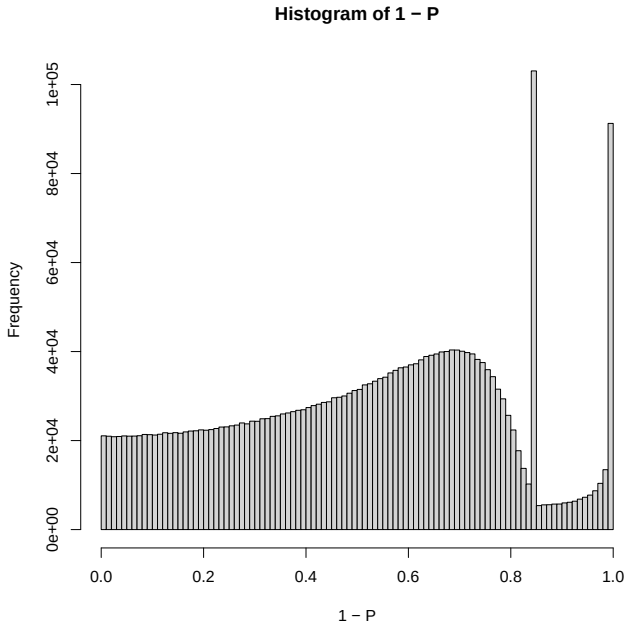

Fig. 6. Histogram of $1-P_i$ computed by the proposed method using the GSE159022 datset (H3K27me3)

tion (Fig. 4). A total of 164466, 37534, and 81249 regions, 14893, 14972, and 13061 Entrez gene IDs, and 14866, 14946, and 13061 gene symbols (2nd PC was used for H3K4me1) were identified, which were uploaded to Enrichr. Table 3 shows the performance evaluation by Enrichr. Again, its superiority toward H3K4me1 modifications was reduced since no enriched H3K4me1 modifications were identified for the "Epigenomics Roadmap HM ChIP-seq" or "ENCODE Histone Modifications 2015" categories. For the other two histone modifications, Enrichr was very successful.

3.2.4 GSE188173

PCA was applied to 16 H3K27ac profiles and v_{1j} was always associated with the independence of j (biological replicates). Next, the corresponding first PC score, u_{1i}, was used for gene selection. SDs were optimized, and the SD = 0.5319398. Figure 5 shows the histogram of $1-P_i$ with optimized SD. A total of 105438 regions, 15579 Entrez gene IDs, and 15548 gene symbols were identified that were uploaded to Enrichr. Table 3 shows the performance evaluation by Enrichr, which clearly works pretty well.

3.2.5 GSE159022

PCA was applied to H3K27me3 ChIP-seq profiles and v_{1j} was always associated with the independence of j (biological replicates). Then, the corresponding first PC score, u_{1i}, was used for gene selection. SDs were optimized and the SD = 0.06544008. Figure 6 shows the histogram of $1-P_i$ with optimized SD. A total of 55923 regions, 5022 Entrez gene IDs, and 4996 gene symbols were identified, which were uploaded to Enrichr. Table 3 shows the performance evaluation by Enrichr, which clearly works pretty well.

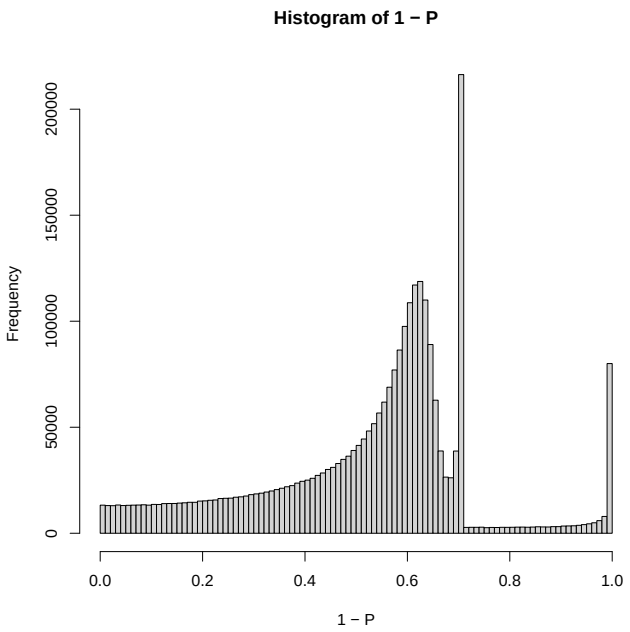

Fig. 7. Histogram of $1 - P_i$ computed by the proposed method using the GSE168971 dataset (H3K9ac)

