JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 16, Number 2, 2009

Pp. 341–355
DOI: 10.1089/cmb.2008.18TT

Targeted Recruitment of Histone Modifications in Humans Predicted by Genomic Sequences

GUO-CHENG YUAN

ABSTRACT

Histone modifications are important epigenetic regulators and play a critical role in development. The targeting mechanism for histone modifications is complex and still incompletely understood. Here we applied a computational approach to predict genome-scale histone modification targets in humans by the genomic DNA sequences using a set of recent ChIP-seq data. We found that a number of histone modification marks could be predicted with high accuracy. On the other hand, the impact of DNA sequences for each mark is intrinsically different dependent upon the target- and tissue-specificity. Diverse patterns are associated with different repetitive elements. Unexpectedly, we found that non-overlapping, functionally opposite histone modification marks could share similar sequence features. We propose that these marks may target a common set of loci but are mutually exclusive and that the competition may be important for developmental control. Taken together, we show that our computational approach has provided new insights into the targeting mechanism of histone modifications.

Key words: DNA sequence, histone modification, human.

1. INTRODUCTION

THE DISTRIBUTION OF HISTONE MODIFICATIONS is not uniform across the human genome. The active genes are often marked with histone acetylation and H3K4me3 (Barski et al., 2007; Bernstein et al., 2006; Liu et al., 2005; Wang et al., 2008), whereas H3K27 trimethylation often marks developmental genes (Bernstein et al., 2006; Boyer et al., 2006; Lee et al., 2006). H3K9 di- and tri-methylations are often associated with heterochromatin (Hall et al., 2002). Certain marks preferentially target the 5′ gene starts, whereas others are biased toward transcribed regions or gene deserts (Bernstein et al., 2006; Hall et al., 2002; Liu et al., 2005; Pokholok et al., 2005). There are also distinct differences among differentially methylated marks within the same group. H3K4me3 punctually marks the transcription start sites (TSS) of most actively transcribed genes (Bernstein et al., 2006; Guenther et al., 2007; Liu et al., 2005; Pokholok et al., 2005). Recently, it has been shown that H3K4me3 can also be recruited to genes that are poised for activation (Core and Lis, 2008; Guenther et al., 2007). On the other hand, H3K4me2 and H3K4me1

Department of Biostatistics, Harvard School of Public Health, and Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Cambridge, Massachusetts.

occupy broader domains and can be associated with either active or repressive genes (Bernstein et al., 2006; Liu et al., 2005; Pokholok et al., 2005).

The regulation of histone modifications is complex. The enzymatic proteins can be recruited by transcription factors to their targets (Robert et al., 2004). They may also recognize specifically modified histones (Kouzarides, 2007). Furthermore, certain histone marks may be brought to the chromatin by transcriptional machineries (Xiao et al., 2003). An initial histone modification mark can be spread over neighboring domains (Talbert and Henikoff, 2006). Histone modification marks are not static but dynamic and removal of existing marks can be regulated by cell division and by specialized enzymes (Kurdistani and Grunstein, 2003; Lan et al., 2008; Shi et al., 2004).

The genomic DNA has been implicated in the targeting of histone modifications. H3K4me3 and H3 acetylation are highly correlated with CpG islands (Bernstein et al., 2006; Roh et al., 2005). In *Drosophila*, the Polycomb complex PRC2 can be recruited to specific DNA elements called the Polycomb response elements (PRE), where H3K27me3 is then synthesized (Ringrose et al., 2003). Although PREs in mammalian organisms have yet been found, in mouse embryonic stem (ES) cells H3K27me3 marked regions are depleted with transposons (Bernstein et al., 2006). Repetitive elements are associated with heterochromatin and H3K9 methylation (Hall et al., 2002; Slotkin and Martienssen, 2007), and diverse histone modification patterns can be associated with different classes of repetitive elements (Martens et al., 2005). Non-coding RNAs also play a role in the regulation of histone marks (Hall et al., 2002; Rinn et al., 2007; Slotkin and Martienssen, 2007). It is thought that the role of DNA sequence may be more important in ES cells but less so in differentiated cells (Bernstein et al., 2007). However, it is not yet known to what extent the DNA sequence affect global histone modification patterns. This probably involves accumulation of many weak interactions rather than distinct short sequence motifs (Straub and Becker, 2008).

Recently genome-wide locations of a number of histone modification marks in humans have been mapped (Barski et al., 2007; Guenther et al., 2007; Lee et al., 2006; Pan et al., 2007; Schones et al., 2008; Wang et al., 2008; Zhao et al., 2007). These data have provided an unprecedented opportunity to evaluate the role of DNA sequence in histone modification locations in humans. We have previously developed a computational model, called the *N*-score model, to predict nucleosome positions, and have found that the model is able to detect discriminative sequence periodic patterns in nucleosome sequences compared to linker sequences in yeast (Yuan and Liu, 2008). Here the *N*-score model is modified by incorporating sequence features identified by other groups (Lee et al., 2007b; Peckham et al., 2007) and applied to predict histone modification locations in humans.

2. RESULTS

2.1. A modified N-score model

The original *N*-score model was developed using a wavelet analysis approach (Yuan and Liu, 2008). A commonly used approach for model improvement is combining the strength of multiple prediction models, and it has been found that such an ensemble-based approach combining often performs better than the best individual model (Hastie et al., 2001; Polikar, 2006). Several groups have developed computational methods to predict genome-wide nucleosome positions by genomic sequences (Ioshikhes et al., 2006; Lee et al., 2007b; Miele et al., 2008; Peckham et al., 2007; Segal et al., 2006). These methods have focused on three aspects of sequence features: (1) sequence periodicity (Ioshikhes et al., 2006; Segal et al., 2006); (2) counts of short nucleotide sequences (Peckham et al., 2007); and (3) structural parameters (Lee et al., 2007b; Miele et al., 2008). We modified our original *N*-score model by integrating the wavelet features in our model together with the sequence features identified in two different studies (Lee et al., 2007b; Peckham et al., 2007). To test whether combining these sequence features improved model performance, we applied this modified *N*-score model to predict nucleosome positions in humans, based on a recent ChIP-seq dataset (Schones et al., 2008). The model was trained by using the highest and lowest nucleosome scoring fragments from chromosomes 1–12 and evaluated based on those similarly selected fragments from the other chromosomes. This new model perform very well (AUC = 0.90, where AUC represents the area under the roc curve), slight better than the best performing individual model (Fig. S1). (See online Supplementary Material at *www.liebertonline.com*.)

We also tested the performance of our model by comparing model prediction and ChIP-seq data over a number of contiguous regions. However, we were disappointed that the model did not perform well. We also repeated the tests for the models that only contained wavelets, word counts, or structural parameters features. All of these models performed similarly. The results from the ROC curves and over contiguous regions do not contradict each other. Notice that an AUC score only measures the model performance in discriminating the sequences at the most enriched and depleted loci, whereas the sequence features associated with lower peaks may be less pronounced. Therefore, an AUC score by itself is not sufficient for evaluating model performance. Previous studies have suggested that while the nucleosome immediate downstream of a TSS is often positioned and predictable from DNA sequence (Ioshikhes et al., 2006; Mavrich et al., 2008; Shivaswamy et al., 2008; Valouev et al., 2008; Yuan and Liu, 2008), the majority of nucleosomes may be statistically positioned (Kornberg and Stryer, 1988; Mavrich et al., 2008; Valouev et al., 2008). Our analysis here suggests a similar situation in humans.

