Published online 25 January 2011 *Nucleic Acids Research, 2011, Vol. 39, No. 10* ***4063–4075***
doi:10.1093/nar/gkr016

Combinatorial chromatin modification patterns in the human genome revealed by subspace clustering

Duygu Ucar[1], Qingyang Hu[2] and Kai Tan[1,3,*]

[1]Department of Internal Medicine, [2]Department of Computer Science and [3]Department of Biomedical Engineering, University of Iowa, Iowa City, 52242 Iowa, USA

Received September 13, 2010; Revised January 4, 2011; Accepted January 5, 2011

ABSTRACT

Chromatin modifications, such as post-translational modification of histone proteins and incorporation of histone variants, play an important role in regulating gene expression. Joint analyses of multiple histone modification maps are starting to reveal combinatorial patterns of modifications that are associated with functional DNA elements, providing support to the 'histone code' hypothesis. However, due to the lack of analytical methods, only a small number of chromatin modification patterns have been discovered so far. Here, we introduce a scalable subspace clustering algorithm, coherent and shifted bicluster identification (CoSBI), to exhaustively identify the set of combinatorial modification patterns across a given epigenome. Performance comparisons demonstrate that CoSBI can generate biclusters with higher intra-cluster coherency and biological relevance. We apply our algorithm to a compendium of 39 genome-wide chromatin modification maps in human $CD4^+$ T cells. We identify 843 combinatorial patterns that recur at $>0.1\%$ of the genome. A total of 19 chromatin modifications are observed in the combinatorial patterns, 10 of which occur in more than half of the patterns. We also identify combinatorial modification signatures for eight classes of functional DNA elements. Application of CoSBI to epigenome maps of different cells and developmental stages will aid in understanding how chromatin structure helps regulate gene expression.

INTRODUCTION

Histone proteins in chromatin are subject to a number of post-translational modifications (PTMs), primarily at their N-terminal tails, including methylation, acetylation, phosphorylation, ubiquitylation and ADP-ribosylation (1). It has been proposed that distinct histone modifications, on one or more nucleosomes, act in combination to form a 'histone code' that is read by other proteins to bring about distinct downstream events (histone code hypothesis) (2). The advent of chromatin immuno precipitation coupled with microarray chip (ChIP-chip)—and recently, ultra high-throughput sequencing (ChIP-seq)—has enabled global and whole-genome histone modification profiling studies. To date, several dozens of histone modifications across multiple human cell types and disease states have been mapped, generating a diversity of epigenomic data sets (1,3,4). A picture is now emerging in which distinct genomic regions such as enhancers, promoters, insulators, gene bodies (both protein coding and non-coding RNA genes), and sub-chromosomal regions have distinct chromatin modification patterns/signatures. For example, high levels of histone 3 lysine 4 (H3K4) methylation and histone 3 and 4 acetylations have been found at gene promoters and enhancers (3–5). Collectively, these observations provide strong support to the histone code hypothesis and suggest that epigenetic signatures could be an effective way to pinpoint functional DNA elements in the genome. However, we are far from deciphering the histone code. From a computational point of view, the current challenge is to develop analytic tools to extract novel and consistent combinatorial patterns and integrate them with various functional genomic data sets.

To date, several computational methods have been developed to identify histone modification patterns from ChIP-Chip/Seq data sets. From a computational perspective, they fall into two categories. The first category uses supervised statistical learning techniques for identifying distinct and predictive histone modification patterns at known classes of functional sites, such as promoters and enhancers (6). Although supervised methods have revealed distinct chromatin signatures, they could not identify novel patterns that are associated with either poorly studied or new classes of functional DNA elements.

For the second category of approaches, Hon *et al.* (7) proposed an unsupervised method, termed ChromaSig, to identify histone modification 'motifs' that are repeated

*To whom correspondence should be addressed. Tel: +1 319 384 4676; Fax: +1 319 384 4785; Email: kai-tan@uiowa.edu

across the human genome. The proposed algorithm uses a progressive alignment approach to identify motifs starting from a seed motif. Although it captures interesting patterns, ChromaSig does not exhaustively search for all combinations of repeated patterns across the genome. Jaschek and Tanay proposed a spatial clustering algorithm that employs hidden Markov model (HMM) to identify a set of common patterns defined over contiguous genomic regions (8). Their probabilistic model describes a set of clusters (i.e. HMM states) with transition probabilities between these states. Their algorithm assumes that consecutive regions in the genome tend to share a functional annotation, which might not necessarily be true. In a more recent work, Ernst and Kellis (9) proposed an alternative HMM algorithm based on the binarization of presence or absence of each histone mark. This approach significantly reduces the number of parameters compared to the spatial clustering algorithm (8). But it still requires a set of non-intuitive parameters to be set. More importantly, for both ChromaSig- and HMM-based algorithms, the final motifs/patterns are forced to include all histone modification marks in the input data. However, multiple studies so far have demonstrated that many combinatorial patterns only involve a few chromatin modifications (10). Therefore, a more powerful approach is to identify sets of genomic loci that are specifically associated with subsets of chromatin modifications.

To address the challenges described above, we propose a new computational algorithm to comprehensively search for combinatorial histone modification patterns across a given epigenome. Benchmarking experiments demonstrate that our algorithm outperforms an existing greedy algorithm in terms of the coherency and biological relevance of inferred biclusters. Application of our algorithm to a compendium of chromatin modification maps in human T cells revealed 843 combinatorial patterns across the genome. We provide supporting evidence for the discovered patterns based on three lines of evidence: combinatorial histone modifications identified using mass spectrometry, location biases of predicted biclusters with respect to known functional DNA elements, and correlations with gene expression data in T cells. The analysis presented here provides a systematic characterization of combinatorial chromatin modifications in a mammalian cell.

MATERIALS AND METHODS

Chromatin modification maps

Genome-wide maps of 18 histone acetylations (3), 20 methylations (11) and a histone variant H2A.Z (3) of human $CD4^+$ T cells have been generated using ChIP-seq (see Supplementary Data for the list of modifications). In this study, for each chromatin modification map, we used the summary tag counts obtained at every 200 bp as our raw data for the pre-processing step described below.

Chromatin modification ChIP-Seq data pre-processing

The genome is split into consecutive non-overlapping windows that we refer to as *genomic loci* throughout the article. Since chromatin modification signals tend to be diffusive, in order to capture the entire signal, we used a window of size 5000 bp. Using the MACS software (12), we then identified signal peaks and mapped these peaks to genomic loci. Peak detection step is used to eliminate genomic loci with no signal for all of the chromatin modifications. Using this strategy, we identified 130 559 genomic loci.

Construction of the GCP matrix

The input to our algorithm is a three-dimensional (3D) matrix of preprocessed chromatin modification data. The three dimensions are genomic locus, chromatin modification, position within a signal peak. Therefore, the matrix is abbreviated as a GCP matrix. We construct such a matrix by using the summary tag count at every 200 bp within each 5000 bp genomic locus. For the genome-wide study, dimensions of the matrix are 39 (number of chromatin modifications) × 25 (number of signals per genomic locus) × 130 559 (number of genomic loci with at least one modification peak).

Computational model

We propose an algorithm to exhaustively search for combinatorial chromatin modification patterns that frequently recur in an epigenome and exhibit similar signals. Before going into details of the algorithm, we start by defining notations and the concept of *coherent bi-cluster*.

Let G be a set of genomic loci each of which has a length of 5kbp, $G^{\mathrm{U}} = \{g_1,\ldots,\mathrm{g}_n\}$, let C be a set of chromatin modifications, $C^{\mathrm{U}} = \{c_1,\ldots,c_m\}$, and let P^{U} be a consecutive set of tag counts from the ChIP-seq experiment that covers the 5 kbp window, $P = \{p_1,\ldots,p_{\mathrm{t}}\}$. A 3D GCP matrix represents a real valued *nxmxt* matrix $\mathbf{GCP} = \mathrm{G}^{\mathrm{U}} \times \mathrm{C}^{\mathrm{U}} \times \mathrm{P}^{\mathrm{U}} = \{\mathrm{d}_{ijk} |\ i \in [1,n],\ j \in [1,m],\ k \in [1,t]\}$, whose dimensions correspond to genomic locus, chromatin modification and signal position accordingly. An entry in this matrix, d_{ijk}, refers to the tag count at position p_k of the genomic locus g_i for chromatin modification h_j. ChIP-seq signal at a genomic locus i for modification j, which is of length t, is referred as GCP[i,j,*] or S_{ij*} for short, throughout the article.

