Journal of Theoretical Biology 380 (2015) 24–39

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Identification of peculiar and common effects of histone modifications on transcription

Yu Liu [a], Xue Liang [a], Rongcun Yang [a,b,c,*]

[a] Department of Immunology, School of Medicine, Nankai University, Tianjin 300071, PR China
[b] State Key Laboratory of Medicinal Chemical Biology, Nankai University, Tianjin 300071, PR China
[c] Key Laboratory of Bioactive Materials Ministry of Education, Nankai University, Tianjin 300071, PR China

HIGHLIGHTS

- Activating and repressing histone modifications have different distribution patterns.
- H3K4me3 can sometimes help its usual opponent H3K27me3 to repress gene expression.
- H3K36me3 can serve as a criterion to identify high and low expressed genes.
- The effect of histone modifications can be summarized by 4 factors underlying them.
- The 4-factor-based model identifies stable RNAs found to be lifespan regulators here.

ARTICLE INFO

Article history:
Received 4 November 2014
Received in revised form
5 February 2015
Accepted 29 April 2015
Available online 20 May 2015

Keywords:
Position effect
Exploratory factor analysis (EFA)
Multiple linear regression
RNA stability
Lifespan

ABSTRACT

Histone modifications (HMs) play an important role in controlling eukaryotic gene expression and next generation sequencing (NGS) has greatly advanced the research on this topic with generating many high-resolution maps for HMs. Here, we use these maps to analyze the relationship between HMs and transcription. By incorporating various segments of genes into analysis without restricting the scope only in the promoter region, we have collected more comprehensive data and captured some details of this process. A position effect of gene regions has been revealed and it can even inverse the property of some HMs from activating to repressing genes such as the cases of H3K4me3, H3K36me3 and H3K14ac. Especially H3K36me3, its dual character on gene transcription makes it able to serve as a criterion to distinguish high and low expressed genes. We also study the general property of different HMs based on the comprehensive data. Using exploratory factor analysis (EFA), we have extracted 4 latent structures underlying the HMs, which are able to represent their activating and repressing effects concisely. These 4 factors have fine properties in the aspects of distinguishing high and low expressed genes, predicting transcription level and identifying genes with unique attributes such as stable RNA generating genes found to have a close relationship with lifespan of organisms here. In summary, while the position effect associated HM peculiarities demonstrate some details of the complex HM regulation network divergently, the common factors catch the nature of the network convergently. This deepens our understanding on the HM-transcription relationship.

1. Introduction

The Central Dogma gives us an elementary blueprint of how the genetic information flows among the 3 kinds of essential biomolecules, DNA, RNA and protein. Although some unusual ways of delivering the information were added to it such as RNA dependent DNA synthesis and RNA dependent RNA synthesis in some viruses, the chain of DNA→RNA→protein is still the foremost information flow in the biological world. Cells possess a sophisticated control system to regulate this important process at different levels, from transcription to post-translation. Because DNA is the start point of the information flow with a top priority over the other 2 molecules, it seems that the regulation on the DNA associated chain of DNA→RNA showed more mechanism diversity. Many elements can be used to control transcription such as gene rearrangement, transcription factors (TFs), histone modifications (HMs) and so on. Among them, HMs are in charge of modulating local chromatin structure and thereby changing the accessibility of TFs, as well as recruiting transcriptional activators or repressors (Kouzarides, 2007).

* Corresponding author at: Department of Immunology, School of Medicine, Nankai University, Tianjin 300071, PR China.
Tel.: +86 22 23509007; fax: +86 22 23502554.
E-mail address: ryang@nankai.edu.cn (R. Yang).

http://dx.doi.org/10.1016/j.jtbi.2015.04.040
0022-5193/

HMs usually occur on the lysine residuals at N terminus of histones, which form octamers winded around by 147 bp of DNA and constitute the basic units of eukaryotic chromosome structure. The post-modifications on histone include methylation, acetylation, phosphorylation and ubiquitination. Their combination is related to distinct chromatin states (Heintzman et al., 2007; Jenuwein and Allis, 2001; Kouzarides, 2007; Li et al., 2007) and by exerting effects on this dynamic condition of chromatin, HMs can control various biological processes. In addition to the aforementioned effect on transcription, they also involved in many other aspects, such as DNA replication (Birney et al., 2007), DNA repair (Li et al., 2013; van Attikum and Gasser, 2005), carcinogenesis (Choi and Lee, 2013), immune response (Mei et al., 2014; Su et al., 2014), and so on. However, because of the central position of transcription in genetic information transfer, the relationship between HMs and gene transcription is still the focus of research and many HMs tightly associated with RNA synthesis are revealed, like H3K4me3, H3K36me3 and H3K27me3. H3K4me3 is identified as a marker of transcription initiation, H3K36me3 is highly correlated with transcription elongation, while H3K27me3 is an inhibitory marker associated with RNA polymerase pausing and elongation repression (Barski et al., 2007). Although there are still many mysteries in the knowledge system on the relationship between HMs and gene transcription, such as the effect of combinations of different HMs, which is referred as the "histone code" hypothesis (Strahl and Allis, 2000), their primary effect on gene expression is clear. This achievement largely benefited from the technological development, especially the appearance of next generation sequencing (NGS).