3.2.6 GSE168971

PCA was applied to six H3K9ac ChIP-seq profiles and two corresponding inputs. v_{1j} was always associated with the independence of j (biological replicates). Next, the corresponding first PC score, u_{1i}, was used for gene selection. SDs were optimized and the SD = 0.3697877. Figure 7 shows the histogram of $1 - P_i$ with optimized SD. A total of 58490 regions, 15460 Entrez gene IDs, and 15452 gene symbols were identified, which were uploaded to Enrichr. Table 3 shows the performance evaluation by Enrichr. Its performance for H3K9ac modifications was reduced since the "ENCODE Histone Modifications 2015" category had to be consulted and it only identified six enriched H3K9ac profiles in the "ENCODE Histone Modifications 2015" category.

3.2.7 GSE159411

PCA was applied to four H3k36me3 ChIP-seq profiles. v_{1j} was always associated with the independence of j (biological replicates). Then, the corresponding first PC score, u_{1i}, was used for gene selection. SDs were optimized and the SD = 0.6005592. Figure 8 shows the histogram of $1 - P_i$ with optimized SD. A total of 25,3326 regions, 12282 Entrez gene IDs, and 12270 gene symbols were identified, which were uploaded to Enrichr. Table 3 shows the Enrichr performance evaluation, which clearly works pretty well.

3.2.8 GSE181596

PCA was applied to four H3K27me3 ChIP-seq profiles. v_{1j} was always associated with the independence of j (biological replicates). Next, the corresponding first PC score, u_{1i}, was used for gene selection. SDs were optimized, and the SD = 1.855494. Figure 9 shows the histogram of $1 - P_i$ with optimized SD. A total of 36972 regions, 3543 Entrez gene IDs, and 3545 gene symbols were identified, which were

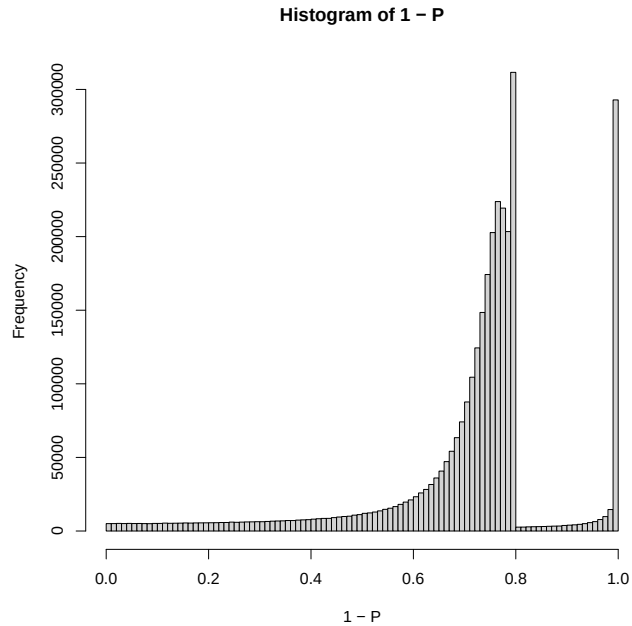

Fig. 8. Histogram of $1 - P_i$ computed by the proposed method using the GSE159411 dataset (H3K36me3)