A *coherent bi-cluster* $\mathbf{B_{[GXC]}}$ is a sub-matrix of **GCP**, i.e. $\mathrm{B} = \mathrm{G} \times \mathrm{C}$ and $\mathrm{G} \subseteq \mathrm{G}^{\mathrm{U}}$ and $\mathrm{C} \subseteq \mathrm{C}^{\mathrm{U}}$ provided that the following two coherency conditions are satisfied:

(i) every pair of chromatin marks in C, C_x and C_y, satisfies $\rho(S_{kx*}, S_{ky*}) > \alpha$ for every locus in G, g_k, where ρ is a measure of correlation between two signal vectors, i.e. S_{kx*} and S_{ky*}, and α is the minimum coherency threshold across the dimension C;

(ii) every pair of genomic loci in G, g_k and g_l, satisfies $\rho(S_{kx*}, S_{lx*}) > \beta$ for every modification in C, say C_x, where ρ is a measure of correlation between two vectors, S_{kx*} and S_{lx*}, and β is the minimum coherency threshold across the dimension G.

We use cross correlation as the correlation measure ρ. It is a standard metric for measuring the similarity between two signals when a time delay is applied to one of the signals (13). This measure enables us to capture correlated patterns that are shifted from each other.

A particular chromatin modification signal at a 5 kb window can be represented as a vector of t consecutive points, which we denote as S_{ij*}. The cross-correlation between two such vectors S_{ij*} and S_{ik*} with delay d can be calculated using the following formula:

$$CC_d(S_{ij*},S_{ik*}) = \frac{\sum S_{ijx} \times S_{ik(x-d)} - \sum S_{ijx} \times \frac{\sum S_{ik(x-d)}}{t}}{\sqrt{\left(\sum S_{ijx}^2 - \frac{\left(\sum S_{ijx}\right)^2}{t}\right) \times \left(\sum S_{ik(x-d)}^2 - \frac{\left(\sum S_{ik(x-d)}\right)^2}{t}\right)}}.$$

To find the best match between these two signals, we choose the delay d that maximizes the correlation between the two signals as follows:

$$\rho(S_{ij*},S_{ik*}) = \max\left(CC_d(S_{ij*},S_{ik*})\right), \text{ where } -t \le d \le t.$$

Algorithm for identifying maximal coherent bi-clusters

We propose an algorithm to identify maximal coherent biclusters in two steps. In the first step, for each genomic locus g_k, we identify the maximal set of chromatin modifications that exhibit a coherent signal at the locus. Each set is named as a *maximal sample set*. To do so, we first construct a binary coherency matrix **CM** of size $|C|\times|C|$. An entry in this matrix is set to 1, if the corresponding pair of chromatin modifications are coherent at g_k, more formally CM[i,j] $=1$ if $\rho(S_{ki*}, S_{kj*}) > \alpha$. Once we construct the **CM** matrix, the problem of finding maximal sample sets is transformed into the problem of enumerating all maximal cliques of size at least $\min_s$ in the graph induced by the **CM** (termed coherency graph). A maximal clique is a subgraph that is fully connected and is not contained in any other such subgraphs. The clique enumeration problem is known to be NP-hard. Fortunately, the maximum size of any **CM** matrix cannot exceed the number of chromatin modifications (39 in our case) since there exist s single node per modification in this graph. Therefore, with efficient search and pruning techniques, we can identify maximal cliques in a scalable manner. For this purpose, we employ a recursive, depth-first search (DFS) of the set enumeration tree of the chromatin modifications. Set enumeration tree provides an efficient and systematic way to enumerate the complete set of combinations (14). For our analysis, we employed the two pruning strategies that are also employed in Jiang *et al.* (15) study. The first strategy lets us to eliminate paths from the tree that will never lead to large enough sample sets. The second strategy enables us to eliminate paths that are subsumed by a maximal subset sample. Using efficient search and pruning strategies on the set enumeration tree, we effectively identify 'maximal sample sets' for every genomic locus at the first step of CoSBI.

In the second step, we identified the maximal **G**×**C** sets that satisfy conditions (i) and (ii) in the 'coherent bicluster' definition. A naïve method would test every possible G and C combination, which is infeasible since there exist >130 K genomic loci in our data. In order to scale this problem, we again employ set enumeration tree of sample sets, which systematically enumerates combinations of histone modifications and prune unpromising combinations before hand. Similar to our previous search, we employ two search strategies that significantly reduces the running time of the algorithm: eliminating unpromising sets and eliminating subsumed sets (16). Every **G** × **C** set computed with this search is a potential maximal coherent bicluster unless a superset of this set already satisfies the coherency conditions. To identify maximal coherent biclusters, we keep track of all **G** × **C** sets that satisfy the two coherency properties. Since we conduct a depth-first search in the set enumeration tree, any coherent superset of **G** × **C** should be reported before its subsets. Therefore, by reporting only **G** × **C** sets that satisfy coherency properties and are not subsumed by any maximal coherent biclusters identified before, we can obtain the final set of maximal coherent biclusters. Since the set enumeration tree systematically enumerates the complete set of combinations, in theory we are guaranteed to find the complete set of coherent biclusters in the data.

Intra-cluster similarity

To quantify the overall quality of a bi-cluster $\mathbf{B_{[G\times C]}}$, that includes $|G|$ genomic loci and $|C|$ chromatin marks we defined the following intra-cluster similarity measure:

$$IS(B_{[G\times C]}) = \frac{\sum_{j\in G}\sum_{i\in G}\rho(\bar{S}_i,\bar{S}_j)}{1/2 \times |G| \times |G-1|},$$

where $\bar{S}_i$ represents the mean signal for genomic locus i over all chromatin modifications in the given bicluster $\mathbf{B_{[G\times C]}}$. G represents the set of genomic loci in the bicluster, and accordingly, $|G|$ represents the number of genomic loci in the cluster. Intra-cluster similarity of a bicluster is the average of all pairwise similarities of mean genomic loci signals. More coherent bi-clusters across both dimensions **C** and **G** score higher with respect to the intra-cluster similarity measure.

Combinatorial histone modifications supported by tandem mass spectrometry data

Combinatorial histone modifications observed in a single histone tail were compiled from references (17–20). For each of the predicted biclusters, we computed the fractions of member histone marks that are also observed in the set of combinatorial histone codes defined by mass spectrometry. We then chose the largest fraction as the fraction of bicluster histone marks supported by a histone code uncovered using mass spectrometry experiments.

Enrichment analysis

To associate identified biclusters with functional DNA elements, we conducted a comprehensive examination of the biclusters for enrichment of the following eight classes of DNA elements (i.e. genomic features): highly conserved sequences (from phastCons analysis across 17 vertebrate genomes obtained from the UCSC Genome Browser), protein-coding genes (from the RefSeq database, version hg18), large intergenic noncoding RNAs (lincRNA) (21), transcription start sites (TSS) of RefSeq genes, CpG islands, DNaseI hypersensitivity sites (DHSs), CTCF binding sites (i.e. insulator) from Cuddapah *et al.* (22), p300 binding sites from Wang *et al.* (23). For conserved sequences, we used genomic regions with a phastCons conservation score of at least 0.5.

To determine the overlap between a genomic feature described above and a predicted bicluster, we calculated the center-to-center distances between a genomic locus in a bicluster and instances of the genomic feature. If this distance is smaller than half of the window size (i.e. |genomic feature center–genomic locus center| $\leq$2500 bp), we regard the genomic locus as an instance of the genomic feature. The only exception to this center-to-center distance rule is TSS overlap. In this case, if a TSS is $<$1000 bp away from the center of a genomic locus, we regarded them as overlapping. This is to ensure a bicluster locus has a large overlap with a promoter region as well as the TSS. Any locus that is not labeled as a TSS but overlaps with the open reading frame of a gene over at least half of the locus length (2500 bp) is regarded as a gene body instance.

Enhancers in human CD4$^+$ T cell

For performance comparison between CoSBI and EDISA, we curated a set of high-confidence enhancers. We first identified distal p300 binding sites in CD4$^+$ T cell mapped by ChIP-seq in a previous study (23). Here, distal is defined as at least 2.5 kbp away from the closest known TSS. We further filtered the set of distal p300 binding sites to include those that overlap at least with one computationally predicted enhancer in the PreMod database (24). This procedure generated 213 high-confidence enhancers. For background genomic sites, we randomly selected loci containing chromatin modification peaks but having no overlap with PreMod predictions. We used the same number of random sites as the number of enhancers.

Promoter regions

To identify chromatin modification signatures of promoter regions, we used RefSeq genes (version hg18) downloaded from the UCSC Genome Browser. After eliminating genes that are unmapped or have alternative TSS, we ended up with 21 123 genes. For promoters, we used 5000 bp regions that span 3500 bp upstream and 1500 bp downstream of a TSS.

Promoter bicluster and gene expression correlation analysis

To identify biclusters that are associated with highly expressed and silent genes, we first sorted all biclusters by their median gene expression levels and then by the standard deviation of the gene expression levels at the bicluster level. We then identified the top and bottom 10 biclusters from the sorted list. This analysis guaranteed to identify biclusters whose genes were highly expressed or silent and have a narrow range of expression levels. We used the gene expression data set generated for CD4$^+$ T cells (25).

