2015

A Stationary Wavelet Entropy-Based Clustering Approach Accurately Predicts Gene Expression

N. Nguyen

A. Vo
Zucker School of Medicine at Hofstra/Northwell

I. Choi

K. J. Won

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 22, Number 3, 2015

Pp. 236–249
DOI: 10.1089/cmb.2014.0221

A Stationary Wavelet Entropy-Based Clustering Approach Accurately Predicts Gene Expression

NHA NGUYEN,[1,2] AN VO,[3] INCHAN CHOI,[1,4] and KYOUNG-JAE WON[1,2]

ABSTRACT

Studying epigenetic landscapes is important to understand the condition for gene regulation. Clustering is a useful approach to study epigenetic landscapes by grouping genes based on their epigenetic conditions. However, classical clustering approaches that often use a representative value of the signals in a fixed-sized window do not fully use the information written in the epigenetic landscapes. Clustering approaches to maximize the information of the epigenetic signals are necessary for better understanding gene regulatory environments. For effective clustering of multidimensional epigenetic signals, we developed a method called Dewer, which uses the entropy of stationary wavelet of epigenetic signals inside enriched regions for gene clustering. Interestingly, the gene expression levels were highly correlated with the entropy levels of epigenetic signals. Dewer separates genes better than a window-based approach in the assessment using gene expression and achieved a correlation coefficient above 0.9 without using any training procedure. Our results show that the changes of the epigenetic signals are useful to study gene regulation.

Key words: algorithms, genetic analysis, genome analysis, gene expression, next-generation sequencing.

1. INTRODUCTION

The epigenetic landscape, represented by DNA methylation, modifications to histones, and other proteins that package the genome, regulates the function of cells by regulating gene activity (Bernstein et al., 2007; Kouzarides, 2007). To understand the complex languages of these epigenetic landscapes, gene clustering has been applied to epigenomic data to study a wide range of biological questions, including development (Mikkelsen et al., 2007, 2010; Lister et al., 2009; Meissner, 2010; Xie et al., 2013; Yu et al., 2013), cancer (Laird, 2003; Jones and Martienssen, 2005; Baylin and Jones, 2011), and aging (Baylin and Jones, 2011; Li et al., 2011; Liu et al., 2011). Clustering approaches provided insights into the dynamics of epigenetic gene changes. Additionally, clustering has been developed to identify co-occurring histone modification marks or ''histone codes'' (Barski et al., 2007; Heintzman et al., 2007, 2009; Won et al., 2008; Xie et al., 2013), as well as to study cell type- or species-specific gene regulation (Ernst et al., 2011; Won

[1]Department of Genetics and [2]Institute for Diabetes, Obesity and Metabolism, School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania.
[3]The Feinstein Institute for Medical Research, Manhasset, New York.
[4]Metabolic Engineering Division, Department of Agricultural Biotechnology, National Academy of Agricutural Science, Suwon Gyeonggido, South Korea.

et al., 2013; Yu et al., 2013). Often, clustering approaches have not fully utilized the characteristics of epigenetic signals because of the difficulties of understanding complex signals that occur in different combinations in epigenetic data. Methodological development for clustering epigenomic landscapes is required for comprehensive understanding of gene regulation.

To cluster epigenomic data more effectively using their statistical characteristics, we developed a new approach called Dewer. Dewer uses the measurement of entropy of epigenetic signals for clustering. Entropy is the expected value of the information contained in a random variable (Shannon, 1948). Entropy has been widely used in many areas, including information theory, signal processing, and biology (Yogesan et al., 1996; Swartz et al., 1999; Li et al., 2004; Shin et al., 2007; Daily et al., 2010; Menayo et al., 2014). Previously, entropy has been applied to genome-wide datasets to check tissue specificity and colocalization of signals (Sun et al., 2011; Won et al., 2013). A quantitative method named QDMR also used entropy to identify differentially methylated regions (DMRs) (Zhang et al., 2011). However, these studies were limited to studying variation across samples or tissues in a defined genomic position. Compared with the previous approaches, Dewer employed sophisticated ways of calculating the entropies in both original and wavelet domain for better discriminative power in clustering genes. Using entropy as a metric, Dewer fully uses the information contained in the distribution of the multidimensional epigenetic data, rather than condensing the data to mere statistics as in traditional window-based approaches.

To apply entropy effectively, Dewer detects areas enriched for multiple histone marks. The stacked histone marks in a region forms a two-dimensional (2D) data. Exploiting 2D spaces is important to fully understand the spatial combinatorial histone modification patterns for gene regulation. Most of border detectors have been developed to use only a single mark (Zang et al., 2009; Rashid, et al., 2011; Song and Smith, 2011; Micsinai et al., 2012). To identify enriched regions from multiple histone modification marks, Dewer upgraded SeqW (Nguyen et al., 2014b), a method developed by our group for better border detection. Entropy is then applied to the identified enriched regions. In this article, we show that both entropy and border detection contribute to the Dewer's discriminative power for gene clustering.

We found that the clusters identified by Dewer have a superior discriminative power over a window-based approach when we compared the gene expression levels among the clusters. Interestingly, gene expressions levels were highly correlated with the entropy levels, suggestive of the importance of the shape of epigenetic signals in studying gene regulation.

2. METHODS

To calculate the entropy from multiple histone modification signals, Dewer uses two-directional information retrieved from epigenomic data such as entropy and stationary wavelet (SW) entropy in the detected areas enriched inside histone modification signals.

2.1. Data preprocessing

For clustering analysis, we used activating marks at around transcription start sites (TSSs) H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K18ac, H3K27ac, and H4K91ac from human fetal lung fibroblasts (IMR90) (Bernstein et al., 2010). Tags were normalized using reads per kilo base per million (RPKM) for 10 bp bins.

We also used histone marks enriched in active gene body: H3K36me3 in murine adipocytes (3T3L1) (Mikkelsen et al., 2010) and H3K36me3, H3K79me1, and H3K79me2 in IMR90 (Bernstein et al., 2010).

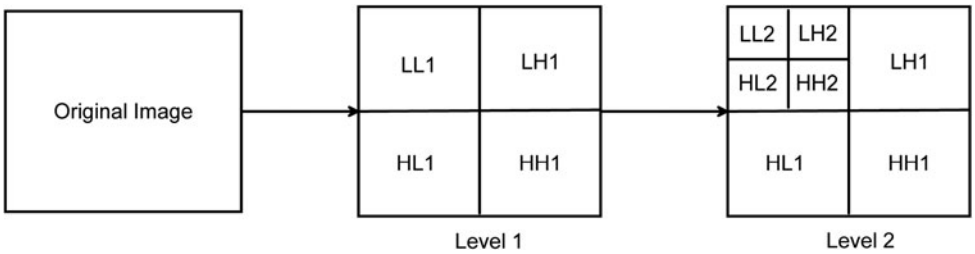

FIG. 1. Stationary wavelet transform with two levels. LL1, LH1, HL1, HH1, LL2, LH2, HL2, and HH2 are the eight scales. L means low frequency. H means high frequency. LH1 means low frequency in horizontal and high frequency in vertical at level 1.

