Evolutionary Bioinformatics

Libertas Academica
FREEDOM TO RESEARCH

ORIGINAL RESEARCH

Identifying the Combinatorial Effects of Histone Modifications by Association Rule Mining in Yeast

Jiang Wang, Xianhua Dai, Qian Xiang, Yangyang Deng, Jihua Feng, Zhiming Dai and Caisheng He

Department of Electronics and Communications Engineering, School of Information Science and Technology,
Sun Yat-Sen University, 135 West Xin'gang Road, Guangzhou, P.R. China.
Corresponding author email: wrdzsu2003@163.com

Abstract: Eukaryotic genomes are packaged into chromatin by histone proteins whose chemical modification can profoundly influence gene expression. The histone modifications often act in combinations, which exert different effects on gene expression. Although a number of experimental techniques and data analysis methods have been developed to study histone modifications, it is still very difficult to identify the relationships among histone modifications on a genome-wide scale.

We proposed a method to identify the combinatorial effects of histone modifications by association rule mining. The method first identified Functional Modification Transactions (FMTs) and then employed association rule mining algorithm and statistics methods to identify histone modification patterns. We applied the proposed methodology to Pokholok et al's data with eight sets of histone modifications and Kurdistani et al's data with eleven histone acetylation sites. Our method succeeds in revealing two different global views of histone modification landscapes on two datasets and identifying a number of modification patterns some of which are supported by previous studies.

We concentrate on combinatorial effects of histone modifications which significantly affect gene expression. Our method succeeds in identifying known interactions among histone modifications and uncovering many previously unknown patterns. After in-depth analysis of possible mechanism by which histone modification patterns can alter transcriptional states, we infer three possible modification pattern reading mechanism ('redundant', 'trivial', 'dominative'). Our results demonstrate several histone modification patterns which show significant correspondence between yeast and human cells.

Keywords: histone code, association rule, yeast

Evolutionary Bioinformatics 2010:6 113–131

This article is available from http://www.la-press.com.

Background

Gene activities in eukaryotic cells are mainly regulated by transcription factors and chromatin structure. Chromatin fibers are composed of polymers of nucleosomes, which are the fundamental units of chromatin. Each nucleosome is composed of approximately 147 base pairs of DNA wrapped ~1.7 turns around a histone octamer consisting of two copies each of the core histones H2A, H2B, H3, and H4.[1–3] On the one hand, the histones cover the DNA and prevent the genetic information from being accessed by many biological machineries, such as transcription, replication and recombination. On the other hand, the histones can be altered by the post-translational chemical modifications, including acetylation, methylation, phosphorylation, ubiquitylation, and sumoylation, through which the interactions between histones and DNA can be altered.[4,5] Among them, histone acetylation leads to a reduction of positive charges on the histone tails, and because of the negative charge of the DNA, less charge compensation leads to an open chromatin state, which is often associated with increased gene expression;[6,7] histone methylation links to both gene activation and gene repression depending on its site and extent of methylation, and different effects may be associated with mono-, di-, or trimethylation of lysine residues.[8]

A large number of residues within the histones are modified. In particular, a single modification can induce the occurrence of one or more subsequent modifications. Therefore, histone modification may form a ‘code’ and then comes histone code hypothesis. According to this hypothesis, covalent posttranslational modification of histone tails act sequentially or in combination to form a ‘histone code’ that is read by other proteins to bring about distinct downstream events.[9] However, this hypothesis has been much debated, some believed that use of histone modifications individually or sequentially cannot be considered a code since the total number of modifications do not necessarily contain more information than the sum of individual modifications;[7] others suggested that the existence of such a code should manifest itself such that different modification combinations lead to distinct outcomes.[1,10] Although different interpretations of ‘histone code’ are possible, we believe that histone modifications function in a manner rated to their combinatorial effects rather than individual effect.

Many methods have been proposed to find the combinatorial effects. On the one hand, a number of experimental techniques have been developed to study histone modifications, such as, chromatin immunoprecipitation (ChIP), ChIP followed by amplification and microarray hybridization (ChIP-chip) and mass spectrometry (MS). On the other hand, some data analysis methods have been introduced to reveal the relationships between histone modification sites, such as, clustering methods[11,12] and correlation analysis.[13,14] But there are several disadvantages of the above methods. Specifically, clustering methods are good at clustering under all sites but difficult to cluster under subset of all sites, hence, they fail to identify whether the sites are functional or not. Meanwhile, correlation analysis can only reveal pairwise correlation between each pair of sites. In order to solve above problems, we proposed a method of association rule mining which not only overcomes aforesaid shortcomings but also successfully identifies the relationship among histone modification sites on a genome-wide scale.

Association rule mining was first proposed by Agrawal et al.[15] Also the Apriori algorithm was published by Agrawal and Srikant, and a method for generating association rules is described in this paper.[16] To improve the efficiency of Apriori, many variations of the Apriori algorithm have been proposed, such as application of hash tables to improve association mining efficiency studied by Park et al,[17] the sampling approach proposed in Hannu[18] and the partitioning technique discussed by Savasere et al.[19] Also, association rule mining has various expansions, such as cyclic association rules mining,[20] negative association rules mining[21] and market basket analysis.[22] Recent work[23,24] has shown that the removal of hierarchically redundant rules from multi-level datasets by using a dataset’s hierarchy is a promising approach to solving redundancy problem. To evaluate statistical significance of an association rules, chi-square test was used for statistical dependency between the antecedent and the consequent of the rule.[25,26]

This method is focused on combinatorial effects of histone modifications which have a significant effect on the gene expression. To identify these histone modifications, we have to rule out other factors which could have an impact on gene expression. Gene activities in eukaryotic cells are concertedly regulated by transcription factors and chromatin structure. Furthermore, since clusters of genes with the same histone modification patterns are enriched with coexpressed genes, the expression coherence (EC) scores are mainly contributed by transcription factors and histone modifications. As a result, if we eliminate the influence of TFs, we will gain transactions whose EC scores of their target genes are mainly attributed to same histone modification combinations. In this paper, through synthetically analysis of histone modification data, gene expression data, gene binding activity data and the relationships between transcription factors and chromatin modifiers, we identified Functional Modification Transactions (FMTs) and employed association rule mining algorithm and statistics methods to identify combinations of histone modifications. After applying this method to two distinct histone modification data, the specific modification combinations as well as global patterns of histone modifications were obtained.

Materials and Methods

Concept descriptions and notations

To better understand this method, we first present some useful concepts and notations used in the paper. First, we give some concept descriptions of transaction, candidate transactions, target genes of a transaction and EC score of a transaction. By definition, each transaction is a set of items (ie, itemset), while in this paper each transaction corresponds to a histone modification combination after discretization of histone modification data; candidate transactions are the specific transactions extracted from all transactions and prepared for data mining; target genes of a transaction is a cluster of genes with a common histone modification described by the transaction; similarly, EC score of a transaction means the EC score of the target genes.

Next, we give some notations used in this paper. For a transaction, each item is denoted as the form of 'XY', where X is the modification site of each gene, $X \in \{1, 2, \ldots, n\}$, where n is the number of histone modification sites (henceforth referred to as just 'sites'), Y is the state of according site, $Y \in \{0, 1, 3\}$, ie, 0: 'NULL', 1: 'under-expressed', 3: 'over-expressed'. Since the items with the 'NULL' state in according sites are eliminated from the transaction, there are only two states in the transaction, 'X1' and 'X3'. Therefore, each transaction can be described as the form of $\{1Y, 2Y, \ldots, nY\}$, $Y \in \{1, 3\}$. The meaning of the notation of 'X' varies with data. Specifically, for Pokholok et al data, the number 'X' from 1 to 8 corresponds to sites of H3K9ac, K3K14ac, H4ac, H3K4me1, H3K4me2, H3K4me3, H3K36me3, H3K79me3 respectively, whereas for Kurdistani et al's data, the number 'X' form 1 to 11 corresponds to sites of H4K8, H4K12, H4K16, H3K9, H3K14, H3K18, H3K23, H3K27, H2AK7, H2BK11, H2BK16 respectively. In addition, we use notation 'X1' and 'X3' to denote 'Under-expressed' and 'Over-expressed' states in transactions and association rules, while in the body of paper, in order to express clearly, we use symbols '[−]' and '[+]' to denote these two states.

