AARHUS UNIVERSITY

MASTER THESIS

Identifying signatures of human epigenetic modifications among tissues

Author:
Alejandro ROCA ARROYO

Supervisor:
Thomas BATAILLON

A thesis submitted in fulfillment of the requirements
for the degree of Master's in Bioinformatics

in the

Bioinformatics Research Centre BiRC

May 21, 2019

AARHUS UNIVERSITY

Abstract

Faculty of Science and Technology
Bioinformatics Research Centre

Master's in Bioinformatics

Identifying signatures of human epigenetic modifications among tissues

by Alejandro ROCA ARROYO

The number of different epigenetic landscapes for a genome may be inestimable, but we can find correlations between specific epigenetic modifications which are typically associated in concrete functions and development states. In such a way, we reduce the dimensionality of the problem making it easier to draw conclusions from the analysis of the epigenetic modifications, as well as being able to use the smaller set of correlated modifications (or signatures) as input for predictive modelling or supervised machine learning analysis.

In order to achieve this, we need to map epigenetic modification reads (from Bisulfite-seq; DNA methylations, or ChIP-seq; histone modifications) into the human genome, specifically into genes and flanking/regulatory regions, which are the ones of interest. In this way we would obtain the counts of each epigenetic modification for different tissues. Non-negative Matrix Factorization (NMF) reveals as an ideal method for the task of finding combinatorial patterns of epigenetic modifications. We can then study the state of each epigenetic modification type in the defined loci of the tissue. From this information we would obtain the different epigenetic signatures which we will use for association and simulation analysis.

Contents

List of Figures

List of Tables

List of Abbreviations

LAH	**L**ist **A**bbreviations **H**ere
WSF	**W**hat (it) **S**tands **F**or

Chapter 1

Introduction

1.1 Epigenetics

The genetic material is known to be modified during the life of an organism, possibly causing modifications in gene behavior. Far is known about mutations being a mayor player in genetic variation and the profiling of these variations has proved to be highly useful when studying diseases and evolutionary theory [GWAS]. In eukaryotes, there are in addition mechanisms in which the DNA can be modified without altering the molecular sequence, so called epigenetic mechanisms. As Conrad Waddington, who coined the term "epigenetics", defined: "it is the branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being" [Waddington]. Such definition led to categorizing as epigenetics all biological phenomena which correlated the genetic material with the genetic products and were not explained entirely by the classic genetic studies. Further studies have revealed that epigenetic mechanisms can be modulated in response to external stimuli [CITATION], entailing an overlay between DNA and environment for the cells and organisms. Moreover, the epigenetic mechanisms behavior can vary for different stages of cell development [CITATION], environmental changes [CITATION] or disease [CITATIONS]. In the same way genetic variability can be profiled, it is possible to decipher shared patterns for the epigenetic modifications in different scenarios and types of tissue.

Due to the latest growth on research efforts and resources about the topic, we were able to characterized the inheritance of gene expression patterns not explained by the encoded information in the DNA sequence but through epigenetic modifications. High-throughput technologies such as Chromatin Inmunoprecipitation next-generation sequencing (ChIP-seq) or Whole-Genome Bisulfite sequencing (WGBS), allowed us to obtain an incredibly vast amount of information on epigenetic marks throughout the genome. It was then possible to determine the epigenome profile consisting of multiple chromatin states that activate or repress the gene expression in a local and cell-specific manner. This "epigenomic profiling" helped the understanding of cell differentiation, where even though the vast majority of cells in a multicellular organism share an identical genotype, the development of the various tissues generates stable but diverse profiles of gene expression, giving rise to the multiple cell types and differentiated cellular functions. In light of this, more specifically epigenetics may be defined as "the study of any potentially stable and, ideally, heritable change in gene expression or cellular phenotype that occurs without changes in Watson-Crick base-pairing of DNA" [CITATION Gol07].

The work on nucleic acids [ ], chromatin [ ] and histone proteins [ ] led to the understanding of the DNA arrangement, as being wrapped around the histone proteins [ ], and furthermore to the cytological distinction between euchromatin and heterochromatin [ ]. It has been proved that post-translational modifications of the histones, such as methylation, acetylation or phosphorylation, can reshape the chromatin structural and functional properties instating the concept of turning "on" and "off" regions of the genetic material [ ]. A challenging yet feasible task is to characterize those configurations responsible for the repression, activation or modulation of the gene expression via epigenetic modifications, both in different tissues and also when affected by diseases as in the case of the cancer cells used in this study. For further understanding, it is essential to know about the diverse ways epigenetic mechanisms work and elucidate the correlation among them and to biological processes.