2.2. *Identifying segments*

We further upgraded SeqW (Nguyen et al., 2014b), our previous border detector, to detect borders of multiple histone modification signals. To reduce computational cost, Haar wavelet moving (HWM) (Nguyen et al., 2014a) borrowed from SeqW is still used in Dewer to detect domains, which represent larger regions with epigenetic signals. Inside each domain, instead of detecting border directly in SeqW, Dewer estimates border indirectly by obtaining many segments by assessing signal enrichment against input. Neighboring segments are combined to form enriched regions. A domain usually has a number of enriched regions. Histone modification signals can be written as

$$g(t, m) = M(m)f(t), \tag{1}$$

where $f(t)$ is a function to normalize tag counts at a genomic position t of a histone modification mark m. Compared with SeqW, we used $M(m)$, which shows relative enrichments across m histone modification marks. With this new representation in Equation 1, Dewer can detect borders from multiple marks.

Assuming a mixture of Gaussians (MoG) for a nucleosome marked by histone modification signals as in (Nguyen et al., 2010, 2014a,b), we obtain

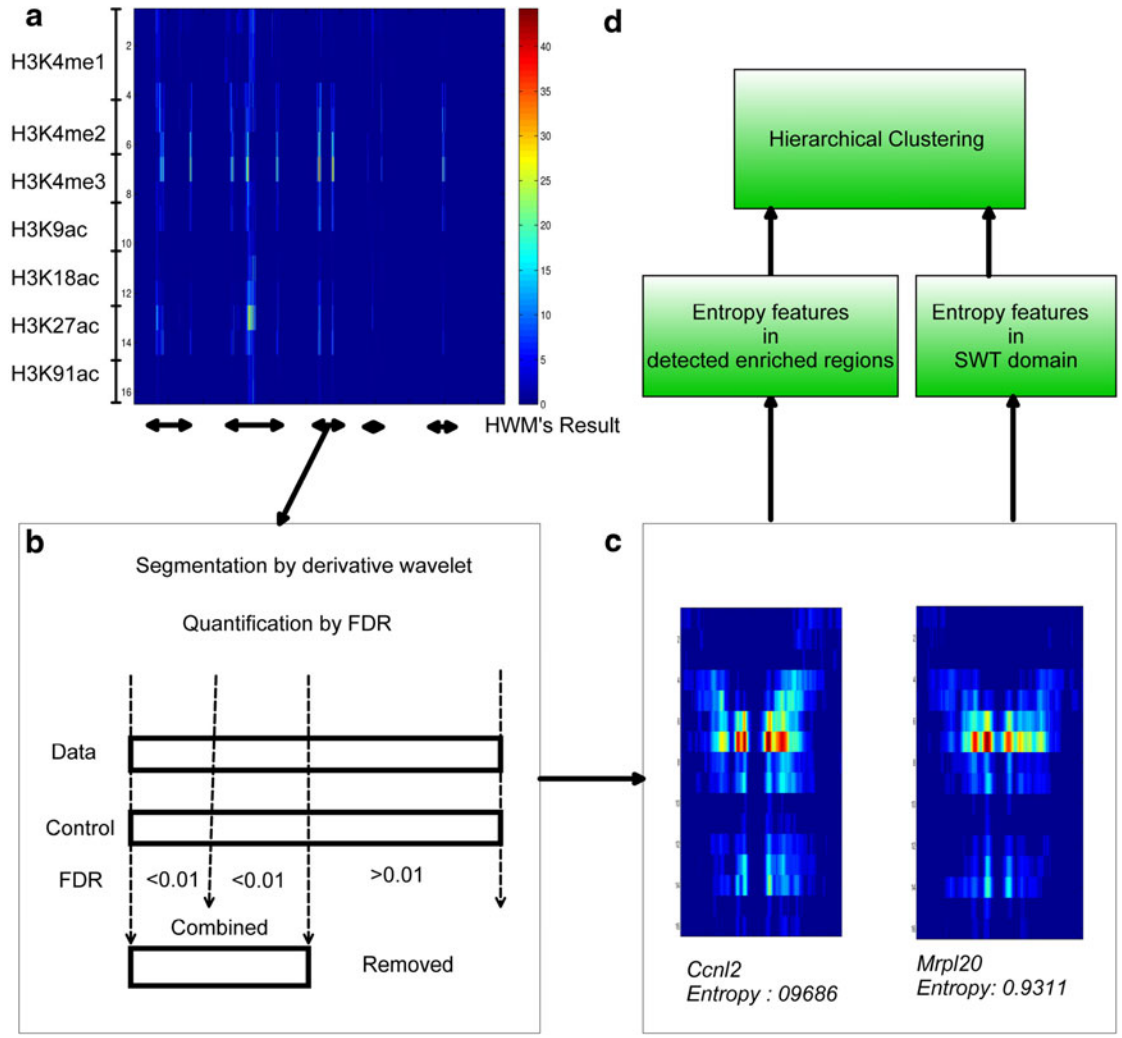

FIG. 2. The Dewer procedure. **(a)** Five domains were identified by Haar wavelet moving (HWM). A domain is composed of a set of segments enriched for histone modification signals. **(b)** Each segment is evaluated using false discovery rate (FDR). Adjacent segments are combined to form an enriched region. **(c)** Entropy features are estimated in the obtained enriched regions. Stationary wavelet entropy features are also calculated in stationary wavelet transform (SWT) domain. **(d)** Both the entropy and stationary wavelet entropy features are used for clustering.

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

where μ_i, and σ_i are the center and the standard deviation of a peak of a nucleosome, respectively. To estimate the borders of MoG signals, Dewer uses the zero-crossing lines across the multiwavelet scale. 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 scale s (Nguyen et al., 2014a).

Wavelet transform converts signals to the WD using a convolution operator.

$$\begin{aligned} Wg(u, m, s) &= M(m) Wf(u, s) \\ &= M(m) \int_{-\infty}^{\infty} f_i(t) \frac{1}{\sqrt{s}} \psi^* \left(\frac{t-u}{s} \right) dt = M(m)(f_i * \widetilde{\psi})(u), \end{aligned} \tag{3}$$

where s denotes the scale, u is the genome position in WD, and $\tilde{\psi}(t) = \frac{1}{\sqrt{s}} \psi * \left(-\frac{t}{s} \right)$.