Steps for the method

Our method first identifies functional histone combinations according to gene expression and then applies association rule mining to candidate transactions. The functional histone modification combinations have two characters (1) significant EC scores (Materials and Methods) (2) no significant TFs in the promoter region or significant TFs which interact with chromatin modifiers. The steps of method can be described as follows: First, we generated 3^n transactions each of which corresponds to a histone modification combination or pattern, where n is the number of sites. Second, we identified M transactions which have significant EC scores among all possible transactions. Third, we determined $N_1 + N_2$ candidate transactions among M transactions which correspond to FMTs (Identifying Functional Modification Transactions, Fig. 1). Finally, we applied association rule mining technology on the $N_1 + N_2$ candidate transactions. Notablely, for the convenience of understanding, a flowchart of this algorithm can be found in Figure S1. When applying aprioir-like algorithm (Christian Borgelt, http://www.borgelt.net/apriori.html), association rules were extracted with absolute minimum support of 5 (which correspond to 0.42% of the whole transactions for Poklok et al's data, 0.37% for Kurdistani et al's data), minimum confidence of 80% and minimum lift (improvement) of 110% (Materials and Methods). This method extracted association rules from transactions whose target genes has higher EC score and they are mainly regulated by histone modification. We applied the proposed methodology to two different histone modification data and identified a number of modification combinations some of which are supported by previous studies. And our

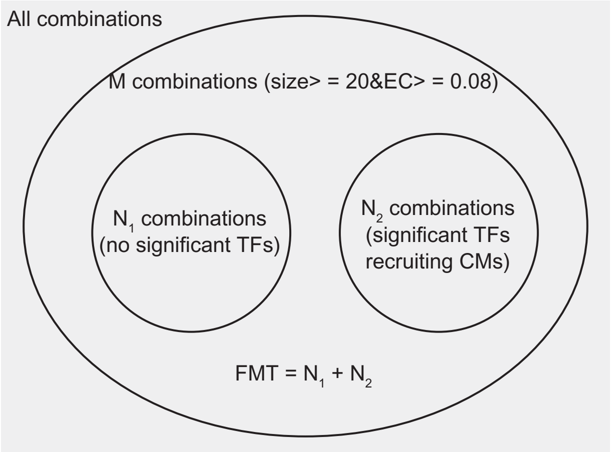
All combinations
M combinations (size> = 20&EC> = 0.08)
N_1 combinations (no significant TFs)
N_2 combinations (significant TFs recruiting CMs)
FMT = N_1 + N_2

Figure 1. The Venn diagram of FMT.

association rules provide two different global views of histone modification landscapes on two datasets (Figs. 2 and 3). These novel patterns we extracted lay a useful foundation for the additional experiments necessary to gain a fuller understanding of the roles of combinations of histone modifications in gene expression.

Data preparation

In this paper, datasets we utilized here mainly include histone modification, gene expression, and transcription factor binding activity data. The first histone modification data employed in this study is from Pokholok et al.[13] We chose 8 sets of histone modifications (H3K9ac, H3K14ac, H4ac, H3k4me1, H3k4me2, H3K4me3, H3K36me3, and H3K79me3), with H4ac referring to non-specific acetylation on any of the four sites, eg, H4K5, H4K8, H4K12, H4K16. The second histone acetylation data is originally from Kurdistani et al.[12] This data set includes measurements of the acetylation levels at 11 different sites (H4K8, H4K12, H4K16, H3K9, H3K14, H3K18, H3K23, H3K27, H2AK7, H2BK11, H2BK16). Since the acetylation levels of histones are affected by the occupancy of nucleosomes in that region, the acetylation data were divided by the average level of H3 and H2A histones from the Bernstein et al data set.[27,28] The major difference between two data sets is that the acetylation in the first data set was measured against nucleosomal DNA while the second data set was measured against genomic DNA. In addition to above two histone modification datasets, expression dataset is necessary to evaluate the influence of modification patterns on expression. The mRNA expression profiles are combined by environmental stresses Gasch et al[29] and cell cycle Spellman et al[30] for 250 total conditions. To identify modification pattern from gene expression, we must eliminate the influence of TFs. Thus we extracted two disjoint subsets from all combinations of histone modification sites. One is the combinations whose promoters have no significant TFs identified by t-tests, of which the transcription factor binding activity dataset is from Young Lab Harbison et al;[31] the other is the combinations whose promoters only have TFs recruiting chromatin modifiers, among which the relations between transcription factors and chromatin modification factors are from Steinfeld et al.[32] Besides, the cell types and experiment conditions in various datasets are match or comparable (Table 1).

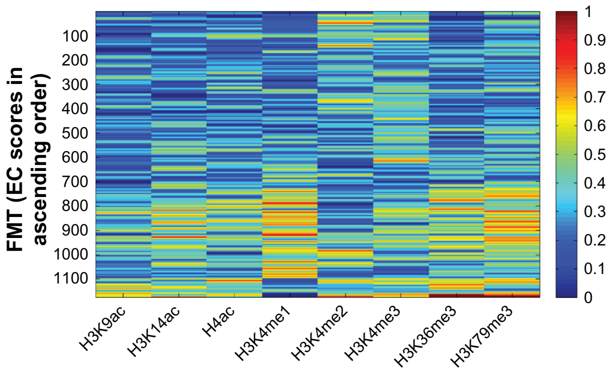

Figure 2. The global view of histone modification of FMTs for Pokholok et al's data.
Notes: Rows represent transactions, and columns represent sites. To obtain global view of histone modification of FMTs, we used a sliding window of 10 transactions to calculate ratio of over-expressed state of sites. The transactions were sorted according to EC scores. The graph showed the over-expressed states of histone modification.

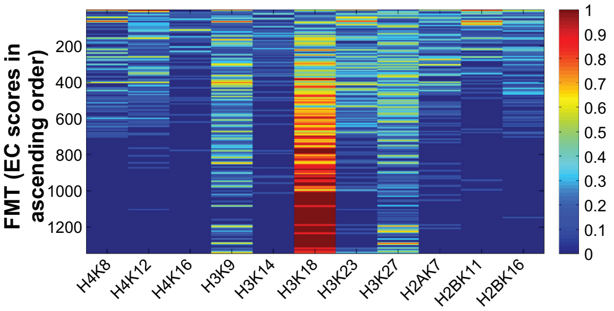

Figure 3. The global view of histone modification of FMTs for Kurdistani et al's data.
Note: Same as in Figure 2 except that Kurdistani et al's data is used.

Histone modification data discretization

To formulate histone modification profiles with transactions, we discretized histone modifica-

Table 1. Summary of various datasets.

Source	Description	Cell type	Experiment condition
Pokholok et al (2005)	A subset of the Pokholok et al's dataset with eight sets of histone modifications under YPD condition	W303	YPD
Kurdistani et al (2004)	A dataset with eleven acetylation sites	YDS2 wt	YEPD (YPD)
Harbison et al (2004)	Transcription factor binding activity	W303	YPD (rich medium)
Gasch et al (2000)	Environmental stresses	Diversity	Diversity
Spellman et al (1998)	Cell cycle	Diversity	YEP medium

Notes: 'YEPD' is often abbreviated as YPD, which corresponds to rich medium. 'YEP medium' is based upon YPD but is without dextrose. 'Diversity' indicates diverse conditions.

tion sites of each gene. The value greater than 60th percentile, less than 40th percentile were regarded as 'over-expressed', 'under-expressed' respectively. To avoid ambiguity due to measurement noise, the middle 20% of genes was regarded as 'NULL'.[33]

Expression coherence score of a transaction

In order to determine whether the target genes of a transaction were significantly correlated in their mRNA expression profiles, we calculated all pairwise Pearson correlation coefficients for the expression profiles of all the genes within a transaction. For a transaction with m target genes, this corresponds to $m(m-1)/2$ correlation values. We defined "Expression Coherence" (EC) as the mean of these values.[12]

Identifying the threshold of EC scores

We first randomly selected 100,000 samples with the size of 20, and selected a threshold (0.08) so that only 5% (95% confidence) of the random samples pass this threshold. Then we iteratively performed above process with different sample sizes, such as 30, 40, 50, etc., and found that the larger the size of sample, the smaller the threshold is (data not shown). Therefore, we identified the strictest threshold with the sample size of 20, and considered the transaction was significant if it's EC score was above this threshold (Fig. S2).

Identifying the histone modification regions of different sites

Since we obtain histone modification patterns according to gene expression (EC score), only the patterns that make a significant contribution to EC score can be found. Therefore, the choices of regions of different sites are based on the contribution of each site to gene expression respectively. Pokholok et al's histone modification data given in genome-wide location format, was separated into there regions, TSS, ORF and promoter. According to genome-wide distribution pattern of histone modifications,[34] acetylation level of H3 and H4 as well as H3K4me3 peak close to TSS region, hence, we assign the mean modification level of TSS region to these sites. Likewise, the remaining sites peak at ORF region, so we assign the mean modification level of ORF region to remaining sites. Moreover, correlations between the mean region modification levels and transcriptional activity are mostly larger than other regions. Besides, transcriptional activities correlate highly with EC scores ($r = 0.53$, for 1183 FMTs). Taken together, the regions we selected have a significant contribution to EC score. For Kurdistani et al's data, we used acetylation data from intergenic regions for two reasons. First, the data was already separated into intergenic region and ORF region according to each gene. Second, histone acetylation correlate positively with transcription levels and highly enriched in promoter regions.[3]

Identifying Functional Modification Transactions (FMT)

As we know, gene expression is mainly regulated by transcription factors and histone modifications. In order to identify combinatorial effects of histone modifications, we should extract transactions whose expressions of their target genes are contributed by their corresponding histone modifications, or rather, these histone modifications are functional, and the extracted transactions are defined as Functional Modification Transactions (FMT). We carried out following steps to select FMTs (Fig. 1). First, for 3^n combinations (where n is the number of histone modification sites), we calculated EC scores for groups of target genes of combinations (transactions) each of which correspond to a transaction with a minimal size of 20. Second, we selected M significant transactions each of which had a minimal threshold, 0.08. Third, in order to identify FMTs, we should exclude the influence to the significant EC scores produced by transcription factors (TFs). Then we applied t-tests to see whether the mean binding activity of transcription factors in each transaction differed from that of the same TFs in the overall genome. We tested all 203 TFs to assess their difference from genomic mean. For all M transactions, the number of hypothesis tests $M \times 203$. Then we used Bonferroni correction to solve the multiple testing problems. We acquired FMTs from two aspects. On the one hand, only N_1 transactions with none of significant TF among 203 TFs were extracted from M transactions. Specifically, the transcription factor binding profile of an example among N_1 transaction set for Pokholok et al's data can be seen in Figure S3. On the other hand, on account of some TFs require the recruitment of a chromatin modifier to facilitate their activity, the N_2 out of M transactions with special significant TFs which recruit chromatin modifiers were considered. From Steinfeld et al's data, we acquired 35 special TFs which could recruit chromatin modifiers,[32] so the TFs among the 35 TFs were considered. Specifically, the transcription factor binding profile of an example among N_2 transaction set for Pokholok et al's data can be seen in Figure S4. Finally, $N_1 + N_2$ FMTs were acquired.

