Study of the Epigenetic Signals in the Human Genome

Susana Ferreira, Vera Afreixo[1, 2], Gabriela Moura[2], Ana Tavares[1]

[1]Department of Mathematics & CIDMA, University of Aveiro, Portugal
[2]Department of Medical Sciences & iBiMED, University of Aveiro, Portugal

Abstract. Epigenetics can be defined as changes in the genome that are inherited during cell division, but without direct modification of the DNA sequence. These genomic changes are supported by three major epigenetic mechanisms: DNA methylation, histone modification and small RNAs. Different epigenetic marks function regulate gene transcription, some of them when altered can trigger various diseases such as cancer. This work is focus on the epigenetic signals in the human genome, studding the dependency between the nucleotide word context and the occurrence of epigenomic marking. We based our study on histone epigenomes available in the NIH Roadmap Epigenomics Mapping Consortium database that contains various types of cells and various types of tissues. We compared genomic contexts of epigenetic marking among chromosomes and among epigenomes. We included a control scenario, the DNA sequence regions without epigenetic marking. We identified significant differences between context occurrence of control and epigenetic regions. The genomic words in epigenetic marking regions present significant association with chromosome and histone modification type.

Keywords: Epigenome, histone modification, epigenetic marking, genome context, data analysis.

1 Introduction

Epigenetics is one of the most promising and intriguing areas of genetics. It is the science that studies the interaction between gene regulation, i.e. how genes are expressed, and its surrounding environment without involving changes in the DNA sequence level, which may still persist in future generations (1–3). The inheritance of epigenetic marks from mother to daughter cells is crucial for the maintenance of a cell differentiation state and could be propagated by various epigenetic mechanisms, such as, DNA methylation, histone modifications and replacement of histone variants. The cell differentiation is a natural event in every organism, which involves no alteration of DNA sequence. However, all cells in an organism share the same genome (except B lymphocytes), each cell type has different kinds of epigenetic signatures, and each

adfa, p. 1, 2011.

has a cell-type specific epigenome (1–3). In other words, epigenetic has to do with changing the whole genome regulatory activity and this can be resumed in the epigenome, which is a kind of map that overlays the map of the genome, with epigenetic means that turn on or off genes, increasing or reducing its activity. The epigenome can be studied through genomics and an important note is that epigenome is not static as the genome, it can be dynamic, influenced by environmental factors and extracellular stimuli, and change rapidly in response to these factors (4,5).

Epigenetics can be regulated by three mechanisms: DNA methylation, histone modification and small RNAs. In this essay, we have studied epigenetic signals presented in the human genome, focusing mainly on the epigenetic regulation related to histone modification. Its importance will be described as follows.

Histone modifications and chromatin structure: The DNA is wrapped around two copies of each of the four core histone proteins H3, H4, H2B, and H2A, to form the nucleosome which is the fundamental repeating unit of chromatin (6–8). The chromatin will be necessary for efficient packaging of the DNA into the nucleus of the cell. However, when DNA is compacted into the chromatin, its accessibility becomes greatly limited, it serves as a mechanism by which the cell protects DNA from external damage but it also regulates DNA mediated processes, such as transcription, DNA replication, DNA repair and chromosome segregation (6–8).

So, these histone proteins can influence chromatin organization and regulate many DNA-templated processes, through their chemical modification patterns (acetylation, methylation, sumoylation, and ubiquitylation) (6,9). Changes on the histone modification status may be associated with active or inactive chromatin. In addition, the combinatorial nature of various histone modifications occurring at different times during development, and at specific sites within histones, provides additional levels of regulation and complexity to the epigenome (8). However, from these modifications can exert some biological effects, and how the addition or removal of many of these modifications is regulated, is still unclear.

The work presented in this paper studied the epigenetic signals of the human genome: the dependence between the context and the occurrence of epigenetic marking, chromosome and histone type; the identification of specific contexts related to epigenomic modification of a specific chromosome or histone.

2 Materials and Methods

We used the NIH Roadmap Epigenomics Mapping Consortium database that was designed, in 2008, to store human epigenomic data in order to encourage research (6,10). This database contains data for 31 histones modifications H2AZ, H2AK5ac, H2AK9ac, H2BK5ac, H2BK12ac, H2BK15ac, H2BK20ac, H2BK120ac, H3H4ac, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me1, H3K9me2, H3K14ac, H3K18ac, H3K23ac, H3K23me2, H3K27ac, H3K27me3, H3K36me3, H3K56ac, H3K79me1, H3K79me2, H3T11ph, H4K5ac, H4K8ac, H4K12ac, H4K20me1, H4K91ac.

The epigenome files contain the sites of epigenomic marking (start and end positions) relative to the reference genome (GRCh37), defining the epigenetic regions. The word (k-mers) counts are obtained by DNA segment regions. The sequences are classified in two subgroups:

Control regions: All regions without epigenetic marking were used as control, consisting of 43916 fragments. The control regions are represented by 19369407 nucleotides, has 5619417 A nucleotides; 5625109 T nucleotides; 4063926 C nucleotides; and 4060955 G nucleotides.

