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

June 12, 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 [Wray et al., 2007]. 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, 1942]. 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 [Liu et al., 2004, Bäckdahl et al., 2009], 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 [Kiefer, 2007], environmental changes [Sutherland and Costa, 2003] or disease [Jessberger et al., 2007]. 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" [Goldberg et al., 2007].

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 [Elgin, 1996]. 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 [Schatz et al., 1964], 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* [Kayne et al., 1988], hypoacetylation of the inactive X chromosome in female mammals [Jeppesen and Turner, 1993], as well as hyperacetylation of the twofold upregulated X chromosome in *D. melanogaster* males [Bone et al., 1994]. 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 [Heintzman and Ren, 2009]. As in methylation patterns, an aberrant histone modification pattern is associated with the development of cancer [Barneda-Zahonero and Parra, 2012]. 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 1948 [Hotchkiss, 1948]. 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 [Deaton and Bird, 2011]. De novo or maintenance DNA methyltransferases (DNMTs) play a

critical role in gene regulation, especially those associated with transposons and imprinted genes [Goll and Bestor, 2005], 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 [Partridge et al., 2002, Jones, 2012]; X chromosome inactivation in female mammals [Weber et al., 2005]; and mammalian imprinting [Bartolomei and Ferguson-Smith, 2011], 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 [Jones and Taylor, 1980, Xie et al., 2013]. 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 [Berdasco and Esteller, 2010]. The abnormal methylation pattern can be either hypomehylated, which usually involves repeated DNA sequences, or hypermethylated which involves CpG islands [Weber et al., 2005]. 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 [Hamilton and Baulcombe, 1999]. 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 [Mochizuki et al., 2002, Borges and Martienssen, 2015]. 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 [Grewal and Moazed, 2003], but also directing programmed DNA elimination [Bernstein and Allis, 2005].

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. 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 [Johnson et al., 2003]. 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 [Wang and Cooper, 2007]. 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 modeling

Mentioned growth on epigenetics data availability, favored by the arise of NGS sequencing techniques, granted the possibility to create extensive descriptions on the different epigenetic marks for various samples. Moreover, the findings on the function of epigenetic modifications to modulate the chromatin arrangement provided a glimpse of the likeliness to characterize the relationships between chromatin states and gene expression profiles. Thanks to the contribution made to the two major public databases, the ENCODE (The ENCyclopedia of Dna Elements) [Feingold et al., 2004] and the NIH Roadmap Epigenomics [Bernstein et al., 2010], this characterization becomes more accessible. Modeling the epigenetic modifications requires the development of computational tools which use the multiple types of data, such as DNA methylation, histone modification or Transcription Factors (TFs), in order to establish a correlation between the epigenomic assays and the biological activity. The main objectives of finding the specified models are both learn more about the biological association of epigenetic modifications and which predictions can be achieved from the models. The nature of the algorithms used to model epigenomics have been diverse but most of the proposed methods fall under unsupervised learning techniques of classification, which seek to devise patterns of chromatin modifications from the datasets alone, relying for it on multiple statistical techniques.

A first proper mathematical attempt to model histone modification dynamics was made in [Dodd et al., 2007], where they characterized the bistable gene expression resulting from the alternative states of the chromosomal regions. They applied a simple stochastic model in which they divided the DNA into $\tilde{1}$.2 kb, corresponding to around 60 nucleosomes, and they consider either modified or unmodified nucleosomes from which they try to infer the state conversion of a region. Further contributions delved into the pattern-finding idea using more complete models such in ChromaSig [Hon et al., 2008]. Nine different types of epigenetic marks were taken into account in order to find strong signals of correlation between the epigenetic signatures profile and the promoter and enhancer activity state. Nevertheless, the analysis worked on a pre-defined set of loci with high grade of epigenetic modifications from the ENCODE project pilot, representing only the 1% of the human genome [ENCODE Project Consortium et al., 2007]. The essence of the method requires to calculate the likelihood for each loci towards being classified in a motif by using an euclidean distance measure. On account of the computational limitations,

these methods could not be applied to study whole-genome marks and on multiple genomic assays.