As described more in detail in (Nguyen et al., 2014a), we obtain

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

The inverse fast Fourier transform of Equation 4 becomes

$$Wg(u, m, s, n) = \beta \frac{d^n}{du^n} e^{-\alpha(u-u_i)^2}, \tag{5}$$

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

To detect the borders of a peak, we use the second derivative wavelet (n = 2) (DOG2) instead of DOG3 in SeqW, where n = 1,2,3 correspond to each order of derivative.

$$Wf(u, s, 2) = -2\alpha\beta[1 - 2\alpha(u-u_i)^2] e^{-\alpha(u-\mu_i)^2}. \tag{6}$$

The zero-crossing points are the parameters of the Gaussians when $Wf\,(u_0,\, s,\, 2) = 0$. Then we have

$$u_0 = \mu_i \pm \sqrt{\sigma_i^2 + s^2}. \tag{7}$$

If s is small compared with σ_i, Equation 7 becomes

$$u_0 \approx \mu_i \pm \sigma_i. \tag{8}$$

In summary, $u_0\,(s)$ in Equation 7 draws a line over the scale s. Equation 8 indicates that the zero-crossing line approximates to σ away from the center of a peak. Equations 7 and 8 were used to detect the border of the peaks. We regarded a segment as the region defined by the borders of an MoG. In the regions where nucleosomes are packed, histone modification signals can be represented with a number of segments. To remove falsely predicted segments, each segment is assessed using false discovery rates (FDRs) after considering enrichment against background (input).

2.3. FDR control

We used the *mattes* and the *mafdr* functions in Matlab R2012a to calculate the statistical significance of the signals in a segment against background (input) after dividing a segment into three subsegments. p-Values were calculated by applying t-test (Huber et al., 2002) on these subsegments. FDRs were obtained from these p-values (Storey, 2002). Segments with an FDR greater than 0.01 were removed. After this step, adjacent segments were combined to form an enriched region.

2.4. Gene clustering using entropy and SW entropy

Dewer used Shannon's entropy (Shannon, 1948) for gene clustering. Shannon's entropy has been widely used in signal processing to measure the expected value of the information contents contained in a signal. Entropy $H(x)$ for a given signal x can be estimated by

$$H(x) = -\sum_i [P(x_i) \log_2 P(x_i)], \tag{9}$$

where P represents the histogram of the normalized intensity of histone modification signals inside the detected enriched regions around TSS (1–5 Kbps, 5 Kbps is default value).

Stationary wavelet transform (SWT) (Mallat, 2009) is a time–frequency analysis that has been widely used in image processing applications such as de-noising (Wang et al., 2003), enhancement (Demirel and Anbarjafari, 2011), segmentation (Deng et al., 2014), and image retrieval (Agarwal et al., 2013). Especially, we used the extracted information from decomposed scales of SWT. SWT (Fig. 1) with two levels has eight scales: LL1, LH1, HL1, HH1, LL2, LH2, HL2, and HH2. The SW entropy is obtained from these eight scales. The SW entropy levels are used as the input of clustering. Decomposition part of SWT can be formatted as follows:

$$\begin{aligned}
LL_{l+1} &= LL_l * L_l(Rows) * L_l(Columns),\\
LH_{l+1} &= LL_l * L_l(Rows) * H_l(Columns),\\
HL_{l+1} &= LL_l * H_l(Rows) * L_l(Columns),\\
HH_{l+1} &= LL_l * H_l(Rows) * H_l(Columns),\\
L_{l+1} &= L_l(\uparrow 2),\\
H_{l+1} &= H_l(\uparrow 2),
\end{aligned} \tag{10}$$

where L_l is low pass filter at level l, H_l is high pass filter at level l, LL_0 is original image, and $\uparrow 2$ is the up-sampling operation by 2.

Dewer used agglomerative cluster algorithm (Szekely and Rizzo, 2005), a hierarchical clustering approach where each object starts from its own cluster, and pairs of clusters are merged as one move up the hierarchy.

3. RESULT

3.1. Dewer overview

Dewer clusters genes based on the entropy features inside enriched regions. Figure 2a shows the five domains identified using seven types of histone modification marks (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K18ac, H3K27ac, and H4K91ac) in IMR90. A domain contains a number of enriched regions. The derivative wavelet method detects the border of the segments after modeling a segment. *p*-Values and FDRs were calculated to measure the enrichment of epigenetic signals against control input. We applied an FDR cutoff of 0.01 (Fig. 2b). The neighboring segments were combined to form a final enriched region. Figure 2c shows two enriched regions around the promoter regions of *Ccnl2* and *Mrlp20*. Entropies were calculated in these regions around the TSS. Also, SW entropies were estimated inside eight scales. A hierarchical clustering method is applied to the obtained entropies as the input features.

3.2. Assessing the performance of enriched region detection

Detecting enriched region is an important component as estimating entropy in Dewer. In this section, we evaluate the performance of enriched region detection.

To evaluate the performance of enriched region detection, we compared Dewer with previous broad region detectors, including SeqW (Nguyen et al., 2014b), Sicer (Zang et al., 2009), RSEG (Song and Smith, 2011), and QESEQ (Micsinai et al., 2012). We also included ChromHMM (Ernst et al., 2010) because it annotates the genome using multiple histone modification marks, even though it was not originally designed for enriched region detection. For this assessment, we used H3K36me3 in murine adipocytes (3T3L1) and evaluated using the annotated genes in GENCODE (Harrow et al., 2012). Only the bodies of the active genes whose gene expressions are higher than averaged value were used.

Figure 3 shows the true-positive rates against false-positive rates. For this test, only a single H3K36me3 (day -2) was used for QESEQ, Rseq, Sicer, and SeqW and four H3K36me3 were used for Dewer and ChromHMM (day -2, 0, 3, and 7). In our test, QESEQ performed better than Rseq and Sicer, in agreement with the previous observations in (Micsinai et al., 2012). QESEQ performed better than ChromHMM using

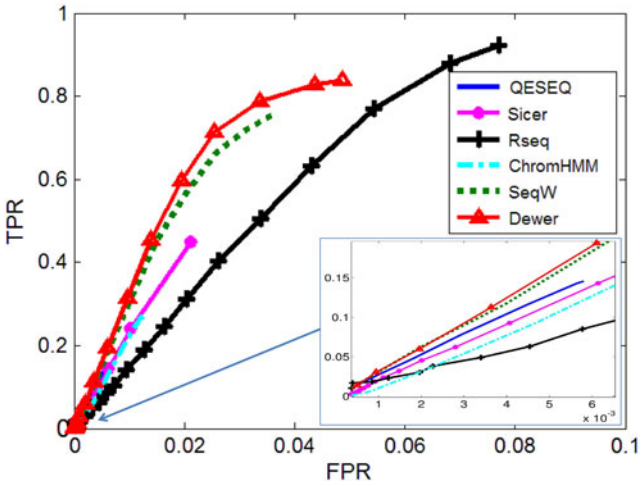

FIG. 3. Performance comparison for enriched region detection using 3T3L1 datasets. H3K36me3 data from 3T3L1 cells were used for each detector. For ChomHMM and Dewer, H3K36me3 in four time points during adipogenesis was used. Dewer outperformed other predictors in this test.

four H3K36me3 marks. It is not surprising because ChromHMM is originally not designed to predict enriched regions.