Epigenomic regions: All regions with at least one epigenetic marking. If two fragments present intersection then we join them into one. Epigenomic regions are represented by 11325056856 nucleotides, has 1672399545 A nucleotides; 1674863703 T nucleotides; 1157288962 C nucleotides; and 1157976218 G nucleotides.

The word context analysis was subdivided essentially in three subanalysis: a global analysis comparing the control and epigenomic regions; a chromosome comparison; and a histone comparison.

We use standard statistical procedures: t-test, chi-square test, Cohen’s d, Cramer’s V, residual analysis and hierarchical clustering methods. To obtain k-mer (k = 1, 2, 3, 4) counts and perform the statistical analysis we use R software.

3 Results

Control and epigenomic regions analysis: In word context, the control and epigenomic regions present significant differences with low effect size difference, for the word lengths under analysis (k=1,…,4). The comparison was performed with chi-square test (p-values<0.001) and complemented with the Cramer's V (0.001<V<0.01).

It is know that the human genome has regions of high C+G content, alternating with regions of low C+G content. To rule out the hypothesis that the C+G contents could be marking the occurrence of epigenetic marking, we explore the differences between CG relative frequencies of control and epigenomic regions. Figure 1, show the CG relative frequency for the epigenomic and the control subset, where the differences between the two groups are globally low. We also applied the t-test and we concluded the C+G content in the two groups of sequences presents significant differences (p-value<0.001). Through the Cohen’s d, we concluded that the size effect of C+G content of our analysis is very small (d=0.039). Thus, we classified the C+G bias between the two groups as negligible.

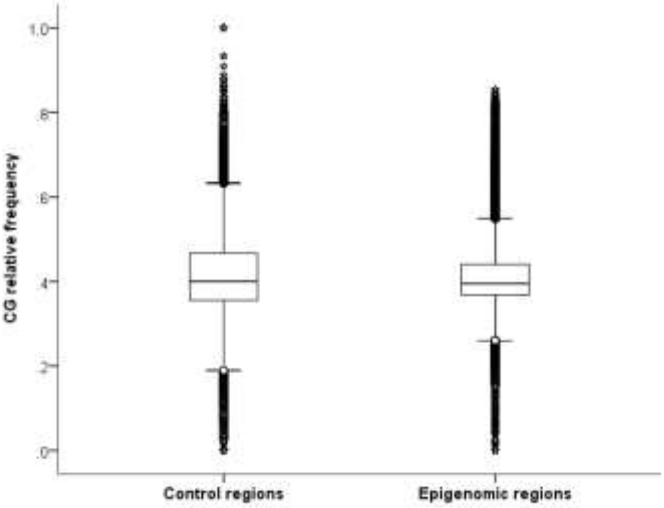

Figure 1. Boxplot of C+G content for control and epigenomic samples. Both regions have several outlier fragments with similar median values, but the set of control regions presents high values dispersion.

Chromosomes analysis: In this analysis, we wanted to evaluate if the genomic context associated with the occurrence of epigenetic marking is homogeneous, among chromosomes. For this, we applied the chi-square test and the Cramer's V value (table 1).

Parameters	X^2	df	p-value	V	N
Nucleotide	10063000	69	*	0.0243	11325056856
Dinucleotide	22290000	345	*	0.0161	11324685040
Trinucleotide	34156000	1449	*	0.0161	11324313240
Tetranucleotide	46218000	5865	*	0.0188	11323941448

Table 1. Chi-square test to evaluate the homogeneity between chromosomes for epigenomic regions word context. X^2 - chi-square test; df - degrees of freedom; V - Cramer's V association measure; N - the sample size. *p-value is <0.001.

For nucleotide, dinucleotide, trinucleotide and tetranucleotide contexts, we concluded that there was a significant heterogeneity between chromosomes. Taking into account the residuals values and the hierarchical analysis, we identified specific k-mers that were able to differentiate the human chromosomes taking into account the epigenomic regions context.

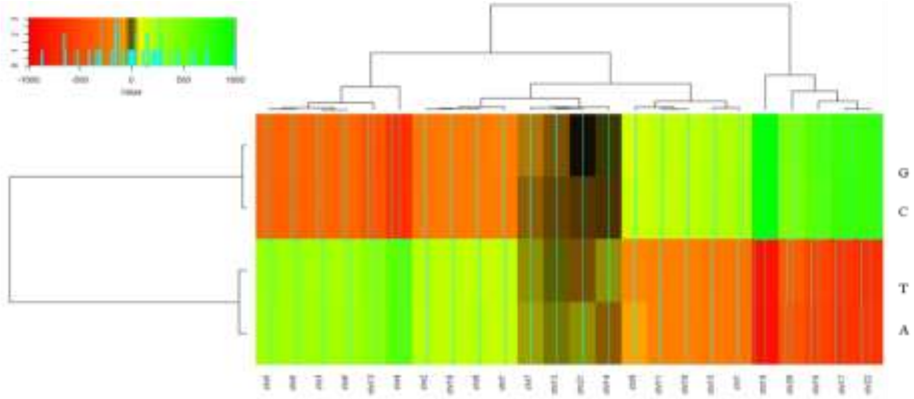

Figure 2. Heatmap of chromosomes vs nucleotides, for epigenomic regions. Three chromosomes clusters were formed, two of which have strong nucleotides preferences (for A/T or G/C, respectively) and another cluster with more similar nucleotide preferences.