RESULTS

The CoSBI algorithm for identifying combinatorial chromatin modification patterns

We propose an unsupervised subspace-clustering algorithm for the analysis of chromatin modification data generated using ChIP-chip/seq technology. The algorithm, termed coherent and shifted bi-cluster identification (CoSBI), aims to identify all recurrent combinatorial patterns with coherent signals over the same set of genomic loci and chromatin modifications. Since the patterns we seek are coherent in two dimensions, our problem is similar to finding biclusters in a compendium of gene expression microarray data, which has been extensively studied (26,27). However, the major difference from previous biclustering algorithms is that individual data entries in the 2D data matrix are not scalar values. Instead, they are vectors representing consecutive measurements of a given chromatin modification across a genomic locus (Figure 1). This third dimension of the data presents additional challenges for the design of the clustering algorithm.

Our algorithm is briefly summarized here and in Figure 1. It first represents a set of chromatin modification ChIP-chip/seq data in the form of a matrix with the following three dimensions: genomic locus, chromatin modification and ChIP-chip/seq signal Position, namely a **GCP** matrix. The algorithm employs techniques from Frequent Itemset Mining and identifies coherent biclusters in two steps. In the first step, for every genomic locus, it identifies maximal subsets of chromatin modifications that exhibit coherent signals among them. Each such subset is termed 'maximal sample set' for the corresponding genomic locus. To effectively identify all 'maximal sample sets', we first construct a coherency graph for every locus. A coherency graph summarizes all pair-wise similarities between different chromatin modifications at the same genomic locus (i.e. one coherency graph for each genomic locus). For a given genomic locus, if two chromatin modification signals are similar enough there exists a link between the corresponding nodes in the coherency graph. Next, by finding maximal cliques in a coherency graph, we obtain the complete list of 'maximal sample sets' for every genomic locus in the input data. In the second step, the algorithm identifies coherent patterns across both genomic locus and chromatin mark dimensions,

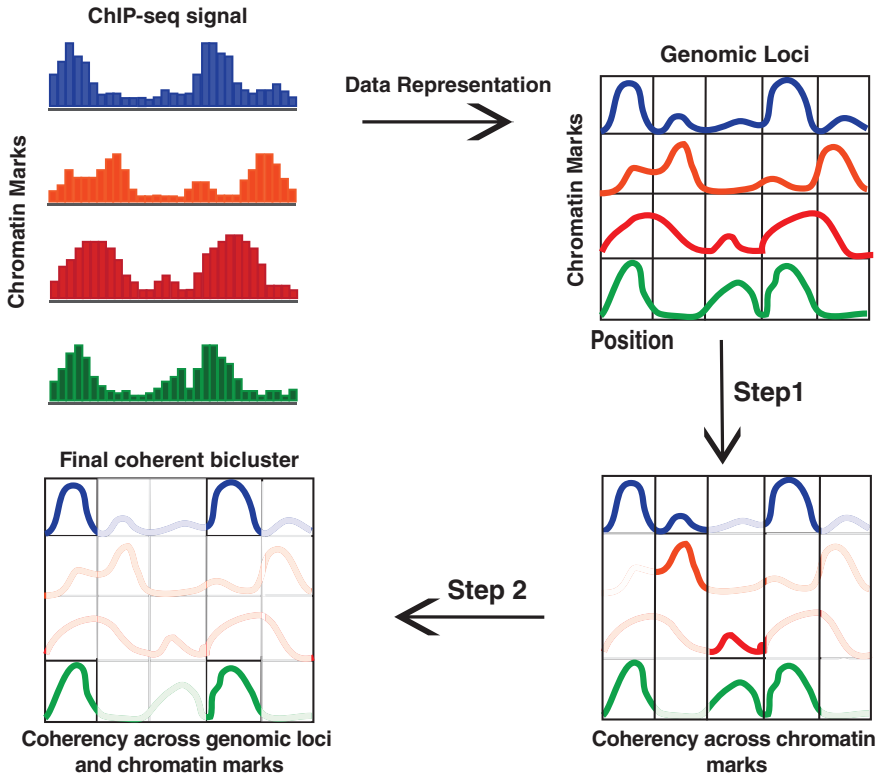

Figure 1. Overview of the CoSBI algorithm. In Step 1, for each genomic locus, we identify sets of maximal coherent chromatin marks. In Step 2, using the results of Step 1, we identify sets of biclusters that are coherent in two dimensions.

generating coherent biclusters. It does so by systematically enumerating all promising chromatin modification combinations using a set enumeration tree. To speed up the enumeration, we use an inverted list (16) of the 'maximal sample sets' computed in Step 1. The final output of our algorithm is a complete collection of biclusters across the given data, each of which contains a set of chromatin modifications that exhibit coherent signals across all genomic loci in the bicluster. The algorithm has the following four parameters: min_g and min_s which specify the minimal numbers of genomic loci and chromatin marks in the identified bicluster, respectively; α which is the minimum coherency threshold for two chromatin modification signals at the same genomic locus; β which is the minimum coherency threshold for the same chromatin modification at two genomic loci. CoSBI is implemented in C language and is freely available at: http://www.medicine.uiowa.edu/Labs/tan/CoSBI.

Performance evaluation of CoSBI using known enhancers

Since CoSBI is designed for subspace clustering of 3D data, we compared its performance with an existing algorithm that is designed for a similar task. Supper *et al.* (28) proposed an iterative greedy algorithm, EDISA, for clustering 3D gene-condition-time microarray data to identify gene sets with coherent expression patterns across a subset of conditions and time points. In this case, time points are analogous to ChIP-chip/seq signal positions in our problem. The coherent biclusters sought by EDISA are analogous to the coherent biclusters that are sought by CoSBI. However, there are two major algorithmic differences between EDISA and CoSBI. First, EDISA is a greedy algorithm, which implies that it does not exhaustively search for all coherent biclusters in the data. CoSBI, in contrast, captures the complete set of maximal coherent biclusters in the data satisfying specified coherency thresholds. Second, instead of using Pearson correlation as a measure of coherency, we use cross correlation to capture correlation between signals that are shifted with respect to each other.

We compared the performance of EDISA and CoSBI in terms of their ability to identify coherent biclusters from a given 3D **GCP** matrix consisting of functional DNA elements and random genomic loci. Based on the histone code hypothesis, we expect genomic loci that share functionality to have a common chromatin modification signature. On the contrary, random loci would lack a consistent signature. Based on this assumption, an effective algorithm should be able to group together genomic loci of the same class along with their signature chromatin modifications. For functional DNA elements, we prepared a set of 213 enhancers in human T cells ('Materials and Methods' section). The chromatin modification data associated with these enhancers and an equal number of random sequences are used as the input to both algorithms.

We ran EDISA using its suggested automatic parameter setting option. Since EDISA is not a deterministic algorithm, we ran EDISA 10 times and used the result of the best run as its final output. We ran CoSBI to identify

biclusters that are repeated across at least 10% of all input sequences and that involve at least three chromatin modification marks. We set the coherency parameters α and β to 0.75 and 0.65, respectively. We chose this parameter setting since it produced almost the same number of biclusters as EDISA. Additional comparisons of the two algorithms using different parameter settings can be found in Supplementary Data.

Using the parameters described above, EDISA produced 234 biclusters that on average include 10 genomic loci and CoSBI produced 249 biclusters that on average include 52 genomic loci. To assess the quality of the identified biclusters, we calculated their intra-cluster similarities. As can be seen from Figure 2A, the average intra-cluster similarities were 0.69 and 0.83 for EDISA and CoSBI biclusters, respectively. Lower intra-cluster similarity of EDISA biclusters suggests that the greedy algorithm had difficulty grouping genomic loci with a coherent modification signal into the same bicluster. In addition, the small size of EDISA biclusters indicates that they do not involve complete coherent patterns.

We also evaluated the quality of the resulting biclusters by exploring the functional 'purity' of the identified biclusters, i.e. the fraction of enhancers present in each bicluster. To do so, we calculated a hypergeometric *P*-value of enhancer enrichment for each bicluster. As can be seen in Figure 2B, CoSBI biclusters have a higher enrichment of enhancers compared to EDISA biclusters.

Taken together, comparison with EDISA demonstrates that CoSBI can infer biclusters with higher intra-cluster similarity and functional coherency. Furthermore, our experiments confirmed the general observation that functional DNA elements tend to have distinct and coherent chromatin modification patterns. This observation suggests that CoSBI biclusters can be effectively used to infer novel epigenetic signatures and associated functional DNA elements.