Dewer performed even better than SeqW, suggesting that Dewer uses multidimensional epigenetic signals effectively. ZINBA (Rashid et al., 2011) was not included in this comparison because of its exhaustive running time.

Additionally, we tested the performance using active marks for gene body (H3K36me3, H3K79me1, and H3K79me2) in IMR90 (Fig. 4). For this test, we compared Sicer, QESEQ, and SeqW as they performed better than other predictors in our previous test. A single mark (H3K36me3) was used for SeqW, Sicer, and QESEQ. Dewer used H3K36me3, H3K79me1, and H3K79me2. The test also confirmed that Dewer has a solid performance in detecting the regions enriched for epigenetic signals.

3.3. Gene clustering using Dewer

3.3.1. The clusters identified by Dewer well discriminate gene expression levels. We assessed the performance of gene clustering by comparing Dewer with a conventional window-based approach. Window sizes have been selected from 1 to 5 Kbps. With Dewer, window sizes help to narrow selected enriched regions. Using seven histone modification marks in IMR90, we performed clustering and generated 6–10 clusters. We used gene expression as a surrogate to check if the identified clusters separated gene groups effectively.

After performing clustering, we evaluated the expression levels of the genes for each cluster and measured if they are different from each other using the Student's *t*-test. For this test, we also investigated the effectiveness of the two algorithms that Dewer employed (detection of enriched regions and calculating entropy). Specifically, to assess the advantages of using enriched regions, we compared the performance of using entropy with/without enriched region detection. We also tested if entropy improves discriminative power in a given window. Figure 5 compares the averaged *p*-values of all clustering pairs in the tested clustering algorithms. For rigorous assessment, we investigated the averaged *p*-values while we increased the size of a window around TSSs and calculated the mean value and the entropy of the signals. Table 1

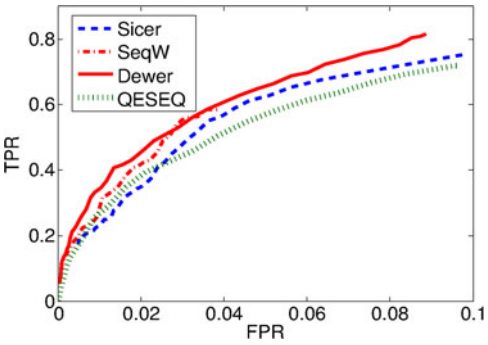

FIG. 4. Assessing the performance of detecting enriched region using IMR90 datasets. SeqW performed Sicer and QESEQ when using H3K36me3. Dewer using H3K36me3, H3K79me1, and H3K79me2 performed best.

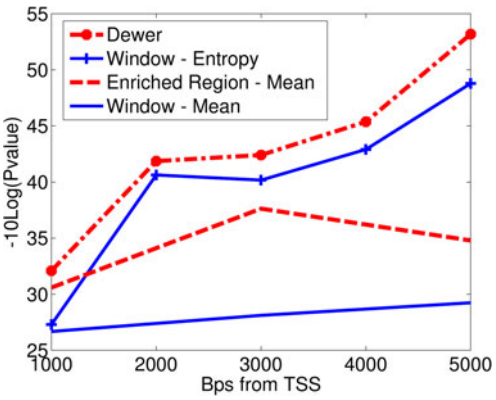

FIG. 5. Comparison of the algorithms Dewer used for its clustering. We compared the algorithms of Dewer (entropy and enriched regions). We used entropy in a window and used the mean value in the enriched region. We also compared the performance using the classical mean value in a window for gene clustering.

TABLE 1. VALUES OF -10 LOG (P-VALUE), EIGHT SWT SCALES, AND MANY WAVELET FILTERS HAVE BEEN CALCULATED

Scale / *Wavelet*	*LL1*	*LH1*	*HL1*	*HH1*	*LL2*	*LH2*	*HL2*	*HH2*
db1	36.2	39.6	42.9	45.0	48.1	46.1	37.9	38.0
db2	27.5	34.7	37.8	44.2	33.8	35.9	36.1	41.7
db3	37.4	35.2	40.5	39.1	30.9	33.3	42.6	39.8
• db4	38.5	48.7	38.9	29.4	40.7	39.1	32.3	39.0
• db5	49.2	47.9	43.4	43.6	36.6	38.1	45.7	40.3
• db6	37.6	38.2	35.8	44.5	31.1	26.9	39.6	34.3
• db7	38.6	39.3	41.9	40.7	29.1	32.5	35.1	37.1
• db8	31.5	36.2	36.2	40.1	28.6	30.6	29.9	36.4
• db9	32.5	30.0	42.8	39.5	34.4	37.6	41.3	36.4
• db10	34.3	32.7	44.9	43.9	21.7	40.2	45.2	44.6
• sym2	27.5	34.7	37.8	44.2	33.8	35.9	36.1	41.7
• sym3	37.4	35.2	40.5	39.1	30.9	33.3	42.6	39.8
• sym4	45.5	38.9	34.3	35.2	45.8	35.5	43.2	31.6
• sym5	32.4	32.3	39.5	43.4	33.1	40.6	42.0	40.1
• sym6	35.5	31.8	34.1	36.8	43.4	37.6	43.0	39.4
• sym7	47.6	39.3	32.4	41.6	33.2	29.1	34.6	32.1
• sym8	32.1	30.0	34.5	41.5	44.5	39.3	48.7	43.9
• sym9	45.3	34.0	38.5	34.0	39.6	32.4	44.7	45.5
• sym10	30.3	41.2	43.8	40.6	25.7	31.0	36.9	61.5
• sym11	30.2	38.9	34.1	42.2	29.5	38.2	42.1	43.8
• coif1	35.5	43.0	34.0	33.3	39.1	29.6	35.2	40.2
• coif2	32.0	35.0	41.4	41.6	43.2	41.3	43.5	41.4
• coif3	35.5	36.6	45.6	35.9	31.3	32.4	43.0	35.3
• coif4	43.0	39.1	37.4	51.5	45.4	34.5	36.6	38.8
• coif5	38.3	38.6	45.9	34.2	34.6	35.1	39.5	40.6
• dmey	45.9	35.7	46.3	33.2	33.5	40.1	36.4	45.8
• bior1.1	36.2	39.6	42.9	45.0	48.1	46.1	37.9	38.0
• bior1.3	39.6	40.7	43.5	45.0	33.8	35.0	37.5	37.8
• bior1.5	43.6	36.9	31.5	45.0	41.9	39.7	50.4	40.0
• bior2.2	41.7	41.6	37.5	42.0	40.3	36.3	52.1	38.5
• bior2.4	36.1	43.6	28.9	42.0	40.4	33.3	38.0	38.3
• bior2.6	34.0	32.7	43.4	42.0	39.4	33.9	51.3	43.5
• bior2.8	37.2	35.4	43.5	42.0	34.3	36.0	35.3	43.6
• rbio1.1	36.2	39.6	42.9	45.0	48.1	46.1	37.9	38.0
• rbio1.3	36.2	38.4	37.9	36.3	48.1	31.3	38.7	44.3
• rbio1.5	36.2	35.6	41.9	35.6	48.1	35.4	34.5	50.9
• rbio2.2	39.4	45.1	41.9	37.3	28.2	27.2	36.3	36.8
• rbio2.4	39.4	36.7	39.3	36.2	28.2	48.4	38.5	42.4
• rbio2.6	39.4	39.6	42.0	36.7	28.2	35.3	31.9	32.8
• rbio2.8	39.4	36.0	36.8	44.0	28.2	25.9	45.8	32.7