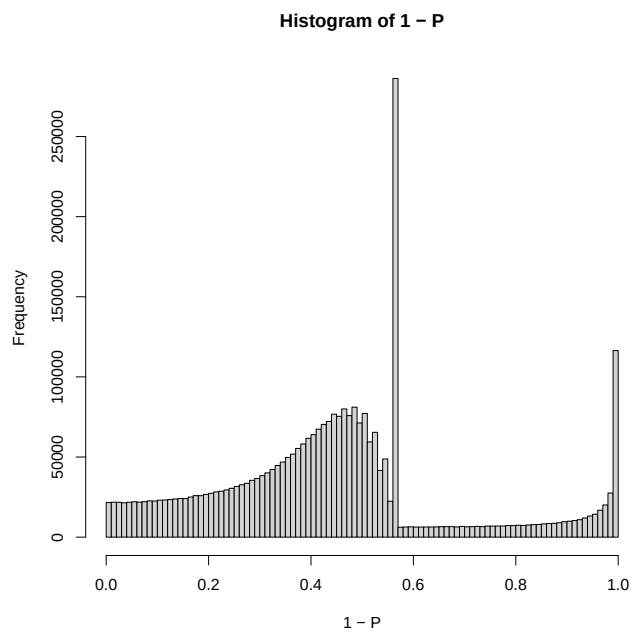

Fig. 9. Histogram of $1 - P_i$ computed by the proposed method using the GSE181596 dataset (H3K27me3)

uploaded to Enrichr. Table 3 shows the Enrichr performance evaluation, which clearly works pretty well.

3.3 TD-based unsupervised FE with optimized SD

All of the above experiments used PCA. To determine whether TD-based unsupervised FE with optimized SD worked as well, TD-based unsupervised FE with optimized SD was used to analyze the GSE74005 dataset, to which PCA-based unsupervised FE with optimized SD did not work very well. Histone modification was formatted as a tensor $x_{ijk} \in \mathbb{R}^{N \times 16 \times 2}$ that represented H3K4me1 modifications in the ith region of the jth sample in the kth treatment ($k = 1$: treated, $k = 2$: input). After obtaining SVV, u_{1j} and u_{2k} were coincidences with the independence of the samples and the distinction between the treated and controlled samples. $G(2, 1, 2)$ had the largest absolute value, thus u_{2i} was used to select regions. As a result, a total of 70187 regions, the associated 14220 Entrez gene IDs, and 14187 gene symbols were identified. Although these numbers are not so different from those in Table 3, performance is improved (Table 3). Thus, when PCA does not work, TD is worth testing as well.

4 Discussion

The proposed method had several advantages. First of all, it is a fully linear method, since it has been done simply by applying PCA or HOSVD to matrices and tensors. The only time-consuming process was SD optimization, which typically ends in a few minutes. The second advantage was to accept the bed or the bigwig file format. In contrast to bam/sam files that retain information about how individual short reads are mapped onto the genome, bed or bigwig files do not keep this information and instead retain to which regions the short read are mapped. The third is robustness. Independent of samples, species, and platforms, the proposed method achieved quite similar performances. For example, the H3K27me3 modification is considered three times in Table 3; two are human ChIP-seq and one is mouse ChIP-seq. In spite of that, these three are associated with a similar number of gene symbols, 5208, 4994, and 3545 and the same number of enriched H3K27me3 profiles in the "Epigenomics Roadmap HM ChIP-seq" category of Enrichr. H3K27ac modifications were considered three times with two distinct protocols, ChIP-seq and CUT&Tag;. A similar number of gene symbols, 11590, 13061, and 15548 and the same number of H3K27ac-enriched profiles, 24 in the "Epigenomics Roadmap HM ChIP-seq" category of Enrichr were associated. Usually, although CUT&Tag requires protocols specific to CUT&Tag [8], the proposed method handled both ChIP-seq and CUT&Tag seamlessly. In addition to this, as seen in Tables 2 and 3, the proposed method successfully deals with all five "core histone marks" (H3K27ac, H3K4me3, H3K36me3, H3K27me3, and H3K9me3. Unfortunately, H3K4me1 cannot be properly dealt with using the proposed method. However, this is not a critical problem, since it can process H3K27ac modifications, which are strongly associated with H3K4me1 modifications [27]. Thus, the proposed method is worth investigating.

