Gene 518 (2013) 171–178

Contents lists available at SciVerse ScienceDirect

Gene

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

Discovering common combinatorial histone modification patterns in the human genome

Changgui Linghu [1], Haiying Zheng [1], Liping Zhang [1], Juhua Zhang *,[2]

School of Life Science, Beijing Institute of Technology, Beijing, China

ARTICLE INFO

Available online 9 December 2012

Keywords:
Epigenetic regulation
Clustering algorithm
Combinational patterns

ABSTRACT

Histone modifications play a crucial role in regulating gene expression and cell lineage determination and maintenance at the epigenetic level. To systematically investigate this phenomenon, this paper presented a statistical hybrid clustering algorithm to identify common combinatorial histone modification patterns. We applied the algorithm to 39 histone modification marks in human CD4+ T cells and detected 854 common combinatorial histone modification patterns. Our results could cover 211 (76.17%) patterns among 277 patterns identified by the tandem mass spectrometry experiments. Based on the frequency statistical analysis, it was found that the co-occurrence frequencies of 20 backbone modifications are greater than or close to 0.2 in the 854 patterns. we also found that 15 modifications (H2BK120ac, H4K91ac, H2BK20ac, etc.), three histone acetylations (H2AK9ac, H4K16ac, and H4K12ac) and five histone methylations (H3K79me1, H3K79me2, 3K79me3, H4K20me1, and H2BK5me1) were most likely prone to coexist respectively in these patterns. In addition, we found that DNA methylation tends to combine with histone acetylation rather than histone methylation.

1. Introduction

Histone modifications are important components of epigenetic regulation, which include histone methylation, acetylation, phosphorylation, ubiquitination, and sumoylation. Such modifications interact with adaptor proteins that recruit other protein complexes in order to maintain the modified state, and to specify cell lineage. The histone modifiers neutralize the charges on the nucleosome, altering its structure and causing the chromatin to open or close (Felsenfeld and Groudine, 2003; Liang, 2010). The varying types of histone modifications can give rise to different biological effects: histone acetylation generally is associated with gene activation (Wang et al., 2008), and histone methylation can lead to either gene repression or activation depending on the modification site (Barski et al., 2007; Mikkelsen et al., 2007; Shi and Dawe, 2006; Zhang et al., 2007). Numerous studies (Kalebic, 2003; Schones and Zhao, 2008; Shahbazian and Grunstein, 2007) also have demonstrated that single histone modification often plays a non-independent role, and the combinations of different modification jointly affect gene expression and transcription through synergism and antagonism, which is the so-called "histone code" hypothesis.

Much of the early work (Schones and Zhao, 2008; Strahl and Allis, 2000; Wang et al., 2008) on the identification of histone modifications for ChIP-Chip/Seq data mainly focused on single pattern and its relationship with the gene expression, which rarely involving the combinatorial patterns of histone modifications. To date, only a small number of histone modification combinatorial patterns have been confirmed. Wang et al. (2008) reported that in human $CD4^+$ T cells 39 types of histone modifications tended to aggregate in certain locations in genome. Modifications of combinational states could be diverse in the enhancers, promoters, and insulators. For example, the active promoters were generally with a high level of H2BK5me1, H3K9me1, H3K27me1, H4K20me1, and H3K36me3. In a more recent work, Ernst and Kellis (2010) proposed an alternative HMM algorithm based on the binarization of presence or absence of each histone mark. They found 51 unique types of chromatin state. Each of these states was associated with a particular function, such as suppressing or increasing the activity of a gene or a class of genes. The authors also found several chromatin states likely to represent different classes of enhancer regions, which can increase gene expression while sitting far away from their target genes. More recently, Ucar et al. (2011) proposed a scalable subspace clustering algorithm, coherent and shifted bicluster identification (CoSBI) based on ChIP-Seq data. By using this algorithm, the authors had identified 843 combinatorial patterns that recurred at >0.1% of the genome. However, multiple studies so far have demonstrated that many combinatorial patterns only involve a few chromatin modifications (Schones and Zhao, 2008; Ucar et al.,

Abbreviations: CD4, cluster of differentiation 4; CoSBI, coherent and shifted bicluster identification; ChIP-chip, chromatin immunoprecipitation coupled with DNA microarray; ChIP-seq, chromatin immunoprecipitation coupled with ultra high-throughput sequencing; DNA, deoxyribonucleic acid; HMM, hidden Markov model; MACS, model-based analysis of ChIP-Seq; MS, mass spectrometry; SOM, Self-Organizing Map.

* Corresponding author.
E-mail address: jhzhang@bit.edu.cn (J. Zhang).

[1] Jointly conceived the studies, designed and performed the experiments and drafted the manuscript.

[2] Proposed, supported and supervised the researches on this paper.

