BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 7 2014, pages 908–914
doi:10.1093/bioinformatics/btt467

Genome analysis

Advance Access publication October 4, 2013

A wavelet-based method to exploit epigenomic language in the regulatory region

Nha Nguyen[1], An Vo[2] and Kyoung-Jae Won[1,*]

[1]Department of Genetics, Institute for Diabetes, Obesity and Metabolism, School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA and [2]Center for Neurosciences, The Feinstein Institute for Medical Research, Manhasset, NY 11030, USA

Associate Editor: Inanc Birol

ABSTRACT

Motivation: Epigenetic landscapes in the regulatory regions reflect binding condition of transcription factors and their co-factors. Identifying epigenetic condition and its variation is important in understanding condition-specific gene regulation. Computational approaches to explore complex multi-dimensional landscapes are needed.

Results: To study epigenomic condition for gene regulation, we developed a method, AWNFR, to classify epigenomic landscapes based on the detected epigenomic landscapes. Assuming mixture of Gaussians for a nucleosome, the proposed method captures the shape of histone modification and identifies potential regulatory regions in the wavelet domain. For accuracy estimation as well as enhanced computational speed, we developed a novel algorithm based on down-sampling operation and footprint in wavelet. We showed the algorithmic advantages of AWNFR using the simulated data. AWNFR identified regulatory regions more effectively and accurately than the previous approaches with the epigenome data in mouse embryonic stem cells and human lung fibroblast cells (IMR90). Based on the detected epigenomic landscapes, AWNFR classified epigenomic status and studied epigenomic codes. We studied co-occurring histone marks and showed that AWNFR captures the epigenomic variation across time.

Availability and implementation: The source code and supplemental document of AWNFR are available at http://wonk.med.upenn.edu/AWNFR.

Contact: wonk@mail.med.upenn.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 18, 2012; revised on May 30, 2013; accepted on August 7, 2013

1 INTRODUCTION

The epigenetic landscapes, represented by modifications to histones, DNA methylation and other proteins that package the genome, regulate the function of cells by activating or repressing gene activity (Bernstein *et al*., 2007; Kouzarides, 2007). The epigenetic landscapes drawn over the genome reflect the commitment of cells to a lineage and/or response to the environmental changes (Maunakea *et al*., 2010). Histone codes representing combinatorial effects of histone modification tell how transcriptional apparatuses are used in a given cell-type/environment. At active promoters, active marks, including mono-, di- and tri-methylation of Lys4 of H3 (H3K4me1/2/3), showed strong signals. Chromatin marks at the enhancers showed strong H3K4me1 and weak H3K4me3 (Hawkins *et al*., 2010). Histone acetylations are enriched both at active promoter and enhancer (Hawkins *et al*., 2010; Wang *et al*., 2008). Epigenetic variations reflected condition-specific binding of transcription factors (TFs) (He *et al*., 2010; Wang *et al*., 2011). Therefore, identifying epigenomic condition in regulatory regions is important to understand condition-specific gene regulation. Algorithmic development to exploit epigenomic landscapes is required for systematic analysis of cell-type specific gene regulation.

Here, we present a method that explores multi-dimensional epigenomic landscapes using the wavelet transforms (WTs). Wavelet compromises between time- and frequency-based views of signal. Because of its property, wavelet has been widely used in signal and image processing (Mallat, 2009). In genome-wide study, wavelet has been applied to analyzing DNA replication profile (Audit *et al*., 2013), studying admixed population (Pugach *et al*., 2011) and analyzing protein or microarray data (Lio, 2003). Also, wavelet has been also applied for epigenomic data analysis (Mitra and Song, 2012; Xiaoquan *et al*., 2004; Zhang *et al*., 2008). We applied wavelet to idntify nucleosome, given the multi-dimensional histone modification data. For accurate and fast performance, we applied down-sampling WT and wavelet footprint in modeling nucleosome.

AWNFR identifies the position and the shape of nucleosomes after modeling a nucleosome with mixture of Gaussians (MoGs). The gapped region between two identified nucleosomes is defined as a nucleosome free region (NFR). Using the shape of histone modification captured in the WD, we classified NFRs based on the epigenomic landscapes and interrogated the combination of histone modification marks.

Previously, a number of algorithms were suggested to detect NFRs from histone modification data. A method named nucleosome positioning from sequencing (NPS) was designed to predict nucleosome position (Zhang *et al*., 2008). NPS used WT for denoising signal after modeling a nucleosome with Laplacian of Gaussian. Homer package (Heinz *et al*., 2010) is equipped with a function to predict NFRs. NOrMAL uses Gaussian mixture model to identify nucleosome position (Polishko *et al*., 2012). However, these frameworks only allow one mark for the input and cannot be applied to detecting comprehensive epigenetic variation or code. An HMM-based approach, called Chromia, predicted regulatory regions using epigenome data (Won *et al*.,

*To whom correspondence should be addressed.

2008). However, it requires carefully selected training set. Compared with supervised methods (Fernandez and Miranda-Saavedra, 2012; Won *et al.*, 2008), AWNFR uses an unsupervised method to exploit epigenomic landscapes. More importantly, AWNFR is capable of detecting epigenetic codes and their variations from the identified NFRs.