In eight SWT scales, L means low frequency and H means high frequency; 1 and 2 are levels of SWT. Daubechies (db1–db10), Coiflets (coif1–coif5), Symlets (sym2–sym11), Discrete Meyer (dmey), Biorthogonal (bior1.1–bior2.8), and Reverse Biorthogonal (rbio1.1–rbio2.8) are the used wavelet filters. SWT, stationary wavelet transform.

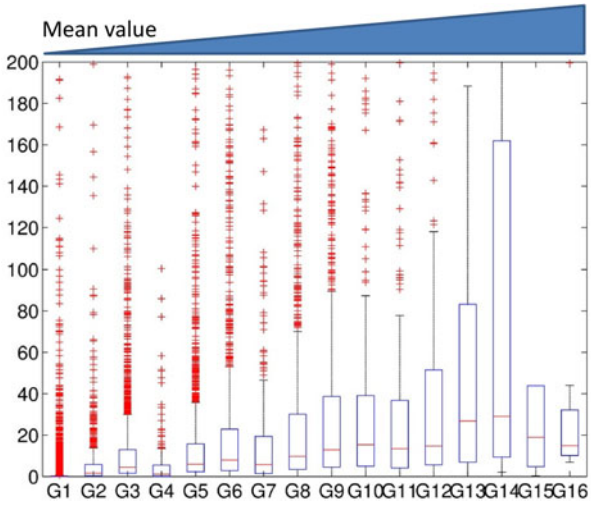

FIG. 6. The association of gene expression levels with the mean levels when using window method. We generated 16 clusters using a window-based approach. Clusters were sorted based on their strength of the features. The expression levels for each cluster were investigated.

compares the performance of various wavelet filters, including Daubechies, Coiflets, Symlets, Discrete Meyer, Biorthogonal, and Reverse Biorthogonal (Mallat, 2009). Eight scales (LL1, HL1, LH1, HH1, LL2, HL2, LH2, and HH2) of SWT for each filter were tested. We found that HH2 of Sym10 performed best. Sym10 is a symmetric filter, suggesting that symmetry and high frequency of Sym10 may be suitable for capturing the characteristics of histone modification signal.

The performances were improved in general as we increased the size of a window around the TSSs except for the case when we used the mean value of the enriched regions (Fig. 5). This may be caused by the high intensity of the signals in some local enriched regions. After 5 Kbps, all performances went down (not shown here). The comparison of using entropy and the mean value in a given window clearly demonstrated that entropy is a good measure for clustering against the window-based method. Calculating the mean value performed worst in our test. Dewer, which used both entropy and the enriched region, performed best in this assessment. This test emphasizes that both enriched region and entropy provide discriminative power for gene clustering.

3.3.2. Entropy levels correlate with gene expression levels. Next, we studied the association of gene expression with entropy in 16 clusters. We also tested the association of gene expression of the clusters obtained using the mean value in a window. The averaged gene expression levels for all the clusters were identified for each configuration: the average value of a window (Fig. 6), the entropy of a window (Fig. 7), and Dewer (the entropy in the enriched regions) (Fig. 8). All clusters were sorted based on their averaged feature values.

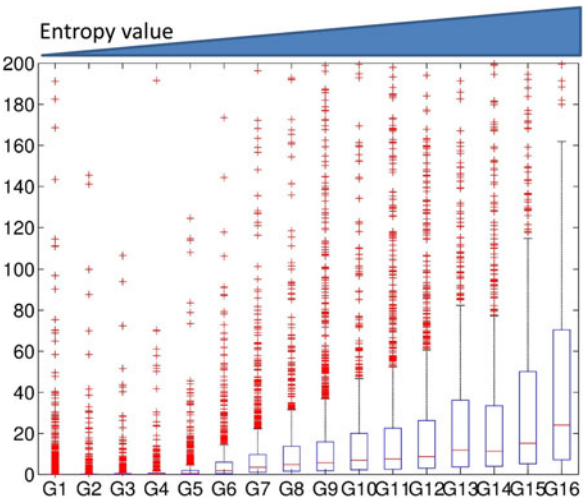

FIG. 7. The association of gene expression levels with the entropy levels when using window method. We generated 16 clusters using a window-based approach. Clusters were sorted based on their strength of the features. The expression levels for each cluster were investigated.

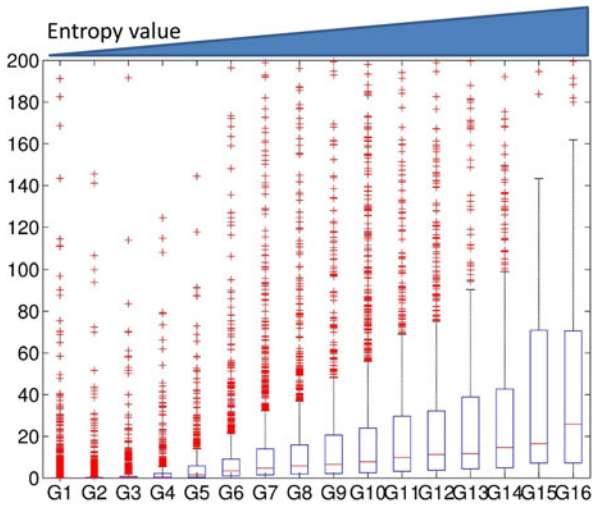

FIG. 8. The association of gene expression levels with the entropy levels when using Dewer method. We generated 16 clusters using Dewer. Clusters were sorted based on their strength of the features. The expression levels for each cluster were investigated.

We observed gradual increase of gene expression levels in association with the entropy levels (Figs. 7 and 8). In Figure 6, the expression level of G14 was higher than that of G15 and G16 in the window-based approach despite its lower mean value level. When we used entropy in a window (Fig. 7), we observed a better correlation than using the averaged values. Dewer showed the best correlation in our test (Fig. 8), and suggested that both entropy and enriched region detection are useful in clustering.