Genome-wide prediction of combinatorial chromatin modifications in human CD4$^+$ T cell

To investigate combinatorial chromatin modifications in the human genome, we applied CoSBI to a set of 39 genome-wide chromatin modification maps in human CD4$^+$ T cells. We set the $\min_g$ and $\min_s$ parameters to 0.1 and 3%, respectively, which allow us to capture patterns that are recurrent across at least 0.1% of the human genome and involve at least three chromatin marks. We set the coherency thresholds α and β to 0.75 and 0.6, respectively, to achieve a reasonable balance between coverage and quality of inferred biclusters (see Supplementary Data for details). With this parameter setting, CoSBI identified 843 biclusters in the CD4$^+$ T cell epigenome. Additional information about the identified biclusters can be found in Supplementary Table S1.

The biclusters predicted by CoSBI are based on histone modification generated using ChIP-Seq technology. An alternative experimental protocol for identifying combinatorial histone modifications is tandem mass spectrometry (MS) (29). Compared to ChIP-based method, the major advantage of MS is that it can detect all modifications associated with a given histone tail simultaneously. As a first means to corroborate our predicted biclusters, we have manually compiled a list of 366 unique combinatorial histone modifications, each of which is observed in a single histone tail using MS (Supplementary Table S3). As shown in Figure 3, for 50% of the predicted biclusters, 40% of their member histone marks are also observed in a single histone tail in the mass spectrometry experiments. Note that this curated list of histone codes only involves histones H3 and H4. Since our biclusters also involve histone H2, the fraction of supported bicluster members will be even higher when MS data on H2 become available in the future.

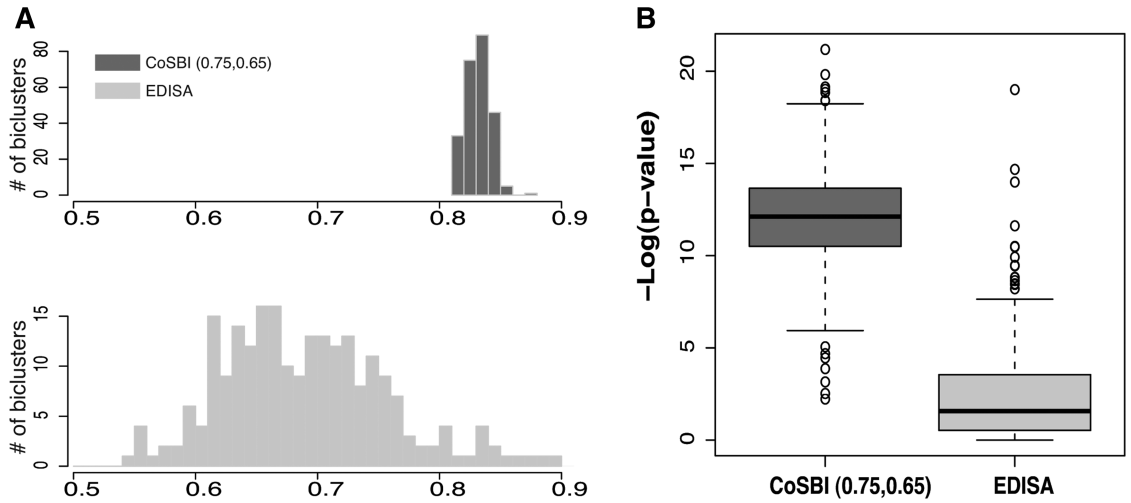

Figure 2. Performance comparison of EDISA and CoSBI. (**A**) Intra-cluster similarity distributions of EDISA and CoSBI biclusters. (**B**) Hyper-geometric *P*-value distributions for enhancer enrichment of EDISA and CoSBI biclusters.

Global features of combinatorial chromatin modifications in CD4$^+$ T cell epigenome

Out of the 39 chromatin modifications we examined, only 19 are present in the identified biclusters. Their occurrence frequencies are depicted in Figure 4A. This set of chromatin marks includes all 17 backbone modifications that were identified by Wang *et al.* (3) (H2A.Z, H2BK5ac, H2BK12ac, H2BK20ac, H2BK120ac, H3K4ac, H3K9ac, H3K18ac, H3K27ac, H3K36ac, H4K5ac, H4K8ac, H4K91ac, H3K4me1, H3K4me2, H3K4me3 and H3K9me1) and two additional acetylations (H4K16ac and H4K12ac). Based on our frequency analysis (Figure 4A), we found that the following 10 chromatin modifications (H2BK5ac, H2BK120ac, H3K4ac, H3K9ac, H3K18ac, H3K27ac, H3K36ac, H4K5ac, H4K8ac and H3K4me3) along with the histone invariant H2AZ are very prone to participate in combinatorial patterns. Each of them occurred in more than half of the biclusters. This high tendency to participate in combinatorial patterns is not due to higher sequencing coverage of these marks. As can be seen in Figure 4A, the occurrence frequencies in CoSBI biclusters and overall occurrence frequencies in the genome are not correlated (Pearson's correlation coefficient = 0.13). Overall occurrence frequency for a histone mark is computed as the ratio of ChIP-seq peaks identified by MACS and the total number of 5 kbp non-overlapping windows in the genome.

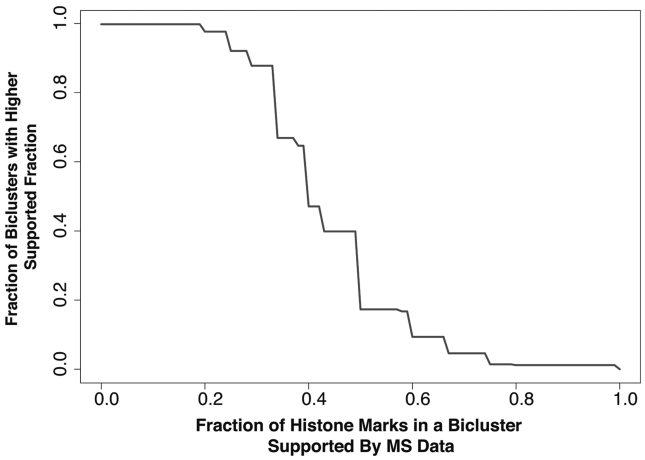

Figure 3. Predicted biclusters supported by mass spectrometry data. Shown is the cumulative distribution of predicted biclusters whose histone modification members are supported by mass spectrometry data.

There were 18 acetylation and 20 methylation marks in our input data. However, only 4 of 19 combinatorial marks in the biclusters were methylations. We analysed the genomic deposition patterns of the 39 chromatin marks in our input data. Many acetylations show a clear deposition bias towards 5′ of TSS. In comparison, methylations tend to display a wider deposition distribution (Supplementary Figure S5). This broader deposition pattern of methylation could explain why we only observe four methylation marks in our biclusters.

Next, we analysed the identified biclusters in order to reveal the set of most frequently co-occurring chromatin marks. For this purpose, we constructed a co-occurrence matrix involving all 19 chromatin marks observed in the biclusters. Values in this matrix are the co-occurrence frequency between a pair of histone marks. Co-occurrence frequency is computed as the ratio of the number of biclusters in which two histone marks appear together to the total number of biclusters in which at least one of the marks appears. Using hierarchical clustering, we clustered the co-occurrence matrix to identify groups of chromatin modifications that frequently co-occur. As can be seen in

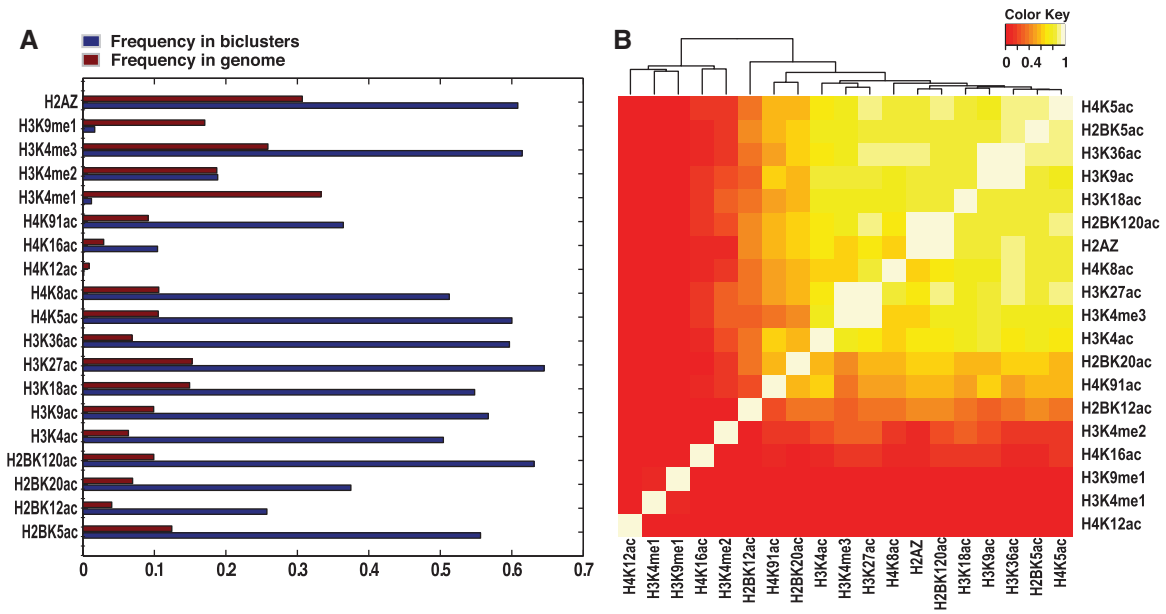

Figure 4. Occurrence and co-occurrence frequencies of chromatin modification marks of CoSBI biclusters. (**A**) Occurrence frequency of chromatin modifications observed in biclusters and across the genome. (**B**) Hierarchical clustering of co-occurrence matrix for all chromatin modification pairs observed in biclusters.