More recent tools are able to directly specify the combination of epigenetic modifications or 'chromatin states' in a genome-wide fashion, making use of integrative models based on Hidden Markov Models (HMM) like in ChromHMM [Ernst and Kellis, 2010, Ernst and Kellis, 2017] and EpicSeg [Mammana and Chung, 2015], or dynamic Bayesian networks like in Segway [Hoffman et al., 2012]. ChromHMM software learns and characterizes the chromatin states from multiple ChIP-seq datasets by creating tracks of presence/absence vectors for the epigenetic mark in a k-binned genome. In such a way, genome-wide annotation is efficient but it misses the quantitative hand of the epigenetic mark reads, using only binary data. EpicSeg uses a similar angle in the analysis, however there is no need to pre-process the data as it uses raw read counts like observations in the analysis, which allows to define a valid discrete multivariate probability distribution and then solve the lost of quantitative data. For its part, Segway refuses the idea of genome segmentation and addresses the most probable sequence of chromatin states by means of a Dynamic Bayerian Network (DBN) at a 1-bp resolution. This approach can elucidate some limitations from previous methods such as missing data handling or finding the correct constrains for segments length.

Yet, there are practical limitations essentially about the precondition to define the number of chromatin states beforehand, since this choice is arbitrary from the model and answers to biological interpretability of the results. Moreover, the algorithms mentioned above are computationally intensive and challenging to use without a sufficiently computational power. A relatively new statistical method called Non-Negative Matrix Factorization (NMF) [Lee and Seung, 1999] has been applied to the problem in question of finding chromatin states from the reads of epigenetic modification marks [Cieślik and Bekiranov, 2014, Gandolfi and Tramontano, 2017]. NMF undertake the dimensionality problem on the genome-wide analysis of epigenetic tracks by approximating the dataset values through a reduced number of meaningful components [Devarajan, 2008]. The method makes it possible to integrate multiple epigenetic marks to the characterization of chromatin combinatory patterns, and also allows to find an optimal number of distinct patterns based on the data. In both studies [Cieślik and Bekiranov, 2014, Gandolfi and Tramontano, 2017], they applied NMF to matrices consisting on multiple epigenetic marks as columns and non-overlapping segments of the genome as rows, being the value of the cell equal to the read counts.

In the first case, the segments used as rows correspond to multiple regions of TSS-proximal gene bodies since they contain epigenetic traces of transcription initiation and elongation. In the second case, they use bins of 200-bp, similar to the case in ChromHMM, in an attempt to approximate a single nucleosome with each bin. In this master thesis, aspects from these papers were used, extending the analysis to multiple cancer cell lines.

1.2 Non-negative Matrix Factorization

Non-negative matrix Factorization reveals as a method to learn parts or components from high-dimensional data. In contrast to other matrix factorization methods such as Principal Component Analysis (PCA) or Vector Quantization (VQ), NMF does not learn holistic but part-based representations. In addition, NMF is constrained to

have positive values as only additive combinations are considered [Lee and Seung, 1999]. Basically, NMF consists on an approximation of an $m \times n$ V matrix by the matrix multiplication of W ($m \times k$) and H ($k \times n$).

$$V \approx W \times H \tag{1.1}$$

This shall be accomplished by iteratively updating the rows and columns of W and H respectively. The constriction of the data to be positive also circumscribe the use of the method, though it has been applied in astronomy [Rén et al., 2017], language processing [Bertin et al., 2010] or image processing [Yang et al., 2007] among others. Gene expression, epigenetic modifications or mutation data have also been subject of analysis using NMF, considering the non-negativity of the data.

Dealing with high dimensionality data using NMF implies taking certain decisions, starting from the choice of NMF dimensionality (number of k signatures). An important measure used for this task is the reconstruction error, calculated by comparing the original data value with the result of the multiplication of the factorized matrices W and H. Some of the studies compare the reconstruction error produced by each defined k using the real data vs the one produced when taking random data, as it is the case in this analysis [Frigyesi and Höglund, 2008]. When examining the Residual Sum of Squares (RSS) as a function of k, the RSS decreases with increasing k within the original dataset, however, this can be less the case for the random dataset. Ideally, for a given interval of k values, there ought to be an optimal k for which the slopes of the real and random data intersect. That is to say, we intend to find the highest k, reducing the reconstruction error and improving the accuracy of the model, before being within the noise. Other studies use the cophenetic correlation coefficient, where the similarity between results from several runs is compared for each k. In other words, the stability of the classification is studied choosing the highest k before the robustness of the results drops [Brunet et al., 2004].