We next applied this modified *N*-score model to predict the locations of histone modification marks in humans, based on two recent datasets that in combination contained genome-wide locations of 18 methylation and 18 acetylation marks (Barski et al., 2007; Wang et al., 2008). For each histone mark, an *N*-score model was trained separately. The predictabilities for different markers are highly variable, where the AUC varies from 0.66 to 0.96 (Table 1). Besides H3K4me3 (AUC = 0.96) there are a number of histone marks that are associated with high predictability, including the well-studied marks H3K4me1, H3K4me2, and H3K9ac (AUC = 0.90, 0.88, and 0.93, respectively; Fig. 1). On the other hand, a number of repressive marks, such as H3K9me3 (AUC = 0.77), are not predicted well.

The high predictability associated with H3K4me3 was expected. In mouse ES cells, it has been found that 95% of the H3K4me3 peaks fall into CpG islands, and 91% CpG islands contain H3K4me3 marks (Bernstein et al., 2006). We compared the performance of our *N*-score model with a simple model using CpG density as the only predictor. Strikingly, the performance of the two models is almost indistinguishable (Fig. 1). However, for most other histone marks, the performance of the *N*-score model is significantly better than using CpG density alone. In addition, our *N*-score model can detect the variation of the histone modification patterns among different marks (Fig. 1). Notice that our model correctly predicts that the

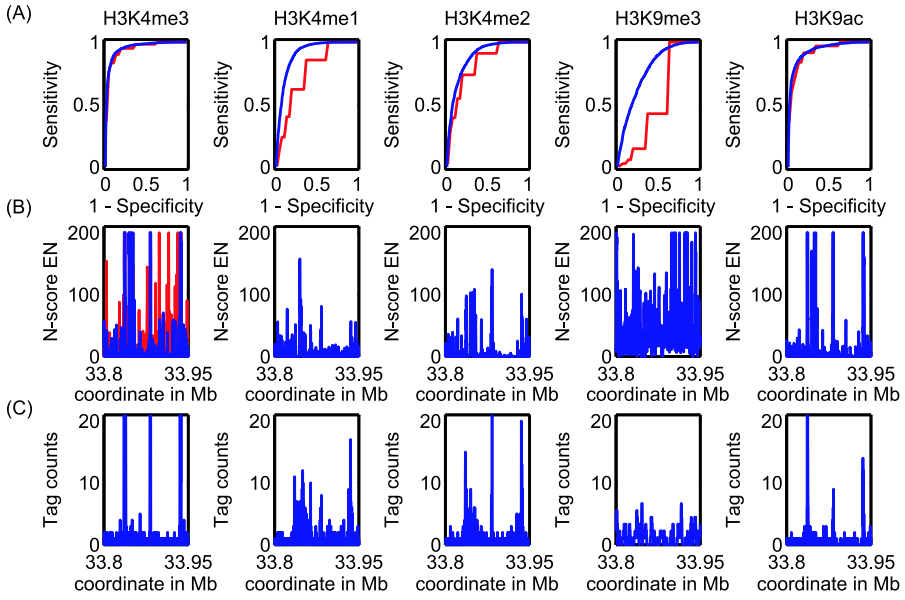

FIG. 1. The performance of the modified *N*-score model for a few representative histone modification marks. **(A)** The receiver operating characteristic (ROC) curves. The red curves correspond to the ROC curves obtained by using the CpG frequency information alone. **(B)** The *N*-score enrichment score (EN). For comparison, the enrichment score corresponding to CpG density is shown as the red curve in the left panel. Values higher than 200 are truncated to aid visualization. **(C)** Sequence tag counts from the ChIP-seq experiment (bin size = 200 bp). Values higher than 20 are truncated to aid visualization.

TABLE 1. MODEL PREDICTABILITY (AUC SCORE) IS CORRELATED WITH SIGNAL STRENGTH AND TARGET SPECIFICITY

Histone mark	*AUC*	*Signal strength score*	*Target specificity score*	*Correlation with gene expression*
H3K4ME3	0.96	0.37	0.37	0.29
POLII	0.95	0.29	0.33	0.26
H4K91AC	0.94	0.53	0.35	0.26
H3K9AC	0.93	0.20	0.37	0.29
H3K18AC	0.93	0.38	0.30	0.26
H2BK5AC	0.93	0.46	0.37	0.28
H2BK120AC	0.92	0.38	0.33	0.27
H3K27AC	0.92	0.45	0.34	0.27
H3K4AC	0.92	0.25	0.28	0.27
H4K20ME1	0.92	0.57	0.30	0.20
H2BK5ME1	0.91	0.34	0.24	0.15
H3K4ME1	0.90	0.47	0.27	0.07
H2BK20AC	0.90	0.34	0.28	0.28
H2A.Z	0.90	0.23	0.23	0.00
H4K5AC	0.88	0.24	0.26	0.25
H3K36AC	0.88	0.19	0.34	0.29
H3K4ME2	0.88	0.33	0.24	0.13
H3R2ME1	0.88	0.05	0.17	0.01
H4K8AC	0.87	0.20	0.25	0.28
H2AK5AC	0.86	0.12	0.18	−0.01
H3K27ME2	0.86	0.05	0.15	−0.14
H3K9ME1	0.86	0.31	0.20	0.24
H4K16AC	0.85	0.06	0.23	0.26
H2BK12AC	0.84	0.18	0.26	0.29
H3K23AC	0.83	0.12	0.17	−0.02
H3K79ME3	0.83	0.10	0.22	0.21
H2AK9AC	0.82	0.12	0.21	0.28
H3K36ME1	0.81	0.03	0.17	−0.07
CTCF	0.80	0.29	0.35	0.26
H3K27ME1	0.80	0.12	0.16	−0.06
H3K36ME3	0.80	0.20	0.21	−0.15
H3K27ME3	0.79	0.06	0.15	0.05
H4R3ME2	0.79	0.03	0.32	−0.11
H3K9ME3	0.77	0.13	0.20	−0.23
H3K14AC	0.77	0.03	0.23	−0.10
H3K9ME2	0.75	0.04	0.16	−0.21
H3K79ME1	0.74	0.02	0.21	0.14
H4K20ME3	0.71	0.22	0.38	0.10
H3R2ME2	0.71	0.03	0.23	−0.17
H3K79ME2	0.69	0.02	0.22	−0.22
H4K12AC	0.66	0.07	0.17	0.21

The last column shows the correlation between the promoter averaged (−1 to +1 kb from TSS) *N*-scores and gene expression levels.

H3K4me1 peak is narrower than and upstream of the H3K4me3 peak (Fig. S3). (See online Supplementary Material at *www.liebertonline.com*.) Such differences cannot be predicted by the CpG density alone.