Figure 4B, three pairs of chromatin marks almost always co-occur in biclusters, including <H3K27ac, H3K4me3>, <H2AZ, H2BK120ac> and <H3K9ac, H3K36ac>. Among these, the two marks, H3K27ac and H3K4me3, have been observed to frequently co-occur by previous studies (5,30,31).

Chromatin modification signatures associated with functional DNA elements

Multiple studies have demonstrated that distinct functional DNA elements (or genomic feature) exhibit characteristic chromatin modification patterns that often involve multiple modifications (3,7). To systematically investigate this phenomenon, we identified biclusters that are enriched for a given type of genomic feature ('Materials and Methods' section). For this purpose, we examined the following eight classes of genomic features: CpG islands, conserved sequences in vertebrates, DNaseI hypersensitivity sites (DHS), enhancers (i.e. distal p300 binding sites), insulators (i.e. CTCF binding sites), large intergenic noncoding RNAs (lincRNAs), promoters (i.e. regions around TSS) and protein-coding genes.

In total, we examined 130 559 genomic loci belonging to the eight classes of genomic features (Table 1). For each identified bicluster, we determined the number of feature loci that overlap with the bicluster. We then calculated a feature enrichment *P*-value for each bicluster using the hypergeometric distribution. Our analysis revealed that 721 of the 843 biclusters (85.5%) were enriched for at least one of these features. Additional information about the full set of biclusters including their chromatin marks, genomic locations, intra-cluster similarity values and feature enrichment *P*-values can be found in the Supplementary Tables S1 and S2.

Next, we focused our analysis on four classes of genomic features that are most abundant in the genome: CpG islands, distal enhancers, insulators and promoters. Figure 5 depicts the co-occurrence maps for the chromatin marks associated with these genomic features. Co-occurrence maps for the remaining four classes of genomic features, i.e. conserved sequences, DHSs, lincRNAs and protein-coding genes, can be found in Supplementary Figures S6–S9.

Table 1. Overlap between identified biclusters and genomic features

Genomic feature	Total mapped loci/feature	No. of enriched biclusters	No. of loci in biclusters	No. of chromatin modifications in biclusters
CpG island	18 249	4	319	12
Conserved sequence	119 442	3	198	11
Insulator	16 902	9	240	10
DHS	44 159	8	332	12
Enhancer	8256	9	236	10
LincRNA	707	5	145	10
Promoter	8737	551	950	17
Protein-coding gene body	53 175	361	1029	18

Overlap *P*-value between a bicluster and a genomic feature is computed using the hypergeometric distribution. A $P < 0.05$ is considered to be enriched. Definitions of features are as following: insulators, CTCF binding sites; enhancers, distal p300 binding sites; promoters, regions that span 3.5 kb upstream and 1.5 kb downstream of a transcription start site. DHS, DNaseI hypersensitivity site; LincRNA, long intergenic non-coding RNA.

Among these four classes, gene promoters have the most complicated combinatorial chromatin modifications, both in terms of the total number of combinatorial patterns and in terms of the total number of chromatin modifications involved. On the contrary, although there are many more annotated CpG islands than promoters in the genome (18 249 versus 8737, Table 1), combinatorial chromatin modification patterns seem to be less prevalent for CpG islands compared to promoters. We would like to point out that many acetylation marks exhibit a biased 5′-deposition patterns in the data set used (Supplementary Figure S5), the observed larger complexity of combinatorial patterns in promoter regions could be due to this bias.

Using dendrograms from the hierarchical clustering of the co-occurrence maps, we identified a set of core modifications for each class of genomic feature. From all subtrees in a dendrogram, we retained only those member whose co-occurrence frequencies are all above 0.5, yielding the strongly co-occurring subsets in the co-occurrence matrices, which we refer to as modification cores. The set of modification cores for each genomic feature are shown in Table 2. In the rest of this section, we discuss supporting evidence for these core chromatin marks from published literature.

We found three modification cores for insulators. Two of them involved only acetylations. A previous study (32) suggested a model that high levels of histone acetylations are maintained by insulators *per se* and over the protected regions that they surround. This model ensures that promoters within insulator-delimited region will be physically more accessible to TF binding. The third core signature involves two methylations (H3K4me2 and H3K4me3). In a recent computational analysis that focused on histone methylation, Jaschek and Tanay (8) identified two clusters that are enriched in CTCF binding and high levels of H3K4me1, H3K4me2 and H3K4me3. Taken together, this study and studies by other groups demonstrated that acetylation is a significant modification mark at insulators and at a subset of insulators both acetylation and H3K4 methylation co-exist.

Several chromatin modifications have been shown to be associated with functional enhancers by various studies. Heintzman *et al.* (5) used ChIP-ChIP to map several histone modifications across the HeLa cell genome. Their analysis revealed that H3K4me1 and H3K27ac are significant marks for enhancers. Roh *et al.* (33) also showed that H3K4me2 and H3 acetylations are enriched at enhancers conserved between human and pufferfish as well as enhancers conserved between human and mouse. Wang *et al.* (3) analysed the same set of chromatin modification data as used in this study. They found the following six modifications that were enriched at >20% of enhancers: H2A.Z, H3K4me1, H3K4me2, H3K4me3, H3K9me1 and H3K18ac. In addition to these experimental studies, a computational analysis of the HeLa cell

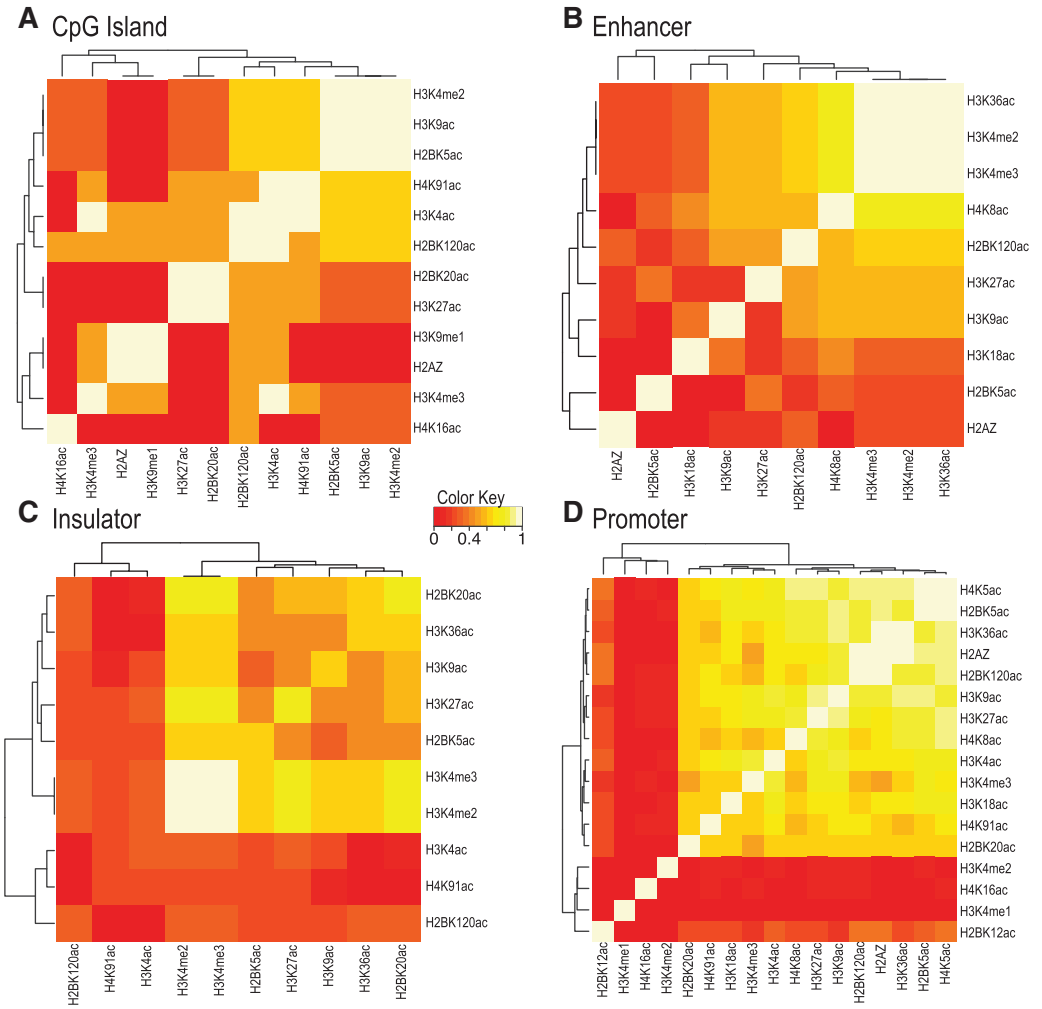

Figure 5. Combinatorial chromatin modification patterns associated with genomic features. (**A**) CpG islands; (**B**) enhancers (distal p300 binding sites); (**C**) insulators (CTCF binding sites); (**D**) promoters. Each cell in the heatmap represents the normalized co-occurrence frequency of a pair of chromatin modifications within the set of biclusters belonging to a specific class of genomic feature. Heatmaps are clustered using hierarchical clustering.