Basic concepts of association rules

Association rules discovery technique finds interesting associations or correlation relationships among a large set of data items. This method extracts sets of items that frequently occur together in the same transaction, and then formulate rules that characterize these relationships.

Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of items and D be a set of database transactions. Each transaction T is a set of items such that $T \subseteq I$. An association rule is an implication of the form, $A \Rightarrow B$, where $A \subset I$, $B \subset I$ and $A \cap B = \phi$. Association rules provide information in the form of 'if-then' statement, and the 'if' part is termed as antecedent (A), whereas the 'then' part is termed as consequent (B). The rule $A \Rightarrow B$ holds in the transaction set D with support s in the transaction set D if s is the percentage of transactions in D containing both A and B. This is equal to the probability, $P(A \cup B)$. The rule $A \Rightarrow B$ holds in the transaction set D with confidence c, where c is the percentage of transactions in D containing A that also contain B. This is equal to the conditional probability, $P(B|A)$. That is,

$$\text{support } (A \Rightarrow B) = P(A \cup B)$$
$$\text{confidence } (A \Rightarrow B) = P(B|A)$$

We take an example to illustrate above concepts. For Pokholok et al's data, I {11, 13, 21, 23, 31, 33, 41, 43, 51, 53, 61, 63, 71, 73, 81, 83}, $D = N_1 + N_2 = 1183$ (transactions), and each transaction T can be formulated like the form of, say, {13, 21, 41, 63, 81}, where $T \subseteq I$. From Table 2, rule1 can be formulated as 81, 13, 63 $\Rightarrow$ 51, of this rule, the support of 1.2% denotes that there are $1183 \times 1.2\%$ transactions which contain{81, 13, 63, 51}; the confidence of 100% is the ratio of the number of transactions that include all items in the antecedent ({81, 13, 63}) as well as the consequent ({51}) to the number of transactions that include all items in the antecedent ({81, 13, 63}).

By the way, we call the confidence of a rule with empty antecedent the priori confidence, and the confidence of a rule with non-empty antecedent we call the posterior confidence. Generally, rules that satisfy both a minimum support threshold and a minimum confidence threshold are called strong. In data mining, associations can be generated using this "support-confidence" framework, but not all of rules satisfying the minimum support and minimum confidence thresholds are interesting to the user. Let's examine the simple example and let us assume that 60% of all customers buy some kind of bread, ie, the rule '$\Rightarrow$ *bread*' with a confidence of 60%. Let's consider the rule '*cheese* $\Rightarrow$ *bread*', which holds with a confidence of, say, 61%. This is not an interesting rule. Because the two values are almost the same and the fact that the event of buying cheese does not have a significant influence on buying bread. If the rule '*cheese* $\Rightarrow$ *bread*' with a confidence of 50%, it is not an interesting rule too, since cheese and bread are negatively associated and the purchase of this item actually decreases the likelihood of purchasing the other. Therefore, our aim is to find rules which antecedent and consequent are positively associ-

Table 2. Rules extracted from Pokholok et al's data.

Rule	Antecedent	Consequent	Supp. (%)	Conf. (%)	Lift (%)	*P*-value	Corrected *P*-value
Rule1	**81 13 63**	**51**	1.2	100	313	5.79E-07	<0.001
Rule2	**81 13 51**	**63**	1.2	100	281.7	1.77E-05	<0.001
Rule3	**81 23 51**	**63**	1.1	100	281.7	1.44E-05	0.001
Rule4	**81 33 51**	**63**	0.9	100	281.7	6.17E-05	0.001
Rule5	13 71 33 63	51	0.7	100	313	7.20E-04	<0.001
Rule6	13 71 23 63	51	0.7	100	313	6.85E-04	<0.001
Rule7	21 41 13	81	0.6	100	410.8	1.99E-05	<0.001
Rule8	41 13 71 33	51	0.4	100	313	1.57E-02	0.001
Rule9	**13 71 51**	**63**	1.4	94.1	265.1	4.97E-05	<0.001
Rule10	**71 23 51**	**63**	1.2	93.3	262.9	5.13E–05	0.007
Rule11	**71 33 51**	**63**	1.2	93.3	262.9	2.96E-05	<0.001
Rule12	**41 13 71 63**	**51**	0.7	88.9	278.2	1.16E-02	0.002
Rule13	81 33 23 63	51	0.7	88.9	278.2	1.25E-03	<0.001
Rule14	41 31 73	83	0.4	83.3	216.2	3.58E-02	0.005
Rule15	**81 71 51 63**	**13**	0.7	80	308.3	2.44E-02	0.038
Rule16	**41 33 73 83**	**53**	0.7	80	249.7	2.91E-03	<0.001
Rule17	61 21 41 31	53	0.7	80	249.7	1.20E-03	<0.001
Rule18	21 73 51 43	63	0.7	80	225.3	2.10E-02	0.003
Rule19	61 21 31 51	71	0.7	80	269.6	1.03E-04	<0.001
Rule20	81 33 43	71	0.7	80	269.6	1.16E-03	0.005
Rule21	**41 23 73 53**	**83**	0.7	80	207.5	1.14E-03	0.004

Notes: The unit position of numbers in the columns of antecedent and consequent are denoted as follows: 1, under-expressed; 3, over-expressed. The other positions of numbers stand for histone modification sites, eg, the number from 1 to 8, which corresponds to sites of H3K9ac, H3K14ac, H4ac, H3K4me1, H3K4me2, H3K4me3, H3K36me3, H3K79me3 respectively. There 11 rules (in boldface) can be explained by previous studies.

ated. In our approach we also have employed the improvement (also named as lift) value to assess rules which is a correlation measure between antecedent and consequent. For the rule $A \Rightarrow B$, the improvement is defined as $corr_{A,B} = \mathrm{P}(A \cup B)/\mathrm{P}(A)\mathrm{P}(B)$, that is, the posterior confidence divided by the prior confidence. If the result of improvement is less than 1, greater than 1, and equal to 1, then the relationship of A and B corresponds to negatively correlated, positively correlated and independent respectively. So the rules that the improvements are greater than 1 are interesting. In our method, rules with improvement greater than 110 percent are selected. In addition, our method produced association rules with a single item in the consequent. The restriction to single item consequents is due to the following reasons: First, association rule mining usually produces too many rules if we have no limitations. Second, the complex rules add almost nothing to the insights about the data set. For example, if you have a rule $a, b \Rightarrow c, d$, you will necessarily also have the rules $a, b \Rightarrow c$ and $a, b \Rightarrow d$ in the output.

Statistical significance of association rules

The above measures do not involve information about statistical significance. The chi-square measure is often used to measure the difference between a supposed independent distribution of two discrete variables and the actual joint distribution in order to determine how strongly two variables depend on each other. A *P*-value of each extracted rule was computed under the assumption that the null hypothesis of the test is true (both the antecedent and consequent part of the rule are independent).[35] In order to obtain exact *P*-value, we adopted Fisher's exact test, because the chi-square statistic became inaccurate when the expected frequency of any cell of contingency tables containing exactly two rows and two columns is less than 5 or the total number is less than 50. Fisher's exact test is the best choice as it always gives the exact *P*-value, while the chi-square test only calculates an approximate *P*-value. We used two-tailed (also called two-sided) *P* values of Fisher's exact test. When simultaneously analyze multiple rules the *P*-values need to be corrected to avoid multiple testing problem. The following steps are taken to solve multiple testing problems. First, we selected M transactions from 3^n transaction, which are selected by two constraints: the number of target genes of each transaction is larger or equal to 20; EC score of target genes of each transaction is larger or equal to 0.08. Second, from M transactions, we stochastically generated 1000 candidate transaction sets, each of which consisted of $N_1 + N_2$ transactions. Third, association rules were extracted from these transaction sets with same confidence and improvement thresholds. Last, corrected *P*-values were calculated for each association rule in real data as the fraction of permutations having any association rule with a *P*-value less than or equal to the observed *P*-value for that association rule.[35,36] Usually the corrected *P*-values less than or equal to 0.05 are considered statistically significant.

Since it is possible to find a biological function associated with a histone modification, we carried out a similar analysis on 1000 randomly generated candidate transaction sets to test the novelty of the combinations detected by our method. As a result, the ratios of explicable rules for these randomized data are much smaller than that for the real data (data not shown). Therefore, the association rules which can be explained by biological function are significant, and then the detected combinations supported by biological function are novel.

Elimination of redundant association rules

From the rules extracted from transactions, we found some redundant rules, such as, 13, 63, 81 ⇒ 51 and 13, 63, 81, 71 ⇒ 51. Because the association rules containing redundancies are difficult to comprehend, we sought methods to deal with this issue of redundancy. According to definition below by Gavin Shaw we can remove hierarchically redundant rules.