To assess the ability of our model to predict histone modification patterns over a contiguous region, we calculated the *N*-score over a 150-kb region on chromosome 21, a region featured in reference (Guenther et al., 2007). Chromosome 21 was not included in our model training. The number of tags mapped to a genomic location is dependent upon the number of total reads, which are not uniform among different histone marks. To highlight the significant peaks predicted from our model, the *N*-scores are converted to an enrichment score as follows. First we calculate the p-values by one-sided z-tests over overlapping sliding windows (window size = 400 bp, step = 20 bp). Then the enrichment score is defined as $-10 \times \log_{10}$ (p-value; Fig. 1B). The predicted peaks indeed agree well with experimental data (Fig. 1C).

As mentioned in the above, our model did not perform well for a number of repressive marks, such as H3K9me3. We suspected that the lower predictability might be correlated with higher noise (lower signal strength), which might be caused by non-specific binding of the ChIP antibodies, insufficient sequencing depth, or sequencing error. We used a Poisson distribution to model the noise background and estimated the signal strength for each histone mark. The signal strength was highly variable (Table 1). Indeed, the model predictability (quantified by AUC score) was strongly correlated with the signal strength ($\rho = 0.71$).

Whereas data noise is introduced by imperfect technology and can be substantially reduced with technology development, it is also possible the distribution of a histone modification mark is neither focal nor DNA sequence dependent. Therefore, we were interested to search for biological reasons related to predictability. Most of the less well predicted markers appeared to occupy broader domains compared to H3K4me3. To test whether there was a general trend, we quantified the target specificity of each histone mark as the ratio of the tag counts for the top 10% significant bins over those in all significant bins. The predictability was indeed significantly correlated with target specificity ($\rho = 0.47$).

2.2. *Predicted promoter and enhancer histone modification patterns*

Histone modifications play an important role in transcription control. The occupancy levels of nucleosomes and several histone modification marks are strongly correlated with gene expression (Barski et al., 2007; Heintzman et al., 2007; Liu et al., 2005; Pokholok et al., 2005), even after controlling for the confounding effect of transcription factor binding (Yuan et al., 2006). Consistent with this regulatory role, distinct histone modification patterns have been found at promoters and enhancers (Barski et al., 2007; Heintzman et al., 2007; Roh et al., 2007) (Fig. S2). (See online Supplementary Material at *www.liebertonline.com*.) To normalize the variation due to sequence depth, the raw tag counts are converted to an enrichment score, calculated as the log2-ratio of the observed tag counts over the number expected at random. Thus, a negative value represents depletion of a histone mark, as is the case for H3K9me3 at promoter regions.

To investigate whether the DNA sequence plays a role in the establishment of the histone modification patterns at promoters, we calculated the *N*-scores at genome-wide promoters and obtained an average profile. The computational findings are generally in good agreement with experimental data (compare Fig. 2A with Fig. S2A). Notably, the predicted H3K4me3 and H3K9ac scores are both concentrated near TSS. In comparison, the peaks for H3K4me2 and H3K4me1 scores are broader and further away from the 5′ ends. Finally, the H3K9me3 scores are negative (less than expected at random) near TSS. To test whether our sequence model was able to correctly predict the observed correlation between gene expression and histone modification marks, we compared the *N*-score patterns for the most active and repressed genes (Fig. 2B). For active marks including H3K4me3 and H3K9ac, higher expressed genes tend to have higher *N*-score, whereas for H3K9me3 the relationship is reversed (Fig. 2B). However, the H3K4me2 and H3K4me1 patterns do not show any obvious trend as suggested by the data. To be quantitative, we summarized the *N*-score over each promoter by averaging over a 2-kb window centered at the TSS and calculated the Pearson correlation between these summary scores and gene expression levels (Table 1). The quantitative results suggest that the H3K4me2 and H3K4me1 scores are also positively correlated with gene expression, refining the qualitative conclusions from visualization. A notable discrepancy is that the score for H3K27me3, a well-studied repressive mark, is slightly positively correlated with gene expression ($\rho = 0.05$). Since the correlation with gene expression is confounded by nucleosome occupancy, which is negative correlated with gene expression ($\rho = -0.18$), we also calculate the partial correlation between

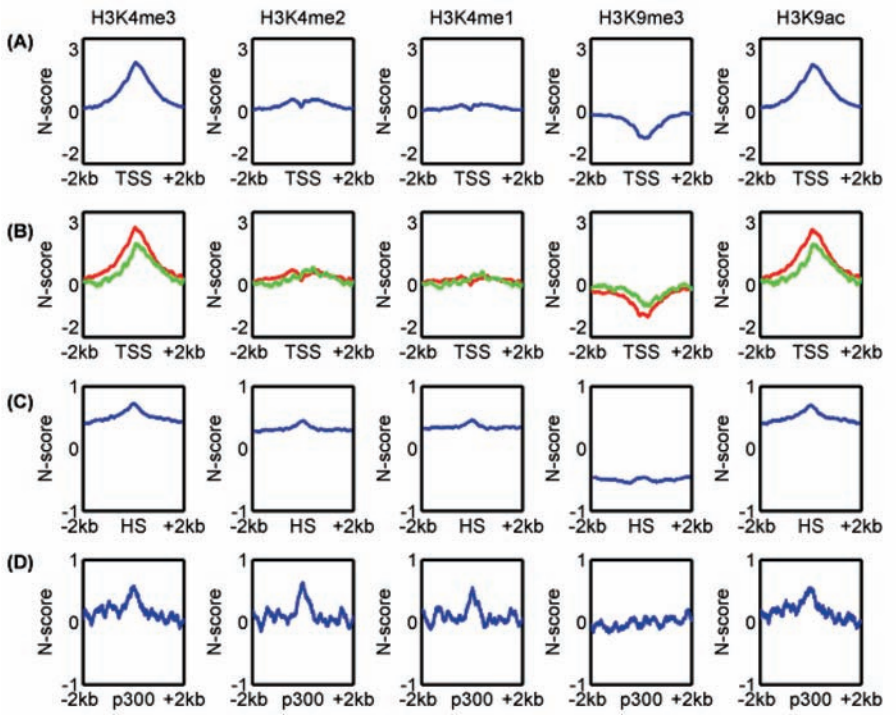

FIG. 2. The average *N*-score profiles over promoters and enhancers. **(A)** The average *N*-score profiles over all promoters. **(B)** The average *N*-score profiles over promoters for highly expressed (expression index > 100; red curves) and lowly expressed (expression index < 20, green curves) genes. **(C)** The average *N*-score profile for the DNase hyper-sensitive sites excluding CTCF and PolII sites. **(D)** The average *N*-score profile for the p300 binding sites in Hela cells.

N-scores and gene expression while controlling for nucleosome occupancy level, similar to our previous work (Yuan et al., 2006). The results are nearly unaffected. A likely explanation for the discrepancy related to H3K27me3 is that its pattern varies significantly from tissue to tissue. Such variation cannot be predicted from sequences alone. The properties for H3K27me3 will be further discussed in the following.

