Computational Biology and Chemistry 50 (2014) 3–10

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

Deciphering histone code of transcriptional regulation in malaria parasites by large-scale data mining

Haifen Chen[a], Stefano Lonardi[b], Jie Zheng[a,c,*]

[a] School of Computer Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Singapore
[b] Department of Computer Science and Engineering, University of California Riverside, 900 University Avenue, Riverside, CA 92521, USA
[c] Genome Institute of Singapore, A*STAR (Agency for Science, Technology, and Research), Biopolis, Singapore 138672, Singapore

ARTICLE INFO

Article history:
Accepted 23 December 2013
Available online 23 January 2014

Keywords:
Histone code
Histone modification
Plasmodium falciparum
Association rule mining
Gene expression prediction

ABSTRACT

Histone modifications play a major role in the regulation of gene expression. Accumulated evidence has shown that histone modifications mediate biological processes such as transcription cooperatively. This has led to the hypothesis of 'histone code' which suggests that combinations of different histone modifications correspond to unique chromatin states and have distinct functions. In this paper, we propose a framework based on association rule mining to discover the potential regulatory relations between histone modifications and gene expression in *Plasmodium falciparum*. Our approach can output rules with statistical significance. Some of the discovered rules are supported by literature of experimental results. Moreover, we have also discovered *de novo* rules which can guide further research in epigenetic regulation of transcription. Based on our association rules we build a model to predict gene expression, which outperforms a published Bayesian network model for gene expression prediction by histone modifications.

The results of our study reveal mechanisms for histone modifications to regulate transcription in large-scale. Among our findings, the cooperation among histone modifications provides new evidence for the hypothesis of histone code. Furthermore, the rules output by our method can be used to predict the change of gene expression.

1. Introduction

1.1. Background

Understanding the epigenetic regulation of transcription by histone modifications is a central problem in molecular biology. The basic unit of chromatin is nucleosome, which consists of an octamer of four core histone proteins around which 147 base pairs of DNA is wrapped. The core histones are subject to many covalent modifications including methylation and acetylation. These histone modifications play essential roles in many biological processes such as transcriptional regulation (Li et al., 2007), X-chromosome inactivation (Brinkman et al., 2006), meiotic recombination hotspots (Wu et al., 2012), and cancer (Esteller, 2007). The activation and repression of gene expression are mediated by histone modifications through direct structural changes of chromatin or recruiting effector proteins (Berger, 2007). In 2001, the hypothesis of 'histone code' was proposed that the combinatorial patterns of histone modifications of different types, positions and times dictate the dynamics of RNA transcription (Jenuwein and Allis, 2001). In the following decade, the histone code has been under intense research (Cheng et al., 2011; Fischle et al., 2003; Kimura et al., 2004; Margueron et al., 2005; Xu et al., 2010). Nevertheless, the existence of the histone code remains an open question. As related high-throughput data (e.g. microarray, RNA-seq, ChIP-chip, and ChIP-seq) become available, it is a good opportunity for computational biologists to crack the histone code by large-scale data analysis and modelling.

In this paper, we aim to study the transcriptional regulation of the most deadly malaria parasite, *Plasmodium falciparum*, which claims about a million human deaths worldwide each year (Nayyar et al., 2012). Despite intense research for decades (Duraisingh et al., 2005; Gupta et al., 2013; van Noort and Huynen, 2006), the mechanism of gene regulation of this species remains a mystery. *Plasmodium* genomes are known for AT-richness, lack of transcription factors, and its unique cyclic patterns of gene expression along its life cycle (Gardner et al., 2002). As there is no evidence of DNA methylation in *P. falciparum*, its histone modifications are likely to be more critical for transcriptional regulation than other species. Recently, there have been efforts to uncover the histone code for *P. falciparum*, which implicate some histone-based mechanisms conserved with other species such as yeast and human (Cabral

* Corresponding author at: School of Computer Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Singapore.
E-mail addresses: hchen009@e.ntu.edu.sg (H. Chen), stelo@cs.ucr.edu (S. Lonardi), zhengjie@ntu.edu.sg (J. Zheng).

1476-9271/$ – see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiolchem.2014.01.002

et al., 2012; Cui and Miao, 2010). However, most of them are experimental works focused on only a few gene families crucial for virulence (e.g. *var*, *rif* genes). There is a need for Bioinformatics approaches to systematically elucidating the histone code for the malaria parasite in large scale.

Here, we apply association rule mining to reveal the regulatory relations between histone modifications and gene expression of *P. falciparum*. As a well-studied technique of data mining, association rule mining was firstly proposed to discover the regularities between products in large transactional database of supermarkets (Agrawal et al., 1993). It has also been applied to problems in Bioinformatics (Chen and Chen, 2006; Morgan et al., 2007; Lopez et al., 2008; Wang et al., 2010). Results from Chen's research (Chen and Chen, 2006) show that association rule mining is a promising technique in biological data analysis, which inspires our work. In Wang et al. (2010), association rule mining was applied on histone modifications data of yeast to identify histone modification patterns that might have effects on transcriptional states. However, the rules they generated were among histone modifications themselves instead of between histone modifications and gene expression. By contrast, in our paper the histone modifications and gene expression have been synchronized for each specific gene and time point, and we identify the association rules directly between histone modifications and expression levels of the same genes. The rules discovered as such are more likely to directly uncover the histone code for gene regulation.

We proposed here a framework based on association rule mining to systematically discover rules for histone code in *P. falciparum*. Our method can generate rules which explicitly show how combinatorial histone modifications dictate gene expression levels. Some of the rules have already been reported in literature. Moreover, our experiments showed that, the prediction of gene expression based on our rules is more consistent with observed gene expression than Bayesian network. Our method is robust to changing distribution in data items. To our knowledge, this is the first computational method tailored for the malaria parasite of *P. falciparum*, which takes into account the dynamic change of both gene expression and histone modifications over time points. Our results are encouraging to decipher the histone code and offer valuable insights into the study of epigenetic regulatory mechanisms in transcription.

1.2. Related works

Bioinformatics approaches have been proposed to study the relationships between histone modifications and gene expression. Those approaches could be classified into three categories: network-based methods, clustering, and regression models.

In 2008, Yu et al. (2008) built a Bayesian network to infer causal and combinatorial relationships between gene expression and histone modifications. Yu's method can generate tree-like rules between histone modifications and gene expression, which can be viewed as the primary shape of histone code. However, they did not predict gene expression using those rules. In 2012, a correlation-based network of histone modifications, DNA methylation and gene expression was constructed by Su et al. (2012). Although this network can explain some aspects of epigenetic regulation of gene expression, their study can only serve as preliminary results for histone code.

A semi-supervised biclustering method was proposed by Teng and Tan (2012) to find the combinatorial chromatin modifications which are correlated with gene expression in human enhancers. Similar application of clustering methods can also be found in Ha et al. (2011). Although the clustering methods are good at capturing "big pictures", they need further analysis to provide detailed information for histone code.