Definition (Hierarchical Redundancy for Approximate Basis): Suppose that two approximate association rules, $R_1 : A_1 \Rightarrow B$ and $R_2 : A_2 \Rightarrow B$, both with exactly the same itemset B as the consequent, differ in that rule R_1 has a confidence of C_1, whereas rule R_2 has a confidence of C_2. Rule R_2 is redundant to rule R_1 if (1) the itemset A_1 is a subset of itemset A_2, and (2) the confidence of R_2 (C_2) is less than or equal to the confidence of R_1 (C_1). On the contrary, Rule R_1 is redundant to rule R_2 if (1) the itemset A_1 is a subset of itemset A_2, and (2) the confidence of R_2 (C_2) is larger than the confidence of R_1 (C_1).[23]

Specifically, we take an example to manifest how to eliminate redundant rules. For rule1 ($A \Rightarrow B$) and rule2 (A, $C \Rightarrow B$), if the confidence of rule2 is less than or equal to rule1, we consider rule2 is redundant to rule1 because rule1 is more general than rule2 and rule2 does not bring any new information to the user. Otherwise, if rule2 have a larger confidence than rule1, we should choose rule2 rather than rule1. Because confidences are very important measure for association rules, which indicate their strength, accuracy and reliability, it is important to keep rules with high confidences.

Results

In this paper we propose a method to identify the combinatorial effect of histone modification by association rule mining. The technology of association rule mining can find interesting association or correlation relationships among a large set of data items. We adopt this data mining technology to explore the relationships or the synergies between these histone modifications. In short, this method first identifies functional histone combinations according to gene expression and then extracts correlation relationships or patterns among these large histone combinations by association rule mining technology. In the following sections, we describe extracted rules and patter ns from Pokholok et al's data and Kurdistani et al's data.

Association rules derived from Pokholok et al's data

We first applied our method to Pokholok et al's data and obtained some intriguing relationships. Specifically, we selected acetylation levels at 3 sites, H3K9, H3K14, and H4 (referring to non-specific acetylation on any of the four acetylable lysines on H4 tails), and 5 methylation levels of H3K4me1, H3K4me2, H3K4me3, H3K36me3, and H3K79me3. The above eight histone modification levels are measured in YPD medium. According to genome-wide distribution patterns of histone modifications,[34] we selected the histone modification levels near TSS (transcription start sites) for H3K9ac, H3K14ac, H4ac, and

H3k4me3, and selected the mean level of the ORF for the rest modifications(Materials and Methods). As shown in Fig. S1, among 3^8 transactions, we filtered out transactions whose number of target genes is less than 20. Then we got 3052 transactions, and after being filtered by EC scores, 0.08(identifying the threshold of EC scores, Fig. S2), 1674 transactions were obtained. We obtained 552 N_1 transactions and 631 N_2 transactions, and then we got $N_1 + N_2 = 1183$ FMTs (Supplementary file2). Applying association rule mining to FMTs, we extracted 21 rules from 8 histone modification levels (H3K9ac, H3K14ac, H4ac, H3k4me1, H3k4me2, H3K4me3, H3K36me3, and H3K79me3) (Table 2).

Overview of the extracted rules

We can see the global view of histone modification of FMTs (Fig. 2). In Fig. 2, the higher the EC scores, the higher the histone modification level is, which confirmed the FMTs we chose were reasonable. From the overall over-expressed state of extracted 21 rules (Fig. 4), we found that a majority of rules (61.9%, 13 of 21) includes item '63' (the modification level of H3K4me3 is over-expressed). It is consistent with the fact that hisone H3K4 methylation is a post-translational modification that is exclusively associated with actively transcribed genes.[37,38] There are 71% (15 of 21) of rules each of which is consisted of any of '63' (H3K4me3), '73' (H3K36me3), or '83' (H3K79me3), and synergized with '13' (H3K9ac), '23' (H3K14ac), or '33' (H4ac), such as 81, 13, 63 ⇒ 51, 41, 33, 73, 83 ⇒ 53, etc. These results indicate that actively transcribed euchromatin has high level of acetylation and is trimethylated at H3K4, H3K36 and H3K79.[5] There are 90% (19 of 21) of rules each of which has an over-expressed histone modification level at sites within H3K9ac, H3K14ac, H4ac, H3K4me2, and H3K4me3. Out of 13 rules which includes '63' item, there are 10 rules(account for 77%) that emerge with '13' or '23', which suggest that H3K4me3 and H3 hyperacetylation occur together.[39] Yet, intriguingly, there are 52% (11 of 21) of rules can be explained by the previous literature. In the following sections, we discuss the patterns extracted from Pokholok et al's data.

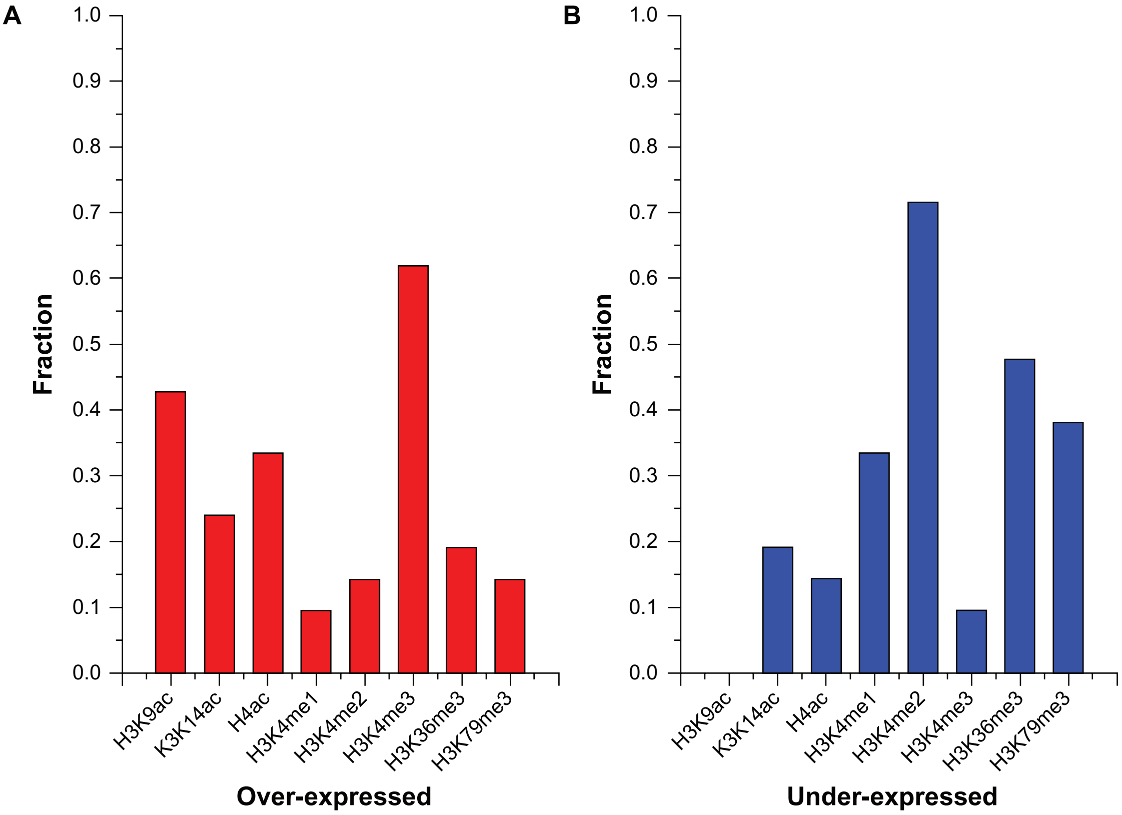

Figure 4. Fraction of over-expressed or under-expressed state at sites from extracted 21 rules for Pokholok et al's data.
Notes: Over-expressed state corresponds to 'X3' and under-expressed state corresponds to 'X1' in the extracted rules, where 'X' is the column number.

Cross-talk between (H3K9ac|H3K14ac|H4ac), H3K4me2, H3K4me3 and H3K79me3

From rule1 to rule4 ([81, 13, 63 ⇒ 51], [81, 13, 51 ⇒ 63], [81, 23, 51 ⇒ 63], [81, 33, 51 ⇒ 63], see Table 2), we found that there was a similar pattern among them, which can be formulated as H3K9ac [+] or H3K14ac [+] or H4ac [+], H3K4me2 [–], H3K4me3 [+] and H3K79 [–]. Since association rules were extracted from FMTs whose target genes have higher EC scores, the above synergic patterns are indispensable to active genes. During gene activation state, TFs at the upstream activator sites recruit positive modifiers, such as histone acetylases (HAT), at the promoter, while DNA-bound RNA polymerase recruits histone methylases at the ORF.[2] In addition, previous studies also showed a correlation between H3K4me3 and RNA polymerase (POL) II initiation.[40] Moreover, acetylation of H3 and H4 or di- or trimethylation (me) of H3K4 are associated with active transcription.[34] Taken together, it is reasonable that the sites of H3 or H4 acetylation and sites of H3K4 methylation coexisted in this pattern, such as H3K9ac, H3K14ac, H4ac and H3K4me2, H3K4me3.

We also found interesting phenomena in above rules. First, H3K79m3 are under-expressed in the above pattern. Since little is known about the function of H3K79me3, the mechanism under this pattern is still unknown. Next we observed that the level of H3K4me3 is over-expressed while H3K4me2 is under-expressed. It is unclear, however, what reasons lead to this exclusive relationship between H3K4m2 and H3K4me3. We speculated that there may be two possibilities behind this phenomenon. First, like their roles in transcription, H3K4me2 is related with potential for gene activity, whereas H3K4me3 is related with gene activity. Second, it is also possible that H3K4me2 and H3K4me3 are at different methylation state,[34] therefore, there are exclusive at active genes. Then we also found that H3K4me3 always was associated with H3 acetylation. The reasons are described as follows: One the one hand, because Gcn5 HAT complex stimulates H3K4me3, H3 acetylation are associated with H3K4 metylation; on the other hand, Chd1 chromodomain links H3K4 methylation with H3 acetylation.[39] The above evidence highlights the interplay that can occur between acetylation of H3 or H4, H3K4me2, H3K4me3 and H3K79me3.

