Neurocomputing 206 (2016) 42–49

Contents lists available at ScienceDirect

Neurocomputing

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

Histone modification patterns in highly differentiation cells

Qiuyang Wu [a], Jihong Guan [a,*], Shuigeng Zhou [b]

[a] Department of Computer Science and Technology, Tongji University, Shanghai 201804, China
[b] School of Computer Science, and Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai 200433, China

ARTICLE INFO

Article history:
Received 9 August 2015
Received in revised form
1 November 2015
Accepted 9 November 2015
Available online 31 May 2016

Keywords:
Epigenetics
Histone modifications
Patterns
Differentiation

ABSTRACT

Nucleosomes and modifications on the histones constitute the fundamental genomic structure of eukaryotes. It was reported that particular combinations of histone modifications occurring on the genome indicate specific chromatin states, which are related to a series of cell states. Although the histone code principle has been found for over a decade, the mechanism of the code is not yet known clearly. In this paper, we first conducted an extensive analysis on 38 histone modifications and 1 histone variant in human CD4+ T cells to reveal possible connections among these modifications. Then, we analyzed the different roles that histone modifications play in highly differentiated cells and undifferentiated cells, and found that the number of histone modifications on genes of differentiated cells is correlated to the expression levels of genes, while there is no direct effect of histone modification on gene expression level in undifferentiated cells. Finally, we compared the distributions of histone modifications of CD4+ T cell line and K562 cell line, and observed that they have similar distribution patterns. So we guess that there may exist certain conservative regions of histone modifications.

1. Introduction

Since the conjecture that eukaryotic histones repress gene transcription [1,2] was proposed in the 1980s, the study of histone code has made great progress. The concept of histone code indicates the combinatorial nature of histone amino-terminal modifications [3].

Histone modifications are considered to control the structure of chromatin fiber [4,5], which further impacts the expression levels of genes. From this perspective, histone modifications can be classified in terms of whether they activate or repress gene expression, but this classification is too simple since the functions of some modifications are not clear or are distinctive in different situations. Thus, more and more research turns to exploring histone modification patterns from the perspective of gene expression. For example, Dirk Schüeler et al. observed the binary pattern of histone modifications [6]; Seligson et al. applied the random forest clustering algorithm to five histone marks in human prostate tissue and found that some patterns of histone modifications are associated with prostate cancer [7]; Ucar et al. proposed a scalable subspace clustering algorithm for histone modification peak signals to detect histone modification patterns and found 843 combinatorial patterns in human CD4+ T cells [8], while Linghu et al. developed a statistical hybrid clustering algorithm for co-occurrence discrete dots histone modification data to mine patterns, then clustered similar patterns and detected 845 combinatorial patterns in human CD4+T cells [9].

In this paper, we tried to mine the spatial patterns of histone modifications in different genome regions: upstream of the transcription start site (TSS), downstream of the transcription termination site (TTS), codingexons, exons, introns, 3' untranslated regions (3'-UTR) and 5' untranslated regions (5'-UTR). Also, it was reported that the number of modifications supports the existence of an epigenetic code on the histone terminal tails [4], thus we further conducted analysis on the number of histone marks, which include histone modifications and histone variant, to see whether the number of different histone marks is correlated to gene expression.

Histone modifications have been identified to impact gene expression, but the mechanism of histone modification regulating gene expression is not yet completely clear. It is commonly acknowledged that histone acetylations may provide DNA access and histone methylations may be modulators of nucleosome stability [10]. According to Leung et al., histone modifications are involved in regulating the embryonic stem cells [11]. This arises the question that whether histone modifications regulate gene expression in the same way in both highly differentiated cells and undifferentiated cells. So in this paper, we also analyzed histone modification patterns in human embryonic stem cells with the control set of human CD4+ T cells, to investigate whether the cells

[*] Corresponding author.

http://dx.doi.org/10.1016/j.neucom.2015.11.108
0925-2312/© 2016 Elsevier B.V. All rights reserved.

at different stages have similar histone marks in terms of expression level.

Histone marks vary from cell line to cell line, and changes in histone modification profiles may imply certain diseases [12–14]. However, it was found that within a single highly-differentiated cell line, histone modification profile is rather stable [15]. We wonder whether there exist relatively stable regions among different highly differentiated cells. For this purpose, we first sought the relatively conservative regions between two highly differentiated cell lines: CD4+T and normal human epidermal keratinocyte, and then checked whether such regions also exist in K562, which is a typical disease cell line.