Regression models are used frequently in studying the relationships between histone modifications and gene expression. Karlić et al. (2010) utilized linear regression techniques to demonstrate systematically that histone modifications levels can predict gene expression accurately. Similar applications of linear regression models can also be found in McLeay et al. (2012), do Rego et al. (2012), Xu et al. (2010). Other regression methods based on machine-learning models have also been proposed in this field, e.g., support vector machine (SVM) (Cheng et al., 2011; Cheng and Gerstein, 2012), Random Forests (RF) and so on (Dong et al., 2012). Although these regression methods can show how well histone modifications can predict gene expression, they are unable to demonstrate how the combinations of histone modifications regulate gene expression in an explicit way. Moreover, these model-based methods produce the relationships between histone modifications and gene expression with a "black box" model which makes it difficult for human to understand and interpret the underlying physical processes and biological meanings.

Overall, most existing methods for the study of histone code are either without prediction of gene expression or lack of interpretability. In the following section, we will present our approach which can generate explicit rules describing how combinatorial histone modifications correspond to particular states of gene expression. Our method provides a straightforward way to decipher the histone code without any assumption of models and the tuning of parameters. Furthermore, our rules can predict the trend of gene expression reasonably well.

2. Methods

2.1. Data of P. falciparum

The input data to our analysis are the transcriptional levels and histone modification (HM) enrichment profiles across the *P. falciparum* genome from Rovira-Graells et al. (2012) and Gupta et al. (2013) (with GEO accession number GSE39238). The transcriptional levels were obtained from long oligonucleotides based microarray experiments and the histone modification profiles were from ChIP-on-chip (also known as ChIP-chip) experiments. The dataset includes 14 variables: gene expression (also referred as RNA or cDNA) and 13 histone modification marks, i.e. H4K5ac, H4K8ac, H4K12ac, H4K16ac, H4ac4, H3K9ac, H3K14ac, H3K56ac, H4K20me1, H4K20me3, H3K4me3, H3K79me3, and H4R3me2. The data are time-series, with six time points for each variable, which show the variation of transcription and histone modification

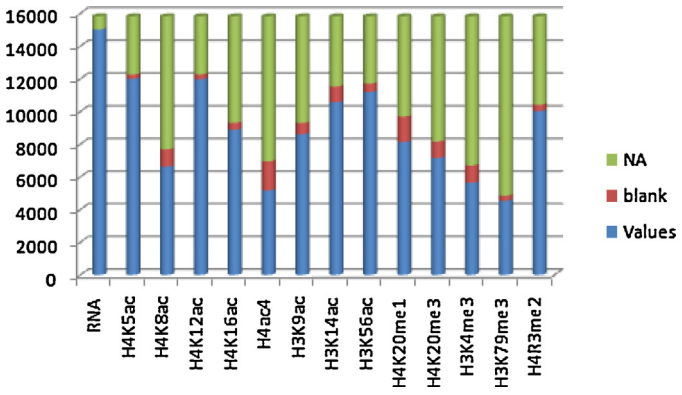

Fig. 1. Distributions of missing data in RNA and different HMs. The vertical axis is the number of probes (for microarray analysis of gene expression and histone modifications), 'NA' means the values are missing for all six time points, 'blank' means there are 2∼5 missing values. Here 'NA' and 'blank' are treated equally as missing data.

profiles along 48 h (six time points, with an interval of 8 h) of *P. falciparum* intra-erythrocytic developmental cycle (IDC). The missing values in this dataset have a distribution as shown in Fig. 1.

2.2. Association rule mining

We apply association rule mining to explore the relationships between combinatorial histone modifications and transcriptional levels. Here we introduce some basics of association rule mining. More details of association rule mining can be found in reviews (Kotsiantis and Kanellopoulos, 2006; Zhao and Bhowmick, 2003).

Association rule mining was firstly proposed to discover interesting relations between variables in large transactional database for market basket analysis. A transactional database is a collection of records each of which is composed of binary values which denote the occurrence of variables, where '1' means the corresponding variable occurs in this transaction and '0' means not. Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of binary attributes (i.e. variables) called *items*, $TD = \{td_1, td_2, \ldots, td_m\}$ be a set of transactions (i.e. records) called *transactional database*. An association rule is represented in the form "$B \Rightarrow A$", where A and B are itemsets (i.e. $A, B \in I$), and $A \cap B = \emptyset$. The concepts of *support* and *confidence* are used the most as constraints for the selection of interesting rules from all possible rules. The *support* of an itemset $supp(A)$ is defined as the proportion of transactions in the dataset TD which contain the itemset A (i.e. A occurs in these transactions). The *confidence* of a rule is defined as $conf(B \Rightarrow A) = supp(AB)/supp(B)$, which is the proportion of the *support* of both A and B to the *support* of B. The *confidence* of "$B \Rightarrow A$" is the estimation of the conditional probability $P(A|B)$, i.e. the probability of finding itemset A in transactions under the condition that those transactions also contain B. In order to identify interesting rules, the thresholds for *support* and *confidence* are usually required, where the lower bound thresholds are specified by the users. It normally consists of two steps: (1) retrieve all frequent itemsets in the database with *support* higher than a threshold; (2) construct rules with *confidence* larger than a threshold from the frequent itemsets found in the previous step.

2.3. Problem formulation

Firstly, we discretize the expression level of each gene at a time point into three states: $-1, 0, 1$, which represent down-regulate ($\downarrow$), unchanged ($-$), up-regulate ($\uparrow$) respectively. We utilize an "*equal-width binning*" technique to discretize the data. The discretization is carried out among the six time points for each variable independently. Then we use a unique integer to represent the state of each variable, as shown in Table 1. By treating each state of each variable as an item, we transform the original dataset into a transactional dataset, with each gene as one transaction. Here we assume that all genes are similar in the sense on how combinations of histone modifications affect the expression level, which is also the assumption of histone code. Let n be the total number of variables in the original dataset, then there would be $3 \times n$ items in the transactional dataset. Since the three states of each variable are mutually exclusive, there would be at most n items in a transaction.

In a transactional database, if one itemset A occurs often enough (with $supp(A)$ larger than a threshold), and when A occurs another itemset B without overlap with A also occurs often (i.e. *confidence* over a threshold), we conclude that A *implies* B (or $A \Rightarrow B$) with significant support. Likewise, for a significant number of genes if there exists a group of histone modifications events, which coincide frequently with a particular state of gene expression, then we conclude that it is a rule of histone code. Note that "*implies*" here does not necessarily mean causal relation. We assume for simplicity that the frequent co-occurrences of HMs with a state of gene expression implicate histone code. Therefore, the problem of deciphering histone code can be solved by association rule mining. The main challenge of association rule mining is to obtain all frequent itemsets from all possible item combinations. Here we use LCM (Uno et al., 2003) to mine all frequent closed item sets which have *support* larger than a pre-set threshold $supp_{min}$, and then obtain all interesting rules with *confidence* higher than our pre-set threshold $conf_{min}$. These rules would reveal regulatory relations between histone modifications and gene expression.