Cross-talk between (H3K4me2|H3K4me3) and (H3K9ac|H3K14ac|H4ac)

This pattern is more universal than above pattern, there are 9 rules (rule1–4, rule9–12, rule15) which satisfy this pattern. In this pattern, H3K36me2 and H3K36me3 are mutually exclusive, while H3K9ac and H3K14ac and H4ac are exclusive with each other. As described above, both H3K4me2 and H3K4me3 are enriched at actively transcribed genes, which is supported by analyses of the different H3K4 methylation states and their distribution in various organisms.[8] Furthermore, H3K4me3 always promote H3 or H4 acetylation. Taken together, this pattern is reasonable and easy to interpret.

Interestingly, this pattern is highly consistent with notable patterns of coexisting histone marks in human cells, such as 'H3K4me2/3 + H4K16ac' and 'H3K4me2/3 + H3K9/14/18/23ac', which were supported by chromatin immunoprecipitatation (ChIP) and mass spectrometry (MS) method.[41] Specifically, from the pattern (H3K4me2|H3K4me3) + (H3K9ac|H3K14ac|H4ac) in yeast, we can see that H3K4me2 or H3K4me3 cooperate with H3K9ac or H3K14ac or H4K5acK8acK12acK16ac (H4ac referring to non-specific acetylation on any of the four sites, eg, H4K5, H4K8, H4K12, H4K16). The above pattern is highly consistent with pattern 'H3K4me2/3 + H4K16ac' in human cells, and apart from H3K18ac and H3K23ac sites, the other pattern 'H3K4me2/3 + H3K9/14/18/23ac' in human cells is also consistent with that yeast pattern. In general, the above yeast pattern nearly corresponds to combination of two patterns of human cells. Therefore, we deduced that there are significant correspondence between yeast pattern and pattern of human cells.

There are several similar histone modification characteristics between human cells and budding yeast, such as, association between H3K4me3 and RNA POL II initiation, association between H3K36me3 and elongation, association between H3K4 methylation and actively transcribed genes, etc.[3,40] Therefore, one possible interpretation is that this pattern can be applicable to human cells. Nevertheless, further experimental methods need to be done to validate this pattern.

Cross-talk between (H3K14ac|H4ac), H3K4me2, H3K36me3 and H3K79me3

We also observed similar pattern form rule16 $(41, 33, 73, 83 \Rightarrow 53)$ and rule21 $(41, 23, 73, 53 \Rightarrow 83)$. This pattern is consistent with activation process of gene expression. During activation state, DNA-bound activator firstly recruits histone acetylases (HAT) at the promoter, whereas DNA-bound RNA POL recruits histone methylases at the ORF. Secondly, early in elongation, phosphorylation of C-terminal domain (CTD) polymerase result in recruiting COMPASS complex, part of which (Set1) methylates H3K4. Last, later during elongation, phosphorylation of the CTD results in recruiting Set2 methyltransferase which methylates H3K36.[2] Besides, H3K36m3 and H3K79me3 are implicated in activation of transcription, and histone H3K36 methylation mediated by Set2 is an important landmark on chromatin during elongation.[5,34] Thus, it is reasonable in above extracted pattern. Since little is known about the function of metylation of H3K79, the exact mechanism by which H3K36me3 cooperates with H3K79me3 remains to be determined.

Association rules derived from Kurdistani et al's data

In addition to Pokholok et al's data, we also tested our method on Kurdistani et al's data, and several obtained relationships are validated. In this dataset there are 11 acetylation sites, including three lysines in histone H4 (K8, K12, K16), four in histone H3 (K9, K14, K18, K23), one in H2A (K7), and two in histone H2B (K11, K16). Here, we utilized acetylation data for intergenic regions (Materials and Methods). Among 3^{11} transactions, we filtered out transactions whose number of target genes is less than 20. Then we got 10941 transactions, and after filtered by EC scores (0.08), 2193 transactions were obtained. We obtained 594 N_1 transactions and 760 N_2 transactions, and then we got $N_1 + N_2 = 1354$ FMTs (Supplementary file 3). Because H3K18 ('6X') accounts for the largest proportion of transactions among FMTs and H3K18 was deemed as a general mark of active transcription (see below), we eliminated rules which contained '6X' (X = 1 or 3) to highlight histone modification patterns of other modification sites. We extracted 188 rules from the 1354 FMTs. After deleting rules including '6X', we obtained 69 rules. Due to space limits, here we only presented 32 rules (Table 3) whose supports are greater than or equal to one, the rest rules are presented in additional file 2.

Overview of the extracted rules

As shown in Figure 3, we found that '63' (H3K18) is the largest abundance with the ratio of 71% (968 of 1354 FMTs). Furthermore, H3K18 was identified as the most widely regulated acetylation site, and its acetylation appears to be a general mark of active transcription,[12,27] thus we deem that H3K18 perform its function in a global manner. On the other hand, from the overall under-expressed state of extracted 69 rules (Fig. 5, after elimination of H3K18), we found that H4K16 accounts for the largest proportion among extracted rules. The reason may be that the number of TFs to regulate H4K16 is the least among 11 acetylation sites, which is only 2 TFs to regulate H4K16 contrast to 15 TFs for H3K18.[27]

Cross-talk involving unacetylation of H4K16 and acetylation at other sites

From Table 3, we found that there was a common item (31) from a large majority of extracted rules. Therefore, the item of '31', ie, unacetylation of H4K16, is important for these histone modifications and plays a special role in transcription. Furthermore, the unacetylated H4K16 has an activating role in euchromatin and unacetylated H4K16 in the context of acetylation at other sites is important for transcription.[6] For example, among 32 extracted rules (supports are greater than or equal to 1, see Table 3), there are 27 rules which are consistent with role of unacetylation of H4K16 in the context of acetylation at other sites, such as, H4K8, H3K23, H3K9 and so forth.

Discussion

We have proposed a new method to identify the combinatorial effects of histone modification by association rule mining. The method is based on association rules discovery technique and can find individual functional sites as well as the combinations of them. Through the analysis of two histone modification datasets, we extracted a number of rules some of which are strongly supported by previous studies. The drawback of association rule mining is that the

Table 3. Rules extracted from Kurdistani et al's data.

Rule	Antecedent	Consequent	Supp. (%)	Conf. (%)	Lift (%)	*P*-value	Corrected *P*-value
Rule1	**13 73**	**31**	6.1	100	209.9	8.95E-30	<0.001
Rule2	**13 43**	**31**	3.8	100	209.9	6.66E-19	<0.001
Rule3	**113 93 73**	**31**	2.5	100	209.9	4.32E-13	<0.001
Rule4	**113 13 23**	**31**	2.4	100	209.9	2.43E-12	<0.001
Rule5	41 21 111	71	2.4	84.2	290.1	9.00E-20	0.002
Rule6	**93 43 31**	**73**	2.4	80	464.9	5.48E-23	<0.001
Rule7	41 11 111	71	2.4	80	275.6	3.16E-17	0.018
Rule8	**113 13 93**	**31**	2.2	100	209.9	1.09E-11	<0.001
Rule9	**93 23 73**	**31**	2.2	100	209.9	9.27E-12	<0.001
Rule10	41 21 11	71	2.2	100	344.5	2.20E-23	<0.001
Rule11	**103 13 23**	**31**	2.1	100	209.9	5.16E-11	<0.001
Rule12	**13 93 23**	**31**	2.1	100	209.9	1.86E-11	<0.001
Rule13	**113 93 23**	**31**	2	100	209.9	1.75E-10	<0.001
Rule14	**13 83**	**31**	1.9	100	209.9	2.83E-10	<0.001
Rule15	**103 13 93**	**31**	1.8	100	209.9	2.17E-09	<0.001
Rule16	**113 23 43**	**31**	1.7	100	209.9	5.10E-09	<0.001
Rule17	**113 93 83**	**31**	1.6	100	209.9	8.28E-09	<0.001
Rule18	**113 93 43**	**31**	1.6	100	209.9	2.34E-08	<0.001
Rule19	**103 23 73**	**31**	1.6	91.3	191.7	3.48E-06	<0.001
Rule20	**103 93 73**	**31**	1.5	100	209.9	5.36E-08	<0.001
Rule21	**93 23 43**	**31**	1.5	100	209.9	3.95E-08	<0.001
Rule22	**103 93 23**	**31**	1.4	100	209.9	1.65E-07	<0.001
Rule23	**103 113 73 31**	**13**	1.3	90	1007.1	3.02E-25	<0.001
Rule24	**103 93 73 31**	**13**	1.3	90	1007.1	1.64E-25	<0.001
Rule25	33 91	41	1.2	100	543.8	6.32E-13	0.002
Rule26	**103 73 43 31**	**13**	1.2	100	1119	3.38E-23	<0.001
Rule27	**103 113 23**	**31**	1.2	88.9	186.6	1.30E-04	<0.001
Rule28	**103 13 43 31**	**73**	1.2	80	464.9	4.08E-13	<0.001
Rule29	**103 23 43**	**31**	1	100	209.9	2.53E-05	<0.001
Rule30	**103 113 93**	**31**	1	92.9	194.9	1.74E-04	<0.001
Rule31	33 81	41	1	82.4	447.8	2.89E-08	0.032
Rule32	**103 113 23 31**	**13**	1	81.3	909.2	1.40E-18	<0.001

Notes: The unit position of numbers in the columns of antecedent and consequent are as follows: 1, under-expressed; 3, over-expressed. The other positions of numbers stand for sites, eg, the number from 1 to 11, which corresponds to sites of H4K8, H4K12, H4K16, H3K9, H3K14, H3K18, H3K23, H3K27, H2AK7, H2BK11, H2BK16 respectively. There are 27 rules (in boldface) are consistent with role of unacetylation of H4K16.

number of generated rules is often very large, even if minimum support and confidence are satisfied. To address this question, besides support and confidence adopted, we also adopted following methods: First, we restricted association rule with a single item in the consequent. Second, a high improvement cut-off (is larger than 110%) must be satisfied for each rule, which guarantees strong dependence between antecedent and consequent (Materials and Methods). Last, the other two measures, *P*-value and corrected *P*-value, must be satisfied.