1.1.1 Epigenetic modification types

Histone Modifications

Chromatin is conformed by the DNA molecule and a range of binding proteins into which histones are included. The DNA molecule winds around an octamer of histones, formed by dimers of four core histones: H2A, H2B, H3 and H4. Histone N-terminal tails, specially long in Histone 3 (H3), are subject of chemical modifications, such as methylation or acetylation, modulating their spatial configuration and with it, the arrangement of the chromatin. First works on the subject, in particular on histone acetylation [CITATION All64], implied a close linkage between the histone modification state and the local gene activity. This assumption was afterwards supported by experiments on histone-tail mutations in *Saccharomyces cerevisiae* [CITATION Kay], hypoacetylation of the inactive X chromosome in female mammals [CITATION Jep93], as well as hyperacetylation of the twofold upregulated X chromosome in *D. melanogaster* males [CITATION Bon94]. These major findings led to the compelling argument that histone modifications, along with DNA methylation, contribute to distinguish between euchromatin state and heterochromatin. Accordingly, depending on the particular histone modification profile, the chromatin can be arranged as euchromatin, which implies gene expression, or as heterochromatin, meaning the repression of the gene activity.

Chromatin state at promoters is largely invariant across diverse cell types, whereas enhancers are marked with highly cell-type-specific histone modification patterns [CITATION Hei09]. As in methylation patterns, an aberrant histone modification pattern is associated with the development of cancer [CITATION Mar01]. Histone deacetylases (HDACs) are implicated expecting both positive and negative effects on oncogenic and oncosuppressive mechanisms. Again, the importance of the histone modification patterns for the gene expression and the diseases associated with an aberrant histone modification profile, call our attention on the topic.

DNA methylation

DNA methylation is perhaps the best characterized chemical modification of chromatin and were detected as early as 1984 [CITATION Hot48]. In mammals, nearly all DNA methylation occurs on cytosine residues of CpG dinucleotides. Regions of the genome that have a high density of CpGs are referred to as CpG islands, and DNA methylation of these islands correlates with transcriptional expression [CITATION Raz80, Bir85]. De novo or maintenance DNA methyltransferases (DNMTs) play a critical role in gene regulation, especially those associated with transposons and imprinted genes [CITATION Gol05], by keeping the genomic patterns of cytosine methylation during embryogenesis and gametogenesis. Moreover, the formation of heterochromatin in several organisms is mediated partly by DNA methylation and its binding proteins together with RNA and histone modifications. DNA methylation takes part in many cellular processes including silencing of repetitive and centromeric sequences from fungi to mammals [CITATION Par02 ł3082]; X chromosome inactivation in female mammals [CITATION Lyo61]; and mammalian imprinting [CITATION Sur84, McG84], all of which can be stably maintained.

As the other epigenetic mechanisms, DNA methylation is reversible and therefore DNA methylation patterns vary in time and space during differentiation [CITATION Bir02]. However, abnormal control of the methylation pattern was detected in cancer cells and may result in the generation of random modification patterns which may serve to unleash new genes for transcription [CITATION Jon86]. The abnormal methylation pattern can be either hypomehylated, which usually involves repeated DNA sequences, or hypermethylated which involves CpG islands [CITATION Ehr02]. In the first case, oncogenes are activated whereas hypermethylation repress the transcription of the promoter regions of tumor suppressor genes, leading to gene silencing.

RNA-Associated Silencing

RNA silencing is another method to turn off genes when it is in form of antisense transcripts, noncoding RNAs or RNA interference. Antisense double-stranded RNA complementary to targeted mRNAs was detected as a method of Post-Transcriptional gene silencing (PTGS) for both cellular and viral genes in a sequence-specific manner [CITATION Ham99]. This kind of process is known as RNA-mediated interference or RNAi. Moreover, small non-coding RNAs were identified as potencial 'templating' molecules for the location-specific epigenetic modifications. Several researches reported the involvement of small nuclear RNAs (snRNAs) in interacting with and presumably directing chromatin-modifying activities [CITATION Vol02, Moc02, Mar]. The snRNAs participate in a nuclear process known as 'transcriptional gene silencing' (TGS) guiding the epigenetic machinery not only for heterochromatin assembly and gene silencing [CITATION Mar] but also directing programmed DNA elimination [CITATION Cha13].

Alternative Splicing

Alternative splicing is one major mechanism that makes the most of the precursor messenger RNAs (pre-mRNAs) by processing the pre-mRNA into a diverse array of mature mRNAs that encode distinct proteins. This phenomenon explains the high complexity of organisms as humans while they have a relatively small number of protein-coding genes. Alternative splicing of RNA leads to a variety of possible mRNA isoforms and proteins, which can have different, and often opposing, functions (Figure 1). Sequences called exons are regions of the pre-mRNA that are included in the mature mRNA, such as the protein-coding sequences and regulatory untranslated regions at either end of the mRNA. Sequences called introns are the portions of the pre-mRNA that are removed during splicing. In alternative splicing, some sequences serve as exons under some conditions and are included in the final mRNA. At other times, however, the alternative splicing process may exclude the same sequence, treating it as an intron and removing it from the mature mRNA.