0378-1119/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2012.11.038

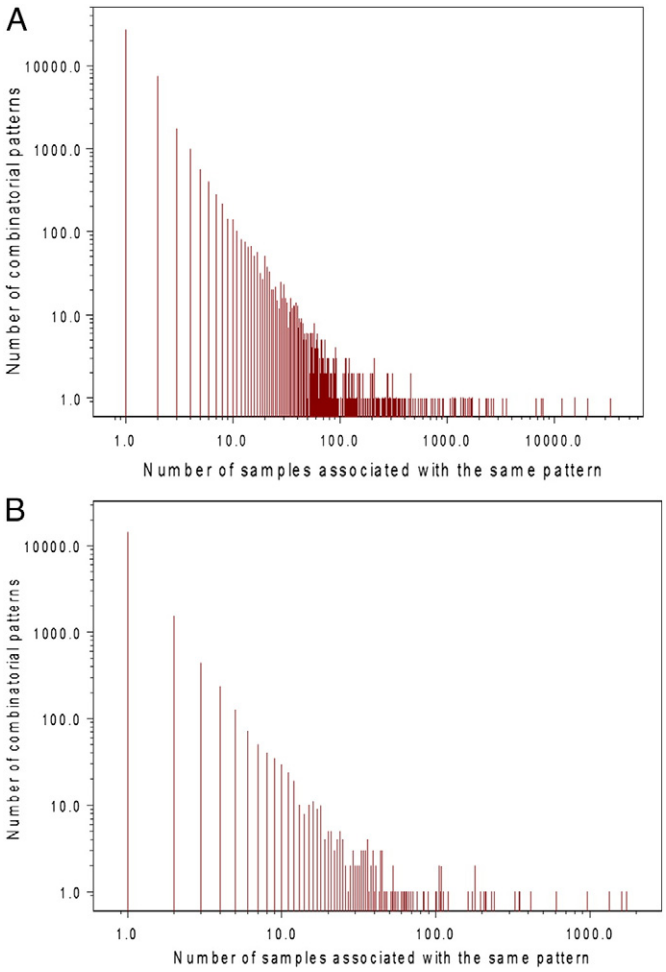

Fig. 1. The relationship between the number of histone combinational patterns and the number of samples. A. The y axis indicates the number of combinational patterns of 39 histone modifications, and the x axis indicates the number of samples associated with the same pattern. We examined the co-occurrence counts for each of the 39,732 samples in the whole genome, and 26,595 samples occurred only once (the size of window is 5000 bp). B. The y axis indicates the number of combinational patterns of 39 histone modifications, and the x axis indicates the number of samples associated with the same pattern. By regarding gene fragments as windows, we detected 17,141 histone modification patterns, while 14,347 patterns occurred only once among them.

2011). We still so far lack suitable and novel computational methods to extract histone modification patterns from ChIP-Chip/Seq datasets, so we need further discovery of the combinational interrelationship among these histone modifications.

According to the current situation, we proposed a statistical hybrid clustering algorithm to discover histone modification patterns, and detected 854 combinatorial patterns. The performance comparison shows that our algorithm is almost consistent with previous reports (Garcia et al., 2007; Strahl and Allis, 2000; Ucar et al., 2011).

2. Materials and methods

2.1. Data source

In this paper, we used genome-wide maps of 39 histone modifications in human $CD4^+$ T cells. Wang et al. (2008) and Barski et al. (2007) gave the links to download the histone modification ChIP-Seq in tag coordinate bed files (BED). Those datasets contain 18 kinds of methylations and 20 kinds of acetylations and a histone variant H2A.Z.

2.2. Data pre-processing

According to Ucar et al. (2011), the genome is split into consecutive non-overlapping windows as *genomic loci*. In this paper we also use the concept of genomic loci on behalf of a small segment on chromosome, and the position of the genomic loci is represented by the midpoint of the window, while the length of the genomic loci is the size of the window. Using MACS (Zhang et al., 2008) with default parameters, we first identified a large amount of signal peaks associated with histone modifications in $CD4^+$ T cells. Then to avoid omitting any of the possible histone modifications, we used a half-overlapping window with a fixed length (such as 5000 bp) for each of the chromosomes to exhaustively capture combinations of histone modification. The so-called half-overlapping window is that the lower half parts of the current window overlap with the upper half of the next window. Using this strategy, we identified more than ten thousand of tag data associated with histone modification patterns in the whole genome.

2.3. Methods

To identify common combinatorial histone modification patterns that frequently co-occur in the human genome, we propose a statistical hybrid clustering algorithm as follows. Before going into details of the algorithm, we start by defining the parameters and the concept of histone modification patterns.

Let g_{ij} be a tag dataset of histone modifications, where i is a set of genomic loci which has a fixed length, and j is a set of histone