Since we aim to discover the relations between histone modifications (HMs) and transcription level, our rules would look like: $f(HMs^{attr}) \Rightarrow GE^{attr}$. The left-hand side of the rule would be combinations of HMs attributes, while right-hand side corresponds to the attribute of gene expression (GE). For time series, it is natural to find rules where HMs are one-time-point ahead of GE: $f(HMs^{attr})|_{t=t_i} \Rightarrow GE^{attr}|_{t=t_{i+1}}$ because the changing of gene expression is supposed to be the effect of HMs. However, we still search for rules like $f(HMs^{attr})|_{t=t_i} \Rightarrow GE^{attr}|_{t=t_i}$ because in our dataset the interval between two consecutive time points is $t_{i+1} - t_i = 8$ h, which is too long for the effect of a particular HM event to last. Therefore, here we assume that the HMs state remains unchanged when its effect on transcription appears, which is reasonable because HMs are relatively stable albeit dynamic.

2.4. p-Values and multiple-testing correction

Using fixed thresholds of *support* and *confidence* to find interesting rule may be a bit arbitrary, because the selection of such thresholds might depend on the data. Thus we use p-value to assess the significance of rules.

For a rule R: $f(HMs) \Rightarrow GE$, which stands for $f(HMs^{attr}) \Rightarrow GE^{attr}$, its p-value is the probability of observing a rule more extreme (which means with higher *support* and *confidence*) than R under the null hypothesis that the HMs and gene expression are independent, which means the HMs imply the three possible changes of gene expression with equal probabilities. Here we used one-tailed Fisher's exact test (Fisher, 1922; Webb, 2007) to calculate the p-value of R. Let $s_R = supp(R)$ be the *support* of the rule R, and $s_M = supp(HMs)$ be the *support* of the combinatorial HMs (i.e. the left-hand side of the rule). Then the p-value of R is given by:

$$\begin{aligned} p(R) &= p(s_R; n, n_g, s_M) \\ &= \sum_{k \le s_M - s_R} H(k; n, n_g, n - s_M) \\ &= \sum_{k \le s_M - s_R} \frac{\binom{n_g}{k}\binom{n - n_g}{s_M - k}}{\binom{n - s_M}{s_M}} \end{aligned} \tag{1}$$

where n is the total number of records in the dataset, n_g is the number of records that contain the item g, and H is *hypergeometric distribution*. Note that $s_M = supp(HMs)$ is the number (rather than proportion) of transactions that contain the HMs. The value of $s_R = supp(R)$ can be calculated as: $supp(R) = confidence * supp(HMs)$.

Since the number of rules we generate from the previous step could be thousands, we might expect many "false positives" if we use a standard cut-off (e.g. 0.05) for p-values. Therefore, we resort to multiple-testing correction techniques to cope with such a problem. Here a permutation-based approach is applied to adjust the p-values of the rules calculated as above. First we randomly generate N permutations for each rule, by randomly shuffling the class labels of items (i.e. the states of gene expression). Let N_0 be the number of rules generated from the original dataset, and

Table 1
Numbers corresponding to events of gene expression and histone modifications.

Variables	Attributes	Integers
Gene Expr.	↓,–, ↑	1, 2, 3
H4K5ac	↓, –, ↑	4, 5, 6
H4K8ac	↓, –, ↑	7, 8, 9
H4K12ac	↓, –, ↑	10, 11, 12
H4K16ac	↓, –, ↑	13,14, 15
H4ac4	↓, –, ↑	16, 17, 18
H3K9ac	↓, –, ↑	19, 20, 21
H3K14ac	↓, –, ↑	22, 23, 24
H3K56ac	↓, –, ↑	25, 26, 27
H4K20me1	↓, –, ↑	28, 29, 30
H4K20me3	↓, –, ↑	31, 32, 33
H3K4me3	↓, –, ↑	34, 35, 36
H3K79me3	↓, –, ↑	37, 38, 39
H4R3me2	↓, –, ↑	40, 41, 42

$RD = \{p_1, p_2, \ldots, p_{N*N_0}\}$ be the p-values of the N_0 rules with N permutations. For each rule, we re-calculate its p-value as follows: let p be the p-value of this rule, then the new p-value is

$$p_{new} = \frac{|\{p_i | p_i \le p, p_i \epsilon RD\}|}{N * N_0}, \quad (2)$$

where the numerator is the number of elements in RD that are smaller than or equal to p.

2.5. Gene expression prediction based on rules

Based on a set of rules, gene expression of each record in the dataset can be predicted as follows. Here we only consider 'up-rules' (i.e., rules with GE ↑) and 'down-rules' (i.e., rules with GE ↓), because no 'flat-rules' (i.e., rules with GE –) are significant in our dataset. For each record in the dataset, we define an *up-score* S_u and a *down-score* S_d, which represent how strongly this record is supported by 'up-rules' and 'down-rules' respectively. To calculate S_u, let A be the occurrence vector of 'up-rules': $A = [a_1, a_2, \ldots, a_n]^T$, where n is the total number of rules 'up-rules' we have obtained from the whole dataset, and a_i $(1 \le i \le 1)$ equals to 1 when the ith up-rule occurs in this record, and 0 otherwise. Then a weight vector of 'up-rules' is defined as follows: $W = [w_1, w_2, \ldots, w_n]$, where w_i $(1 \le i \le 1)$ equals to the *confidence* of the ith up-rule. Then, the *up-score* S_u is calculated using $S_u = W^* A$. The *down-score* S_d of this record is calculated similarly. After that, gene expression can be predicted in this way: When the *up-score* of one record is larger than its *down-score*, its corresponding gene expression is predicted as 'up-regulated'; otherwise, its gene expression is predicted as 'down-regulated'.

For each gene we classify its original gene expression levels of six time points into six groups: $GE = 1$, $GE = 2$, …, $GE = 6$, which correspond to the lowest, second lowest, …, largest expression value of the gene among the six time points. For each of the six groups, we calculate the average of the observed gene expression and normalize them into the range between 0 and 1. The normalized average expression value is regarded as the *observed* gene expression level of each of the six groups. For the predicted gene expression, we calculate the percentage of genes which are predicted as 'up-regulated' for each of the six groups, and take this percentage value as the *predicted* gene expression level of this group.

3. Results

Applying our method on the transactional dataset of *P. falciparum*, we obtain 2015 rules with *support* and *confidence* larger than 1000 and 0.4 respectively. After calculation and correction of p-values, we extract 60 rules by applying a threshold 0.01 on the corrected p-values (see Additional file 1).