1.2.1 NMF algorithms

Within NMF method we find several ways of solving the factorization problem of form $V \approx WH$. Again, the non-negativity constrain makes previous factorization approaches not applicable and then new algorithms needed to be developed. First approaches [Lee and Seung, 2001] were based on minimizing either the euclidean distance between V and WH such that

$$||V - WH||^2 = \sum_{ij}(V_{ij} - (WH)_{ij})^2 \tag{1.2}$$

or minimizing the divergence between V and WH,

$$D(V||WH) = \sum_{ij}\left(V_{ij}\log\frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij}\right) \tag{1.3}$$

as in both cases, the measure is lower bounded by zero and minimizes when $V = WH$. Once the cost function for the minimization was defined, multiplicative rules were proposed as the procedure to update W and H. A compelling property of the approximation is that (1.1) can be reformulated in a column or row-wise manner:

- Update columns in H: $v \approx Wh$ when v represents a column in V,

$$v = (V_{1n}, \ldots, V_{Mn}) \in \mathbb{R}^M$$

, and h a column in H,

$$h = (H_{1n}, \ldots, H_{Kn}) \in \mathbb{R}^K$$

- Update rows in W: $v \approx wH$ when v represents a row in V,

$$v = (V_{1m}, \ldots, V_{Nm}) \in \mathbb{R}^N$$

and w a row in W

$$w = (W_{1m}, \ldots, W_{Km}) \in \mathbb{R}^K$$

Therefore, it is possible to iteratively update rows in W and then columns in H. The multiplicative updates for H and W in the euclidean distance minimization were described with the next form [Lee and Seung, 2001]:

$$H_{kj} \leftarrow H_{kj} \frac{(W^T V)_{kj}}{(W^T W H)_{kj}} \tag{1.4}$$

$$W_{ik} \leftarrow W_{ik} \frac{(V H^T)_{ik}}{(W H H^T)_{ik}} \tag{1.5}$$

and the multiplicative updates for the divergence, based on Kullback-Leibler divergence, and subsequently applied to retrieve meaningful patterns from cancer gene expression data [Brunet et al., 2004], were described as follows:

$$H_{kj} \leftarrow H_{kj} \frac{\sum_l \frac{W_{lk} V_{ij}}{(WH)_{ij}}}{\sum_l W_{lk}} \tag{1.6}$$

$$W_{ik} \leftarrow W_{ik} \frac{\sum_l \left[\frac{H_{kl} V_{ij}}{(WH)_{il}} \right]}{\sum_l W_{kl}} \tag{1.7}$$

Proofs of these theorems and convergence rules are further explained in [Lee and Seung, 2001]. After these principle algorithms, some approaches were developed based on them in order to get sparser results [Pascual-Montano et al., 2006] or include an intercept into the NMF fit [Badea, 2008].

1.2.2 NMF and gene expression

NMF was promptly introduced in bioinformatics as a method of dimensionality reduction of large-scale gene expression data [Kim and Tidor, 2003]. Here, functional relationships are yield from the analysis of 300 genome-wide expression measurements of yeast and compared with previous expertise. The 5346 genes analyzed in 300 samples were summarized in 50 patterns from which 12 were annotated with MIPS [Mewes et al., 2002] functional categories, based on the frequency which genes from each category appeared in each of the signatures or patterns. In a similar way, NMF has also been applied to categorize tumor subtypes using 4651 human genes in 108 cases [Frigyesi and Höglund, 2008], showing that NMF signatures may correspond specific disease gene expression patterns.