We observed similar phenomena when using 32 clusters. We investigated the correlation coefficient (CC) of the clusters when we used the epigenetic data 1–5 Kbps around TSSs (Figs. 9 and 10). The CC between expression levels and the entropy levels reached 0.9 when using Dewer (Fig. 10). We further analyzed the contributions of each histone modification mark for gene expression (Fig. 9). When mean values were used, the contributions of H3K4me1 and H4K91ac for gene expression prediction were very weak (Figs. 11 and 12). This is mainly because of the low intensity of the signals. When all marks were combined, the window-based method resulted in a correlation of approximately 0.6. The same comparison using the entropy levels resulted in much stronger correlation in general even when using H3K4me1 only. Even though entropies of H3K4me2 and H3K4me3 were the highest, their associations with gene expression were relatively low, especially for H3K4me3. This may be because H3K4me3 is also enriched for transcriptionally paused genes (Guenther et al., 2007). We observed that the correlation was the highest when using H3K27ac, consistent with previous studies (Karlic et al., 2010). H3K9ac and H3K18ac also had high correlation though the signal of H3K18ac was very weak (Figs. 11 and 12). Using all marks giving the highest CC may suggest that Dewer makes use of multiple marks effectively by employing entropy.

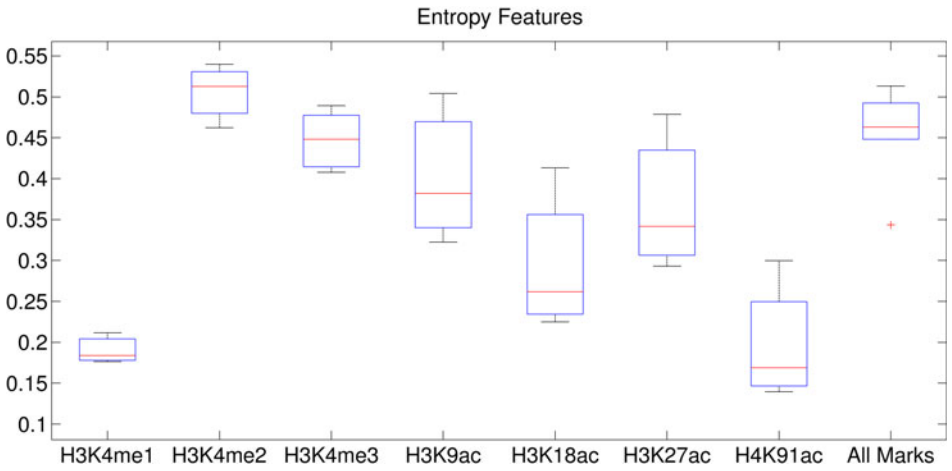

FIG. 9. The entropy value of each histone modification mark using Dewer.

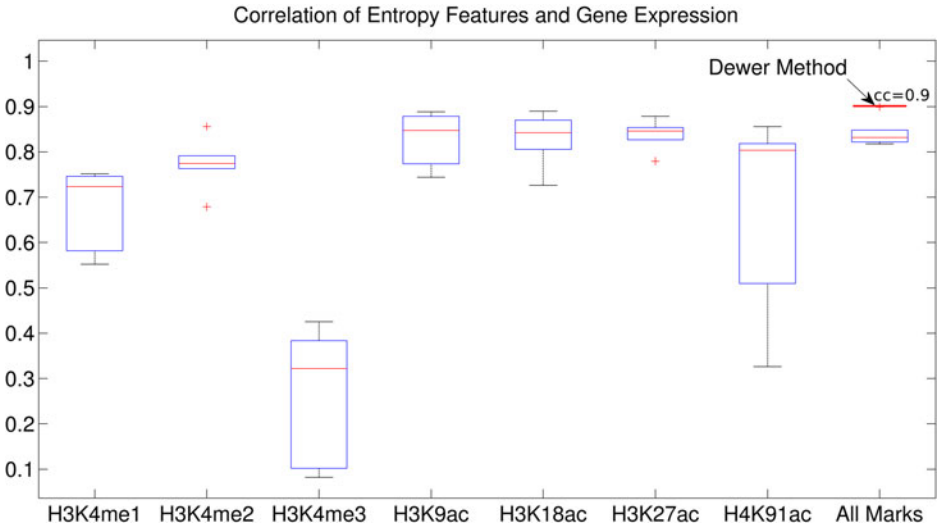

FIG. 10. The entropy value and its association for each histone modification mark. Correlation coefficients (CCs) between expression levels and the entropy levels for the clusters were calculated when using the epigenetic data 1–5 Kbps around the transcription start sites (TSSs). The CC of Dewer reached to 0.9.

Figure 13 shows an example of the epigenetic landscapes around the TSS of *Cd151* and *Orc1* in IMR90. Though *Orc1* had a higher mean value around the TSS, its expression was much lower compared with *Cd151*. The entropy of *Cd151* was much higher, which correlated well with their gene expression levels. These examples as well as our results suggest that entropy is a good marker for evaluating gene expression

4. DISCUSSION

Understanding of the complex nature of epigenomic signals is still a challenging problem. Gene clustering has been successfully applied to study epigenetic regulation. In this article, we introduce a method to improve the clustering performance by using the signal processing approaches. Dewer effectively uses the nature of epigenetic signals for clustering by employing the entropy from the SW. We found that the genes clustered by Dewer better dissected the gene groups in terms of gene expression. Interestingly, gene expression levels well correlated with the entropy levels of epigenetic data. Our results showed that signal processing approaches effectively use the characteristics of epigenetic signals for gene clustering.

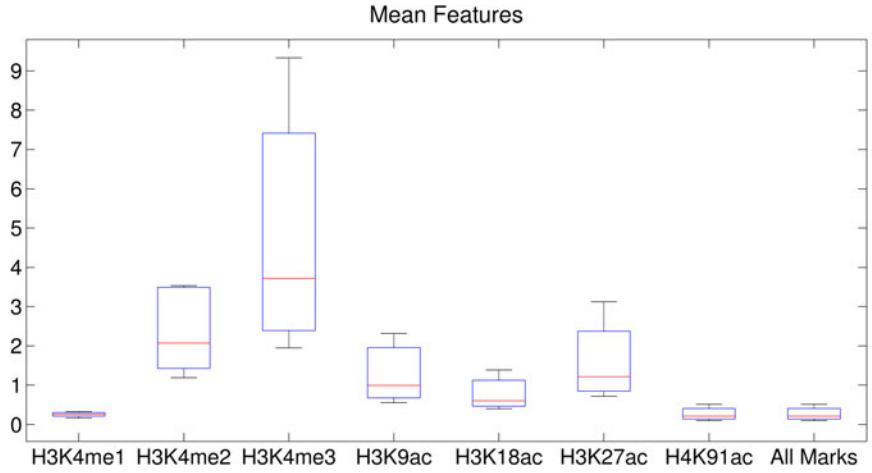

FIG. 11. The mean value of each histone modification mark using the window method.