The study obtained two sets of association rules and patterns from two different datasets with little overlap. We consider that the selection of modification sites is the major cause of this non-overlapping. Specifically, for Pokholok et al's data, there are eight sets of histone modifications (H3K9ac, H3K14ac, H4ac, H3K4me1, H3K4me2, H3K4me3, H3K36me3, and H3K79me3) whereas for Kurdistani et al's data, there are 11 acetylation sites (H4K[8,12,16], H3K [9,14,18,23,27], H2AK7, H2BK [11,16]) without methylation sites. Strictly speaking, there are only two common sites (H3K9ac, H3k14ac) between two datasets (H4ac denotes H4K5ac8ac12ac16ac). In addition to addressing the above questions implicated in this method, we sought to understand how these patterns

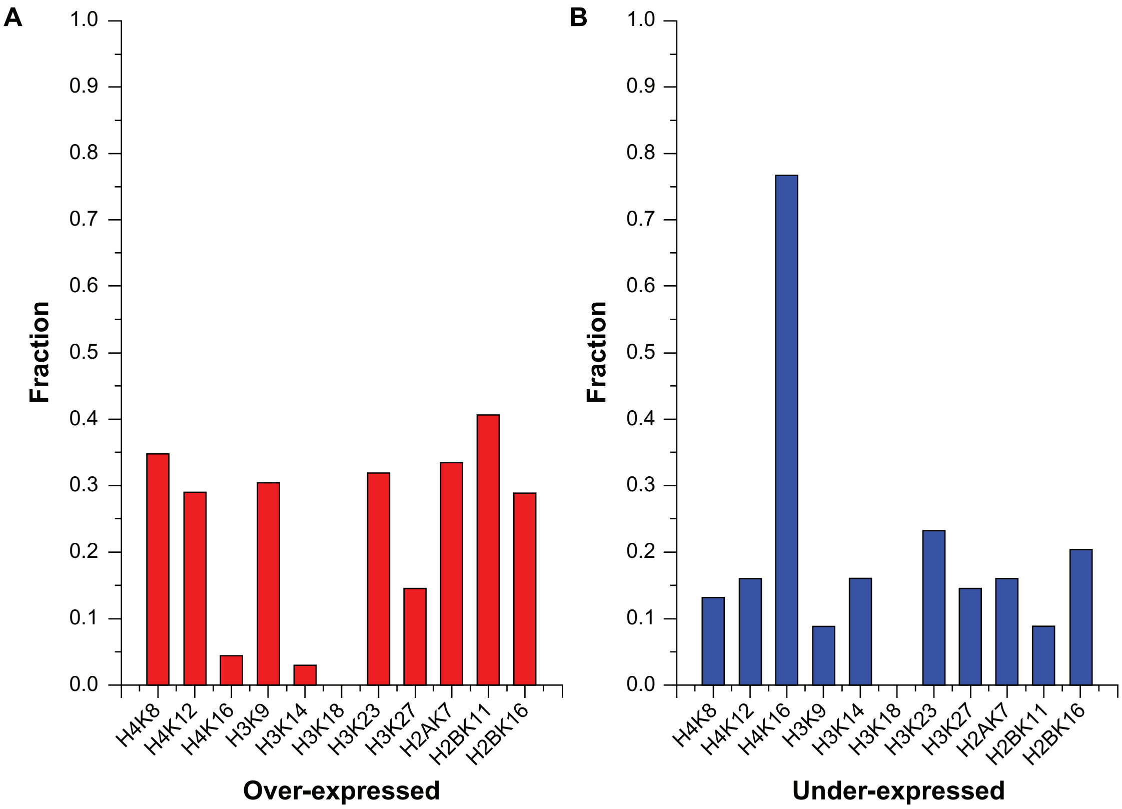

Figure 5. Fraction of over-expressed or under-expressed state at sites from extracted 69 rules for Kurdistani et al's data.
Notes: It's same as in Figure 3. Because of eliminating '6X' (X = 1 or 3) among the extracted rules, there are no value on H3K18 sites ('6X' site).

could influence gene expression, and the mechanisms underlying these combinatorial effects. To our knowledge, histone modification can change transcriptional states through two mechanisms.[42] The first mechanism implicates recruitment of proteins or complexes through binding of their specific protein domains to acetylated or methylated hisones. For example, the bromodomain recognizes acetylated lysine residue whereas chromodomain recognizes methylated lysine residue, which are found in many chromatin-regulating proteins. The second mechanism involves the formation of a less compact nucleosome structure. Because positively charged histone tail interacts with the negatively charged DNA in a nucleosome, acetylation neutralizing the positive charge of the lysine results in a destabilized nuelosome structure. For Pokholok et al's data, we extracted 777 genes from 21 rules, while 857 genes from 32 rules from Kurdistani et al's data. We observed very little overlap (174 genes, hypergeometric $P = 0.001$). These genes (1460 genes) whose expression EC scores are largely attributed to histone modification patterns (the union of above two gene sets) show significant nucleosome fuzziness[43] in their promoters ($P < 3 \times 10^{-4}$, Kolmogorov-Smirnov Test, between these genes and the rest of genes). We also test turnover rates of histone H3,[44] but they do not show significant level ($P = 0.72$, Kolmogorov-Smirnov Test). In order to identify the influence of chromatin modifiers on chromatin structure, we examined six chromatin regulators[32] on these genes, such as, SWI/SNF (Swi1 Snf2), SWR (Swr1 Swc2 Swc6 Yaf9 Arp4), INO80 (Ino80 Arp4 Arp8 Ies6), RSC (Rsc3 Rsc8 Rsc30), Isw2, ISWI (Ioc4 Isw1), and these regulators do not have a significant influence on chromatin structure of these genes (data not shown). Taken together, these genes whose expression are largely dependent on histone modification patterns, with significant nucleosome fuzziness, nonsignificant turnover rates of histone H3, and few chromatin regulators involved, suggesting their chromatin structure are mainly altered through nucleosome sliding rather than nucleosome displacement, belong to the second mechanism.

In addition, the genes (1460 genes) mentioned above contain a high percentage of genes (hypergeometric $P < 0.02$) with TATA boxes in their promoters,[45] consistent with a report showing that TATA-containing genes are often regulated by chromatin. Importantly, we could infer the possible reading mechanism of a variety of modification states from deduced modification patterns: (1) the redundant class in which some modification sites are functionally redundant. For instance, among the relationships between (H3K9ac|H3K14ac|H4ac), H3K4me2, H3K4me3 and H3K79me3, the acetylation of H3K9, H3K14, and H4ac seem to be redundant from one another. One possible interpretation is that the either the acetylation of H3 or H4 could be sufficient to combine the H3K4me2, H3K4me3 and H3K79me3 for active genes; (2) the trivial class, where gene expression is independent of its modification states, is consistent with previous report. From the method we proposed, we could observed that only 1460 genes have significant influence on gene expression, while gene expression of the rest of genes are uninfluenced by their modification states; (3) the dominative class, where several distinctive modification sites are indispensable for a group of modification patterns. Such as, a bunch of association rules which mainly composed of over-expressed state of H3K18 or unacetylation of H4K16 belong to this class.

Interestingly, there are also a few novel rules which contradict with extracted patterns. For example, rule5 (13, 71, 33, 63 ⇒ 51) and rule6 (13, 71, 23, 63 ⇒ 51) have either H3K9ac and H4ac or H3K9ac and H3K14ac in the antecedent, which contradict with the above statement that these acetylation sites are exclusive for each other. One possible explanation could be that they are mutually exclusive in some patterns, but not in others; another possible explanation is that they are exclusive, and the contradiction is due to noise. Which statement is correct remains to be elucidated. Interestingly, there exist common modification patterns between yeast and human cells, which could guide further experimental methods to explore histone modification patterns. Besides Chip-chip, Chip-SAGE and Chip-Seq experimental approaches, knockout experiment is a good choice to study modification patterns.