2. Materials and methods

2.1. Datasets

We downloaded the widely used human CD4+T cell histone modification dataset from the work of Barsk and Wang [16,17], which includes 38 histone modifications (20 methylations and 18 acetylations) and 1 histone variant, for the analysis of histone modification patterns and the study of histone marks' impact on gene expression of highly differentiation cells. We got the transcription start and end sites, noncoding exons, exons, introns, 5' untranslated region and 3' untranslated region data from the website of UCSC[1]. To investigate histone marks' impact on gene expression of undifferentiated cells, we got the dataset of human embryonic stem cell (hESC) line H1, which includes 26 histone modifications (10 methylations and 16 acetylations) and 1 histone variant, from the website of NCBI [18]. The expression data for both CD4+T cell and hESC cell line H1 were downloaded from NCBI [18], with the expression data of CD4+ T cell line represented by PolII assayed by ChIP-seq sequencing technology and hESC cell line H1 represented by RNA-seq. The analysis of conservative regions covers nine histone modifications of normal human epidermal keratinocyte (Nhek) and K562 from the HHMD database [19].

2.2. Generating histone modification domains

The histone modification level is a discrete distribution along chromosomes, one value for each base position. However, the raw data of histone modification levels may contain false positives, so we applied a clustering method [20] as a preprocessing step over the raw data. In this method, to improve specificity and sensitivity, a threshold is set for detecting histone modification *gaps* and *domains* on the chromosomes. A *basic gap* refers to a certain base position along the chromosomes where the enrichment value (read count) of a certain histone modification is below the preset threshold. Consecutive basic gaps are merged to a *larger gap* (simply *gap* in the sequel). And a *basic domain* is any base position on the chromosomes that is not a basic gap. Two basic domains with at most one basic gap between them are merged to a *larger domain* (simply *domain*).

For each histone modification, a score is calculated for each domain based on the histone modification's level (read count) within the domain. Fig. 1 illustrates the histone modification domains and gaps for a certain histone modification. Take H3K4me3 for example, after applying the cluster method, the average domain length is about 2301 bps. Here, we used the default threshold value of the method proposed in Zang [20], which is 1. It means that if one locus has a value greater than zero, it should be taken into account to form a domain.

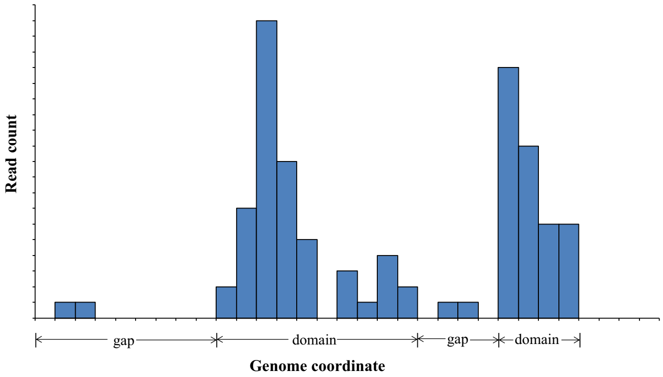

Fig. 1. Illustration of histone modification domains and gaps on the genome for a certain histone modification. Horizontal axis is the position on the genome (i.e., genome coordinate); vertical axis indicates read count. The threshold is set to 2.

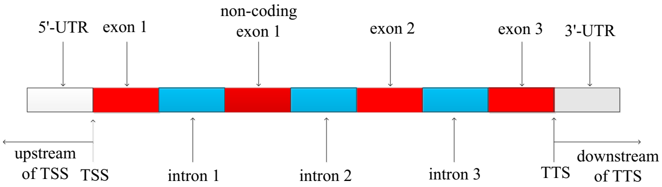

Fig. 2. Gene structure. The red regions represent exons, the blue regions represent introns, and the gray regions represent untranslated regions. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

2.3. Mining histone modification patterns

To find histone modification patterns, we first detected histone modification domains (using the method above) in the seven targeted regions: upstream of the transcription start site (TSS), downstream of the transcription termination site (TTS), codingexons, exons, introns, 3' untranslated regions (3'-UTR) and 5' untranslated regions (5'-UTR), as shown in Fig. 2.

Then we checked whether a region is valid for a certain histone modification. To do this, we use a coverage threshold. If a region's coverage by a histone modification's domains surpasses the threshold, the region is valid. For upstream and downstream regions, the threshold is set to 50%; and for the other regions, the threshold is set to 80%. The reason is that upstream and downstream regions (up to 5 kbps) are longer than the others. Concretely, if the domains of histone mark *Hx* cover over 50% of the TSS or TTS region on a gene, the gene is considered *Hx* modified on TSS or TTS. Similarly, if the domains of histone mark *Hx* cover over 80% of the codingexons/exons/introns/3'-UTR/5'-UTR region on a gene, the gene is considered *Hx* modified on codingexons/exons/introns/3'-UTR/5'-UTR.

For each type of regions, its modification pattern can be represented as a matrix G (so we have seven matrices corresponding to the seven types of regions), where $G[i,k]=1$ means that histone modification i modifies gene k, and $G[i,k]=0$ indicates that histone modification i does not modify gene k. Furthermore, we counted the number of co-modified genes for each pair of histone marks (modifications or variant) within a target region, the result was also represented by a matrix R. So we have

$$R[i,j]=\sum_{k=1:n}(G[i,k]\wedge G[j,k]) \qquad (1)$$

[1] http://genome.ucsc.edu