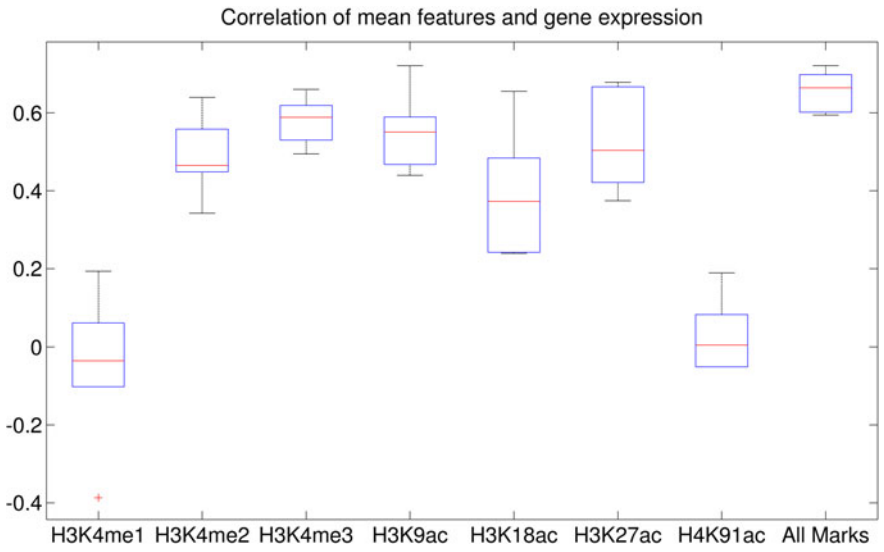

FIG. 12. The mean value and its association for each histone modification mark. CCs between expression levels and the mean levels for the clusters were calculated when using the epigenetic data 1–5 Kbps around the TSSs.

Our tests showed a number of advantages of using entropies. Particularly, entropy can be applied even when the signal intensity is low. This suggests that using entropy or SW entropy is a more robust way to study gene regulation than using mean-based approaches. Also, entropy can be obtained from multiple histone modification marks. We found that using the entropy performed best when using all marks together. This is because the relationships among histone modification marks are well stored in the entropy of the 2D epigenetic images.

Dewer was not designed to predict gene expression using histone modification (Karlic et al., 2010; Dong et al., 2012; Kumar et al., 2013). We used gene expression as a surrogate to show the discriminative power of Dewer in clustering. Our results, at least, suggest that statistical features such as entropy can be a good measure to predict gene expression.

To apply entropy effectively, we restricted entropy to the enriched regions for histone modification signals. SW entropy in HH2 scale produced the best performance, suggesting that the information in high-frequency scale is important for gene clustering. This further suggests that the shape of epigenetic signal is more important than the averaged signals (low frequency) for gene clustering.

Clustering analyses have identified various histone codes, including bivalent promoters (Bernstein et al., 2006), poised enhancers (Creyghton et al., 2010; Rada-Iglesias and Wysocka, 2011), and alternative splicing (Luco et al., 2010; Luco and Misteli, 2011). Computational approaches have been applied to

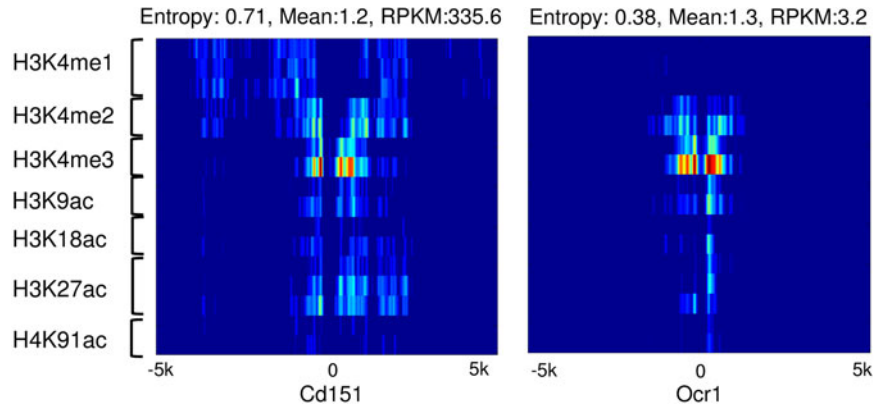

FIG. 13. The epigenetic landscapes around the promoter of *Cd151* and *Ocr1*. The entropy, mean value, and expression levels (reads per kilo base per million, RPKM) are compared in association with epigenetic landscapes.

identify co-enriched histone modifications (Ucar et al., 2011; Rajagopal et al., 2013; Santoni, 2013; Nguyen et al., 2014a). While previous approaches identified combinatorial patterns, it is hard to interpret the clusters obtained using entropy using the same manner. Instead, we use the information contents residing in epigenetic signals for clustering. The results provide us with a good measurement to detect gene expression. Though we restricted our study to promoter regions to investigate the relationships with genes, our approach can easily be applied to distal regulatory regions.

ACKNOWLEDGMENTS

This work was supported by R21-DK098769 and P30-DK19525 from the National Institutes of Diabetes, as well as Digestive and Kidney Diseases and the Diabetes Research Center at the University of Pennsylvania.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Agarwal, S., Verma, A.K., and Singh, P. 2013. Content based image retrieval using discrete wavelet transform and edge histogram descriptor. 2013 International Conference on Information Systems and Computer Networks (ISCON), pp. 19–23.

Barski, A., Cuddapah, S., Cui, K., et al. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* 129, 823–837.

Baylin, S.B., and Jones, P.A. 2011. A decade of exploring the cancer epigenome—biological and translational implications. *Nat. Rev. Cancer* 11, 726–734.

Bernstein, B.E., Meissner, A., and Lander, E.S. 2007. The mammalian epigenome. *Cell* 128, 669–681.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., et al. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* 125, 315–326.

Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., et al. 2010. The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* 28, 1045–1048.

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

Daily, K., Rigor, P., Christley, S., et al. 2010. Data structures and compression algorithms for high-throughput sequencing technologies. *BMC Bioinform.* 11, 514.

Demirel, H., and Anbarjafari, G. 2011. IMAGE resolution enhancement by using discrete and stationary wavelet decomposition. *IEEE Trans. Image Process.* 20, 1458–1460.

Deng, J., Ban, Y.F., Liu, J.S., et al. 2014. Hierarchical segmentation of multitemporal RADARSAT-2 SAR data using stationary wavelet transform and algebraic multigrid method. *IEEE Trans. Geosci. Remote Sens.* 52, 4353–4363.

Dong, X., Greven, M.C., Kundaje, A., et al. 2012. Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol.* 13, R53.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473, 43–49.

Ernst, J., Plasterer, H.L., Simon, I., et al. 2010. Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Res.* 20, 526–536.