Table 2
Significant association rules of HMs and gene expression (GE).

Rules	Support	Conf.	p-Values
H3K56ac ↑ ⇒ GE ↑	22,108	0.4314	0.000353
H3K56ac ↓ ⇒ GE ↓	22,115	0.4287	0.000938
H4K8ac ↓ ⇒ GE ↓	12,538	0.4471	0.000614
H4K8ac ↑ ⇒ GE ↑	12,409	0.4435	0.000672
H3K4me3 ↓ ⇒ GE ↓	11,699	0.4462	0.001265
H3K4me3 ↑ ⇒ GE ↑	11,648	0.4414	0.00207
H3K9ac ↑ ⇒ GE ↑	16,543	0.4328	0.001355
H3K9ac ↓ ⇒ GE ↓	16,167	0.4332	0.00307
H4K20me1 ↓ ⇒ GE ↓	16,003	0.4288	0.00398
H4K20me1 ↑ ⇒ GE ↑	16,079	0.425	0.005183
H3K9ac ↑, H3K56ac ↑ ⇒ GE ↑	9062	0.4542	0.001008
H3K9ac ↓, H3K56ac ↓ ⇒ GE ↓	8927	0.4527	0.002151
H4R3me2 ↑, H3K56ac ↓ ⇒ GE ↓	6634	0.4543	0.004927
H4R3me2 ↓, H3K56ac ↑ ⇒ GE ↑	6260	0.4594	0.002766
H4K20me1 ↓, H3K56ac ↓ ⇒ GE ↓	5764	0.4585	0.00584
H4K20me1 ↑, H3K56ac ↑ ⇒ GE ↑	5922	0.4589	0.004159
H4K20me1 ↓, H4K5ac ↑ ⇒ GE ↓	5559	0.4588	0.00617
H4K20me1 ↑, H4K5ac ↓ ⇒ GE ↑	5094	0.4628	0.004852
H3K56ac ↓, H4K5ac ↑ ⇒ GE ↓	5030	0.4811	0.001549
H3K56ac ↑, H4K5ac ↓ ⇒ GE ↑	4468	0.4836	0.001719
H3K9ac ↓, H4R3me2 ↑ ⇒ GE ↓	4692	0.4641	0.007133
H3K9ac ↑, H4R3me2 ↓ ⇒ GE ↑	4887	0.4585	0.006956
H3K4me3 ↓, H3K56ac ↓ ⇒ GE ↓	4271	0.4874	0.001855
H3K4me3 ↑, H3K56ac ↑ ⇒ GE ↑	4173	0.468	0.005847
H3K4me3 ↓, H4K5ac ↑ ⇒ GE ↓	3930	0.4666	0.009993
H3K4me3 ↑, H4K5ac ↓ ⇒ GE ↑	3593	0.4823	0.003789
H3K4me3 ↓, H4K8ac ↓ ⇒ GE ↓	3825	0.5071	0.000309
H3K4me3 ↑, H4K8ac ↑ ⇒ GE ↑	3794	0.4815	0.003444
H4K20me1 ↓, H3K9ac ↓ ⇒ GE ↓	3840	0.4731	0.006805
H4K20me1 ↑, H3K9ac ↑ ⇒ GE ↑	4098	0.4729	0.004502
H3K9ac ↓, H4K5ac ↑ ⇒ GE ↓	3776	0.4907	0.002756
H3K9ac ↑, H4K5ac ↓ ⇒ GE ↑	3626	0.4884	0.002425
H3K56ac ↓, H4K12ac ↑ ⇒ GE ↓	3760	0.4702	0.008417
H3K56ac ↑, H4K12ac ↓ ⇒ GE ↑	3999	0.4693	0.006235
H3K4me3 ↓, H3K9ac ↓ ⇒ GE ↓	3436	0.4921	0.00338
H3K4me3 ↑, H3K9ac ↑ ⇒ GE ↑	3394	0.4675	0.009406
H4K8ac ↓, H4K5ac ↑ ⇒ GE ↓	2688	0.5111	0.002441
H4K8ac ↑, H4K5ac ↓ ⇒ GE ↑	2396	0.4974	0.005507
H3K9ac ↓, H3K56ac ↓, H4K5ac ↑ ⇒ GE ↓	1798	0.5189	0.005544
H3K9ac ↑, H3K56ac ↑, H4K5ac ↓ ⇒ GE ↑	1705	0.5026	0.009744

3.1. Rules of HMs and gene expression

From the 60 rules which have corrected p-values less than 0.01, we select those having both directions in regulating gene expression and show them in Table 2. For example, the rule "$H3K56ac\uparrow \Rightarrow GE\uparrow$" is included, if and only if "$H3K56ac\downarrow \Rightarrow GE\downarrow$" also exists in the 60 rules. We did this selection because bi-directions would give more certainty to the rules about the role of combinatorial HMs in regulating gene expression.

First let us look at the relations between a single HM and gene expression (i.e. the first ten rules in Table 2). These rules show that H3K56ac, H4K8ac, H3K4me3, H3K9ac and H4K20me1 positively associate with gene expression globally with statistical significance. Among these five HMs, H4K20me1 and H3K4me3 are histone methylation marks that are well known as activating histone modification marks (Berger, 2007; Barski et al., 2007; Bernstein et al., 2002; Wang et al., 2008), which is also confirmed by Yu et al. (2008) and Xu et al. (2010) with computational methods. Our method on *Plasmodium* data also show the two HMs are strongly correlated with gene activation. The other three HMs are histone acetylation marks, which are often observed to be associated with gene activation since acetylation removes the positive charges on histones and relaxes the affinity between histones and DNA. However, only H3K56ac, H4K8ac and H3K9ac are found to be significantly linked to transcription activation in our dataset. Although other histone

acetylation marks such as H4K16ac and H3K14ac may also have activating trends, their trends are not significant enough to be detected by our approach.

We have observed three special HMs from our rules: H4K5ac, H4K12ac, and H4R3me2. H4K5ac and H4K12ac as histone acetylation marks are generally linked to gene activation. However in our dataset, they are observed to be negatively correlated with gene expression (in the rules with multiple HMs). Similar observations have been found in literature. H4K12ac was observed negatively correlated with gene expression in yeast (Kurdistani et al., 2004), which was confirmed in *P. falciparum* by Chaal et al. (2010). Although H4K5ac is generally considered as a mark for activating transcription which was tested in many species such as human (Wang et al., 2008), it has a clear sign of globally transcription suppression in our dataset of *P. falciparum*. It is therefore interesting to test if H4K5ac and H4K12ac are transcriptional suppressors (maybe by cooperating with other proteins) in *P. falciparum*, in contrast to other species. In addition, our method discovers that H4R3me2 is negatively associated with gene expression (in the rules with multiple HMs). H4R3me2 was revealed to globally repress gene expression by Xu et al. (2010). Hence it is worth further study to confirm the role of H4R3me2 in the gene regulation of *P. falciparum*.

