Gene 575 (2016) 90–100

Contents lists available at ScienceDirect

Gene

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

Research paper

Association analysis between the distributions of histone modifications and gene expression in the human embryonic stem cell

Wen-Xia Su [a], Qian-Zhong Li [a,*], Yong-Chun Zuo [a,b], Lu-Qiang Zhang [a]

[a] Laboratory of Theoretical Biophysics, School of Physical Science and Technology, Inner Mongolia University, Hohhot 010021, China
[b] The Key Laboratory of Mammalian Reproductive Biology and Biotechnology of the Ministry of Education, College of Life Sciences, Inner Mongolia University, Hohhot, 010021, China

ARTICLE INFO

Article history:
Received 6 February 2015
Received in revised form 13 May 2015
Accepted 20 August 2015
Available online 22 August 2015

Keywords:
Histone modification
Gene expression
Regional distribution
Embryonic stem cells

ABSTRACT

It is well known that histone modifications are associated with gene expression. In order to further study this relationship, 16 kinds of Chip-seq histone modification data and mRNA-seq data of the human embryonic stem cell H1 are chosen. The distributions of histone modifications in the regions flanking transcription start sites (TSSs) for highly expressed and lowly expressed genes are computed, respectively. And four types of distributions of histone modifications in regions flanking TSSs and the spatial patterning of the correlations between histone modifications and gene expression are detected. Our results suggest that the correlations between the regions overlapped by peaks are higher than the non-overlapped ones for each histone modification. In addition, to obtain the effect of the cooperative action of histone modification on gene expression, five histone modification clusters are found in highly expressed and lowly expressed genes, histone modification and gene expression interaction network is constructed. To further explore which region is the main target region for the specific histone modification, the human genes are divided into five functional regions. The results indicate that histone modifications are mostly located in the promoters of highly expressed genes versus the exons of lowly expressed genes, and exons have a smaller range of normalized tag counts than other gene elements in the two groups of genes. Finally, the type specificity and regional bias of histone modifications for 11 key transcription factor genes regulating the stem cell renewal are analyzed.

1. Introduction

Epigenetic modification, which participates in directing spatial- and temporal-specific gene expression, is a hot topic in recent years. It plays a crucial role in cellular differentiation (Mikkelsen et al., 2007; Wei et al., 2009; Young, 2011; Xie et al., 2013) and diseases (Shen and Laird, 2013; Roos-Araujo et al., 2014; Sharma et al., 2014). Epigenetics consists of DNA methylation, histone modifications, chromatin remodeling, and RNA interference, etc. Histone modifications, being intimately related to the control of the gene expression, are one of the most important parts of epigenetics (Kouzarides, 2007; Eberhart et al., 2013). Histones can be modified by numerous covalent modifications, including methylation, acetylation, phosphorylation and ubiquitylation. Those modifications mainly occur at their N-terminal tails and can affect the expression of genes (Kouzarides, 2007). For example, histone acetylation is generally associated with gene activity; histone lysine methylation can either repress or activate gene expression according to which site is methylated or the site is methylated by how many methyl groups (Martin and Zhang, 2005), and H3-K9 and H4-K20 trimethylations are generally associated with mammalian heterochromatin regions which have low expression levels (Schotta et al., 2004). Specifically, H3-K4 and H3-K27 trimethylations which are associated with gene activity and repression have been used as bivalent modifications in cellular differentiation, respectively (Mikkelsen et al., 2007).

With the rapid development of sequencing technique, many important researches can be executed. As pointed out in a recent review (Chou, 2015), the explosive growth of biological sequences generated in the Postgenomic Age has stimulated an unprecedented revolution in medicinal science (Zhong and Zhou, 2014). For instance, a series of recent studies have demonstrated that many useful clues for drug development and medical treatment can be acquired by conducting various kinds of genome analyses (Chen et al., 2013, 2014a,b,c; Guo et al., 2014; Lin et al., 2014; Liu et al., 2015a,b,c,d; Qiu et al., 2014) by means of the approach of pseudo nucleotide composition or PseKNC (Chen et al., 2014a,b,c, 2015; Liu et al., 2015a,b,c,d).

The rapid development of sequencing technique also promotes the researches related to histone modifications. The data of histone modifications have grown exponentially by the next-generation sequencing techniques including chromatin immunoprecipitation (ChIP) coupled with microarray (ChIP-chip) (Ren et al., 2000), and high-throughput

Abbreviations: 3′UTR, 3′ untranslated regions; 5′UTR, 5′ untranslated regions; ES, embryonic stem; RPKM, the reads per kilobase of exon model per million mapped reads; TF, transcription factor; TSS, transcription start sites.

* Corresponding author.
E-mail address: qzli@imu.edu.cn (Q.-Z. Li).

http://dx.doi.org/10.1016/j.gene.2015.08.041
0378-1119/

sequencing (ChIP-seq) (Barski et al., 2007). These data can be used to determine the genome-wide locations of histone modifications (Beck et al., 2012). Based on these locations, the profiles of the modifications can be obtained; and the relationships between histone modifications and gene expression may be also further investigated. But dealing with such big data is quite difficult and time-consuming, which really is a challenge for computational biology.

Recently, several studies of histone modifications have been published (Wang et al., 2008; Karlic et al., 2010; Pekowska et al., 2010; Cheng et al., 2011; Zhang and Zhang, 2011; Cheng and Gerstein, 2012; Jung and Kim, 2012; Su et al., 2012; Dong et al., 2012). Barski et al. have obtained the genome-wide distributions of 20 histone methylations in $CD4^+$T cells, and their results demonstrated that the histone methylations are related to gene activation and gene repression (Barski et al., 2007). Further, Zhibin Wang et al. detected a module consisting of 17 modifications from 39 histone modifications for $CD4^+$ T cells, suggesting that histone modifications may act cooperatively to affect gene expression (Wang et al., 2008). Karlic et al. used histone modification levels to predict the gene expression by developing a quantitative model. Their work indicated that histone modification levels are related to gene expression for $CD4^+$T cells (Karlic et al., 2010). Cheng et al. indicates also that H3-K4 and H3-K27 trimethylations are important to gene expression in mouse embryonic stem cells (Cheng and Gerstein, 2012).