An enhancer is a region of DNA that can regulate gene expression from a distance. An interesting question is whether the function of enhancers is related to distinct histone modification patterns. However, enhancer activities are cell-specific and it remains difficult to experimentally identify active enhancers at a genomic scale. The histone modification patterns associated with enhancers have been recently investigated by two groups (Barski et al., 2007; Heintzman et al., 2007). Heintzman et al. use p300 locations as the key mark for putative enhancers and have investigated their associated signatures in Hela cells (Heintzman et al., 2007), whereas Barski et al. (2007) choose DNase I hypersensitive (HS) sites excluding CTCF and PolII sites and have investigated in the context of $CD4^+$ T cells (Barski et al., 2007). These two studies have identified a number of similar patterns, including nucleosome depletion, hyper-acetylation, and high level of H3K9me1, H3K4me1, and H3K4me2. On the other hand, the two studies have also resulted important differences. Notably, H3K4me3 is found to be depleted in Heintzman et al.'s (2007) study but enriched in Barski et al.'s (2007) study. Two possibilities have been suggested to explain this discrepancy. First, there may exist different classes of enhancers associated with different histone modification patterns. Second, the discrepancy may be related to the technical differences between ChIP-chip versus ChIP-seq.

We were interested to predict histone modification patterns associated with enhancers by using the DNA sequence. To this end, we calculated the *N*-scores at the putative enhancers in both studies. For both sets of putative enhancers, the H3K9ac, H3K4me1, and H3K4me2 scores are higher than random, whereas the H3K9me3 score is depleted near the center (Figs. 2C and 2D), consistent with experimental data (Figs. S2C and S2D). Interestingly, we found that the H3K4me3 score at the putative enhancers was also elevated for both sets of putative enhancers, although more moderate compared to the promoters, suggesting that local DNA sequence was favorable for H3K4me3.

Our *N*-score analysis suggests significant differences between the two sets of putative enhancers. For example, the H3K9me3 data show depletion in the HS sites but this bias is less obvious for the p300 sites. The baseline for H3K4 methylations and H3K9 acetylation is higher at the HS sites than at the p300 binding sites, suggesting HS sites are more open. Interestingly, these subtle differences in experimental data can be predicted from the DNA sequences.

2.3. Histone modification patterns associated with repetitive elements

About half of the human genome consists of repetitive elements. It is increasingly clear that these repetitive elements are not junk DNA, but rather play important roles in gene regulation (Slotkin and Martienssen, 2007). Current technologies can identify the DNA sequences marked by a specific histone modification, but it is very difficult to identify the locations of the marked loci as these DNA sequences cannot be uniquely mapped to the genome. In fact, repetitive sequences are routinely excluded in standard ChIP-chip and ChIP-seq data analyses. An advantage of our computational approach is that it is not affected by sequence mappability or any other technical limitations, thus providing a useful tool for analyzing functions of histone modifications in repetitive regions.

We applied our model to calculate the *N*-scores associated with eight types of repetitive elements: the AT-rich and GC-rich elements, the Alu elements, the long-terminal repeats (LTR), the LINE1 and LINE2 repeats, and the Charlie and Mariner DNA transposons. For Charlie and Mariner, we analyzed their genome-wide sites, whereas for the other repetitive sequences, we only analyzed the ENCODE regions because they were more abundant. The distributions of *N*-scores are diverse among different classes of repeated elements (Fig. 3A). For the Alu elements, the H3K9me2, H3K9me3, and H4K20me3 scores are all low, but the H3K4me2 score is slightly above random. LINE1 and the DNA transposons Charlie and Mariner are associated with high scores for both H3K9me2 and H3K9me3. On the other hand, LINE2 and LTR elements are associated with high levels of H3K27me3 and H3K9me2. These predictions are qualitatively similar to the experimental data in mouse (Martens et al., 2005). The most significant disagreement occurs for H4K20me3, for which our model does not work well (Table 1).

We next analyzed the histone modification pattern at the juncture between the centromere and pericentromere, containing highly repetitive sequences characterized mainly by α-satellite sequences. The centromere is the assembly site for kinetochore during mitosis and its function is closely related to its unique chromatin structure called the CEN chromatin (Schueler and Sullivan, 2006). However, high resolution location analysis difficult because of the mappability issue mentioned above. Current studies rely on the florescent *in situ* hybridization (Lam et al., 2006), which is limited by resolution.

We applied our *N*-score model to predict the histone modification pattern at a 10 bp resolution over a 100-kb region on the X-chromosome immediately adjacent to the centromere gap in the current genome assembly. The higher-order α-satellite array DXZ1 marks the centromere territory, which ends at around 40 kb (relative to the genome assembly gap). Our *N*-score analysis shows that the entire region was associated with high H3K9me2/me3 and low H3K4 methylation scores, consistent with heterochromatin conformation (Fig. 3B). While the H3K9me prediction is consistent with experimental data, we found disagreement between the predicted and observed H3K4me2 pattern. The experimental data show that the H3K4me2 mark is enriched in the CEN chromatin (Lam et al., 2006); however, our model predicts low H3K4me2 occupancy throughout centromeric and pericentromeric regions. Interestingly, a closer look at the *N*-score pattern suggests subtle transitions at the edge of the DXZ1 boundary. On the inner side of the boundary, the H3K4me2 score is slightly elevated and the H3K9me3 score is slightly decreased. Although small in magnitude, these results suggest that the DNA sequence may be partially involved in establishing the boundary between the CEN chromatin and heterochromatin. It appears that the predicted pattern can be partially explained by using the α-satellite sequence alone. However, our model training step used only ChIP-seq data in uniquely mappable regions, and that the association between *N*-score and α-satellite sequence could not have been expected.

2.4. Common sequence features shared by different histone modification marks

Analysis of high-resolution ChIP-chip data has shown that the distributions of different histone marks are not independent with each other, but they rather form relatively few clusters (Liu et al., 2005; Wang