The ability to discover rules with multiple HMs is one of the main advantages of our method. This kind of rules shows how the combinations of HMs cooperatively regulate gene expression, as a clear sign of 'histone code' on transcription. However, compared to the rules with single HMs, the experimental supports of such rules can hardly be found in the literature. To compare these multiple-HMs rules with causal relations, we construct a Bayesian network as in Yu et al. (2008) based on our dataset, as shown in Fig. 2. Red edges are compelled edges, which retain their directions in all equivalent DAGs (directed acyclic graphs). Green edges are reversible edges, which may change their directionality among different DAGs in the same equivalence class. According to Chickering (1995), compelled edges denote causal relationships, whereas reversible edges might be just correlation but not necessarily causality. In Fig. 2, green edges all occur among the nodes on the top level. Here we only focus on the red edges. From Fig. 2, we can see that H4K5ac, H3K56ac, and H3K4me3 have direct causal relations with RNA expression, which implies the three HMs are closely related with gene expression. Note that almost all the association rules with multiple HMs in Table 2 include at least one of the three HMs: H3K56ac, H4K5ac and H3K4me3. This indicates that the rules we retrieved reveal some causal relations between histone modifications and gene expression. Some, but not all, rules include HMs with causal relations inferred by the Bayesian network. These HMs with causal relations are "H4K20me1 and H4K5ac", "H3K4me3 and H4K8ac", "H3K56ac and H4K12ac" and "H4K8ac and H4K5ac". This suggests that HMs could regulate gene expression cooperatively.

3.2. Robustness of rules

The rules we discovered might be biased by the missing data (see Fig. 1). Because missing data would lead to different percentages of different features (e.g. HMs) in the records, rules consisting of the HMs with more missing data would have lower *support* and hence less significant *p*-values. To test if our rules are robust to the different percentages of HMs in the records, we apply bootstrapping to the original dataset. First we randomly generate 20 datasets from the original dataset by sampling with replacement. Each of these bootstrapping datasets contain 90,168 rows, as the original dataset (15, 028probes × 6timepoints). The process of bootstrapping could bring randomization to the percentage of HMs in the records. Next we are going to test if our rules are robust to such resampling.

We apply our method on the 20 bootstrapping datasets and obtain 20 sets of significant rules respectively. From the 20 sets of significant rules, we count the occurrences of the significant rules from the original dataset (i.e. rules in Table 2). As shown in Fig. 3 most significant rules from the original dataset also appear as significant rules in the bootstrapping datasets. Then we calculate the *p*-values of the rules in these 20 datasets. Fig. 4 shows the distributions of *p*-values of the rules in Table 2 from the original dataset (left box) and 20 bootstrapping datasets (right box). The difference between the two groups is not significant (p-value = 0.7709), which indicates that the resampling may not affect the detection of significant rules.

For comparison, we generate a set of random rules which have the same number as the true rules, and calculate *p*-values of these random rules in both the original dataset and the 20 bootstrapping datasets. The two sub-figures in Fig. 5 show the comparisons of *p*-values from these two groups of rules, in the original dataset (left) and in the 20 bootstrapping datasets (right) respectively. In both original dataset and the bootstrapping datasets, true rules

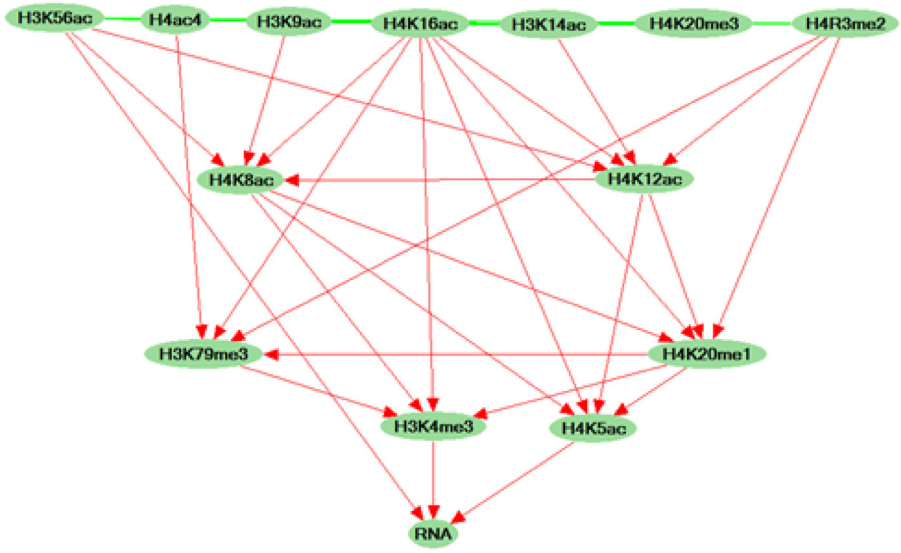

Fig. 2. Bayesian network of HMs and RNA expression. Red edges are compelled edges, and green edges are reversible edges. Here we only focus on the red edges, which denote causal relationships. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

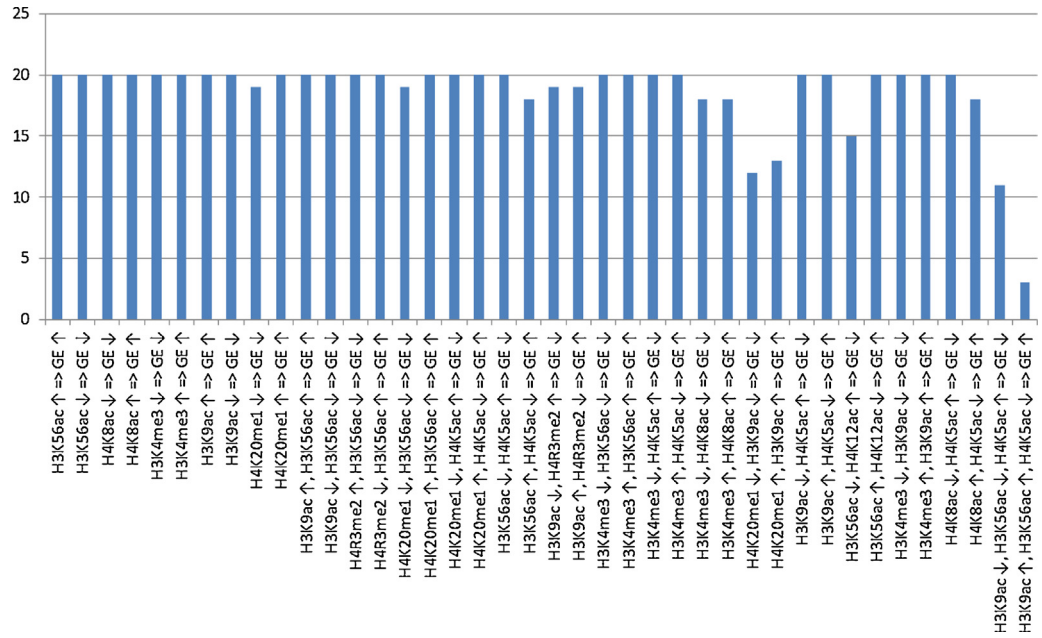

Fig. 3. The numbers of occurrences of the significant rules in the bootstrapping datasets.

have *p*-values far more significant than the random rules, which indicate that the rules we discovered are not obtained by chance.