Conclusions

We have developed a method for the analysis of histone modification data, which can identify the combinatorial effects of histone modifications. This approach is based on the association rules discovery technique and statistical hypothesis testing method. We applied this algorithm to two different datasets and generated many interesting rules and patterns. From Pokholok et al's data, we acquired 21 rules. From the landscape of extracted rules, we found the dominant signal of H3Kme3, which could be supported by many literatures. Also, we extracted three typical patterns, such as, '(H3K9ac|H3K14ac|H4ac) + H3K4me2 + H3K4me3 + H3K79me3', '(H3K4me2|H3K4me3) + (H3K9ac|H3K14ac|H4ac)', and '(H3K14ac|H4ac), H3K4me2 + H3K36me3 + H3K79me3', which also can be explained by previous studies. Most interestingly, our results demonstrated a very good correspondence between yeast and human cells regarding the dominant role of H3K4me3 and the pattern '(H3K4me2|H3K4me3)+(H3K9ac|H3K14ac|H4ac)'. We also obtained several global factors from Kurdistani et al's data. Because over-expressed state of H3K18 occurs in large majority of rules and its acetylation appears to be a general mark of active transcription, we deem that H3K18 perform its function in a global manner. Due to the global role of H3K18, we eliminated rules which contained '6X' (X = 1 or 3) to highlight histone modification patterns of other modification sites. After elimination of H3K18, we found synergy between unacetylation of H4K16 and acetylation at other sites, which also be supported by literatures. Furthermore, these novel patterns we extracted lay a useful foundation for the additional experiments necessary to gain a fuller understanding of the roles of combinations of histone modifications in gene expression. More importantly, the method used here to identify the combinatorial effects of histone modifications can also be used to gain insights into relationships between histone modification sites across the genome in other higher eukaryotes.

Acknowledgements

JW and XD designed the study, analyzed the results and drafted the manuscript, and JW also implemented the algorithms, carried out the experiments. XQ, YD, JF, ZD and CH participated in the analysis and discussion.

Abbreviations

MS, Mass Spectrometry. TF, Transcription Factor; EC, Expression Coherence, the mean of all pair-wise Pearson correlation coefficients for the expression profiles of all the genes within a transaction; FMT, Functional Modification Transactions, for which histone modification combinations are functional and have influence on their gene expression; ChIP-chip, ChIP experiments (chromatin immunoprecipitation) followed by chip (DNA microarrays) to profile protein targeting or chromatin modifications over large genomic regions;[3] ChIP-Seq, the combination of ChIP experiments with high-throughput sequencing to analyse chromatin modifications;[3] ChIP-SAGE, Chromatin immunoprecipitation (ChIP) combined with serial analysis of gene expression.[3]

Disclosures

This manuscript has been read and approved by all authors. This paper is unique and is not under consideration by any other publication and has not been published elsewhere. The authors and peer reviewers of this paper report no conflicts of interest. The authors confirm that they have permission to reproduce any copyrighted material.

References

1. Schreiber SL, Bernstein BE. Signaling network model of chromatin. *Cell*. 2002 Dec 13;111(6):771–8.
2. Berger SL. The complex language of chromatin regulation during transcription. *Nature*. 2007 May 24;447(7143):407–12.
3. Schones DE, Zhao K. Genome-wide approaches to studying chromatin modifications. *Nat Rev Genet*. 2008 Mar;9(3):179–91.
4. Rando OJ. Global patterns of histone modifications. *Curr Opin Genet Dev*. 2007 Apr;17(2):94–9.
5. Kouzarides T. Chromatin modifications and their function. *Cell*. 2007 Feb 23;128(4):693–705.
6. Shahbazian MD, Grunstein M. Functions of site-specific histone acetylation and deacetylation. *Annu Rev Biochem*. 2007;76:75–100.
7. Kurdistani SK, Grunstein M. Histone acetylation and deacetylation in yeast. *Nat Rev Mol Cell Biol*. 2003 Apr;4(4):276–84.
8. Martin C, Zhang Y. The diverse functions of histone lysine methylation. *Nat Rev Mol Cell Biol*. 2005 Nov;6(11):838–49.
9. Strahl BD, Allis CD. The language of covalent histone modifications. *Nature*. 2000 Jan 6;403(6765):41–5.
10. Fischer JJ, Toedling J, Krueger T, Schueler M, Huber W, Sperling S. Combinatorial effects of four histone modifications in transcription and differentiation. *Genomics*. 2008 Jan;91(1):41–51.
11. Guo X, Tatsuoka K, Liu R. Histone acetylation and transcriptional regulation in the genome of Saccharomyces cerevisiae. *Bioinformatics*. 2006 Feb 15;22(4):392–9.
12. Kurdistani SK, Tavazoie S, Grunstein M. Mapping global histone acetylation patterns to gene expression. *Cell*. 2004 Jun 11;117(6):721–33.
13. Pokholok DK, Harbison CT, Levine S, et al. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell*. 2005 Aug 26;122(4):517–27.
14. Liu CL, Kaplan T, Kim M, et al. Single-nucleosome mapping of histone modifications in S. cerevisiae. *PLoS Biol*. 2005 Oct;3(10):e328.
15. Rakesh A, Tomasz I, ski, Arun S. Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*. Washington, D.C., United States: ACM; 1993.
16. Rakesh A, Ramakrishnan S. Fast Algorithms for Mining Association Rules in Large Databases. *Proceedings of the 20th International Conference on Very Large Data Bases*: Morgan Kaufmann Publishers Inc.; 1994.
17. Jong Soo P, Ming-Syan C, Philip SY. An effective hash-based algorithm for mining association rules. *Proceedings of the 1995 ACM SIGMOD international conference on Management of data*. San Jose, California, United States: ACM; 1995.
18. Hannu T. Sampling Large Databases for Association Rules. *Proceedings of the 22th International Conference on Very Large Data Bases*: Morgan Kaufmann Publishers Inc.; 1996.
19. Ashoka S, Edward O, Shamkant BN. An Efficient Algorithm for Mining Association Rules in Large Databases. *Proceedings of the 21th International Conference on Very Large Data Bases*: Morgan Kaufmann Publishers Inc.; 1995.
20. Banu, zden, Sridhar R, Abraham S. Cyclic Association Rules. *Proceedings of the Fourteenth International Conference on Data Engineering*: IEEE Computer Society; 1998.
21. Ashoka S, Edward O, Shamkant BN. Mining for Strong Negative Associations in a Large Database of Customer Transactions. *Proceedings of the Fourteenth International Conference on Data Engineering*: IEEE Computer Society; 1998.
22. Sridhar R, Sameer M, Abraham S. On the Discovery of Interesting Patterns in Association Rules. *Proceedings of the 24rd International Conference on Very Large Data Bases*: Morgan Kaufmann Publishers Inc.; 1998.
23. Gavin S, Yue X, Shlomo G. Deriving non-redundant approximate association rules from hierarchical datasets. *Proceeding of the 17th ACM conference on Information and knowledge management*. Napa Valley, California, USA: ACM; 2008.
24. Shaw G, Xu Y, Geva S. Eliminating redundant association rules in multi-level datasets. *Proceedings of the 4th International Conference on Data Mining (DMIN' 08)*. Las Vegas, USA: 2008.
25. Sergey B, Rajeev M, Craig S. Beyond market baskets: generalizing association rules to correlations. *Proceedings of the 1997 ACM SIGMOD international conference on Management of data*. Tucson, Arizona, United States: ACM; 1997.
26. Bing L, Wynne H, Yiming M. Pruning and summarizing the discovered associations. *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. San Diego, California, United States: ACM; 1999.
27. Pham H, Ferrari R, Cokus SJ, Kurdistani SK, Pellegrini M. Modeling the regulatory network of histone acetylation in Saccharomyces cerevisiae. *Mol Syst Biol*. 2007;3:153.
28. Bernstein BE, Liu CL, Humphrey EL, Perlstein EO, Schreiber SL. Global nucleosome occupancy in yeast. *Genome Biol*. 2004;5(9):R62.
29. Gasch AP, Spellman PT, Kao CM, et al. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*. 2000 Dec;11(12):4241–57.
30. Spellman PT, Sherlock G, Zhang MQ, et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*. 1998 Dec;9(12):3273–97.
31. Harbison CT, Gordon DB, Lee TI, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*. 2004 Sep 2;431(7004):99–104.
32. Steinfeld I, Shamir R, Kupiec M. A genome-wide analysis in Saccharomyces cerevisiae demonstrates the influence of chromatin modifiers on transcription. *Nat Genet*. 2007 Mar;39(3):303–9.
33. Yuan GC, Ma P, Zhong W, Liu JS. Statistical assessment of the global regulatory role of histone acetylation in Saccharomyces cerevisiae. *Genome Biol*. 2006;7(8):R70.
34. Li B, Carey M, Workman JL. The role of chromatin during transcription. *Cell*. 2007 Feb 23;128(4):707–19.
35. Carmona-Saez P, Chagoyen M, Rodriguez A, Trelles O, Carazo JM, Pascual-Montano A. Integrated analysis of gene expression by Association Rules Discovery. *BMC Bioinformatics*. 2006;7:54.
36. Boyle EI, Weng S, Gollub J, et al. GO: TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*. 2004 Dec 12;20(18):3710–5.
37. Shilatifard A. Molecular implementation and physiological roles for histone H3 lysine 4 (H3K4) methylation. *Curr Opin Cell Biol*. 2008 Jun; 20(3):341–8.