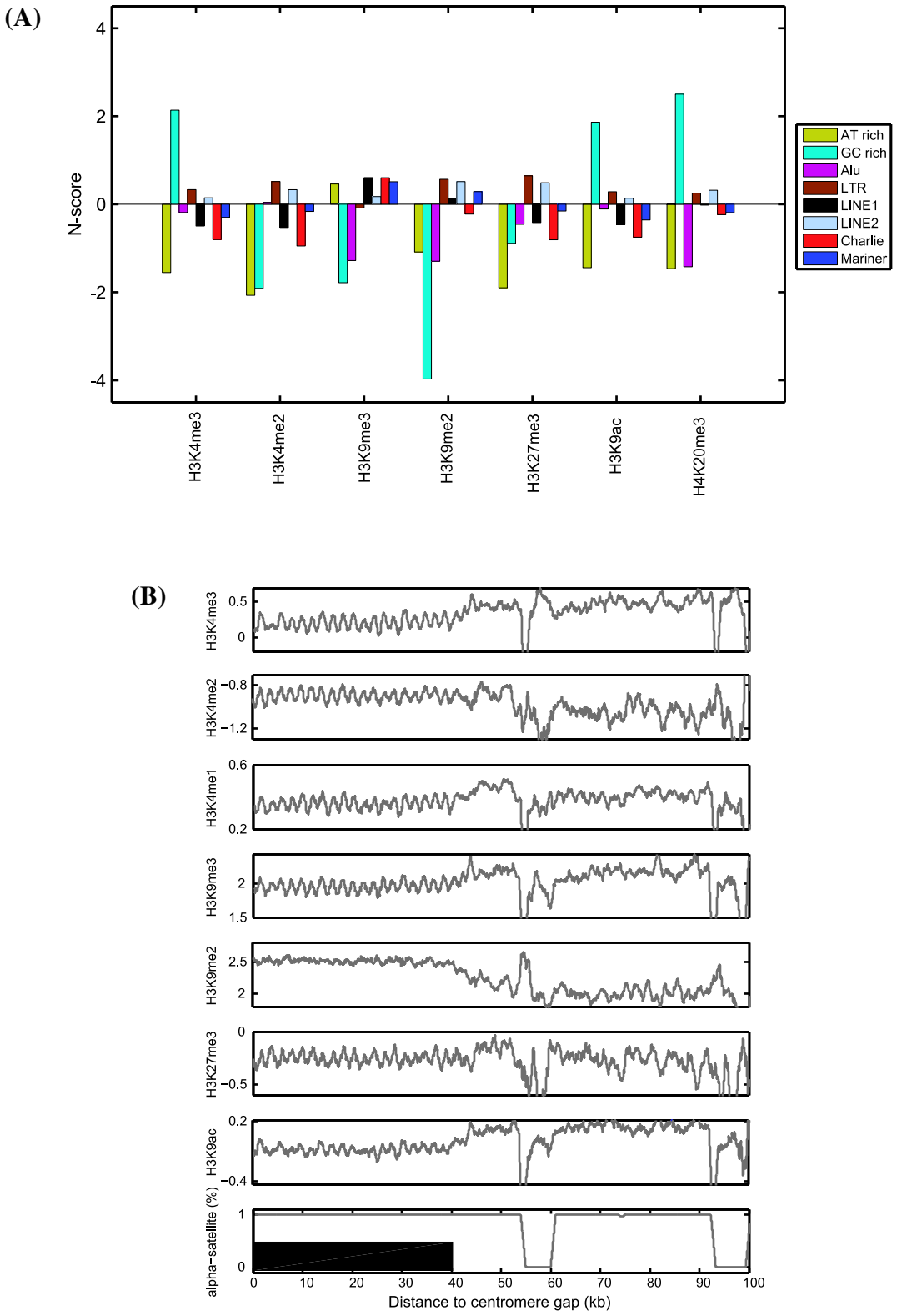

FIG. 3. The *N*-score analysis over repetitive regions. **(A)** The average *N*-score values for different groups of repetitive elements. **(B)** The *N*-score profiles over a 100-kb region striding the DXZ1 array boundary. Coordinate zero marks the end position of the centromere gap on chromosome X. The position of the DXZ1 array elements is marked by the black rectangle. The α-satellite percentage was calculated by averaging over 1-kb windows.

et al., 2008). By clustering analysis, Wang et al. (2008) identified a histone modification backbone pattern containing 17 different marks occupying the promoters of more than three thousand genes, and these genes tend to have higher expressions than others. This backbone pattern strongly suggests high degree of cooperativity among different histone marks.

We were interested to test whether sequence analysis could identify correlations that were not obvious from simple comparison of the ChIP-seq data. We focused on a subset of 27 histone marks for which our sequence had at least modest prediction power. The ChIP-seq data were clustered based on pairwise Pearson correlation as in Wang et al. (2008), and the similarity between two sequence features was quantified by the Pearson correlation between their corresponding *N*-scores. As a simple test for consistency, colocalized marks shared similar sequences (Fig. 4). Interestingly, we also observed striking differences between predicted and experimentally identified clustering patterns. Notably, H3K4me2 and H3K27me3 are anti-correlated in the ChIP-seq data but their corresponding *N*-scores are highly correlated ($\rho = 0.88$). Although less significantly, the correlation between H3K4me3 and H3K27me3 is also high ($\rho = 0.67$).

To explain the apparent discrepancy between ChIP-seq data and our model predictions, we hypothesized that functionally opposite marks might be recruited to the same region but interact antagonistically. Such a competition mechanism, if it exists, would provide a source for epigenetic variability under different conditions. Indeed, we analyzed additional ChIP-chip data in two different studies and found support for such a competition mechanism, as explained below. A recent ChIP-chip study has identified striking differences in the H3K27me3 profiles at the HoxA gene cluster among different cell-types in humans (Rinn et al., 2007), in a manner consistent with tissue-specific gene expression patterns. In particular, the H3K27me3 data in the lung is very different from that in the foot (Fig. 5A), and their Pearson correlation is negative ($\rho = -0.3$). Strikingly, the H3K4me2 data in the lung are highly correlated with the H3K27me3 data in the foot ($\rho = 0.6$), suggesting that the two histone marks share a common pool of potential targets.

As indicated in Figure 5, the HoxA cluster is divided into two large-scale epigenetic domains, each being activated in a specific tissue type. Constrained on a single domain, such as the left one, we found good agreement between the predicted *N*-score and tiling array data ($\rho = 0.33$; Fig. 5B). However, the correlation coefficient decreases sharply when both domains were included ($\rho = 0.05$). Within each domain, we calculated the *N*-scores for the 10% most enriched and depleted probes (for H3K27me3) and found that they indeed have very different distributions (AUC = 0.85). Thus, while our model is unable to predict tissue-specific histone modification patterns, it appears to be useful for predicting the overall targetable loci.

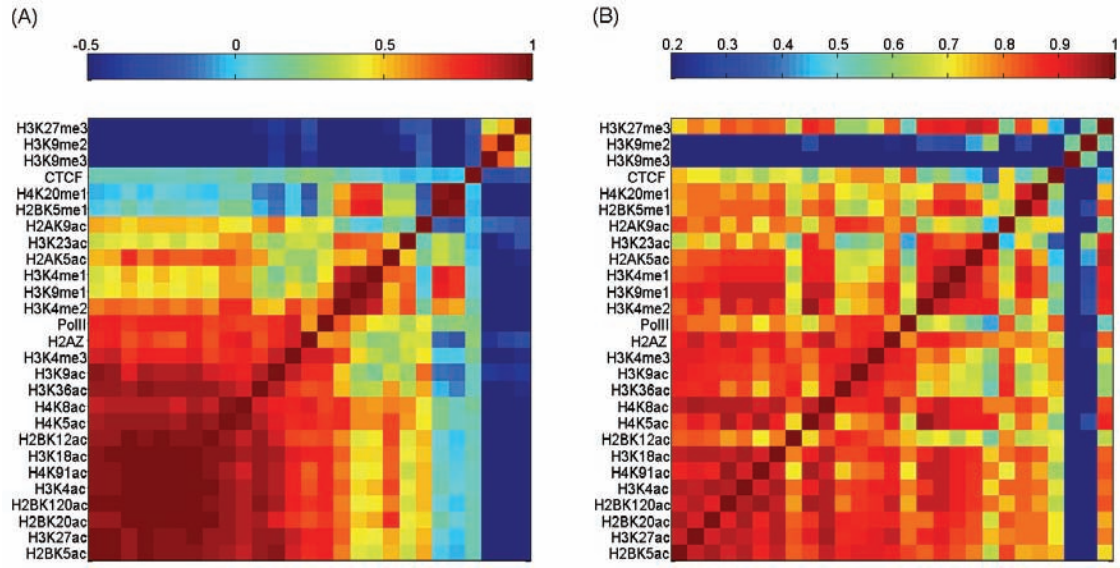

FIG. 4. A heatmap showing the Pearson correlation coefficients between different histone modification patterns. **(A)** ChIP-seq data. **(B)** Sequence features. Notice the color scale difference between the two plots.