1.2.3 NMF and mutations

Besides finding shared patterns between diseases and gene expression, it has been shown that it is possible to characterize specific mutations produced in varied types

of cancer [Ramakrishna et al., 2012]. A catalogue of complete genomes for 21 primary breast cancer samples was sequenced and compared to the normal DNA of those same individuals. From these, 183,916 somatically acquired mutations were called and classified into 96 possible trinucleotide variations, composing the V matrix of 183916×21 cells. NMF was then applied on this dataset and from the results they identified contribution of each of the mutational signatures in each of the patients, which yield similar arrangement for similar cancer types.

Although the number of identified signatures was five, further on this number got reduced to only four signatures [Alexandrov et al., 2013]. In such a way, mutational processes operative in cancer genomes were modeled with great accuracy by a combination of this four mutational signatures. This allowed to make predictions based on the signatures and the data, as well as understand more deeply the biological processes involved on the different types of cancer.

1.2.4 NMF and epigenetics

The epigenetic modeling techniques explained before have shown the possibility to find genome-wide chromatin states based on the partition of the DNA. Nevertheless, they make the assumption that a small set of these chromatin states would be sufficient to describe the genomic expression, whereas several hundreds chromatin states have been estimated even with a small set of epigenetic marks used [Ucar et al., 2011]. In addition, chromatin modifications tend to be highly correlated, hampering the task of assessing the importance of the chromatin marks and relating them to the biological mechanisms. NMF can be used in order to overcome these downsides by identifying combinatorial patterns of chromatin states.

This application was initially presented as a way to get combinatorial patterns of epigenetic marks from integrated epigenetic data sets [Cieślik and Bekiranov, 2014]. They characterized a small amount of combinatorial patterns, which could be displayed and interpreted, were statistically capable of regression and classification tasks. Each row of the V matrix represents 2 kbp of regions flanking Transcription Start Sites (TSS) and columns represent the epigenetic marks used. In a case study for regression of *Pol2* binding, ten epigenetic marks were used to identify seven quantitative epigenetic patterns. Using the 7 chromatin patterns in a regression model yield a performance of $r^2 = 0.85$, matching the one obtained by using 10 epigenetic marks but solving multicollinearity problems.

Later on, functional classification of the epigenome was performed by adding more epigenetic marks and in this case, segmenting the DNA into 200 bp regions as an attempt to resemble one nucleosome with each 200-bp bin [Gandolfi and Tramontano, 2017]. NMF was here applied to 13 different epigenetic marks over 833,738 significant bins in human embryonic stem cells, finding 7 epigenetic signals or chromatin profiles. These seven epigenetic signals were then labeled based on the related biological process and then used for a wide range of tasks: study the genomic distribution of the signatures, investigate their recovery power on genomic features, association with the gene expression ...

1.3 Objectives

The main aim of this master thesis was to develop a computational pipeline capable of finding combinatorial patterns in the epigenomic data.

The absence of mapping information for the different epigenetic marks implied that a large proportion of the work consisted of processing the data. First, tools were developed for downloading epigenetic raw data and preparing the data for the analysis. The final employed data contained per-bin-counts for every epigenetic modification included in the analysis. Data from three cell types was used, hence three per-bin-counts matrices were produced.

With the correct data, NMF algorithm was used on the assumption that we are able to find combinatorial patters or "signatures". The underlying idea of the present analysis is the ability of these signatures to summarize the epigenetic modifications states in two angles: (1) the interaction between the several epigenetic modification types, comprised by the H matrix, and (2) the interaction between the genomic regions, held by the W matrix. This been fulfilled, we can seek to relate the signatures with biological processes, as well as compare their differences among tissues.

Over the present analysis, aspects from previous related analysis have been adopted, reproducing the arrangement of the data information for the processing step [Gandolfi and Tramontano, 2017], or segmenting the genome into 200-bp bins. Results show that there are similar chromatin signatures found for these tissues, suggesting a possible generalization of the chromatin profiles in normal and cancer cells. Moreover, some differences were found and are discussed in the upcoming sections.