3.3. Comparison with existing methods

We have compared our method with Bayesian network (Yu et al., 2008), which is a pioneering approach in the field. Similar to our method, Bayesian network applies discretization on data, and then it infers causality among variables. Bayesian network can output rules just as our method does. One advantage of Bayesian network is that it can infer hierarchical and causal interactions among HMs, while right now our rules only include the associations between HMs and GE. However, our method can obtain not only rules with causal relations among HMs (which are like the rules output by Bayesian network) but also the rules in which HMs have other kind of relationships such as cooperation. It is known that HMs regulate transcription in a cooperatively way, so our method is more suitable for the decoding of histone code. We have compared our rules with Bayesian network in Section 3.1. Next we will compare the performance of our method with Bayesian network in the prediction of gene expression. We have introduced the process of predicting gene expression given a set of rules in the last part of Section 2. For comparison, Bayesian network is also used to predict gene expression by applying Libra (http://libra.cs.uoregon.edu/). We have built a Bayesian network on our dataset, shown in Fig. 2. The structure of Bayesian network and the discrete values of HMs are input to Libra, which would then output the predicted gene expression level for each record. Then for each of the six groups of gene expression (see Section 2), the percentage of genes which are predicted as 'up-regulated' is calculated and treated as the *predicted* gene expression level of this group.

The plot of "predicted gene expression vs. observed gene expression" is shown in Fig. 6, where the two straight lines are obtained by linear regression. The detailed results of linear regression are shown in Table 3. As seen, our method has significantly higher R^2 than Bayesian network, which indicates that our method is able to predict gene expression more consistently. The coefficients and *p*-values show that our method can predict the trend of gene expression more accurately. In addition, Pearson correlation coefficients were calculated between the observed and predicted gene expression, which are 0.7407 (*p*-value = 0.0921) for Bayesian network and 0.9879 (*p*-value = 0.0002) for our method. These results indicate that our method has better performance than

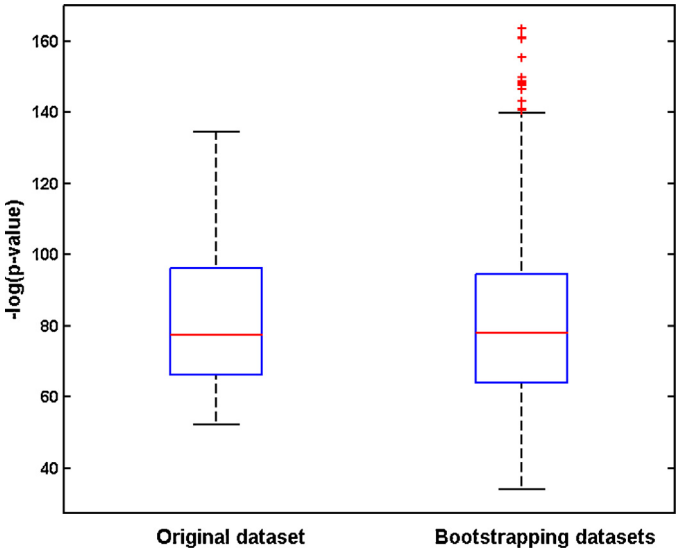

Fig. 4. *p*-values (in log scale) distributions of the significant rules in original dataset and bootstrapping datasets. The significant rules are the rules in Table 2. Here two boxplots are used to depict the distributions of *p*-values of these rules in the original dataset and the bootstrapping datasets. Applying Wilcoxon test on these two distributions, the *p*-value is 0.7709.

Table 3
Comparison between Bayesian network and our approach in gene expression prediction, where both linear regression and Pearson correlation are applied to two data vectors: the *predicted* and *observed* expression levels.

		Bayesian network	Our approach
Linear regression	R^2	0.5487	0.9760
	Coefficient	0.0142	0.1432
	p-Value	0.0921	0.0002
Pearson correlation	PCC	0.7407	0.9879
	p-Value	0.0921	0.0002

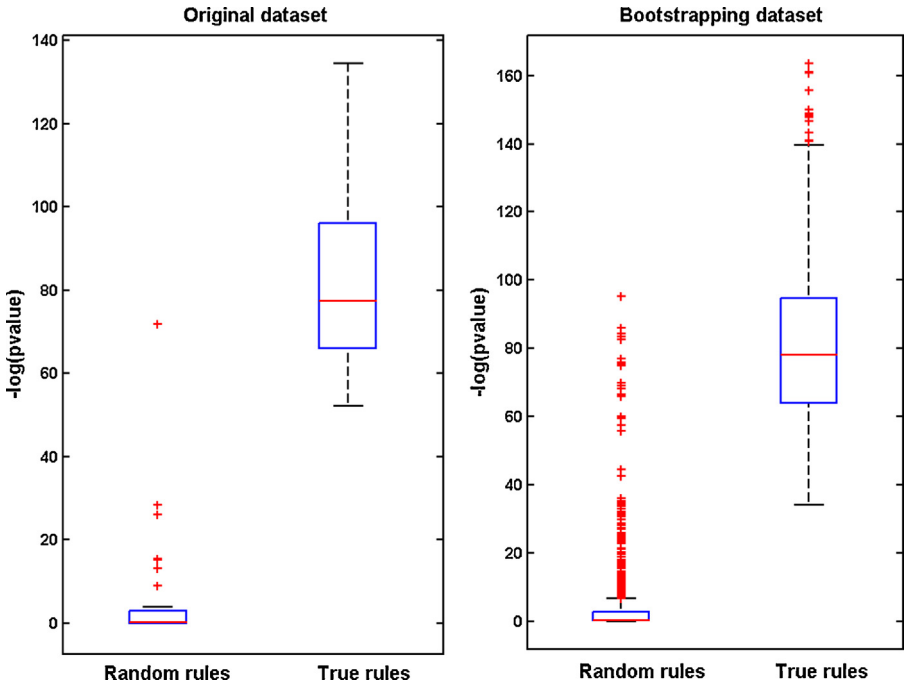

Fig. 5. Comparison of p-values (−log) between random rules and true rules The comparison is done on both original dataset (left) and bootstrapping datasets (right). Wilcoxon test p-values are 2.5889e-14 (left) and 6.0629e-253 (right) respectively.