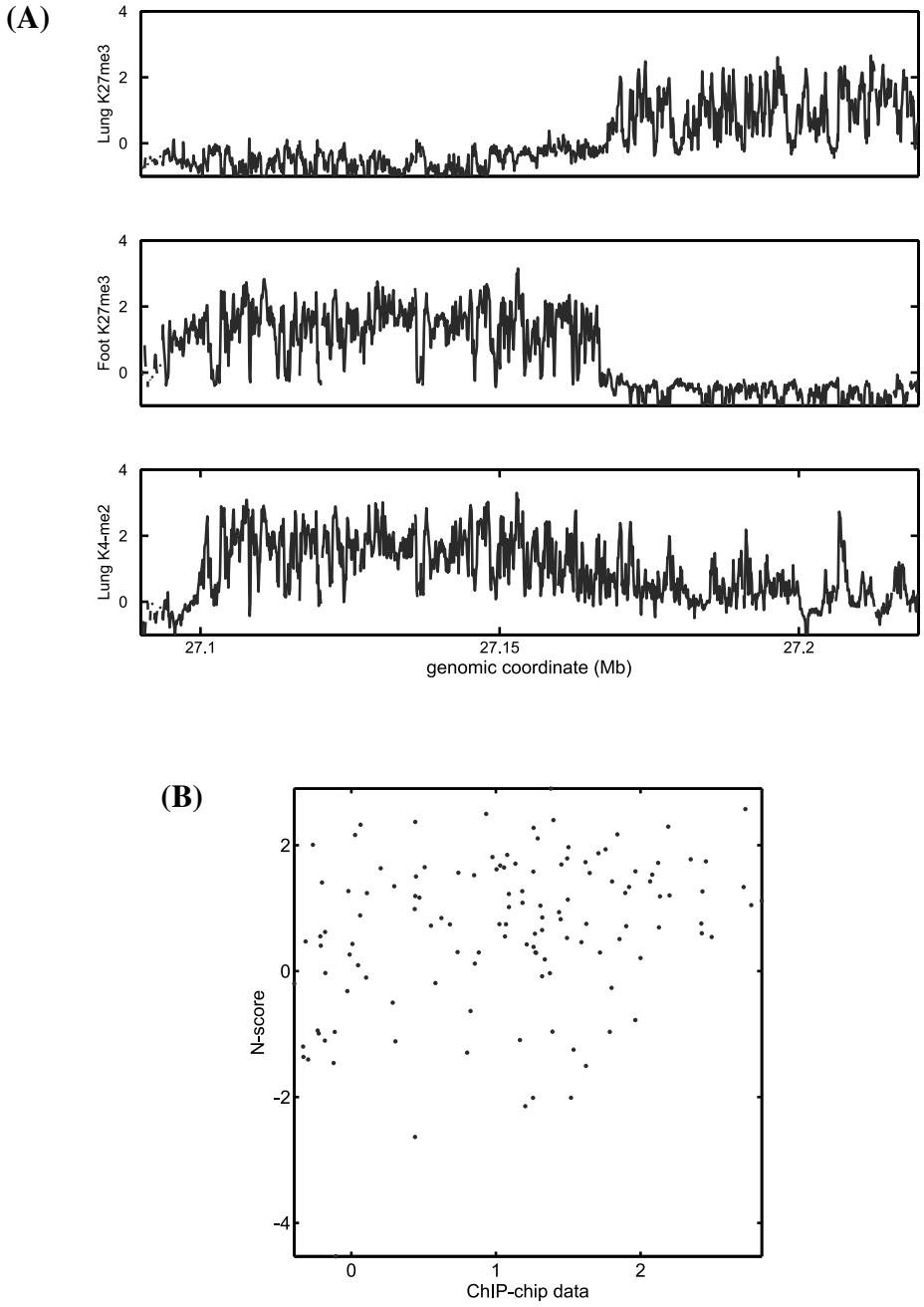

FIG. 5. **(A)** The ChIP-chip data for H3K4me2 and H3K27me3 in different tissues over the HoxA cluster. **(B)** The scatter plot of the predicted H3K27me3 *N*-score versus the H3K27me3 ChIP-chip data in foot over the sub-region with [27090000 27170000].

Another pair of functionally opposite histone marks is H3K4me3 and H3K27me3, while the two marks usually target different regions in differentiated cells, they co-occupy in ES cells at a number of genes, many of which are developmental regulators whose expression is poised for activation (Bernstein et al., 2006; Guenther et al., 2007; Lee et al., 2006; Pan et al., 2007; Zhao et al., 2007). By merging the H3K4me3 and H3K27me3 targets in two studies (Lee et al., 2006; Guenther et al., 2007), we identified a total number of 1354 bivalent genes, that is, whose promoters were cooccupied by H3K4me3 and H3K27me3. For each promoter, we defined a summary *N*-score by averaging over a window near TSS. We chose

a wider window size for H3K27me3 (−5 to +5 kb from TSS) than for H3K4me3 (−1 to 1 kb from TSS), as the H3K27me3 peaks were in general broader than H3K4me3. We found that 525 (or 39%) of the bivalent genes were associated with a high H3K4me3-score (>3), compared to 18% expected at random, 521 (or 38%) were associated with a high H3K27me3-score (>1), compared to 27% expected at random, and 254 (or 19%) of the bivalent genes were associated with high *N*-scores for both marks, compared to 7% expected at random. These differences are not substantial but highly statistically significant (p-value $< 10^{-20}$ in all three cases).

3. DISCUSSION

We have developed a computational model to predict histone modification patterns in humans from genomic sequences. Our model is effective for the prediction of a number of histone marks, as evidenced by the ROC curves, regional predictions, and promoter and enhancer analysis. In addition, the computational method has also enabled us to explore the histone modification pattern at highly repetitive regions such as the centromere and pericentromere.

On the other hand, our model performs poorly for a number of repressive marks such as H4K20me3. Whereas new development of computational models and increased data quality are expected to further improve the prediction power, it is also likely that DNA sequences may not be important for some of these marks. We have identified two biological factors affecting the predictability of sequence-based models: target-specificity and tissue-specificity. Whereas data noise is also an important limiting factor, its impact can be removed in principle as technology advances.

We have detected similar sequence patterns between functionally opposite histone modifications exemplified by the H3K4me2 and H3K27me3 pair and hypothesized that these two marks are recruited toward common genomic loci and compete against each other for targeting. Direct interactions between the H3K4 and H3K27 methylation and demethylation machineries have been detected and they form a positive feedback loop. In particular, the H3K4 methyltransferase MLL can recruit the H3K27 demethylase UTX (Lee et al., 2007a), whereas the H3K27 methyltransferase Ezh2 can recruit the H3K4 demethylase Rpb2 (Pasini et al., 2008). Such feedback loops may be sufficient for establishing stable and heritable binary chromatin states (Dodd et al., 2007). Our analysis suggests that there is an extra layer of cooperativity, that is, the epigenetic factors are recruited to common targets by the DNA sequences, thus facilitating interaction efficiency.