In the above, $u_{\ell i}$ does not obey a Gaussian distribution, but instead fits a combination of two Gaussian distributions. To confirm these points, the analysis of the Gaussian mixture to u_{1i} of the GSE24850 dataset was applied (whose histogram $1 - P_i$ is shown in Fig. 1). Then, the Bayesian information criterion of the Gaussian mixture drastically improves when the number of clusters increases from one to two and does not improve for more numbers of clusters. Table 4 shows the confusion matrix between clustering and selection of the proposed method. It is obvious that they are in agreement. All regions not selected by the proposed method are in one cluster and the majority of the selected regions belong to another cluster. Thus, as expected, $u_{\ell i}$ did not obey a Gaussian distribution, but a mixture of two Gaussian distributions. Since the optimization of SD assumes a single Gaussian distribution, it is not expected to work well, but it works empirically pretty well (Table 3). The limitations of the proposed methods applied to histone modification must be investigated in the future. FALSE TRUE 1 2

TABLE 4
Confusion matrix between clustering by a mixture of Gaussian distributions(row) and selection by the proposed method (column)

		Proposed Method	
		Adjusted P-value	
	Cluster	> 0.01	≤ 0.01
Gaussian Mixture	1	104902	428
	2	0	874

However, the proposed method has some weaknesses. If it is applied to non-core histone marks, it does not work well. For example, if it is applied to H3K9ac modifications, its performance is reduced. More detailed conditions by which the proposed method may be successfully applied to histone modification should be investigated.

5 Conclusions

In this paper, a recently proposed PCA-based unsupervised FE with optimized SD was applied to identify histone modifications. It successfully counted five core histone marks and outperformed the state-of-the-art methods. The proposed method is expected to be a new state-of-the-art method for histone modification.

Acknowledgments

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

References

[1] R. Nakato and T. Sakata, "Methods for chip-seq analysis: A practical workflow and advanced applications," *Methods*, vol. 187, pp. 44–53, 2021, advance Epigenetics Methods in Biomedicine. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1046202320300591

[2] Y.-h. Taguchi and T. Turki, "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. [Online]. Available: https://www.biorxiv.org/content/early/2022/02/22/2022.02.18.481115

[3] Y. H. Taguchi and T. Turki, "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," *bioRxiv*, 2022. [Online]. Available: https://www.biorxiv.org/content/early/2022/04/05/2022.04.02.486807

[4] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, no. 1, pp. 207–210, 01 2002. [Online]. Available: https://doi.org/10.1093/nar/30.1.207

[5] I. Maze, J. Feng, M. B. Wilkinson, H. Sun, L. Shen, and E. J. Nestler, "Cocaine dynamically regulates heterochromatin and repetitive element unsilencing in nucleus accumbens," *Proceedings of the National Academy of Sciences*, vol. 108, no. 7, pp. 3035–3040, 2011. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1015483108

[6] Y. Kanki, M. Muramatsu, Y. Miyamura, K. Kikuchi, Y. Higashijima, R. Nakaki, J. ichi Suehiro, Y. Sasaki, Y. Kubota, H. Koseki, H. Morioka, T. Kodama, M. Nakao, D. Kurotaki, H. Aburatani, and T. Minami, "Bivalent-histone-marked immediate-early gene regulation is vital for VEGF-responsive angiogenesis," *Cell Reports*, vol. 38, no. 6, p. 110332, Feb. 2022. [Online]. Available: https://doi.org/10.1016/j.celrep.2022.110332

[7] J. Yan, S.-A. A. Chen, A. Local, T. Liu, Y. Qiu, K. M. Dorighi, S. Preissl, C. M. Rivera, C. Wang, Z. Ye, K. Ge, M. Hu, J. Wysocka, and B. Ren, "Histone h3 lysine 4 monomethylation modulates long-range chromatin interactions at enhancers," *Cell Research*, vol. 28, no. 2, pp. 204–220, Jan. 2018. [Online]. Available: https://doi.org/10.1038/cr.2018.1