Bayesian network in predicting gene expression. This is probably due to the fact that our rules are selected from all possible combinations of different states of HMs and gene expression (i.e. global searching), while Bayesian network is constructed based on local structure searching.

However, it is generally difficult to compare our method with most existing methods such as regression models because they are different at the technical level. For example, although SVM has good performance in non-linear regression, it is not suitable for large-scale data analysis. Our dataset includes 90,168 records, which is too big to use SVM. Moreover, there are a lot of missing data in our dataset, which is not unusual for microarray data. The missing data in synchronized multi-variables dataset make it difficult for regression using SVM or Random Forests. Because there are missing values of at least one variable for almost all records, the accuracy of regression could easily be affected. In contrast, our method is relatively robust to missing data, partly because data discretization is carried out independently for each variable. Even if we are using a new dataset which is suitable for both our method and those regression models such as SVM, it is still difficult to compare these two kinds of methods. It could be unfair to compare their prediction accuracy on gene expression because our method requires data discretization which would cause loss of certain information. But our method can provide results with interpretability, which those regression models lack.

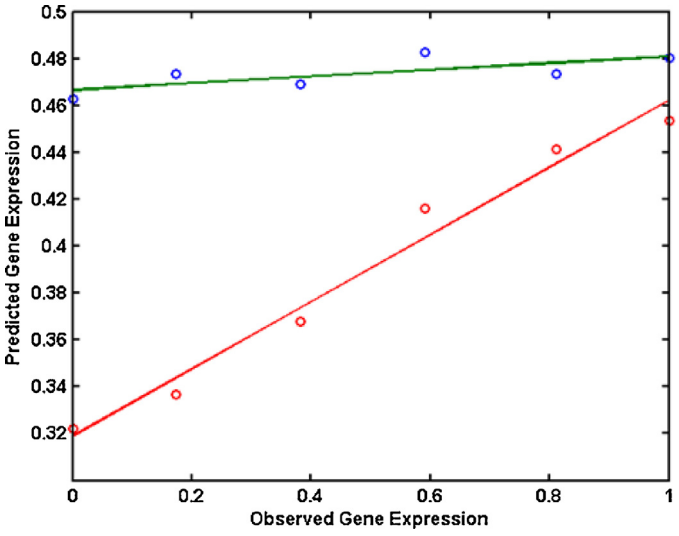

Fig. 6. Comparison between Bayesian network and our method in predicting gene expression. Blue dots and the green line denote prediction of Bayesian method, while red dots and red line denote prediction of our method. The six dots (blue or green) from left to right correspond to six groups of genes with expression levels from highest to lowest respectively among the six time points (see Section 2.5). The lines are obtained from linear regression. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4. Discussion and conclusions

In this paper, we proposed a framework based on association rule mining to discover in large-scale the regulatory relations between histone modifications and gene expression. Based on the technique of association rule mining, we designed a method to identify interesting relations by the measures of *support* and *confidence*. Then we used Fisher's exact test and multiple testing correction to calculate p-values for all the rules obtained from the previous step. A threshold for p-values (here we use the widely accepted value of 0.01) was used to select the most significant rules. Applied to the data of *P. falciparum*, our method is able to discover rules that are consistent with literature of experimental biology. We demonstrated that our rules are not obtained by chance and they are robust to random resampling. Compared with Bayesian network, our model based on the significant association rules could predict gene expression more accurately and consistently.

The proposed method can generate rules with interpretable biological meanings. Driven by the power of data mining, our method is able to detect significant patterns of histone modifications regulating transcription, from all possible combinations. The direct mining from data could provide a more reliable way to study and

interpret the underlying biological processes. Because our method is data-driven and unsupervised, it can uncover novel regulations rules between histone modification and gene expression to guide further biological experiments and discoveries. To the best of our knowledge, this is the first paper to apply association rule mining to reveal regulation relations between gene expression and histone modifications in *P. falciparum*, while taking into account the dynamic epigenetic regulation of gene expression.

One interesting future work is to apply our method to data of other species, with the aim to find evolutionarily conserved regulation patterns. Another future work is to extend simple rules with only a few histone modifications on the left hand side to incorporate hierarchical and combinatorial interactions among histone modifications.

Acknowledgements

We would like to thank Zbynek Bozdech from School of Biological Sciences, Nanyang Technological University, Singapore, for providing the time-course datasets of gene expression and histone modifications of *Plasmodium falciparum*.

This project is supported in part by NTU Start-up Grant, (COE_SUG/RSS_1FEB11_1/8) and Singapore Ministry of Education Academic Research Tier 2 Grant (ARC 09/10).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.compbiolchem.2014.01.002.

References

Agrawal, R., Imieliński, T., Swami, A., 1993. Mining association rules between sets of items in large databases. ACM SIGMOD Record, 22. ACM, pp. 207–216.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., Zhao, K., 2007. High-resolution profiling of histone methylations in the human genome. Cell 129 (4), 823–837.

Berger, S.L., 2007. The complex language of chromatin regulation during transcription. Nature 447 (7143), 407–412.

Bernstein, B.E., Humphrey, E.L., Erlich, R.L., Schneider, R., Bouman, P., Liu, J.S., Kouzarides, T., Schreiber, S.L., 2002. Methylation of histone H3 Lys 4 in coding regions of active genes. Proceedings of the National Academy of Sciences of United States of America 99 (13), 8695–8700.

Brinkman, A.B., Roelofsen, T., Pennings, S.W., Martens, J.H., Jenuwein, T., Stunnenberg, H.G., 2006. Histone modification patterns associated with the human x chromosome. EMBO Reports 7 (6), 628–634.

Cabral, F.J., Fotoran, W.L., Wunderlich, G., 2012. Dynamic activation and repression of the *Plasmodium falciparum* rif gene family and their relation to chromatin modification. PLoS ONE 7 (1), e29881.

Chaal, B.K., Gupta, A.P., Wastuwidyaningtyas, B.D., Luah, Y.-H., Bozdech, Z., 2010. Histone deacetylases play a major role in the transcriptional regulation of the *Plasmodium falciparum* life cycle. PLoS Pathogens 6 (1), e1000737.

Chen, Q., Chen, Y.-P.P., 2006. Mining frequent patterns for Amp-activated protein kinase regulation on skeletal muscle. BMC Bioinformatics 7 (1), 394.

Cheng, C., Gerstein, M., 2012. Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. Nucleic Acids Research 40 (2), 553–568.

Cheng, C., Yan, K.-K., Yip, K.Y., Rozowsky, J., Alexander, R., Shou, C., Gerstein, M., et al., 2011. A statistical framework for modeling gene expression using chromatin features and application to modencode datasets. Genome Biology 12 (2), R15.