A major computational challenge for detection of sequence signals associated with epigenetic factors is that they are likely to involve the accumulation of many weak features (Sekinger et al., 2005; Straub and Becker, 2008). Traditional methods for transcription factor binding motif discovery are not suitable for this task because they rely on the assumption that there exist distinct short sequence patterns. New approaches, such as those described in this paper, will undoubtedly provide new insights into the targeting mechanisms for epigenetic factors.

4. METHODS

4.1. Data source

Genome-wide nucleosome scores in resting $CD4^+$ T cells were downloaded from Schones et al. (2008). The nucleosome scores were calculated on a 10-bp sliding window as the number of sequenced reads mapping to the sense strand 80-bp upstream of the window and anti-sense strand 80-bp downstream of the window. The histone modification location data for $CD4^+$ T cells were downloaded from Barski et al. (2007) and Wang et al. (2008). These datasets contain the number sequenced reads within 200- or 400-bp summary windows. The DNA sequences were downloaded from Genome Browser based on based NCBI Build 36 (hg18). The genome-wide TSS annotations were based on Refseq genes, including a total of 26,007 genes. Gene expression data were downloaded from the GNF Atlas 2 (Su et al., 2004). The repeat elements were based on *RepeatMasker* (Smit, 1999). The Charlie and Mariner transposon locations were downloaded from the RepBase (*www.girinst.org/*). The bivalent genes in human ES cells were determined

by combining two genome-wide set of ChIP-chip data (Guenther et al., 2007; Lee et al., 2006). We called a gene to be bivalent if its promoter was bound by both H3K27me3 and H3K4me3. The HoxA cluster ChIP-chip data were obtained from Rinn et al. (2007). The HS sites in $CD4^+$ T cells were obtained from Crawford et al. (2006). The 124 high-confidence enhancers in Hela cells were downloaded from Heintzman et al. (2007). The histone modification data at repeated elements in mouse were downloaded from Martens et al. (2005).

4.2. Training sets

For nucleosome positions, we selected the 1% highest and lowest nucleosome scoring windows in Schones et al. (2008). For each window, a 131-bp DNA sequence centered at the midpoint of the window was extracted. These sequences were divided into a training set, containing those from chromosomes 1–12, and a testing set, containing the other sequences. For histone modifications, we selected the 1% non-empty bins with the highest tag counts and also 10,000 random locations in the genome. Again, we extracted a 131-bp DNA sequence from each target or random location and trained our model using the sequences from chromosomes 1–12.

4.3. A modified N-score model

We modified the original *N*-score model by incorporating wavelet features, word counts, and structural parameters examined in three different studies (Lee et al., 2007b; Peckham et al., 2007; Yuan and Liu, 2008). A model for each histone mark was trained independently. The wavelet features were defined as before (Yuan and Liu, 2008). Each training sequence was converted to 16 numerical series representing the frequencies for each of the 16 dinucleotides. These numerical series were then transformed to wavelet coefficients, which for a sequence S and a dinucleotide D were defined by

$$c_k^j(S, D) = \sum_i f_{S,D}(i/128)\psi_k^j(i/128)$$

where $f_{S,D}(i/128)$ represented the average frequency of D at the ith, $(i+1)$th, and $(i+2)$th positions, and ψ_k^j represented the wavelet functions at the jth level and kth position. We used the Haar basis, defined as

$$\psi(x) = \begin{cases} 1, & \text{for } 0 \le x < 1/2 \\ -1, & \text{for } 1/2 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

All the wavelet functions were related to ψ via rescaling and translation:

$$\psi_k^j(x) = 2^{j/2}\psi(2^j x - k), \qquad k = 0, \ldots, 2^j - 1, \quad j = 0, \ldots, 7.$$

The wavelet energy coefficients were defined as the total variance of the wavelets at each level and used as predictive variable in our model.

The word counts were derived as in Peckham et al. (2007). For each sequence, the frequencies of all overlapping k-mers, where k from 1 up to 6, were enumerated. Complementary words were counted together. In total there were 2772 such words, which were then ranked based on their discriminative power evaluated by t-tests. The most significant 200 words were kept. The 20 DNA structural features (e.g., twist, tilt) were calculated as in Lee et al. (2007b). Our final model could be written as follows:

$$q(S) \equiv \log\left(\frac{p(S)}{1-p(S)}\right) = \beta_0 + \sum_l \beta_{xl}x_l(S) + \sum_l \beta_{yl}y_l(S) + \sum_l \beta_{zl}z_l(S)$$

where $x_l(S)$, $y_l(S)$, $z_l(S)$ represent the wavelet energies, word counts, and structural parameters for the sequence S respectively. We used a stepwise procedure to select informative predictors as before (Yuan and Liu, 2008). We defined the *N*-score as the normalized version of $q(S)$, given by $N\text{-}score \equiv (q(S) - E(q(S)))/\sqrt{Var(q(S))}$.

4.4. Noise estimation

To estimate the noise level, the tag counts were first binned of width 200 bp. We used a Poisson distribution to model the noise background, assuming a sequence tag was randomly picked from the genome. A limitation of this simple noise model was that it ignored the fact that only uniquely mappable sequence tags were retrieved in a ChIP-seq experiment. A more realistic noise model needs to exclude unmappable regions. However, since such information was not publicly available, we only used a cruder model instead. Due to concern of multiple hypothesis testing error, we used a stringent cutoff at $p = 0.0001$. A rough estimate suggested that the corresponding false discovery rate was less than 15%. The bins containing more tags than the cutoff value (which is different for each histone mark) were called significant. The signal strength was quantified as the fraction of total tag counts that fell into the significant bins.

4.5. Correlation analysis for sequence features

To compare sequence features associated with different histone modification marks, we randomly selected a set of DNA sequences from the human genome and apply each model to evaluate the corresponding *N*-score. The similarity between a pair of models was quantified by the Pearson correlation of the associated *N*-score values. Alternatively, the models could be compared based on the difference in regression coefficients. However, this latter approach was not favored since the estimated parameter values were numerically unstable due to inter-correlation among the predicting variables. Mathematically speaking, our approach can be viewed as a weak-topology for linear operators in a Banach space (Naylor and Sell, 1982). In order to estimate statistical significance, we generated 20 null models by replacing the training marker sequences with random sequences independently generated from the control sequences. The null distribution of the Pearson correlation coefficients was estimated by correlations values among the null models. By using a Gaussian distribution to approximate the null distribution, we determined a correlation cutoff at 0.27, corresponding to a p-value of 0.05.

ACKNOWLEDGMENTS

We are grateful to Yujiang Shi and Rui Fang for helpful discussions. We also thank Dustin Schones for information about their ChIP-seq data. This research was supported by the Claudia Adams Barr program.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

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

Bernstein, B.E., Kamal, M., Lindblad-Toh, K., et al. 2005. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* 120, 169–181.

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.

Boyer, L.A., Plath, K., Zeitlinger, J., et al. 2006. Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* 441, 349–353.