NGS has given us an unprecedented chance to gain massive data and greatly promoted the research process of analyzing the mechanism of epigenetic regulation. For example, by the technique of ChIP-seq, a genome-wide map of 20 histone methylation and 18 histone acetylation sites in human resting CD4 T cell has been presented at an individual nucleosome level (Barski et al., 2007; Wang et al., 2008). This HM map not only confirms the known associations of HMs with gene expression, and discovered novel ones, but also provides an invaluable resource to perform elaborate quantitative analysis on this relationship, such as the Bayesian inference model demonstrating causal relationships among different HMs (Yu et al., 2008) and the linear regression model predicting transcription level from HM levels (Karlic et al., 2010). They have successfully constructed quantitative relationships between HMs and gene expression. However, actually they could have done more things using the abundant information provided by this dataset. For example, it is obvious that to a specific HM, different gene regions have different influences on its function and thus a position effect exists during the process of HM-gene interaction. However, in most cases, only the gene promoter region is involved in performing computation and the position effect based on considering as many regions as possible is largely neglected. In addition, while deciphering the complex regulation network constituted by the diverse HMs, the effort on exploring latent structures underlying them to summarize their function, and then understand the nature of the HMs briefly, is relatively insufficient.

Here, to remedy these defects, we perform statistical analyses on this CD4 T cell ChIP-seq map again. We have not restricted our analysis in the region of gene promoter and have extended our research range to more wide space of the gene, it has led us to illuminate the position effect of gene regions on the function of HMs that can even inverse the function of the HMs sometimes. We have also used exploratory factor analysis (EFA) to excavate the nature of their functions and 4 common factors underlying them have been found with the ability to summarize the influence of the 38 HMs. These factors displayed fine properties in the aspect of identifying high and low expressed genes, predicting gene expression and finding genes with unique attributes such as stable RNA generating genes here found to have a close relationship with lifespan of organisms and what's more, play a potential role in regulating the progression of some age-associated neurodegenerative disorders like Parkinson's disease, Huntington's disease and Alzheimer's disease, so have a significance in studying their pathological mechanisms and developing corresponding clinical therapies. All the results here deepen our understanding on the HM-transcription relationship.

2. Materials and methods

2.1. Datasets

The 10899 human genes involved in this study were from an expression microarray data for human resting CD4 T cells conducted on the platform of Affymetrix Human Genome U133 Plus 2.0 Array with a GEO accession ID of GSE10437 (Schones et al., 2008) and the 13327 mouse genes were from an microarray data for mouse embryonic stem cell Bruce4 (ES-Bruce4) performed on the Affymetrix GeneChip Mouse Genome 430 2.0 Array platform with a GEO accession ID of GSE10246 (Lattin et al., 2008). These data were processed by the limma package in the R program (Smyth Gordon, 2004). All of these genes were intron-containing RefSeq genes from different Unigene clusters and could be uniquely mapped to an Affymetrix probe set. Their expression values were averaged over all replicates.

The RefSeq gene annotations for hg18 and mm9 were downloaded from the UCSC genome browser (http://genome.ucsc.edu/). The fastq files for human ChIP-seq were from genome wide studies on the distributions of 20 histone methylations and 18 histone acetylations in human resting CD4 T cells (Barski et al., 2007; Wang et al., 2008) and the ones for mouse ES-Bruce4 ChIP-seq were from a study involved in 7 histone modifications with a GEO accession ID of GSE29184 (Shen et al., 2012). They were mapped to human (hg18) or mouse (mm9) genome by bowtie (Langmead et al., 2009) and MACS was used to generate wiggle files with a default p-value cutoff of 1e-5 (Zhang et al., 2008).The resulting tags were then mapped to the 13 regions of the 10,899 human genes or the 13,327 mouse genes defined in the section "Results" and the coverage of each nucleotide (counts per bp) for each histone modification (HM) within each region of a gene was calculated.

2.2. Hierachical clustering

For each preprocessed matrix, the Euclidean distance between both rows and columns was calculated and then the complete linkage method was used to compute hierarchical clustering. In some cases where genes were sorted by their expression levels, only the column dimension was clustered.

2.3. Exploratory factor analysis (EFA)

After a preliminary screening on the initial dataset to reserve only 1 region value for each HM, EFA based on iterated principle axis method was applied to this smaller dataset to find common factors underlying the variables. Then, oblimin rotation was used to make the factors more interpretable.

2.4. Multiple regression model

Predictor variables were selected from the 4 common factors extracted by EFA and the log transformed gene expression value was used as response variable to construct a multiple regression