Using simulated data, we showed the advantages of new peak and edge detection algorithms which are used in our method. We applied AWNFR to the epigenomic data in mouse embryonic stem cells (mESC) and IMR90 and showed the outperforming performance over the previous approaches in detecting potential regulatory regions. Also, we show that AWNFR can study the epigenetic variations using the histone modification data during adipogenesis (Mikkelsen *et al.*, 2010).

2 METHODS

To detect the epigenetic variation and its code, we developed a wavelet-based method called AWNFR. AWNFR is composed of three steps.

Step1: Detecting enriched regions (Section 2.1)

Haar wavelet moving (HWM) is implemented to identify enriched region with activating histone marks (active regions) by scanning the genome across multiple histone marks.

Step 2: Identifying nucleosome positions (Section 2.2)

After assuming MoGs for a nucleosome, we identified the parameters of MoGs using zero-crossing. NFRs are defined as the region between two imaginary MoGs.

Step 3: Clustering NFRs (Section 2.3)

A clustering method was applied to the identified NFRs to study the combination of epigenetic marks.

2.1 AWNFR detects epigenomic enriched regions

We developed a shift-invariant and down-sampling method called HWM. In WT, shift variance has been a major problem (Kingsbury, 2001). Lack of shift invariance can cause major variation in the output by small shifts in the input signal. Several methods such as stationary WT and dual tree complex wavelet transform (DT-CWT) were suggested to overcome shift variance, but at the cost of computation time (Kingsbury, 2001) (see Supplementary Documents for more detail about the previous WTs). To deal with large-scale data efficiently, we developed HWM, a shift invariant and down-sampling wavelet-based method.

An approximate scale, representing the low frequency (L) of the signal after WT, can be represented as

$$y_{i\_app} = (..(x * g_{1L}) * g_{2L}) * .....) * g_{iL}),$$

Where $*$ denotes a convolution operation, and x is an input signal or histone modification data with a length N. g_{iL} is the low pass filters at level i. Applying down-sampling, we get

$$y_{i\_app} = (...(x * g_{1L}) \downarrow 2) * g_{2L} \downarrow 2) * ...) * g_{iL}) \downarrow 2),$$

where $\downarrow 2$ is the down-sampling operation.

Using Haar WT, g_{iL} can be written as $g_{iL} = \left[\frac{1}{\sqrt{2}}; \frac{1}{\sqrt{2}};\right]$ (Mallat, 2009). Then, at a level l,

$$y_l[n] = \sqrt[-2]{2}^{1} \sum_{i=0}^{2^1-1} x[2^1 n - i], \quad (1)$$

where, n is the index of the output. Derived from Haar WT, Equation (1) is called HWM because of its similarity with the moving average algorithm. At level l, the genomic position is then

$$p_1[n] = p_x[n_m], \quad (2)$$

where p_x is genomic position of x, and n_m is the position at maxima of $x[k]$ with $(2^1(n-1)-1) < k < 2^1 n$. The wavelet level 1 then can be estimated

$$1 = \log_2(bs_n/bs_o), \quad (3)$$

where bs_o is the size of a bin of the data and bs_n is the new bin size.

To handle multiple marks, we stacked the genomic data. We applied HWM across the genome (HWM1d) and the stacked histone modification data (HWM2d). We denoted that n histone marks are represented as $\mathbf{X} = (x_1, \ldots, x_n)$.

HWM2d is then represented by the series of HWM1d.

$$Y_h = [y_1 y_2 ... y_n] = HWM1d(X) = HWM1d[x_1 x_2 ... x_n] \quad (4)$$

Appling HWM1d to the transposed matrix of y_h, we get

$$Y_v = HWM1d(Y_h^T) = HWM1d[y_1 y_2 ... y_n]^T, \quad (5)$$

where T is the transpose of a matrix. h and v were used for horizontal row and vertical column.

Figure 1 demonstrates how AWNFR detects active regions using 18 histone modification marks and 3 DNaseI hypersensitivity data in IMR90 (Hawkins *et al.*, 2010; Lister *et al.*, 2009) (Supplementary Table S1 lists the marks we used). The epigenome data (Fig. 1a) were transformed into the WD (Fig. 1b). HWM2d found four peaks that correspond to four active regions in Y_v (Fig. 1c).

2.2 AWNFR detects NFRs after modeling a nucleosome with MoGs

AWNFR detects the shape of histone modifications and the position of nucleosomes. This is an important step for the subsequence procedure of exploiting epigenetic landscapes and classifying regulatory regions. Previously, Laplacian of Gaussian was used to model a nucleosome (Zhang *et al.*, 2008). However, diverse histone modification patterns are observed in the genome. We used MoG as it can model more general signal than Laplacian of Gaussian. Because any signal can be represented as a sum of multiple Gaussians (Mallat, 2009), we used MoG to model a nucleosome. To detect the parameters of MoG, AWNFR used zero-crossing, which was successfully applied for parameter estimation using continuous wavelet transform (CWT) (Nguyen *et al.*, 2010). Zero-crossing is a method to identify wavelet footprint that captures the

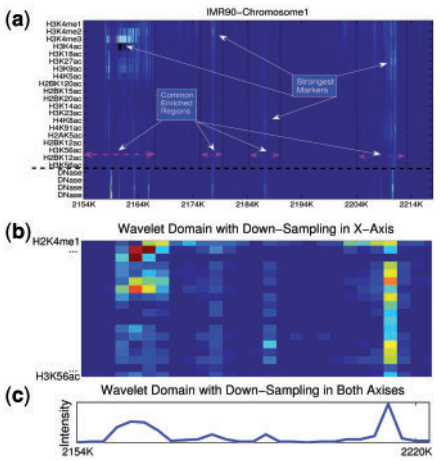

Fig. 1. Applying HWM to detect enriched regions. (**a**) 18 histone modification marks and 3 DNaseI data in IMR90. (**b**) the transformed signals in the WD. (**c**) Four enriched regions were obtained by applying HWM2d

characteristics of signals against noise (Kwon and Oweiss, 2011) (See Supplementary text for detail). In this article, zero-crossing of the derivatives of Gaussians was applied for parameter identification.