Chen, X., Xu, H., Yuan, P., et al. 2008. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 133, 1106–1117.

Core, L.J., and Lis, J.T. 2008. Transcription regulation through promoter-proximal pausing of RNA polymerase II. *Science* 319, 1791–1792.

Crawford, G.E., Holt, I.E., Whittle, J., et al. 2006. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.* 16, 123–131.

Dodd, I.B., Micheelsen, M.A., Sneppen, K., et al. 2007. Theoretical analysis of epigenetic cell memory by nucleosome modification. *Cell* 129, 813–822.

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.

Hall, I.M., Shankaranarayana, G.D., Noma, K., et al. 2002. Establishment and maintenance of a heterochromatin domain. *Science* 297, 2232–2237.

Hastie, T., Tibishirani, R., and Friedman, J. 2001. *The Elements of Statistical Learning*. Springer, New York.

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.

Ioshikhes, I.P., Albert, I., Zanton, S.J., et al. 2006. Nucleosome positions predicted through comparative genomics. *Nat. Genet.* 38, 1210–1215.

Kornberg, R.D., and Stryer, L. 1988. Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. *Nucleic Acids Res.* 16, 6677–6690.

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

Kurdistani, S.K., and Grunstein, M. 2003. Histone acetylation and deacetylation in yeast. *Nat. Rev. Mol. Cell Biol.* 4, 276–284.

Lam, A.L., Boivin, C.D., Bonney, C.F., et al. 2006. Human centromeric chromatin is a dynamic chromosomal domain that can spread over noncentromeric DNA. *Proc. Natl. Acad. Sci. USA* 103, 4186–4191.

Lan, F., Nottke, A.C., and Shi, Y. 2008. Mechanisms involved in the regulation of histone lysine demethylases. *Curr. Opin. Cell Biol.* 20, 316–325.

Lee, M.G., Villa, R., Trojer, P., et al. 2007a. Demethylation of H3K27 regulates polycomb recruitment and H2A ubiquitination. *Science* 318, 447–450.

Lee, T.I., Jenner, R.G., Boyer, L.A., et al. 2006. Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* 125, 301–313.

Lee, W., Tillo, D., Bray, N., et al. 2007b. A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.* 39, 1235–1244.

Liu, C.L., Kaplan, T., Kim, M., et al. 2005. Single-nucleosome mapping of histone modifications in *S. cerevisiae*. *PLoS Biol.* 3, e328.

Martens, J.H., O'Sullivan, R.J., Braunschweig, U., et al. 2005. The profile of repeat-associated histone lysine methylation states in the mouse epigenome. *EMBO J.* 24, 800–812.

Mavrich, T.N., Ioshikhes, I.P., Venters, B.J., et al. 2008. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.* 18, 1073–1083.

Miele, V., Vaillant, C., d'Aubenton-Carafa, Y., et al. 2008. DNA physical properties determine nucleosome occupancy from yeast to fly. *Nucleic Acids Res.* 36, 3746–3756.

Naylor, A.W., and Sell, G.R. 1982. *Linear Operator Theory in Engineering and Science*. Springer, New York.

Pan, G., Tian, S., Nie, J., et al. 2007. Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. *Cell Stem Cell* 1, 299–312.

Pasini, D., Hansen, K.H., Christensen, J., et al. 2008. Coordinated regulation of transcriptional repression by the RBP2 H3K4 demethylase and polycomb-repressive complex 2. *Genes Dev.* 22, 1345–1355.

Peckham, H.E., Thurman, R.E., Fu, Y., et al. 2007. Nucleosome positioning signals in genomic DNA. *Genome Res.* 17, 1170–1177.

Pokholok, D.K., Harbison, C.T., Levine, S., et al. 2005. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* 122, 517–527.

Polikar, R. 2006. Ensemble-based systems in decision making. *IEEE Circuits Syst.* 24, 21–45.

Ringrose, L., Rehmsmeier, M., Dura, J.M., et al. 2003. Genome-wide prediction of polycomb/trithorax response elements in *Drosophila melanogaster*. *Dev. Cell* 5, 759–771.

Rinn, J.L., Kertesz, M., Wang, J.K., et al. 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* 129, 1311–1323.

Robert, F., Pokholok, D.K., Hannett, N.M., et al. 2004. Global position and recruitment of HATs and HDACs in the yeast genome. *Mol. Cell* 16, 199–209.

Roh, T.Y., Cuddapah, S., and Zhao, K. 2005. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev.* 19, 542–552.

Roh, T.Y., Wei, G., Farrell, C.M., et al. 2007. Genome-wide prediction of conserved and nonconserved enhancers by histone acetylation patterns. *Genome Res.* 17, 74–81.

Schones, D.E., Cui, K., Cuddapah. S., et al. 2008. Dynamic regulation of nucleosome positioning in the human genome. *Cell* 132, 887–898.

Schueler, M.G., and Sullivan, B.A. 2006. Structural and functional dynamics of human centromeric chromatin. *Annu. Rev. Genomics Hum. Genet.* 7, 301–313.

Schuettengruber, B., Chourrout, D., Vervoort, M., et al. 2007. Genome regulation by polycomb and trithorax proteins. *Cell* 128, 735–745.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., et al. 2006. A genomic code for nucleosome positioning. *Nature* 442, 772–778.

Sekinger, E.A., Moqtaderi, Z., and Struhl, K. 2005. Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. *Mol. Cell* 18, 735–748.

Shi, Y., Lan, F., Matson, C., et al. 2004. Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. *Cell* 119, 941–953.

Shivaswamy, S., Bhinge, A., Zhao, Y., et al. 2008. Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol.* 6, e65.

Slotkin, R.K., and Martienssen, R. 2007. Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.* 8, 272–285.

Smit, A.F. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* 9, 657–663.

Straub, T., and Becker, P.B. 2008. DNA sequence and the organization of chromosomal domains. *Curr. Opin. Genet. Dev.* 18, 175–180.

Su, A.I., Wiltshire, T., Batalov, S., et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA* 101, 6062–6067.

Talbert, P.B., and Henikoff, S. 2006. Spreading of silent chromatin: inaction at a distance. *Nat. Rev. Genet.* 7, 793–803.

Valouev, A., Ichikawa, J., Tonthat, T., et al. 2008. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.* 18, 1051–1063.

Wang, Z., Zang, C., Rosenfeld, J.A., et al. 2008. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.* 40, 897–903.

Xiao, T., Hall, H., Kizer, K.O., et al. 2003. Phosphorylation of RNA polymerase II CTD regulates H3 methylation in yeast. *Genes Dev.* 17, 654–663.

Yuan, G.C., and Liu, J.S. 2008. Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Comput. Biol.* 4, e13.

Yuan, G.C., Ma, P., Zhong, W., et al. 2006. Statistical assessment of the global regulatory role of histone acetylation in Saccharomyces cerevisiae. *Genome Biol.* 7, R70.

Zhao, X.D., Han, X., Chew, J.L., et al. 2007. Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. *Cell Stem Cell* 1, 286–298.

Address reprint requests to:
Dr. Guo-Cheng Yuan
Department of Biostatistics
Harvard School of Public Health
Cambridge, MA 02466

E-mail: gcyuan@jimmy.harvard.edu