Chapter 2

Methods

2.1 NMF

Non-Negative Matrix Factorization is the statistical framework in which this analysis is based on. Given a non-negative matrix V, NMF is an unsupervised learning method which tries to find non-negative matrix factors W and H such that

$$V \approx W \times H. \tag{2.1}$$

The epigenetic data used, described in a later section 2.2, fulfills the precondition of non-negativity of the data. The V matrix used in the analysis is composed by 200-bp bins as rows and the epigenetic marks as columns. We are investigating the possibility of finding k signatures which will summarize combinatorial patterns of the data to the epigenetic marks by means of the H matrix and to the genome-wide bins by means of the W matrix.

In order to perform the analysis, the V matrix was loaded into R environment and concretely the Rpackage *NMF* was used [Gaujoux and Seoighe, 2010]. The package used was select for consistency with previous similar studies [Gandolfi and Tramontano, 2017] and seeing that attempts to implement the algorithm resulted in more time-consuming alternatives. Rpackage *NMF* offers a framework with several NMF algorithms, including the ones explained in a previous section 1.2.1, from which `brunet` option was chosen. This algorithm is based on Kullback-Leibler divergence and matches the one described in [Lee and Seung, 2001] and used in [Brunet et al., 2004], enhanced to avoid numerical overflow.

2.1.1 Algorithm

Passed on a non-negative matrix V of $m \times n$ size and a chosen number of k signatures, `brunet` implementation will find the approximation $V \approx WH$.

1. First, both W and H matrices are randomly initialized.
2. Then, every row in W is updated according to the correspondent multiplicative update rules mentioned in (1.7).
3. Every column in H is updated according to (1.6).
4. Repeat steps 2 and 3 until the default convergence criteria is reached.

The stopping or convergence criteria for the NMF algorithm can be based on a fixed number of iterations or the invariability of the target values

$$(||WH||)_t = (||WH||)_{t+1}.$$

Since no fixed number of iterations was assumed, the stopping criteria for the analysis was the invariability of the WH matrix multiplication, which means there were different number of iterations when NMF was applied to the alternative tissues, varying between 350-600 iterations. The implementation in *NMF* package includes parallel computations to speed up the process.

2.1.2 Choosing number of signatures

In order to apply NMF to the data we foremost need to choose a suitable k number of signatures, also called rank. The number of clusters defined will largely influence the results and the explanation of them, therefore it is highly important to find the optimal number to produce "meaningful" clusters [Brunet et al., 2004]. As explained in previous sections, there is no standard procedure to find the best k number. Therefore, an additional matrix was created with randomly permuted values from the original V matrix, which we will refer to as "random" V_R matrix. The V_R matrix is composed of the same values as V but column-wise permuted, meaning that the presumed chromatic patterns would not be able to be recognized.

To compare the performance of NMF using different k values, the analysis is performed using values from 2 to 10 and then compared the reconstruction errors of the resultant models. As the NMF factorized matrix can vary from one run to another, the process is repeated 30 times and the results can be seen in [INSERT CROSS-REFERENCE]. The reconstruction error was calculated for comparison by residuals sum of squares (RSS),

$$||V - WH||^2 = \sum_{ij}(V_{ij} - (WH)_{ij})^2 \tag{2.2}$$

[INSERT CHOOSE N PLOT]

As we can see from the results, for the same k value, the NMF model produces worst results for the V_R matrix as well as more variable results. Nevertheless, there is a k value for which the plots intersect due to the model being over-fitted. We chose $k = 7$ because it is the value for which the reconstruction error is the lowest possible before modeling the noise in the data. The results found similar in all the three cell types.

2.2 Data

The data used in this study was obtained entirely from the ENCODE project database [Feingold et al., 2004], 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) [Aden et al., 1979], myelogenous leukemia cell line (K562) [Andersson et al., 1979] and cells derived from HeLa cancerous cervical tumor line (HeLa-S3) [Douglas, 1973, Chen, 2008].

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 for this analysis.
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`.

2.3 Pipeline

In order to integrate the different scripts used in the project, a bash command line tool was created. The command line requires of a conda environment which can be set up using the `createCondaEnv.sh` script. This command line includes options to 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`).

All the scripts used for the analysis as well as for creating the report are available as a Bitbucket repository.