Assuming MoG for a nucleosome, we get

$$f(t) = \sum_i f_i(t) = \sum_i A_i e^{-(t-\mu_i)/(2\sigma_i^2)}, \quad (6)$$

where μ_1 and σ_i are the center and the standard deviation of a nucleosome. AWNFR uses the zero-crossing lines across wavelet scales. A zero-crossing is a point where the sign of a function changes. A zero-crossing line is obtained by connecting the zero-crossing points obtained over the wavelet scales.

CWT converts signals to the WD using a convolution operator.

$$Wf(u,s) = \int_{-\infty}^{\infty} f_i(t)\frac{1}{\sqrt{s}}\Psi^*(\frac{t-u}{s}dt = (f_i * \widetilde{\Psi}_s)(u), \quad (7)$$

where s denotes a wavelet scale, u is a position in the genome and $\widetilde{\Psi}(t) = \frac{1}{\sqrt{s}}\Psi^*\left(-\frac{t}{s}\right)$. Let $WF(w,s)$,$F_i(w)$ and $\widetilde{\Psi}(w)$ be the fast Fourier transform (FFT) of $Wf(u,s)$, $f_i(t)$ and $\widetilde{\Psi}_s(t)$, respectively.

$$Wf(u,s) = (f_i * \widetilde{\Psi}_s)(u) = \mathcal{F}^-\left\{WF(w,s)\right\} = \mathcal{F}^-\left\{F(w).\widetilde{\Psi}(w)\right\}, \quad (8)$$

where $\mathcal{F}^-$ denotes the inverse FFT (iFFT). Instead of using convolution, we use FFT ($\mathcal{F}$) and iFFT ($\mathcal{F}^-$) to estimate the wavelet coefficients ($Wf(u,s)$)

$$F_i(w,s) = A_i\sigma_i e^{-i\mu_i\omega}e^{-\frac{w^2\sigma_i^2}{2}}, \quad (9)$$

The peaks and the edges of the MoGs can be obtained by taking derivative

$$\widetilde{\Psi}(w) = -\frac{1}{\sqrt{\Gamma(n+\frac{1}{2})}}(jws)^n e^{-\frac{(sw)^2}{2}}, \quad (10)$$

Where Γ is a gamma function (Mallat, 2009), and n is the order of derivative of the Gaussian wavelet. Replacing $\widetilde{\Psi}$ and F from Equations (10) and (9) to Equation (8), we get

$$WF(w,s) = \beta(jw)^n\frac{1}{\sqrt{2\alpha}}e^{-\frac{w^2}{4\alpha}}e^{-i\mu_i\omega}, \quad (11)$$

where $\beta\frac{A_i\sigma i s^n\sqrt{2\alpha}}{\sqrt{\Gamma(n+\frac{1}{2})}}$ and $\alpha = \frac{2}{s^2+\sigma_i^2}$

The iFFT of Equation (11) becomes

$$Wf(u,s,n) = \beta\frac{d^n}{du^n}e^{-a(u-\mu_i)^2} \quad (12)$$

$n = 1, 2, 3$ correspond to each order of derivative.

$$Wf(u,s,1) = -2\alpha\beta(u-\mu_i)e^{-a(u-\mu_i)^2}, \quad (13)$$

$$Wf(u,s,2) = -2\alpha\beta[1-2\alpha(u-\mu_i)^2]e^{-a(u-\mu_i)^2}, \quad (14)$$

$$Wf(u,s,3) = -4\alpha^2\beta(u-\mu_i)[3+2\alpha(u-\mu_i)^2]e^{-a(u-\mu_i)^2}. \quad (15)$$

The parameters of the Gaussian for each derivative can be obtained using

$$Wf(u_0,s,1) = 0, \text{we have } u_0 = \mu_i, \quad (16)$$

$$Wf(u_0,s,2) = 0, \text{we have } \mu_0 = \mu_i \pm \sqrt{\sigma_1^2+s^2} \quad (17)$$

$$Wf(u_0,s,3) = 0, \text{we have } u_0 = \mu_i \text{ or } \mu_0 = \mu_i \pm \sqrt{3}\sqrt{\sigma_1^2+s^2}. \quad (18)$$