Table 2. Core chromatin modification signatures associated with functional DNA elements

Genomic feature	Core chromatin modification signatures
Insulator	<H3BK20ac, H3K36ac, H3K9ac> <H3K27ac, H2BK5ac> <H3K4me2, H3K4me3>
Enhancer	<H2BK120ac, H3K27ac, H3K36ac, H4K8ac, H3K4me2, H3K4me3>
Promoter	<H2AZ, H2BK5ac, H2BK120ac, H3K36ac, H4K5ac, H3K9ac, H3K27ac, H4K8ac, H3K4ac, H3K18ac, H3K4me3, H2BK20ac, H4K91ac>
CpG island	<H2BK5ac, H3K9ac, H3K4me2, H4K91ac> <H2BK120ac, H3K4ac> <H2BK20ac, H3K27ac> <H2AZ, H3K4me3, H3K9me1>

Core modification sets represent highly co-occurring regions of the co-occurrence matrix. These are identified based on the hierarchical clustering dendrogram of the corresponding co-occurrence matrix.

histone modification data set identified two clusters that are associated with enhancers (7). These clusters involve the signature <H3ac, H4ac, H3K9ac, H3K18ac, H3K27ac, H3K4me1, H3K4me2>. As shown in Table 2, the enhancer modification core we found is consistent with previous findings (34,35). In addition, our analysis also identified a novel core enhancer histone mark, i.e. H2BK120ac, which provides a testable hypothesis for future studies.

Since ~60% of human gene promoters are associated with CpG islands, it is worth comparing the modification cores of promoters and CpG islands (36). The set of modifications associated with CpG islands was a subset of those associated with promoters except for H3K4me2 and H3K9me1. H3K9me1 is a repressive chromatin mark itself and is also involved in the recruitment of DNA methyltransferase for *de novo* DNA methylation (37). Since many CpG islands are methylated during development in a tissue-specific manner (38), H3K9me1 might be an important player for this process. Besides H3K9me1,

H3K27me3 and H4R3me2 are known histone marks to recruit DNA methyltransferase. Neither of these two histone marks was present in our CpG island biclusters. This could due to the wider genomic deposition patterns of H3K27me3 and H4R3me2 compared to H3K9me1 (Supplementary Figure S5).

In summary, our method recovered many known chromatin modification patterns associated with different classes of functional DNA elements. We also identified additional novel chromatin marks associated with these DNA elements. Furthermore, our analysis has revealed a set of 122 novel combinatorial modifications that do not overlap with any of the eight genomic features (Supplementary Table S2). These patterns represent potential epigenetic signatures of yet undiscovered DNA elements.

Relationship of promoter chromatin modifications and gene expression

To better understand the relationship between promoter chromatin modification and gene expression, we applied CoSBI to promoter regions of 21 123 RefSeq genes in the human genome. We ran CoSBI to capture biclusters that occur at least on 0.5% of all promoters and include at least three chromatin modification marks. We increased the min_s parameter from 0.1 to 0.5% because promoter regions are much smaller than the whole genome. This ensures we capture patterns that recur at similar number of loci as the genome-wide analysis (105 versus 130). The coherency cut-offs were set at 0.75 and 0.625, respectively. A higher β parameter value was used than that of the genome-wide analysis since the promoter regions in general contain more tag counts compared to genome-wide data. With this parameter setting, our algorithm identified 2206 biclusters.

Next we focused on promoter biclusters whose target genes are either highly expressed or silent in human $CD4^+$ T cells based on the gene expression profiles generated by Schones *et al.* (25). To do so, for each bicluster, we computed the median expression level of all genes associated with the promoters in the bicluster. We then chose top 10 biclusters with highest median expression levels and bottom 10 biclusters with lowest median expression level (see 'Materials and Methods' section for details). They were regarded as being associated with highly expressed and silent genes in T cells. By examining the chromatin modification patterns of these two groups of promoters, we made several interesting observations regarding the relationship of gene expression and combinatorial patterns of chromatin modifications at gene promoters. Most strikingly, we observed that silent genes can also be associated with acetylations, despite the fact that acetylation is generally regarded as an activating modification (Figure 6). Previously, activating methylation but not acetylation marks have been observed at the promoters of silent genes poised for activation (39–41). Our finding with acetylation is consistent with the result of a more recent study by Barski *et al.* (42). By examining genome-wide histone modification profiles and gene expression during $CD4^+$ T-cell activation, Barski *et al.* found that activating acetylations were already in place for a majority of inducible genes, even though the genes were silent in resting cells. Similarly, genes that were silenced upon T-cell activation retained positive chromatin modifications even after being silenced. Two mechanisms have been proposed by the authors to explain the presence of activating acetylation marks at silent genes: *de novo* poising of silent genes for future expression or as a memory of past transcription. Additional experiments will be needed to determine if either or both mechanisms are responsible for this phenomenon.

Upon further examination of the co-occurrence maps in Figure 6, we noticed that acetylation marks could associate with either activating or repressive methylation marks. Interestingly, promoters of silenced genes have a combinatorial modification pattern composed of a repressive methylation mark, H3K27me3 and several acetylation marks. On the other hand, highly expressed genes have promoters that are decorated by an activating methylation mark, H3K36me1, along with a few acetylations. Therefore, depending on the nature of the methylation that co-occurs with acetylation marks, the regulatory outcome of acetylations in local chromatin environment may be either inducing or repressing the gene expression. This observation and results reported by Barski *et al.* (42) provided a generalization to the bivalent domain concept (promoter regions with overlapping H3K4me3 and H3K27me3 modifications) first proposed by Bernstein *et al.* (39), i.e. the presence of both acetylation and repressive methylation marks may poise silent genes for activation.

Finally, our analysis also revealed that two different repressive methylation marks could be present at silent gene promoters, one involving lysine methylation (H3K27me3) and the other involving arginine methylation (H4R3me2) (Figure 6). Interestingly, these two repressive marks do not co-occur at high frequency and they form modification cores with different modification marks, i.e. H3K27me3 mostly with acetylation whereas H4R3me2 with H3K36me1. Having different modification mark partners and low frequency of co-occurrence suggests that these two repressive marks might be involved in different pathways for maintaining the silent state of target genes. It has also been shown that both modifications can recruit DNA methyltransferases for *de novo* DNA methylation (43,44). Therefore, it appears that at least three repressive marks could be present at these silent gene promoters, two histone methylation marks and DNA cytosine methylation.

DISCUSSION

Subspace clustering is an extension of traditional clustering that seeks to find clusters in different subspaces given a data set (45). Subspace-clustering algorithms have been developed to analyse 2D microarray data (also known as biclustering in microarray data analysis literature), associating subsets of genes whose expression are coherent under a subset of conditions (46–50). Compared to the microarray data, epigenomic data has a unique