38. Wysocka J, Swigut T, Milne TA, et al. WDR5 associates with histone H3 methylated at K4 and is essential for H3K4 methylation and vertebrate development. *Cell*. 2005 Jun 17;121(6):859–72.
39. Latham JA, Dent SY. Cross-regulation of histone modifications. *Nat Struct Mol Biol*. 2007 Nov 5;14(11):1017–24.
40. Mendenhall EM, Bernstein BE. Chromatin state maps: new technologies, new insights. *Curr Opin Genet Dev*. 2008 Apr;18(2):109–15.
41. Ruthenburg AJ, Li H, Patel DJ, Allis CD. Multivalent engagement of chromatin modifications by linked binding modules. *Nat Rev Mol Cell Biol*. 2007 Dec;8(12):983–94.
42. Dion MF, Altschuler SJ, Wu LF, Rando OJ. Genomic characterization reveals a simple histone H4 acetylation code. *Proc Natl Acad Sci U S A*. 2005;102(15):5501–6. Epub 2005 Mar 5528.
43. Mavrich TN, Ioshikhes IP, Venters BJ, et al. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res*. 2008;18(7):1073–83. Epub 2008 Jun 1012.
44. Dion MF, Kaplan T, Kim M, Buratowski S, Friedman N, Rando OJ. Dynamics of replication-independent histone turnover in budding yeast. *Science*. 2007;315(5817):1405–8.
45. Basehoar AD, Zanton SJ, Pugh BF. Identification and distinct regulation of yeast TATA box-containing genes. *Cell*. 2004;116(5):699–709.

Supplementary Material

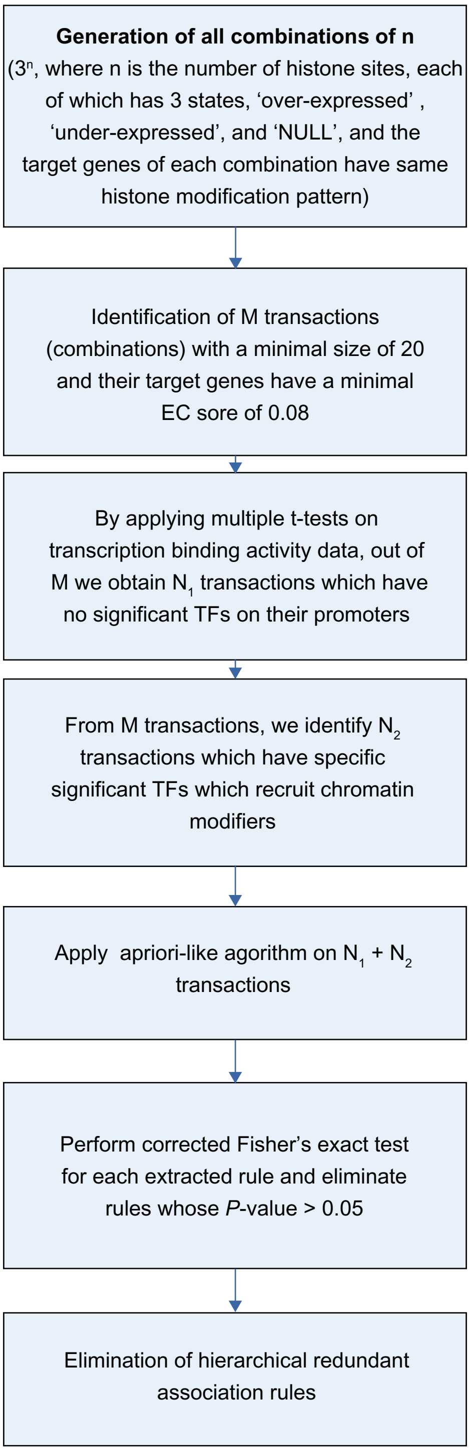
Generation of all combinations of n (3ⁿ, where n is the number of histone sites, each of which has 3 states, 'over-expressed' , 'under-expressed', and 'NULL', and the target genes of each combination have same histone modification pattern)
Identification of M transactions (combinations) with a minimal size of 20 and their target genes have a minimal EC sore of 0.08
By applying multiple t-tests on transcription binding activity data, out of M we obtain N₁ transactions which have no significant TFs on their promoters
From M transactions, we identify N₂ transactions which have specific significant TFs which recruit chromatin modifiers
Apply apriori-like agorithm on N₁ + N₂ transactions
Perform corrected Fisher's exact test for each extracted rule and eliminate rules whose P-value > 0.05
Elimination of hierarchical redundant association rules

Figure S1. Flowchart of the method.

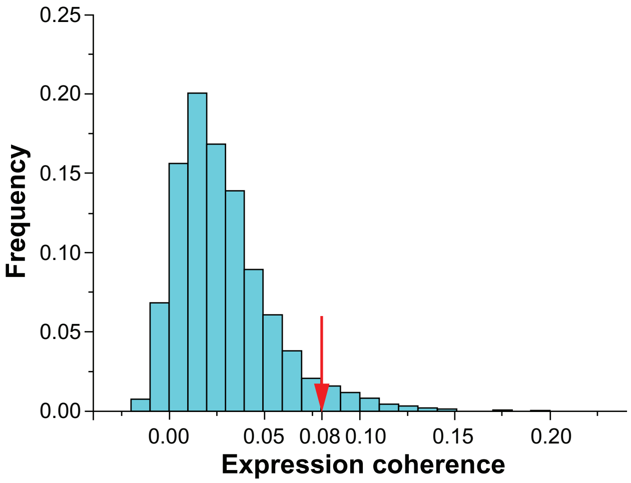

Figure S2. Identifying the threshold of EC scores.
Notes: The histogram is obtained based on 100,000 samples with the size of 20. The arrow on the right marks the threshold of EC scores, which is statistically significant (95% confidence).

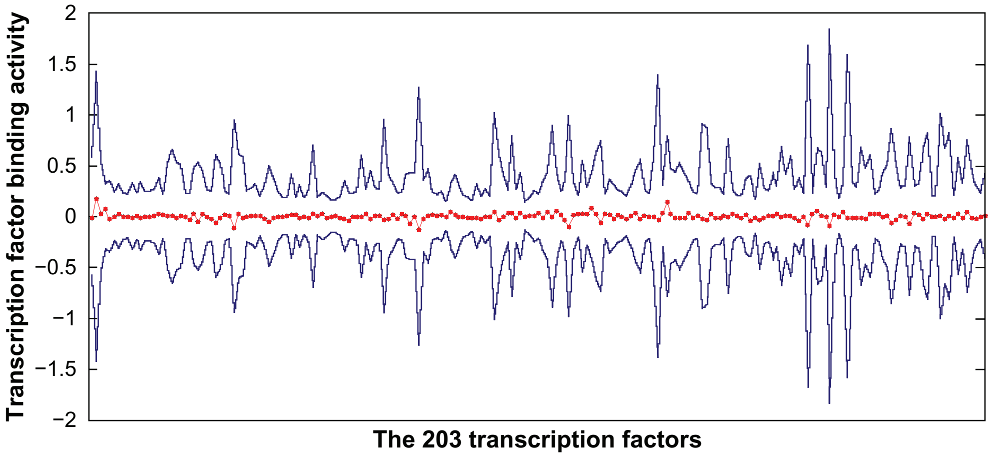

Figure S3. The transcription factor binding profile in transaction {61, 73, 83} among N_1 transaction set for Pokholok et al's data.
Notes: The transaction {61, 73, 83} belongs to N_1 transaction set, which have no significant TFs in the promoter regions through multiple t-tests. The y-axis shows the TF binding activity which is relative to their mean in the genome. The genomic mean of binding activity for all 203 TFs is normalized to zero. The upper blue curve is genomic mean binding activity plus its standard deviation, while the lower curve is genomic mean binding activity minus its stand deviation. The red line is the mean of transcription factor binding activity for transaction {61, 73, 83}.

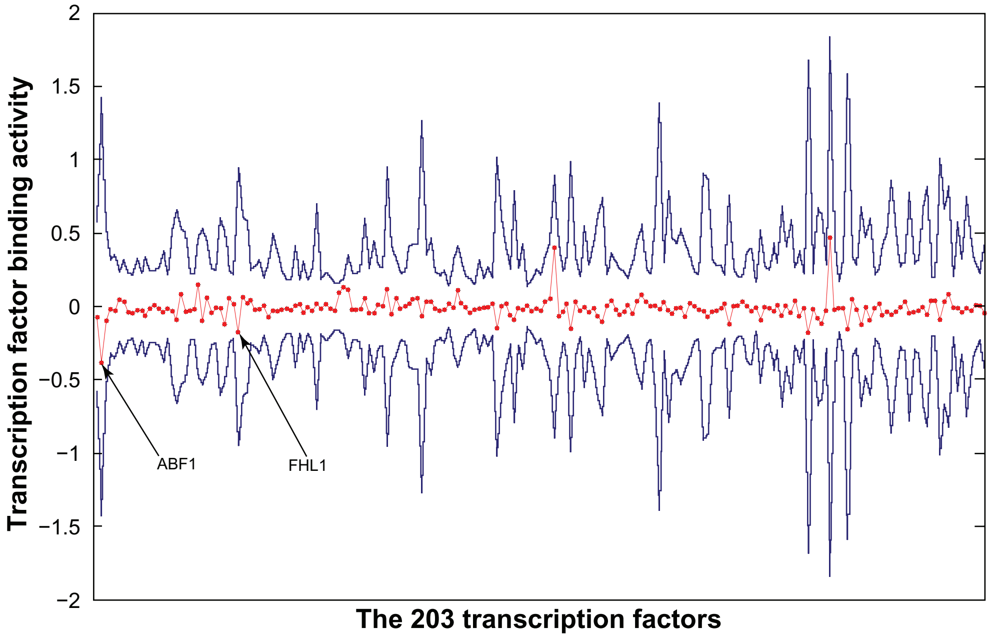

Figure S4. The transcription factor binding profile in transaction {33, 43, 53, 71, 81} among N_2 transaction set for Pokholok et al's data.
Notes: Same as Figure S3, except that the transaction {33, 43, 53, 71, 81} within N_1 transaction set is illustrated. This transaction has two significant TFs (ABF1, FHL1) in the promoter regions.