We call the first, the second and the third derivative of Gaussian wavelet as DOG1, DOG2 and DOG3, respectively. They are selectively used to estimate the position, the height and the standard deviation of Gaussians. Supplementary Table S2 and the Supplementary document summarize how the parameters of MoGs were calculated using each derivative. In conclusion, AWNFR selected DOG1 to calculate the peak and the height and DOG2 to obatin standard deviation of MoGs because they showed the best estimation performance in the simulation using artifical histone modification data (See Section 3.2). To summarize,

$$\text{peak position} : \mu_i = \frac{1}{N}\sum_{S=1}^{N} u_0(s) \quad (19)$$

$$\text{peak height} : A_i = f_i(\mu_i) - b \quad (20)$$

$$\text{standard deviation} : \sigma_{i\_left}(s) = \sqrt{(u_{0\_left}-\mu_i)^2 - s^2} \quad (21)$$

$$\sigma_{i\_right}(s) = \sqrt{(u_{0\_right}-\mu_i)^2 - s^2} \quad (22)$$

Figure 2 demonstrates how NFRs are detected in the wavelet domain (WD) using the zero-crossing lines in the WD. After the conversion into the WD, DOG1 (upper panel) and DOG2 (lower panel) are applied to obtain zero-crossing lines, respectively. DOG1 is applied to detect the peak positions, and DOG2 is applied to detect the standard deviation of the Gaussians.

2.3 Clustering epigenetic patterns

NFRs provide access to sequence specific TFs and basal transcription machinery (Bai and Morozov, 2010; Lu *et al.*, 1994). Identifying epigenomic landscapes around NFRs is important to understand the epigenetic condition for gene regulation. We defined an NFR as the gapped region between the two estimated nucleosomes. We clustered NFRs using *K*-means algorithm based on the epigenomic conditions (binding score) around them.

For n data ($x_1, \ldots, x_n$), HWM2d detects active regions and the strongest marks among them. The binding scores are calculated as follows using the strongest mark.

From Equation (20), the height of the left and the right peak of sample x_j at the enriched region number i is

$$A_{j\_i\_left\_peak} = x_j(\mu_{i\_left\_peak})^{-b} \quad (23)$$

$$A_{j\_i\_right\_peak} = x_j(\mu_{i\_right\_peak})^{-b} \quad (24)$$

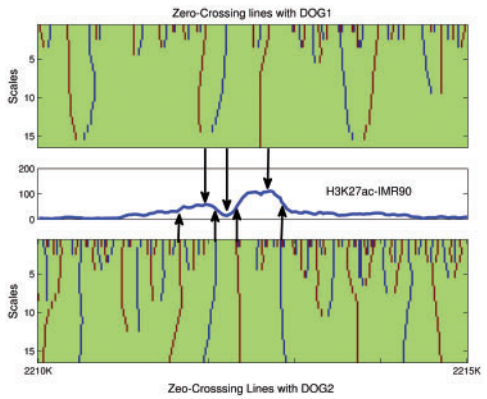

Fig. 2. Parameter estimation of AWNFR. Histone modification data (middle panel) is converted into WD by DOG1 (upper panel) and DOG2 (lower panel) over wavelet scales (vertical axis). The red and blue lines in the WD are the zero-crossing lines, representing positive (concave) and the negative Gaussian peaks (convex) using DOG1, respectively. The zero-crossing lines in the lower panel correspond to the edge of the Gaussians using DOG2

We defined the binding score as

$$\begin{aligned} Binding\ Score_{j\_i} &= A_{j\_i\_left\_peak} + A_{j\_i\_rightt\_peak} \\ &= x_j(\mu_i\_left\_peak) + x_j(\mu_i\_right\_peak) - 2b. \end{aligned} \quad (25)$$

Binding score matrix from Equation (25) is used as the input feature of the *K*-means clustering algorithm. From the initial *K* clusters, we reduced the number of clusters by combining clusters correlated each other [Pearson's correlation coefficient (CC) >0.75]. Therefore, the number of output cluster is equal or smaller than the number of the cluster given initially. Using the clustering results, we identified epigenomic codes and variations.

3 RESULTS

To show the algorithmic advantages of AWNFR, we evaluated the performance of HWM in both parameter estimation and NFR detection. First, we compared HWM with DWT and DT-CWT (Kingsbury, 2001) for enriched region detection. Also, we compared DOG1/2/3 with other competitors such as Gass1/2/3 (Nguyen *et al*., 2010), NPS (Zhang *et al*., 2008) and Canny (Canny, 1986) for the parameter estimation of MoGs.

For real data, we compared the performance using histone modifications in mESC and IMR90. Supplementary Table S1 summarizes the data we used to predict NFRs.

3.1 Assessment of HWM in detecting enriched regions

We compared the performance of HWM with DWT and DT-CWT (Kingsbury, 2001) using a simulating model. For the simulation, we used $f_s(t) = |\sin(wt)|f_o(t) + \mathcal{N}$, where $f_o(t) = \sum_i A_i e^{-(t-\mu_i)/(2\sigma i^2)}$. μ_i is the middle position of an active region, and 2 σ_i is the width of the active region. We assume that μ_i follows a uniform distribution, and N is Poisson noise.

We evaluated each algorithm by calculating the CC between $f_o(t)$ and the estimated model. We also calculated running time (RT) of 100 independent tests (Table 1). The simulation showed that that CC of HWM is significantly better than DTW-CWT and DWT. Notably, HWM implemented >4 and 8 times faster than DWT and DT-CWT, respectively. This demonstrates that HWM is an accurate and effective way to identify enriched regions.