The epigenetics modification is cell-type specific. In recent years, many projects including the ENCODE Project and the US National Institutes of Health (NIH) Roadmap Epigenomics Program have used the embryonic stem (ES) cell as a principal cell type for epigenetics researches (Meissner, 2010). Embryonic stem (ES) cells have the ability to self-renew and maintain the potential to differentiate into every adult tissue type (Smith, 2001). It has been suggested that different chromatin structure states or histone modifications are related to different adult tissue types differentiated from embryonic stem cells (Azuara et al., 2006). And stem cell renewal can be controlled by a network consisting of transcription factors and epigenetic modifications (Zhou et al., 2011).

However, the potential of differentiating into every tissue type and stem cell renewal are still not completely understood, and the major histone modifications studied in the embryonic stem (ES) cells are H3-K4 and H3-K27 trimethylations (Pan et al., 2007; Zhao et al., 2007; Van Heeringen et al., 2014). So we particularly choose this important cell type and more histone modifications as research content in this paper to further obtain the relationships between the distributions of histone modifications in different regions and gene expression for embryonic stem (ES). The Pearson correlation coefficients between gene expression levels and histone modification signals are firstly obtained to quantify the relationship between them. The results show that most of the histone modifications are positively associated with gene expression, only two modifications show negative association. So the human genes are separated into two groups, the highly expressed genes and the lowly expressed genes, according to the reads per kilobase of exon model per million mapped reads (RPKM) (Mortazavi et al., 2008). The distributions of 16 kinds of histone modifications in the regions flanking transcription start sites (TSSs) for highly expressed and lowly expressed genes are computed, respectively. Four types of distributions of different histone modifications are found by calculating the p-value of the peaks and the spatial patterning of the correlations between histone modifications and gene expression are also obtained.

For the purpose of discovering the cooperative effect of different sites for the same histone modification on gene expression, the correlations between different bins for each histone modification are calculated. Our results suggest that the correlations between the bins overlapped by peaks are higher than the non-overlapped ones for each histone modification. In addition, we study the cross-talking of histone modifications by detecting histone modification clusters in highly and lowly expressed genes, respectively, five histone modification clusters are found in the two groups of genes.

In order to further explore which region is the main target region for the specific histone modification, the human genes are divided into promoter, 5′UTR, exon, intron, 3′UTR regions based on the annotation of the RefSeq genes. The distributions of histone modifications in these regions are analyzed. The results indicate that histone modifications are mostly located in the promoters of highly expressed genes versus the exons of lowly expressed genes, and exons have a smaller range of normalized tag counts than other gene elements in the two groups of genes. Finally, 11 transcription factors that are critically important to stem cell renewal, are chosen to study the type specificity and regional bias of histone modifications.

2. Materials and methods

2.1. The data of reference human genes and histone modifications

The RefSeq genes of human genome are downloaded from UCSC (http://genome.ucsc.edu/), which contain the gene names, names, chromosomes, strands, transcription starts, transcription ends, translation starts, translation ends, exon counts, exon starts, and exon ends. The genes which begin with NM (the mature messenger RNA) are remained. In order to avoid the case that some of the RefSeq genes are actually the alternative transcripts of the same gene, only one of the genes which have the same transcription start site (TSS) is remained for all RefSeq genes. At last, total 19,120 genes are chosen.

The ChIP-seq data of 16 kinds of histone modifications (H2BK12ac, H3K4ac, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me3, H3K14ac, H3K18ac, H3K23ac, H3K27ac, H3K27me3, H3K36me3, H3K79me1, H4K8ac and H4K91ac) for the human H1 cell line are chosen from the NIH Roadmap Epigenomics project (http://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/).

2.2. Partition of highly and lowly expressed genes

The gene expression data of the human H1 cell line, measured by mRNA-seq, are also chosen from NIH Roadmap Epigenomics project (http://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/). To quantitatively measure the expression of genes, we use RPKM as the expression values of genes. Then the genes are ranked by RPKM, the top fifteen percent and the bottom fifteen percent of all genes are selected as the highly and lowly expressed genes. Finally, each group consists of 2868 genes, respectively.

2.3. Calculation of histone modification signal profiles

2.3.1. The tag counts of histone modifications in the five functional regions

It has been proposed that different histone modifications are preferred to locate in different functional elements in mammalian genomes (Takahashi and Yamanaka, 2006). So we divide each gene into promoter, 5′UTR, exon, intron, 3′UTR regions based on the annotation of the RefSeq genes. In order to get the signal profiles of 16 kinds of histone modifications in these regions, the normalized tag counts are calculated based on the ChIP-seq data. In order to eliminate the influence of the length of regions and the sequencing depth, we define the computational formulas of the normalized tag counts as Eqs. (1) and (2) according to computing method of RPKM. The tag counts of histone modifications are normalized by the length of the functional regions and the total tag counts in the measurement, showing as Eq. (1), in which $N_r(tag\ counts)$ represents the normalized tag counts, n_{region} is the total tag counts that located in this functional region, l_{region} is the length of this functional region, and n_{tag} is the total tag counts in the measurement.

$$N_r(tag\ counts) = n_{region}/(l_{region} \times n_{tag}) \times 10^9 \quad (1)$$