For example, in the nucleotide context, through a residual analysis we can observe that there are identical profiles in various chromosomes, with similar nucleotides preference (see figure 2). Table 2 presents in simultaneous the trinucleotide words and chromosomes with the highest residual values (>20) identifying specific preferred genomic contexts.

Chromosome with favored genomic contexts	**Identify specific genomic contexts in some chromosomes**
Chr3; Chr4; Chr5; Chr6; Chr13; ChrX	TAT; ATA
Chr16; Chr17; Chr19; Chr20; Chr22	GGG; CCC; GCC; GGC

Table 2. Identification most favored genomic contexts in some chromosomes.

Histone modifications analysis: This analysis was performed in order to compare the word context between different histone modification types. Specific contexts are associated with specific histone modification (p-value<0.005, qui-square test).

Histone modifications	**Identify specific genomic contexts**
H2AZ; H2AK5ac; H2BK5ac; H2Bk15ac; H3K4ac: H3K4me1; H3K4me2: K3K4me3; H3K9ac; H3K9me3; H3K23ac; H3K27ac; H3K27me3; H3K36me3; H3K79me1; H3K79me2	TTA; TAA
H2AK9ac	GGG; GCC; GGC
H2BK12ac; H2BK20ac; H3K14ac; H3K18ac; H3K56ac; H4K12ac	GCT; CTC
H2BK120ac	TTA; TAA; CTC; GCT
H3K9me1	AGG; GCT; CTG; CCT
H3K23me2	AGG; CTC; GCT
H3T11ph	GCT
H4K5ac	GCT; CTC;TTA
H4K12ac	GGG; GGC; CCC; GCC; GAG
H4K20me1	CAG; CAC; AGG; CTC; GCT; CCT
H4K91ac	GCT; CTT; AGG

Table 3. Identify specific genomic contexts in histone modifications.

For example, in the trinucleotide context, we concluded from the analysis of residues that each modifications has specific preferences. Table 3 presents in simultaneous the trinucleotide words and histones with the highest residual values (>20).

4 Discussion

In this study we globally study the human epigenome, and the main objective was to identify motifs that could be associated with histone modification to further understand the relationship between DNA sequences and the occurrence of epigenetic marking.

Through heatmaps and the hierarchical clustering analysis, we could identify specific genomic contexts associated to each histone modification. One of the strongest contexts was TTA and TAA trinucleotides that are present mainly in regions of H2 and H3 histone modification, for both acetylation and methylation. However, there are other histone modifications that have other enriched motifs, as shown in the results. So, with these results, it may be possible to predict the occurrence of a modification from the nucleotide context of the region. Our epigenomic data is obtained from healthy cells, so with these profiles and the identification of the words with the greatest effect on the modifications, a comparison should be made between healthy and unhealthy cells and evaluate what differentiates them. It was also possible to create groups according to the type of histones (H2, H3 and H4) and the type of modification (acetylation or methylation) (6,8,9). Curiously, it was observed two distinct groups: one including transcription-activating histone modifications (normally acetylations) and other including transcription-inactivating ones (normally methylation).

The trinucleotide and tetranucleotides contexts were the most informative ones differentiate chromosomes and histones. From the results, we can speculate that increasing the word size of contexts, more information and conclusions could be addressed. Because the computational complexity, we did not study higher word length, which is a limitation of this analysis.

Acknowledgment

This work was partially supported by Portuguese funds through the FCT (Portuguese Foundation for Science and Technology), CIDMA and iBiMED, within projects: UID/MAT/04106/2013, UID/BIM/04501/2013.

References

1. Ng RK, Gurdon JB. Epigenetic inheritance of cell differentiation status. Cell Cycle. 2008;7(9):1173–7.
2. Roloff TC, Nuber UA. Chromatin, epigenetics and stem cells. Eur J Cell Biol. 2005;84(2–3):123–35.
3. Probst A V., Dunleavy E, Almouzni G. Epigenetic inheritance during the cell cycle. Nat Rev Mol Cell Biol. 2009;10(3):192–206.
4. Bernstein BE, Meissner A, Lander ES. The Mammalian Epigenome. Cell. 2007;128(4):669–81.
5. WHO. Genetics, genomics and the patenting of DNA: review of potential implications for health in developing countries. World Heal Organ. 2005;
6. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. Analysis of Dynamic Changes in Posttranslational Modifications of Human Histones during Cell Cycle by Mass Spectrometry. NIH Public Access. 2013;28(10):1045–8.
7. Scholz B, Marschalek R. Epigenetics and blood disorders. Br J Haematol. 2012;158(3):307–22.
8. Chen T, En L. Structure and Function of Eukaryotic DNA Methyltransferases. Curr Top Dev Biol. 2004;60:55–9.
9. Bird A. DNA methylation patterns and epigenetic memory. Genes Dev. 2002;16:6–21.
10. Consortium RE, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317–30.