3.2 Assessment of parameter estimation algorithm

To assess the MoG parameter estimation, we again used a simulation model.

$$f(t) = 10e^{-(t-5)/(2\times 0.5^2)} + 5 + \mathcal{N} \quad (27)$$

For N, we used Gaussian noise ($\mathcal{N} = \mathcal{G}(\mu = 0, \sigma = 1,)$) as well as Poisson noise [$\mathcal{N} = \mathcal{P}(\lambda = \infty)$]. We estimated the position, the height and the variance of Gaussians. After repeating the test 200 times, we calculated the error rate

$$error\ rate = \frac{|true\ value - estimated\ value|}{true\ value} \times 100. \quad (28)$$

Besides DOG1/2/3, we compared Gauss1/2/3 (Nguyen *et al*., 2010), NPS (Zhang *et al*., 2008) and Canny (Canny, 1986). Among them, NPS (Zhang *et al*., 2008) is a method designed to identify nucleosome positions.

Table 1. Performance assessment of HWM

DWT	DT-CWT	HWM			
RT (s)	CC	RT (s)	CC	RT (s)	CC
12.19	0.58	24.24	0.55	**2.97**	**0.67**

We compared DWT, DT-CWT and HWM using a simulating model. We calculated the RT and the CC between $f_o(t)$ from 100 independent parameter estimation tests. Bold value means the best performance or best value.

Table 2 compares the averaged error rates of peak position and standard deviation estimation. DOG1 performed best in peak detection. We did not include NPS and Canny, as they are not designed to detect peak position. For standard deviation, DOG2 outperformed other methods. Using the same model, we also evaluate the performance in estimating the peak height. In this test, Gaus1 and DOG1 performed better than other methods (Table 3). The performance of NPS was worse than DOG2/3. This is possibly because DOG2/3 uses MoGs that is more general function than Laplacian of Gaussian that NPS uses. Based on the results shown in Tables 2 and 3, AWNFR chose DOG1 for peak and height and DOG2 for standard deviation estimation.

3.3 AWNFR identifies regulatory regions accurately

Capturing the shape of histone modification around NFRs, AWNFR has a function to predict potential regulatory regions. To assess the performance in identifying NFRs, we compared the performance of AWNFR with Chromia (Won *et al*., 2008), Homer (Heinz *et al*., 2010) and NPS (Zhang *et al*., 2008). First, we compared the performance after detecting NFRs using H3K4me1/2/3 in mESC (Mikkelsen *et al*., 2007). For this test, we used the binding sites of 13 TFs as a measure to assess the prediction (Supplementary Table S1). We defined a prediction as a true positive (TP) if a prediction is within 1 kb of a TF binding site (TFBS). We calculated true predicative rate against false predicative rate while changing the threshold for the binding score. We evaluated the performance of AWNFR using a single mark (AWNFR4s) as well as multiple marks (AWNFR4m). Chromia was trained using the same set of histone marks (H3K4me1/2/3) using active promoter and enhancer. To obtain active promoter, we used the top 200 genes based on gene expression. To obtain the training set for enhancer, we used distal TFBSs in chromosome 1.

AWNFR4s using H3K4me3 outperformed Homer and Chromia. AWNFR4s showed comparable results with NPS. Using multiple histone modification marks, AWNFR4m outperformed all other predictors (Fig. 3a).

As an additional test, we predicted NFRs using 18 marks in IMR90 (Hawkins *et al*., 2010; Lister *et al*., 2009) (Supplementary Table S1). To evaluate the performance, we used DNaseI hypersensitive sites (DHSs) (Bernstein *et al*., 2010). For a single mark, we used H3K4me3. Compared with NPS, AWNFR4s showed higher true predicative rate when false predicative rate <0.02 (Fig. 3b). Homer predicted NFRs only when the nucleosomes around them are clearly observed and made only limited number of predictions. The HMM-based method (Chromia) was trained using the epigenome data at active promoter (top 200) and distal

Table 2. Assessment of the peak position and standard deviation estimation

Methods	Peak position		Standard deviation	
	Poisson (%)	Gaussian (%)	Poisson (%)	Gaussian (%)
Gauss1	0.1868	0.1838		
DOG1	**0.1424**	**0.1414**		
Gauss2	0.3866	0.3871	3.7344	4.1534
DOG2	0.2917	0.3412	2.5967	3.1079
Gauss3	0.8966	0.9486	6.6410	6.2918
DOG3	0.7308	0.7777	5.5199	5.1086
NPS			8.6300	8.8950
Canny			7.0700	7.9250

The average of the error rates from the 200 independent tests was calculated. Bold values show that DOG1 outperformed than others.

Table 3. Assessment of the peak height estimation

Methods	Poison (%)	Gaussian (%)
Gauss1 (C1 Option)	**1.4**	**1.4**
DOG1 (C1 Option)	**1.4**	**1.4**
Gauss1, Gauss2 (C2 Option)	4.5	4.6
DOG1, DOG2 (C2 Option)	4.0	4.2

The average of the error rates from the 200 independent tests was calculated. Bold values show that DOG1 outperformed than others.