A critical finding regarding the prevalence of alternative splicing was that a majority of human genes produce a wide variety of messenger RNAs (mRNA) that in turn encode distinct proteins [CITATION Joh03]. Scientists estimate that 15–60 percent of human genetic diseases involve splicing mutations, either through direct mutation of the splice-site signals or through disruption of other components of the splicing pathway [CITATION Wan07 ł3082]. Therefore, understanding how the splicing machinery distinguish between exons, which are part of the mature mRNA, and introns, which are removed from the pre-mRNA, is of critical importance. Alternative splicing adds an extra layer of complexity, because regulatory sequences that sometimes designate an exon's inclusion into the mature mRNA dictate the exclusion of that exon under other conditions.

1.1.2 Epigenomic modelling

Explain previous attempts on modelling the epigenetic profiles and where did they focus

1.2 Non-negative Matrix Factorization

Explain briefly the NMF principles and explain previous applications.

1.2.1 NMF and gene expression

Application of NMF to gene expression

1.2.2 NMF and mutations

Application of NMF to mutational processes

1.2.3 NMF and epigenetics

Application of NMF to epigenetics,

Explain the use of Gandolfi, 2017 as guidance an explain differences

Gandolfi, Francesco, and Anna Tramontano. "A computational approach for the functional classification of the epigenome." Epigenetics & chromatin vol. 10 26. 15 May. 2017, doi:10.1186/s13072-017-0131-7

Chapter 2

Methods

2.1 NMF description

Explain the statistical framework in detail

2.1.1 Algorithms

Explain the different types of algorithms to solve the problem and the one used.

2.1.2 Choosing number of signatures

Explain different ways to choose N, the way used, the N used and why.

2.2 Bio Justification

[Change title]

Explain the relation between signatures found in NMF and biological processes

2.3 Data

The data used in this study is obtained entirely from the ENCODE project database [ ], where multiple epigenetic marks and reference epigenomes are available. Data sets from 11 types of epigenetic marks were collected, including histone modifications (H3K4me3, H3K4me1, H3K27ac, H3K36me3, H3K9me3, H3K27me3 and H3K9ac), chromatin remodeling proteins (EP300 and H2A.Z) and transcription regulation factors (CTCF, POLR2A). Replicate samples for each of the 11 epigenetic marks were used in the three cell lines of study, which includes a human liver cancer cell line (HepG2) [ ], myelogenous leukemia cell line (K562) [ ] and cells derived from HeLa cancerous cervical tumor line (HeLa-S3).

In order to standardize the input and facilitate the data sets processing, the information for all the samples was arranged in a tab separated file with the following fields as columns:

1. **Cell line type**. Either HepG2, K562 or Hela-S3.
2. **Epigenetic modification category**. Where does the modification apply. Either histone modification ('histonemod'), chromatin modulation ('openchromatin') or transcription factor ('TFBS').
3. **Epigenetic modification name**. The label of the mark which the sample corresponds to, as mentioned above.

4. **Accession ID**. The accession name for the sample in the database.
5. **File name**.
6. **Processing status**. Either 'raw' or 'process'. Raw samples need to be preprocessed.
7. **Replicate number**. Integer enumerating the various replicates for a particular epigenetic mark.
8. **Database name**. In the present case, 'ENCODE'.
9. **Download link**. Used to download the sample reads file.

As a matter of convenience, '`downloadData.sh [DataInfo.tsv]`' or '`epigeNMF.sh -d [DataInfo.tsv]`' can be used to automatically download the data sets into the appropriate directory tree: `CELL_LINE/SIGNAL_TRACK/SAMPLE_ID`.

2.4 Pipeline

Explain the command line tool, the pipeline followed for the analysis and explain functioning of each script

You should follow the next pipeline for the analysis:

1. Download the data (option `-d` or `--download`)
2. Prepare BED alignment files (option `-p` or `--process`)
3. Create V matrix of counts for marks (cols) by bins (rows) (option `-c` or `--counts`)
4. Filter V matrix bins to remove noise (option `-f` or `--filter`)
5. Choose the optimal 'n' number of signatures for NMF (option `-k` or `--chooseN`)
6. NMF analysis (option `-n` or `--nmf`)

2.5 Performance

Perform analysis with percentage of the data and get performance (time) plots

Chapter 3

Results

3.1 Coverage

Study the epigenetic marks along the different chromosomes

3.2 NMF Signatures

H matrix

Study the NMF signatures obtained and compare between different tissues. Relate to biology.

3.3 Association Study

[MAYBE]

W Matrix

Get hotspots and study association and relation to biology.

Chapter 4

Discussion

- Comparison
    - Convenience
    - Outcome
    - Future perspectives

Appendix A

Appendix A