[8] H. S. Kaya-Okur, S. J. Wu, C. A. Codomo, E. S. Pledger, T. D. Bryson, J. G. Henikoff, K. Ahmad, and S. Henikoff, "CUT&Tag for efficient epigenomic profiling of small samples and single cells," *Nature Communications*, vol. 10, no. 1, Apr. 2019. [Online]. Available: https://doi.org/10.1038/s41467-019-09982-5

[9] X. Wei, M. Lienhard, A. Murgai, J. Franke, S. Pöhle-Kronawitter, G. Kotsaris, H. Wu, S. Börno, B. Timmermann, R. Glauben, and S. Stricker, "Neurofibromin 1 controls metabolic balance and notch-dependent quiescence of juvenile myogenic progenitors," *bioRxiv*, 2021. [Online]. Available: https://www.biorxiv.org/content/early/2021/09/30/2021.09.28.462168

[10] G. V. Sarode, K. Neier, N. M. Shibata, Y. Shen, D. A. Goncharov, E. A. Goncharova, T. A. Mazi, N. Joshi, M. L. Settles, J. M. LaSalle, and V. Medici, "Wilson disease: Intersecting DNA methylation and histone acetylation regulation of gene expression in a mouse model of hepatic copper accumulation," *Cellular and Molecular Gastroenterology and Hepatology*, vol. 12, no. 4, pp. 1457–1477, 2021. [Online]. Available: https://doi.org/10.1016/j.jcmgh.2021.05.020

[11] B. Gonzalez-Teran, M. Pittman, F. Felix, R. Thomas, D. Richmond-Buccola, R. Hüttenhain, K. Choudhary, E. Moroni, M. W. Costa, Y. Huang, A. Padmanabhan, M. Alexanian, C. Y. Lee, B. E. Maven, K. Samse-Knapp, S. U. Morton, M. McGregor, C. A. Gifford, J. Seidman, C. E. Seidman, B. D. Gelb, G. Colombo, B. R. Conklin, B. L. Black, B. G. Bruneau, N. J. Krogan, K. S. Pollard, and D. Srivastava, "Transcription factor protein interactomes reveal genetic determinants in heart disease," *Cell*, vol. 185, no. 5, pp. 794–814.e30, Mar. 2022. [Online]. Available: https://doi.org/10.1016/j.cell.2022.01.021

[12] H. Yuan, S. Suzuki, H. Terui, S. Hirata-Tsuchiya, E. Nemoto, K. Yamasaki, M. Saito, H. Shiba, S. Aiba, and S. Yamada, "Loss of IκBζ drives dentin formation via altered H3K4me3 status," *Journal of Dental Research*, vol. 0, no. 0, p. 00220345221075968, 0, pMID: 35193410. [Online]. Available: https://doi.org/10.1177/00220345221075968

[13] Y.-H. Taguchi, *Unsupervised Feature Extraction Applied to Bioinformatics*. Springer International Publishing, 2020. [Online]. Available: https://doi.org/10.1007/978-3-030-22456-1

[14] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature Protocols*, vol. 4, no. 1, pp. 44–57, Dec. 2008. [Online]. Available: https://doi.org/10.1038/nprot.2008.211

[15] ——, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Research*, vol. 37, no. 1, pp. 1–13, 11 2008. [Online]. Available: https://doi.org/10.1093/nar/gkn923

[16] Z. Xie, A. Bailey, M. V. Kuleshov, D. J. B. Clarke, J. E. Evangelista, S. L. Jenkins, A. Lachmann, M. L. Wojciechowicz, E. Kropiwnicki, K. M. Jagodnik, M. Jeon, and A. Ma'ayan, "Gene set knowledge discovery with enrichr," *Current Protocols*, vol. 1, no. 3, p. e90, 2021. [Online]. Available: https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/cpz1.90