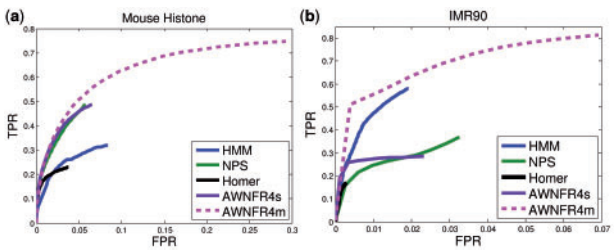

Fig. 3. The performance comparison using the histone modification data (**a**) in mESC (**b**) in IMR90. We ran AWNFR for a single mark (AWNFR4s) and multiple histone modification (AWNFR4m)

DHS in chromosome 1. Using multiple histone modification marks as well as a supervised learning approach, Chromia performed much better than NPS or Homer. AWNFR4m outperformed Chromia when they used the same datasets, even though AWNFR4s uses an unsupervised learning approach.

NPS showed comparable performance with AWNFR4m when using a single mark. NPS was developed based on Laplacian of Gaussian as a single filter to identify nucleosome positions (Zhang *et al.*, 2008). Compared with NPS, AWNFR used multiple wavelet scales. Supplementary Figure S1 shows an example that compared the results of AWNFR and NPS. Three peaks of H3K4me3 formed two valleys. Although NPS identified one NFR, AWNFR identified two NFRs correctly. The most right peak might not be modeled well with the Laplacian of Gaussian. All the NFRs identified by NPS were also predicted by AWNFR.

It is also interesting that AWNFR performed better than the supervised learning method, Chromia (Won *et al.*, 2008). Chromia used HMM to capture the characteristic patterns of histone modification data. Histone modification data are highly stochastic. Noise in the training dataset could deteriorate the performance. Training set of Chromia was collected at active promoter and potential enhancers (TFBSs in mESC, DHS in IMR90). Though Chromia can capture the profiles of histone modification, there could be some members in the training set that do not follow the average profile. Noise in the training data and the shift of the signal due to the low resolution of TFBSs could also affect the performance of Chromia. Homer uses the intensity of signal simply to predict NFRs. Though it performed worst, it runs fast, as it only applies simple enrichment calling to identify NFRs. Supplementary Table S3 compares the characteristics of the methods we tested.

3.4 Clustering using AWNFR finds diverse combinations of histone modification

Applying AWNFR, we classified the NFRs based on the combination of histone modifications. Figure 4 and Supplementary Table S4 show 64 clusters when we used eight histone marks: H2BK12ac, H2BK120ac, H3K4me1/2/3, H3K18ac, H3K27ac and H4K91ac (Supplementary Table S1). The clustering results showed diverse combinations of histone modifications. For example, Groups 11 and 12 were with acetylation but without H3K4me1/3. Group 15 was with H3K27ac, H3K18ac and H3K4me2/3. Supplementary Table S4 summarizes the histone codes for each group.

Expanding this research, we studied overlapping histone modifications in IMR90 (Fig. 5). We found H3K4me3 poorly co-occupied with H3K4me1, H2BK20ac or H2BK12ac. H2BK20ac was highly co-occupied with H2BK12ac but not with H3K4me1. It is because H2BK20ac and H2BK12ac are enriched in the promoter and the transcribed region of active genes (Wang *et al.*, 2008). Also, we observed co-enrichment of H2BK15ac with H2BK20ac and H2BK12ac. Interestingly, the same results were observed by coherent and shifted bicluster identification (CoSBI) in CD4+T cells (Ucar *et al.*, 2011). Contrary to the results of CoSBI, we observed a high co-occurrence of H3K4me2 with H3K18ac, H3K27ac and H3K9ac. However, our observation is supported by other genome-wide study in CD4+T cells where H3K4me2 is one of the backbone histone marks along with H3K27ac and H3K9ac (Wang *et al.*, 2008).

We also observed that H3K4me3 co-occurs with H3K9ac (80.9%) more frequently than any other histone acetylation marks. Though H3K27ac overlapped with both H3K4me1 and H3K4me3, its closest marks were H3K18ac (82.3%) and H4K91ac (74.9%).

We systematically investigated whether a histone mark can be represented with a combination of other marks. For this, we used the union of the marks in the order of overlapping ratio (Supplementary Table S5). Both H3K14ac and H3K23ac can be represented with the combination of H4K9ac and H3K27ac >95%. For other marks, it required at least three other marks to

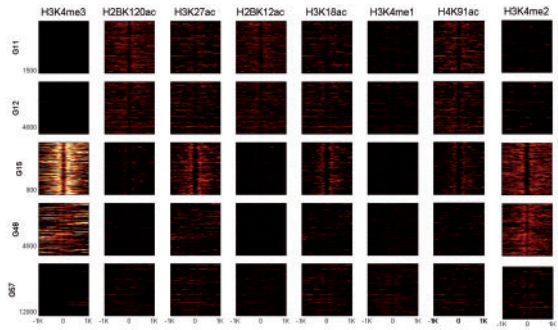

Fig. 4. Clustering results in IMR90. Groups 11 and 12 are with acetylation but without methylation. Group 15 is with H3K27ac, H3K18ac, H4K91ac and H3K4me2/3. Group 48 is with H3K4me2/3 but not with others. The entire groups are found at the Supplementary Document

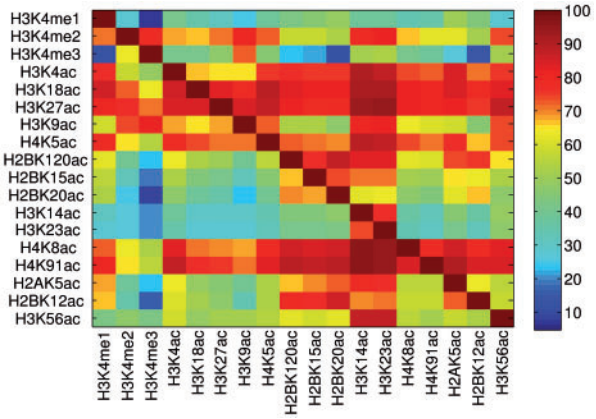

Fig. 5. Overlapping analysis using 18 histone marks. Number of overlapping NFR predictions over total number of NFR predictions on the histone marks were calculated

cover >95%. For example, six marks were needed to represent 95% of H3K4me3 (Supplementary Table S5).