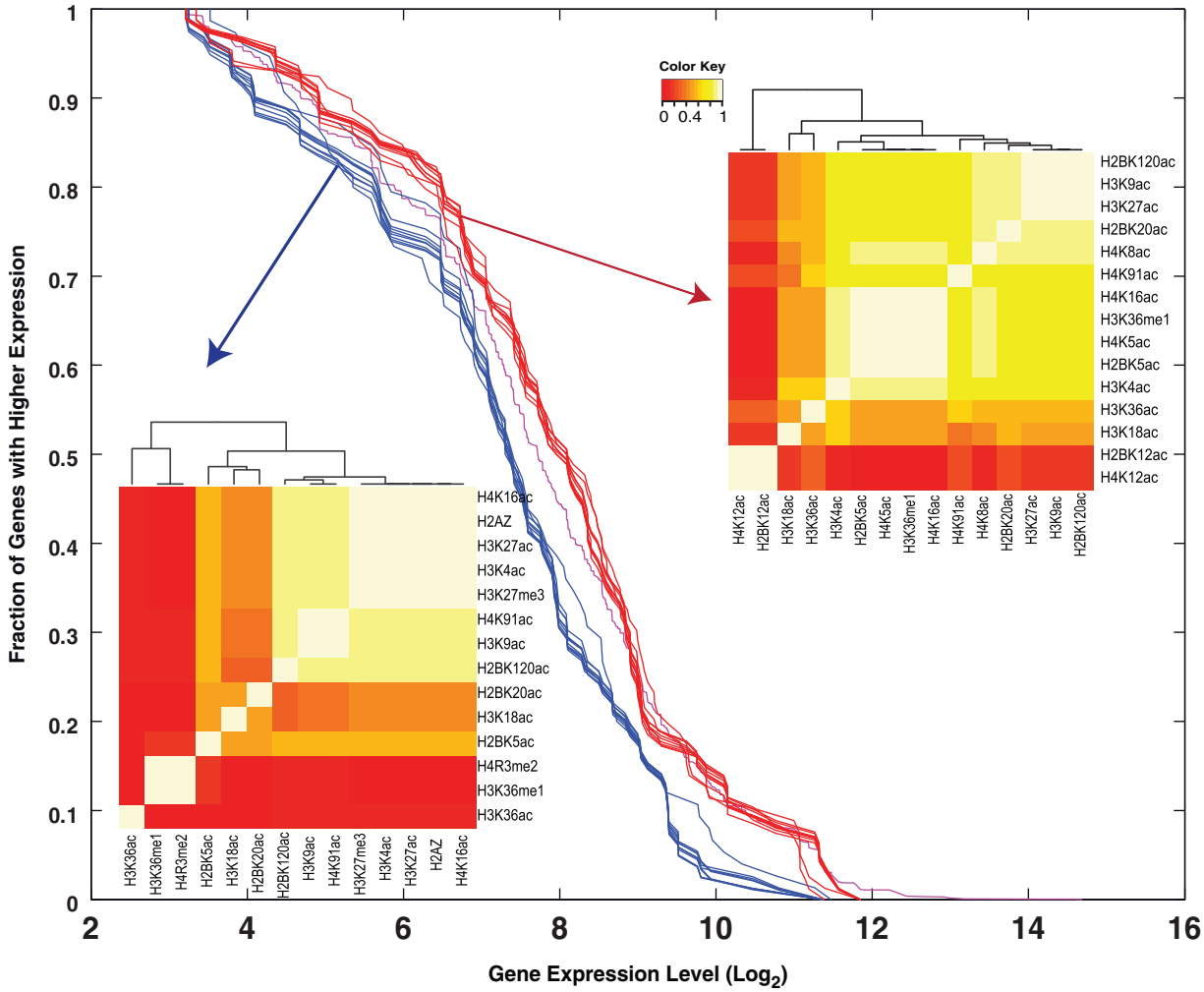

Figure 6. Combinatorial chromatin modifications of promoters associated with highly expressed and silent genes in human T cells. Biclusters associated with highly expressed and silent genes are shown at top right and bottom left, respectively. Each curve represents the cumulative distribution of expression levels of a set of genes whose promoters are associated with chromatin modification biclusters. Gene expression levels in log scale is represented by *x*-axis and *y*-axis represents the fraction of genes in the biclusters that are expressed at higher levels than the corresponding *x*-axis values.

feature that must be accounted for, i.e. the data sets are 'spatially arranged' over chromosomes. This feature can be regarded as the third dimension of the data with the first two being genomic locus and type of epigenetic modification. The CoSBI algorithm is designed specifically for this task and it is capable of clustering 3D epigenomic data and grouping them into defined clusters of common epigenomic behaviour.

Our proposed algorithm aims to identify 2D patterns from the 3D **GCP** matrix. Other algorithms have been proposed for similar analysis. Zhao and Zaki (26) proposed a tri-clustering approach to analyse multiple time-series data sets, which extends the biclustering concept into three dimensions. However, identified 'tri-clusters' by their algorithm are composed of subsets from three dimensions, which implies partial chromatin modification signals/peaks from ChIP-seq reads. Therefore, this algorithm is not suitable for the analysis of the epigenomic data. In addition, the tri-clustering algorithm is computationally very expensive and did not scale to our data.

As discussed in the Introduction, there is a fundamental difference between a subspace-clustering-based algorithm such as CoSBI and previous algorithms such as ChromaSig (7) and HMM-based algorithms (8,9). The latter category of algorithms seek patterns that involve all chromatin marks in the input data whereas the former category of algorithms seek patterns involving only subsets of chromatin modifications in the input data. Numerous studies so far have demonstrated that many combinatorial patterns only involve a few of the many chromatin modifications available in a typical ChIP-Seq data set nowadays. Therefore, subspace-clustering algorithm is better suited for tasks of identifying subsets of re-occurring modifications given a large compendium of chromatin modification data. It also helps to identify more precise patterns. As a comparison, we ran ChromaSig on the same histone modification data used in our comparison with EDISA, i.e. 39 histone modification data (5 kb windows) at 213 enhancers and 213 random sequences. Using the parameter settings suggested by the

ChromaSig authors, however, ChromaSig failed to find any significant clusters. We believe the reason for the failure is that enhancer histone modification patterns do not involve all 39 marks but ChromaSig is designed to find patterns involving all histone marks in the input data. On the other hand, unlike HMM-based algorithms, since CoSBI is designed to identify a set of chromatin marks restricted to the same genomic location (overlapping marks), it cannot identify patterns that involve spatially separated (i.e. sequential) chromatin modifications. Known examples of sequential chromatin modification patterns include H3K79me3–H3K4me3–H3K36me3 associated with promoter and gene bodies (10) and H3K4me3–H3K36me3 associated with lincRNAs. HMM-based algorithms by design are better suited for identifying these kind of patterns.

Although CoSBI was developed for epigenomic data analysis, it can also be used to analyse 3D microarray data (e.g. expression profiles from different patients/samples over a time course). In this setting, the goal of the analysis is to infer clusters that are coherent in two dimensions in the 3D microarray data. We also envision several different ways that CoSBI can be applied to epigenomic data sets. For instance, to examine the spatial and temporal dynamics of combinatorial chromatin modifications, users could apply CoSBI to the chromatin modification maps from different cell types or different developmental stages. Comparative analysis of the inferred biclusters could reveal common and condition-specific chromatin modification patterns. Fuelled by several large-scale epigenomic projects [Epigenome Roadmap, ENCODE (51), modENCODE (52)], epigenomic data sets are becoming increasingly abundant. Comparative analysis of different chromatin modification maps using CoSBI could lead to novel insights into the histone code hypothesis.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank members of the Tan laboratory and three anonymous reviewers for their invaluable comments that help to improve the article.