[17] G. Sun, D. Chung, K. Liang, and S. Keleş, "Statistical analysis of ChIP-seq data with MOSAiCS," in *Methods in Molecular Biology*. Humana Press, 2013, pp. 193–212. [Online]. Available: https://doi.org/10.1007/978-1-62703-514-9_12

[18] W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K. D. Hansen, R. A. Irizarry, M. Lawrence, M. I. Love, J. MacDonald, V. Obenchain, A. K. Ole's, H. Pag'es, A. Reyes, P. Shannon, G. K. Smyth, D. Tenenbaum, L. Waldron, and M. Morgan, "Orchestrating high-throughput genomic analysis with Bioconductor," *Nature Methods*, vol. 12, no. 2, pp. 115–121, 2015. [Online]. Available: http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html

[19] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[20] V. Kumar, M. Muratani, N. A. Rayan, P. Kraus, T. Lufkin, H. H. Ng, and S. Prabhakar, "Uniform, optimal signal processing of mapped deep-sequencing data," *Nature Biotechnology*, vol. 31, no. 7, pp. 615–622, Jun. 2013. [Online]. Available: https://doi.org/10.1038/nbt.2596

[21] N. Zhao and A. P. Boyle, "F-Seq2: improving the feature density based peak caller with dynamic statistics," *NAR Genomics and Bioinformatics*, vol. 3, no. 1, 02 2021, lqab012. [Online]. Available: https://doi.org/10.1093/nargab/lqab012

[22] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass, "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities," *Molecular Cell*, vol. 38, no. 4, pp. 576–589, May 2010. [Online]. Available: https://doi.org/10.1016/j.molcel.2010.05.004

[23] Q. Song and A. D. Smith, "Identifying dispersed epigenomic domains from ChIP-Seq data," *Bioinformatics*, vol. 27, no. 6, pp. 870–871, 02 2011. [Online]. Available: https://doi.org/10.1093/bioinformatics/btr030

[24] L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery, "mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models," *The R Journal*, vol. 8, no. 1, pp. 289–317, 2016. [Online]. Available: https://doi.org/10.32614/RJ-2016-021

[25] J. Morales, S. Pujar, J. E. Loveland, A. Astashyn, R. Bennett, A. Berry, E. Cox, C. Davidson, O. Ermolaeva, C. M. Farrell, R. Fatima, L. Gil, T. Goldfarb, J. M. Gonzalez, D. Haddad, M. Hardy, T. Hunt, J. Jackson, V. S. Joardar, M. Kay, V. K. Kodali, K. M. McGarvey, A. McMahon, J. M. Mudge, D. N. Murphy, M. R. Murphy, B. Rajput, S. H. Rangwala, L. D. Riddick, F. Thibaud-Nissen, G. Threadgold, A. R. Vatsan, C. Wallin, D. Webb, P. Flicek, E. Birney, K. D. Pruitt, A. Frankish, F. Cunningham, and T. D. Murphy, "A joint NCBI and EMBL-EBI transcript set for clinical genomics and research," *Nature*, vol. 604, no. 7905, pp. 310–315, Apr. 2022. [Online]. Available: https://doi.org/10.1038/s41586-022-04558-8

[26] A. Kundaje, , W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y.-C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K.-H. Farh, S. Feizi, R. Karlic, A.-R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. D. Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. M. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L.-H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, and M. Kellis, "Integrative analysis of 111 reference human epigenomes," *Nature*, vol. 518, no. 7539, pp. 317–330, Feb. 2015. [Online]. Available: https://doi.org/10.1038/nature14248

[27] Y. Kang, Y. W. Kim, J. Kang, and A. Kim, "Histone H3K4me1 and H3K27ac play roles in nucleosome eviction and eRNA transcription, respectively, at enhancers," *The FASEB Journal*, vol. 35, no. 8, p. e21781, 2021. [Online]. Available: https://faseb.onlinelibrary.wiley.com/doi/abs/10.1096/fj.202100488R

PLACE
PHOTO
HERE

Michael Shell Biography text here.

John Doe Biography text here.

Jane Doe Biography text here.