3.5 Cell type specific histone code

We compared the epigenomic code in IMR90 against H1 (human embryonic stem cells). We investigated the overlap between each histone marks (Supplementary Fig. S2). In general, we observed the decreased level of overlaps. This is because the small number of overlaps between histone marks at the p300 binding sites in H1. Similarly, overlaps were shown decreased at the p300 binding sites in H1 (Rajagopal *et al.*, 2013). For example, the overlaps of H3K23ac and H3K4ac were significantly reduced at p300 binding sites, which are reflected by our observation. Interestingly, however, the random forest method did not catch it and even showed higher overlaps in H1 than in IMR90 (Rajagopal *et al.*, 2013). The overlaps of H3K4me1 with other marks were smaller in H1 compared with IMR90, which is not surprising because H3K4me1 marks poised enhancers in ESCs (Creyghton *et al.*, 2010; Rada-Iglesias *et al.*, 2011). H3K4me2, H3K9ac and H3K27ac showed strong overlapping with other histone marks in IMR90 as well as in H1.

3.6 AWNFR identifies epigenetic variations

AWNFR clusters NFRs based on the binding score. The clusters provided epigenetic codes. In the same way, we clustered H3K27ac across multiple time points during murine adipogenesis (Mikkelsen *et al.*, 2010). Instead of histone codes, it provides epigenetic variation across time. H3K27as was measured at 4 time points during white adipocyte cells differentiation: proliferating (day −2), confluent preadipocytes (day 0), immature adipocyte (day 3) and mature adipocytes in mouse (day 7) (Mikkelsen *et al.*, 2010). Investigating epigenetic variations, we identified eight clusters with distinct variations (Supplementary Fig. S3). For example, H3K27ac in Group 8 gradually increased, whereas Group 4 showed gradual decrease of H3K27ac during the adipogenesis. This result shows that AWNFR detects epigenetic variations across time.

4 DISCUSSIONS

Identifying epigenomic condition is important to understand the conditions for gene regulation. Our knowledge about the combination of histone modifications and the role for gene regulation is still limited. Computational approaches that exploit the complex landscapes and collect useful information out of them are highly in need. In this article, we presented a wavelet-based approach to explore multi-dimensional epigenomic landscapes. Using this approach, we identified nucleosome position and the epigenetic combination. Though there are several computational algorithms to predict nucleosome, our approach is uniquely designed to explore the epigenetic landscapes. Besides its outperforming performance in identifying nucleosome, we believe AWNFR suggests a new way of studying epigenetic data.

One of the problems in studying large-scale multi-dimensional data is the computational cost. To deal with large amount of epigenmic data efficiently, AWNFR is equipped with down-sampling wavelet. As shown in the simulation, HWM that AWNFR uses identified enriched regions faster than previous wavelet-based method with better accuracy. In the enriched regions, AWNFR explored more deeply to the landscapes by identifying nucleosome position. This framework using multiple epigenetic marks provided the condition for outperforming performance in predicting NFRs and exploring epigenetic codes.

We tested the algorithmic advantages of AWNFR using simulated models as well as real data. AWNFR outperformed previous competitors (Heinz *et al.*, 2010; Won *et al.*, 2008; Zhang *et al.*, 2008) in identifying NFRs in mESC and IMR90.

We also compared our histone codes with the previous predictors. Previously, Chromasig (Hon *et al.*, 2008) was developed to study histone code using correlation of histone signals. ChromHMM used histone code to annotate the genome (Ernst *et al.*, 2011). CoSBI exhaustively searched for histone code based on correlation within a 5 kb window (Ucar *et al.*, 2011). ChromHMM also used the enriched information of histone marks rather than studying the pattern itself. More recently, a random forest based enhancer identification method from chromatin states (RFECS) was developed based on random forests (Rajagopal *et al.*, 2013). Unique to AWNFR is the approach to study histone codes from the identified NFRs.

Compared with the CoSBI's results, we observed consistent (H3K4me3 partners poorly with H2BK120ac and H2BK12ac) as well as inconsistent results (a high co-occurrence of H3K4me2 with H3K18ac, H3K27ac and H3K9ac). However, our results were further supported by another genome-wide study in CD4+T cells (Wang *et al*., 2008). Using our histone codes, we studied whether a histone mark can be represented by other marks. Although H3K14ac and H3K23ac can be replaced easily by the combination of two marks, H3K4me3 required more combination to be replaced (Supplementary Table S5). We also studied cell-type specific histone codes. We found much less co-occurrences of histone modification in H1. These results, however, were similar with the results from RFECS (Rajagopal *et al*., 2013).