FUNDING

Pharmaceutical Research and Manufacturers of America Foundation Informatics Research Starter Grant (to K.T.); National Science Foundation under Grant #0937060 to the Computing Research Association for the CIFellows Project (subaward CIF-239 to D.U.). Funding for open access charge: Pharmaceutical Research and Manufacturers of America Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Zhang,L.W., Eugeni,E.E., Parthun,M.R. and Freitas,M.A. (2003) Identification of novel histone post-translational modifications by peptide mass fingerprinting. *Chromosoma*, **112**, 77–86.
2. Strahl,B.D. and Allis,C.D. (2000) The language of covalent histone modifications. *Nature*, **403**, 41–45.
3. Wang,Z.B., Zang,C.Z., Rosenfeld,J.A., Schones,D.E., Barski,A., Cuddapah,S., Cui,K.R., Roh,T.Y., Peng,W.Q., Zhang,M.Q. *et al.* (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genetics*, **40**, 897–903.
4. Heintzman,N.D., Stuart,R.K., Hon,G., Fu,Y., Ching,C.W., Hawkins,R.D., Barrera,L.O., Van Calcar,S., Qu,C., Ching,K.A. *et al.* (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genetics*, **39**, 311–318.
5. Heintzman,N.D., Hon,G.C., Hawkins,R.D., Kheradpour,P., Stark,A., Harp,L.F., Ye,Z., Lee,L.K., Stuart,R.K., Ching,C.W. *et al.* (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**, 108–112.
6. Won,K.J., Chepelev,I., Ren,B. and Wang,W. (2008) Prediction of regulatory elements in mammalian genomes using chromatin signatures. *BMC Bioinformatics*, **9**, 547.
7. Hon,G., Ren,B. and Wang,W. (2008) ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. *PLoS Comput. Biol.*, **4**, e1000201.
8. Jaschek,R. and Tanay,A. (2009) Spatial clustering of multivariate genomic and epigenomic information. *Proceedings of the 13th Annual International Conference on Research in Computational Molecular Biology*. Springer-Verlag, pp. 170–183.
9. Ernst,J. and Kellis,M. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.*, **28**, 817–825.
10. Schones,D.E. and Zhao,K. (2008) Genome-wide approaches to studying chromatin modifications. *Nat. Rev. Genet.*, **9**, 179–191.
11. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
12. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nussbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-seq (MACS). *Genome Biol.*, **9**, R137.
13. Knapp,C. and Carter,G. (1976) The generalized correlation method for estimation of time delay. *Acoustics, Speech Signal Proc. IEEE Trans.*, **24**, 320–327.
14. Coenen,F., Leng,P. and Ahmed,S. (2004) Data structure for association rule mining: T-trees and P-trees. *EEE Trans. Knowledge Data Eng.*, **16**, 774–778.
15. Jiang,D., Pei,J., Ramanathan,M., Tang,C. and Zhang,A. (2004) Mining coherent gene clusters from gene-sample-time microarray data. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 430–439.
16. Knuth,D.E. (1973) *The Art of Computer Programming*. Addison-Wesley, Reading, MA, p. 3.
17. Garcia,B.A., Pesavento,J.J., Mizzen,C.A. and Kelleher,N.L. (2007) Pervasive combinatorial modification of histone H3 in human cells. *Nat. Methods*, **4**, 487–489.
18. LeRoy,G., Weston,J.T., Zee,B.M., Young,N.L., Plazas-Mayorca,M.D. and Garcia,B.A. (2009) Heterochromatin protein 1 is extensively decorated with histone code-like post-translational modifications. *Mol. Cell Proteomics*, **8**, 2432–2442.
19. Pesavento,J.J., Bullock,C.R., LeDuc,R.D., Mizzen,C.A. and Kelleher,N.L. (2008) Combinatorial modification of human histone H4 quantitated by two-dimensional liquid chromatography coupled with top down mass spectrometry. *J. Biol. Chem.*, **283**, 14927–14937.
20. Phanstiel,D., Brumbaugh,J., Berggren,W.T., Conard,K., Feng,X., Levenstein,M.E., McAlister,G.C., Thomson,J.A. and Coon,J.J. (2008) Mass spectrometry identifies and quantifies 74 unique histone H4 isoforms in differentiating human embryonic stem cells. *Proc. Natl Acad. Sci. USA*, **105**, 4093–4098.
21. Khalil,A.M., Guttman,M., Huarte,M., Garber,M., Raj,A., Morales,D.R., Thomas,K., Presser,A., Bernstein,B.E.,

van Oudenaarden,A. *et al.* (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl Acad. Sci. USA*, **106**, 11667–11672.
22. Cuddapah,S., Jothi,R., Schones,D.E., Roh,T.Y., Cui,K. and Zhao,K. (2009) Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res.*, **19**, 24–32.
23. Wang,Z.B., Zang,C.Z., Cui,K.R., Schones,D.E., Barski,A., Peng,W.Q. and Zhao,K.J. (2009) Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell*, **138**, 1019–1031.
24. Ferretti,V., Poitras,C., Bergeron,D., Coulombe,B., Robert,F. and Blanchette,M. (2007) PReMod: a database of genome-wide mammalian cis-regulatory module predictions. *Nucleic Acids Res.*, **35**, D122–D126.
25. Schones,D.E., Cui,K.R., Cuddapah,S., Roh,T.Y., Barski,A., Wang,Z.B., Wei,G. and Zhao,K.J. (2008) Dynamic regulation of nucleosome positioning in the human genome. *Cell*, **132**, 887–898.
26. Zhao,L. and Zaki,M.J. (2005) Tricluster: an effective algorithm for mining coherent clusters in 3d microarray data. *ACM SIGMOD International Conference on Management of Data*. ACM, pp. 694–705.
27. Madeira,S.C. and Oliveira,A.L. (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE/Acm Trans. Comput. Biol. Bioinform.*, **1**, 24–45.
28. Supper,J., Strauch,M., Wanke,D., Harter,K. and Zell,A. (2007) EDISA: extracting biclusters from multiple time-series of gene expression profiles. *BMC Bioinformatics*, **8**, 334.
29. Garcia,B.A., Mollah,S., Ueberheide,B.M., Busby,S.A., Muratore,T.L., Shabanowitz,J. and Hunt,D.F. (2007) Chemical derivatization of histones for facilitated analysis by mass spectrometry. *Nat. Protoc.*, **2**, 933–938.
30. Tie,F., Banerjee,R., Stratton,C.A., Prasad-Sinha,J., Stepanik,V., Zlobin,A., Diaz,M.O., Scacheri,P.C. and Harte,P.J. (2009) CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing. *Development*, **136**, 3131–3141.
31. Karlic,R., Chung,H.R., Lasserre,J., Vlahovicek,K. and Vingron,M. (2010) Histone modification levels are predictive for gene expression. *Proc. Natl Acad. Sci. USA*, **107**, 2926–2931.
32. Mutskov,V.J., Farrell,C.M., Wade,P.A., Wolffe,A.P. and Felsenfeld,G. (2002) The barrier function of an insulator couples high histone acetylation levels with specific protection of promoter DNA from methylation. *Genes Dev.*, **16**, 1540–1554.
33. Roh,T.Y., Wei,G., Farrell,C.M. and Zhao,K. (2007) Genome-wide prediction of conserved and nonconserved enhancers by histone acetylation patterns. *Genome Res.*, **17**, 74–81.
34. Hon,G.C., Hawkins,R.D. and Ren,B. (2009) Predictive chromatin signatures in the mammalian genome. *Human Mol. Genet.*, **18**, R195–R201.
35. Hon,G., Wang,W. and Ren,B. (2009) Discovery and annotation of functional chromatin signatures in the human genome. *PLoS Comput. Biol.*, **5**, e1000566.
36. Bernstein,B.E., Meissner,A. and Lander,E.S. (2007) The mammalian epigenome. *Cell*, **128**, 669–681.
37. Martin,C. and Zhang,Y. (2005) The diverse functions of histone lysine methylation. *Nat. Rev. Mol. Cell Biol.*, **6**, 838–849.
38. Illingworth,R.S. and Bird,A.P. (2009) CpG islands–'a rough guide'. *FEBS Lett.*, **583**, 1713–1720.
39. Bernstein,B.E., Mikkelsen,T.S., Xie,X., Kamal,M., Huebert,D.J., Cuff,J., Fry,B., Meissner,A., Wernig,M., Plath,K. *et al.* (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, **125**, 315–326.
40. Ng,H.H., Robert,F., Young,R.A. and Struhl,K. (2003) Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. *Mol. Cell*, **11**, 709–719.
41. Raisner,R.M., Hartley,P.D., Meneghini,M.D., Bao,M.Z., Liu,C.L., Schreiber,S.L., Rando,O.J. and Madhani,H.D. (2005) Histone variant H2A.Z marks the 5' ends of both active and inactive genes in euchromatin. *Cell*, **123**, 33–248.
42. Barski,A., Jothi,R., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E. and Zhao,K. (2009) Chromatin poises miRNA- and protein-coding genes for expression. *Genome Res.*, **19**, 1742–1751.
43. Fuks,F. (2005) DNA methylation and histone modifications: teaming up to silence genes. *Curr. Opin. Genet. Dev.*, **15**, 490–495.
44. Zhao,Q., Rank,G., Tan,Y.T., Li,H., Moritz,R.L., Simpson,R.J., Cerruti,L., Curtis,D.J., Patel,D.J., Allis,C.D. *et al.* (2009) PRMT5-mediated methylation of histone H4R3 recruits DNMT3A, coupling histone and DNA methylation in gene silencing. *Nat. Struct. Mol. Biol.*, **16**, 304–311.
45. Parsons,L., Haque,E. and Liu,H. (2004) Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explorations Newslett.*, **6**, 90–105.
46. Bergmann,S., Ihmels,J. and Barkai,N. (2003) Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, **67**, 031902.
47. Cheng,Y. and Church,G.M. (2000) Biclustering of expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 93–103.
48. Tanay,A., Sharan,R. and Shamir,R. (2002) Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, **18(Suppl. 1)**, S136–S144.
49. Gu,J. and Liu,J.S. (2008) Bayesian biclustering of gene expression data. *BMC Genomics*, **9(Suppl. 1)**, S4.
50. Prelic,A., Bleuler,S., Zimmermann,P., Wille,A., Buhlmann,P., Gruissem,W., Hennig,L., Thiele,L. and Zitzler,E. (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, **22**, 1122–1129.
51. Birney,E., Stamatoyannopoulos,J.A., Dutta,A., Guigo,R., Gingeras,T.R., Margulies,E.H., Weng,Z., Snyder,M., Dermitzakis,E.T., Thurman,R.E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
52. Celniker,S.E., Dillon,L.A., Gerstein,M.B., Gunsalus,K.C., Henikoff,S., Karpen,G.H., Kellis,M., Lai,E.C., Lieb,J.D., MacAlpine,D.M. *et al.* (2009) Unlocking the secrets of the genome. *Nature*, **459**, 927–930.