Guenther, M.G., Levine, S.S., Boyer, L.A., et al. 2007. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* 130, 77–88.

Harrow, J., Frankish, A., Gonzalez, J.M., et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 22, 1760–1774.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., et al. 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*. 459, 108–112.

Heintzman, N.D., Stuart, R.K., Hon, G., et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* 39, 311–318.

Huber, W., von Heydebreck, A., Sultmann, H., et al. 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 18 Suppl 1, S96–S104.

Jones, P.A., and Martienssen, R. 2005. A blueprint for a Human Epigenome Project: the AACR Human Epigenome Workshop. *Cancer Res.* 65, 11241–11246.

Karlic, R., Chung, H.R., Lasserre, J., et al. 2010. Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci. USA* 107, 2926–2931.

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

Kumar, V., Muratani, M., Rayan, N.A., et al. 2013. Uniform, optimal signal processing of mapped deep-sequencing data. *Nat. Biotechnol.* 31, 615–622.

Laird, P.W. 2003. The power and the promise of DNA methylation markers. *Nat. Rev*. Cancer 3, 253–266.

Li, H., Zhang, K., and Jiang, T. 2004. Minimum entropy clustering and applications to gene expression analysis. Proceedings of the IEEE Computational Systems Bioinformatics Conference, pp. 142–151.

Li, Y., Daniel, M., and Tollefsbol, T.O. 2011. Epigenetic regulation of caloric restriction in aging. *BMC Med.* 9, 98.

Lister, R., Pelizzola, M., Dowen, R.H., et al. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462, 315–322.

Liu, L., van Groen, T., Kadish, I., et al. 2011. Insufficient DNA methylation affects healthy aging and promotes age-related health problems. *Clin. Epigenet.* 2, 349–360.

Luco, R.F., and Misteli, T. 2011. More than a splicing code: integrating the role of RNA, chromatin and non-coding RNA in alternative splicing regulation. *Curr. Opin. Genet. Dev.* 21, 366–372.

Luco, R.F., Pan, Q., Tominaga, K., et al. 2010. Regulation of alternative splicing by histone modifications. *Science* 327, 996–1000.

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

Meissner, A. 2010. Epigenetic modifications in pluripotent and differentiated cells. *Nat. Biotechnol.* 28, 1079–1088.

Menayo, R., Encarnacion, A., Gea, G.M., et al. 2014. Sample entropy-based analysis of differential and traditional training effects on dynamic balance in healthy people. *J. Motor Behav.* 46, 73–82.

Micsinai, M., Parisi, F., Strino, F., et al. 2012. Picking ChIP-seq peak detectors for analyzing chromatin modification experiments. *Nucleic Acids Res*. 40, e70.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., et al. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448, 553–560.

Mikkelsen, T.S., Xu, Z., Zhang, X., et al. 2010. Comparative epigenomic analysis of murine and human adipogenesis. *Cell* 143, 156–169.

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

Nguyen, N., Vo, A., and Won, K.J. 2014a. A wavelet-based method to exploit epigenomic language in the regulatory region. *Bioinformatics*. 30, 908–914.

Nguyen, N., Vo, A., and Won, K.J. 2014b. A wavelet approach to detect enriched regions and explore epigenomic landscapes. *J. Comput. Biol.* Accepted.

Rada-Iglesias, A., and Wysocka, J. 2011. Epigenomics of human embryonic stem cells and induced pluripotent stem cells: insights into pluripotency and implications for disease. *Genome Med*. 3, 36.

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

Rashid, N.U., Giresi, P.G., Ibrahim, J.G., et al. 2011. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biol.* 12, R67.

Santoni, F.A. 2013. EMdeCODE: a novel algorithm capable of reading words of epigenetic code to predict enhancers and retroviral integration sites and to identify H3R2me1 as a distinctive mark of coding versus non-coding genes. *Nucleic Acids Res.* 41, e48.

Shannon, C.E. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423.

Shin, J.H., Park, C.H., Yang, Y.J., et al. 2007. Entropy-based analysis of the non-linear relationship between gene expression profiles of amplified and non-amplified RNA. *Int. J. Mol. Med.* 20, 905–912.

Song, Q., and Smith, A.D. 2011. Identifying dispersed epigenomic domains from ChIP-Seq data. *Bioinformatics* 27, 870–871.

Storey, J.D. 2002. A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 64, 479–498.

Sun, H., Wu, J., Wickramasinghe, P., et al. 2011. Genome-wide mapping of RNA Pol-II promoter usage in mouse tissues by ChIP-seq. Nucleic acids research 39, 190–201.

Swartz, J.B., Rothenberg, S.J., Teklehaimanot, S., et al. 1999. Comparison of the entropy technique with two other techniques for detecting disease clustering using data from children with high blood lead levels. *Am. J. Epidemiol.* 149, 750–760.

Szekely, G.J., and Rizzo, M.L. 2005. Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method. *J. Classification* 22, 151–183.

Ucar, D., Hu, Q., and Tan, K. 2011. Combinatorial chromatin modification patterns in the human genome revealed by subspace clustering. *Nucleic Acids Res.* 39, 4063–4075.

Wang, X.H., Istepanian, R.S.H., and Song, Y.H. 2003. Microarray image enhancement by denoising using stationary wavelet transform. *IEEE Trans. Nanobiosci.* 2, 184–189.

Won, K.J., Chepelev, I., Ren, B., et al. 2008. Prediction of regulatory elements in mammalian genomes using chromatin signatures. *BMC Bioinform.* 9, 547.

Won, K.J., Zhang, X., Wang, T., et al. 2013. Comparative annotation of functional regions in the human genome using epigenomic data. *Nucleic Acids Res.* 41, 4423–4432.

Xie, R., Everett, L.J., Lim, H.W., et al. 2013. Dynamic chromatin remodeling mediated by polycomb proteins orchestrates pancreatic differentiation of human embryonic stem cells. *Cell Stem Cell* 12, 224–237.

Yogesan, K., Jorgensen, T., Albregtsen, F., et al. 1996. Entropy-based texture analysis of chromatin structure in advanced prostate cancer. *Cytometry* 24, 268–276.

Yu, P., Xiao, S., Xin, X., et al. 2013. Spatiotemporal clustering of the epigenome reveals rules of dynamic gene regulation. *Genome Res.* 23, 352–364.

Zang, C., Schones, D.E., Zeng, C., et al. 2009. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* 25, 1952–1958.

Zhang, Y., Liu, H., Lv, J., et al. 2011. QDMR: a quantitative method for identification of differentially methylated regions by entropy. *Nucleic Acids Res.* 39, e58.

Address correspondence to:
Dr. Kyoung-Jae Won
Department of Genetics
School of Medicine
University of Pennsylvania
3400 Civic Center Boulevard
Philadelphia, PA 19104

E-mail: wonk@mail.med.upenn.edu