Chickering, D.M.,1995. A transformational characterization of equivalent Bayesian network structures. In: Proceedings of the Eleventh conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., pp. 87–98.

Cui, L., Miao, J., 2010. Chromatin-mediated epigenetic regulation in the malaria parasite *Plasmodium falciparum*. Eukaryotic Cell 9 (8), 1138–1149.

do Rego, T.G., Roider, H.G., de Carvalho, F.A., Costa, I.G., 2012. Inferring epigenetic and transcriptional regulation during blood cell development with a mixture of sparse linear models. Bioinformatics 28 (18), 2297–2303.

Dong, X., Greven, M.C., Kundaje, A., Djebali, S., Brown, J.B., Cheng, C., Gingeras, T.R., Gerstein, M., Guigó, R., Birney, E., et al., 2012. Modeling gene expression using chromatin features in various cellular contexts. Genome Biology 13 (9), R53.

Duraisingh, M.T., et al., 2005. Heterochromatin Silencing and Locus Repositioning Linked to Regulation of Virulence Genes in *Plasmodium falciparum*. Cell 121 (1), 13–24.

Esteller, M., 2007. Cancer epigenomics: DNA methylomes and histone-modification maps. Nature Reviews Genetics 8 (4), 286–298.

Fischle, W., Wang, Y., Allis, C.D., 2003. Binary switches and modification cassettes in histone biology and beyond. Nature 425 (6957), 475–479.

Fisher, R.A., 1922. On the interpretation of $x2$ from contingency tables, and the calculation of p. Journal of the Royal Statistical Society 85 (1), 87–94.

Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., et al., 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419 (6906), 498–511.

Gupta, A.P., Chin, W.H., Zhu, L., Mok, S., Luah, Y.-H., Lim, E.-H., Bozdech, Z., 2013. Dynamic epigenetic regulation of gene expression during the life cycle of malaria parasite *Plasmodium falciparum*. PLoS Pathogens 9 (2), e1003170.

Ha, M., Ng, D.W., Li, W.-H., Chen, Z.J., 2011. Coordinated histone modifications are associated with gene expression variation within and between species. Genome Research 21 (4), 590–598.

Jenuwein, T., Allis, C.D., 2001. Translating the histone code. Science 293 (5532), 1074–1080.

Karlić, R., Chung, H.-R., Lasserre, J., Vlahoviček, K., Vingron, M., 2010. Histone modification levels are predictive for gene expression. Proceedings of the National Academy of Sciences of United States of America 107 (7), 2926–2931.

Kimura, H., Tada, M., Nakatsuji, N., Tada, T., 2004. Histone code modifications on pluripotential nuclei of reprogrammed somatic cells. Molecularand cellular biology 24 (13), 5710–5720.

Kotsiantis, S., Kanellopoulos, D., 2006. Association rules mining: a recent overview. GESTS International Transactions on Computer Science and Engineering 32 (1), 71–82.

Kurdistani, S.K., Tavazoie, S., Grunstein, M., 2004. Mapping global histone acetylation patterns to gene expression. Cell 117 (6), 721–733.

Li, B., Carey, M., Workman, J.L., 2007. The role of chromatin during transcription. Cell 128 (4), 707–719.

Lopez, F.J., Blanco, A., Garcia, F., Cano, C., Marin, A., 2008. Fuzzy association rules for biological data analysis: a case study on yeast. BMC Bioinformatics 9 (1), 107.

Margueron, R., Trojer, P., Reinberg, D., 2005. The key to development: interpreting the histone code? Current opinion in genetics & development 15 (2), 163–176.

McLeay, R.C., Lesluyes, T., Partida, G.C., Bailey, T.L., 2012. Genome-wide in silico prediction of gene expression. Bioinformatics 28 (21), 2789–2796.

Morgan, X., Ni, S., Miranker, D., Iyer, V., 2007. Predicting combinatorial binding of transcription factors to regulatory elements in the human genome by association rule mining. BMC Bioinformatics 8 (1), 445.

Nayyar, G.M., Breman, J.G., Newton, P.N., Herrington, J., 2012. Poor-quality antimalarial drugs in Southeast Asia and Sub-Saharan Africa. The Lancet Infectious Diseases 12 (6), 488–496.

Rovira-Graells, N., Gupta, A.P., Planet, E., Crowley, V.M., Mok, S., de Pouplana, L.R., Preiser, P.R., Bozdech, Z., Cortés, A., 2012. Transcriptional variation in the malaria parasite *Plasmodium falciparum*. Genome Research 22 (5), 925–938.

Su, J., Qi, Y., Liu, S., Wu, X., Lv, J., Liu, H., Zhang, R., Zhang, Y., 2012. Revealing epigenetic patterns in gene regulation through integrative analysis of epigenetic interaction network. Molecular Biology Reports 39 (2), 1701–1712.

Teng, L., Tan, K., 2012. Finding combinatorial histone code by semi-supervised biclustering. BMC Genomics 13 (1), 301.

Uno, T., Asai, T., Uchida, Y., Arimura, H., 2003. LCM: an efficient algorithm for enumerating frequent closed item sets. FIMI, vol. 90. Citeseer.

van Noort, V., Huynen, M.A., 2006. Combinatorial gene regulation in *Plasmodium falciparum*. Trends in Genetics 22 (2), 73–78.

Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Peng, W., Zhang, M.Q., et al., 2008. Combinatorial patterns of histone acetylations and methylations in the human genome. Nature Genetics 40 (7), 897–903.

Wang, J., Dai, X., Xiang, Q., Deng, Y., Feng, J., Dai, Z., He, C., 2010. Identifying the combinatorial effects of histone modifications by association rule mining in yeast. Evolutionary Bioinformatics 6, 113.

Webb, G.I., 2007. Discovering significant patterns. Machine Learning 68 (1), 1–33.

Wu, M., Kwoh, C.-K., Przytycka, T.M., Li, J., Zheng, J., et al., 2012. Epigenetic functions enriched in transcription factors binding to mouse recombination hotspots. Proteome Science 10 (Suppl 1), S11.

Xu, X., Hoang, S., Mayo, M.W., Bekiranov, S., 2010. Application of machine learning methods to histone methylation chip-seq data reveals h4r3me2 globally represses gene expression. BMC Bioinformatics 11 (1), 396.

Yu, H., Zhu, S., Zhou, B., Xue, H., Han, J.-D.J., 2008. Inferring causal relationships among different histone modifications and gene expression. Genome Research 18 (8), 1314–1324.

Zhao, Q., Bhowmick, S.S., 2003. Association Rule Mining: A Survey, Technical Report Centre for Advanced Information Systems. School of Computer Engineering, Nanyang Technological University, Singapore, No. 2003116.