We did not find differences in motif enrichment between groups. However, it clearly shows that AWNFR can capture histone codes and the epigenomic variations across time. As genome-wide analyses of epigenetic regulations have become more popular, the demand for analysis tools such as the one described in this study will be high.

As the clustering is based on *K*-means algorithm, the number of clusters depends on the initial number of clusters. For a binary call for n samples, 2^n clusters are needed. However, we required more than 2^n clusters, as we used floating values for the binding score. For simplicity, we used 64 clusters when we used eight histone marks in IMR90.

We also tested the implementation speed. In processing 18 marks (Supplementary Table S1) at chromosome 1, NPS took >72 min, whereas our method took around 64 min (Intel core CPU 870@2.93 GHZ, 16 GB RAM), though AWNFR uses multiple wavelet scales.

Funding: [R21-DK098769 and P30-DK19525] from the National Institutes of Diabetes, and Digestive and Kidney Diseases and the Diabetes Research Center at the University of Pennsylvania.

Conflict of interest: none declared.

REFERENCES

Audit,B. *et al*. (2013) Multiscale analysis of genome-wide replication timing profiles using a wavelet-based signal-processing algorithm. *Nat. Protoc*., **8**, 98–110.

Bai,L. and Morozov,A.V. (2010) Gene regulation by nucleosome positioning. *Trends Genet*., **26**, 476–483.

Bernstein,B.E. *et al*. (2007) The mammalian epigenome. *Cell*, **128**, 669–681.

Bernstein,B.E. *et al*. (2010) The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol*., **28**, 1045–1048.

Canny,J. (1986) A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell*., **8**, 679–698.

Creyghton,M.P. *et al*. (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl Acad. Sci. USA*, **107**, 21931–21936.

Ernst,J. *et al*. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.

Fernandez,M. and Miranda-Saavedra,D. (2012) Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines. *Nucleic Acids Res*., **40**, e77.

Hawkins,R.D. *et al*. (2010) Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell*, **6**, 479–491.

He,H.H. *et al*. (2010) Nucleosome dynamics define transcriptional enhancers. *Nat. Genet*., **42**, 343–347.

Heinz,S. *et al*. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.

Hon,G. *et al*. (2008) ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. *PLoS Comput. Biol*., **4**, e1000201.

Kingsbury,N. (2001) Complex wavelets for shift invariant analysis and filtering of signals. *Appl. Comput. Harmon. A*., **10**, 234–253.

Kouzarides,T. (2007) Chromatin modifications and their function. *Cell*, **128**, 693–705.

Kwon,K.Y. and Oweiss,K. (2011) Wavelet footprints for detection and sorting of extracellular neural action potentials. In: *Proceedings of 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic*. IEEE, pp. 609–612.

Lio,P. (2003) Wavelets in bioinformatics and computational biology: state of art and perspectives. *Bioinformatics*, **19**, 2–9.

Lister,R. *et al*. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.

Lu,Q. *et al*. (1994) Nucleosome positioning and gene regulation. *J. Cell. Biochem*., **55**, 83–92.

Mallat,S.G. (2009) *A Wavelet Tour of Signal Processing : The Sparse Way*. Elsevier/Academic Press, Amsterdam, Boston.

Maunakea,A.K. *et al*. (2010) Epigenome mapping in normal and disease States. *Circ. Res*., **107**, 327–339.

Mikkelsen,T.S. *et al*. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.

Mikkelsen,T.S. *et al*. (2010) Comparative epigenomic analysis of murine and human adipogenesis. *Cell*, **143**, 156–169.

Mitra,A. and Song,J. (2012) WaveSeq: a novel data-driven method of detecting histone modification enrichments using wavelets. *PLoS One*, **7**, e45486.

Nguyen,N. *et al*. (2010) Mass spectrometry data processing using zero-crossing lines in multi-scale of Gaussian derivative wavelet. *Bioinformatics*, **26**, i659–i665.

Polishko,A. *et al*. (2012) NORMAL: accurate nucleosome positioning using a modified Gaussian mixture model. *Bioinformatics*, **28**, i242–i249.

Pugach,I. *et al*. (2011) Dating the age of admixture via wavelet transform analysis of genome-wide data. *Genome Biol*., **12**, R19.

Rada-Iglesias,A. *et al*. (2011) A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, **470**, 279–283.

Rajagopal,N. *et al*. (2013) RFECS: a random-forest based algorithm for enhancer identification from chromatin state. *PLoS Comput. Biol*., **9**, e1002968.

Ucar,D. *et al*. (2011) Combinatorial chromatin modification patterns in the human genome revealed by subspace clustering. *Nucleic Acids Res*., **39**, 4063–4075.

Wang,Z. *et al*. (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet*., **40**, 897–903.

Wang,D. *et al*. (2011) Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature*, **474**, 390–394.

Won,K.J. *et al*. (2008) Prediction of regulatory elements in mammalian genomes using chromatin signatures. *BMC Bioinformatics*, **9**, 547.

Xiaoquan,L. *et al*. (2004) Maximum spectrum of continuous wavelet transform and its application in resolving an overlapped signal. *J. Chem. Inf. Comput. Sci*., **44**, 1228–1237.

Zhang,Y. *et al*. (2008) Identifying positioned nucleosomes with epigenetic marks in human from ChIP-Seq. *BMC Genomics*, **9**, 537.