2.3.1 Download Data

Command: `scripts/epigeNMF.sh -d [Datasets.tsv]`

Makes use of `downloadData.sh` script. It requires the file with the datasets info as explained in the Data section 2.2. For each sample, the dataset will be downloaded and saved in the `data` directory using the next file tree: `data/[cell_line]/[epigenetic_mark_name]/[sample_id]/`.

2.3.2 Processing BAM files

Command: `scripts/epigeNMF.sh -p [Datasets.tsv]`

Makes use of `prepareBedAlignment.sh` script, based on the processing pipeline used in [Gandolfi and Tramontano, 2017], following the scheme and updating outdated methods. The script process raw BAM files into filtered BED files by (1) removing duplicate reads, (2) filter reads by quality, (3) add a 200-bp tag extension in both 3' ends to transcription factors and histone modification signals (resembling half of the average ChIP-seq fragments) and (4) convert BAM files into BED files. It outputs processed BED files to the same path as the BAM file.

2.3.3 Generate V matrix

Command: `scripts/epigeNMF.sh -c [cell-lines]`

Calls `bedToNormCounts.sh` script, based on the processing pipeline used in [Gandolfi and Tramontano, 2017], following the scheme and updating outdated methods. Takes the cell lines desired to generate the V matrix from their samples. It requires the datasets info file, human genome segmented into 200-bp fragments as well as a uniqueness mappability track file (using the reference *Duke Uniqueness Regions*) to be in the data directory. For each sample in each cell line, the script assigns the reads to a bin in the segmented genome and calls `bedCountsToV.R` script which adds the columns of the corresponding cell line V matrix.

The bin counts are normalized with an scaling factor based on the uniqueness mappability positions present in each bin and the total number of reads in the sample. The technical replicates available for a epigenetic sample were combined by taking the mean of the normalized counts for each bin. The result of the execution is a V matrix csv file for each of the cell lines of interest, saved in the file tree directory `results/genomic_survey/[cell_line]/V_matrix.csv`.

2.3.4 Filter V matrix

Command: `scripts/epigeNMF.sh -f [input-V.csv] [output-V.csv]`

Uses `summariseFilterData.R` script. Takes as arguments the input and output directories, being the input an unfiltered V matrix in CSV format and the output a V filtered matrix also in CSV format. The bins are filtered taking only those for which at least one of the epigenetic marks fall into the top 2.5% percentile. The threshold was set to the 0.975 quantile in order to filter out the counts considered as noise for our analysis.

2.3.5 Choose number of signatures

Command: `scripts/epigeNMF.sh -k [input] [output] [n-repeats]`

With `chooseN.R` script used to choose the optimal k number of signatures for the NMF analysis in the filtered V matrix data. Generates a plot comparing real and random data reconstruction errors (choose k based on that). It also includes an option to define the number of times NMF and reconstruction error are calculated for each k value. Outputs a residual error plot as the one described in 2.1.2 section.

2.3.6 NMF analysis

Command: `scripts/epigeNMF.sh -n [input-file] [output-file] [n-signatures]`

Calls `NMFanalysis.R` with a filtered V matrix in CSV file as input, and produces several informative plots describing the H and W matrices generated, saving them to the output directory. The number of signatures can be passed as an argument, with a default of $k = 7$.

2.4 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

Bibliography

[Aden et al., 1979] Aden, D. P., Fogel, A., Plotkin, S., Damjanov, I., and Knowles, B. B. (1979). Controlled synthesis of HBsAg in a differentiated human liver carcinoma-derived cell line.

[Alexandrov et al., 2013] Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J., and Stratton, M. R. (2013). Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Reports*.

[Andersson et al., 1979] Andersson, L. C., Nilsson, K., and Gahmberg, C. G. (1979). K562—A human erythroleukemic cell line. *International Journal of Cancer*.

[Bäckdahl et al., 2009] Bäckdahl, L., Bushell, A., and Beck, S. (2009). Inflammatory signalling as mediator of epigenetic modulation in tissue-specific chronic inflammation.

[Badea, 2008] Badea, L. (2008). Extracting gene expression profiles common to colon and pancreatic adenocarcinoma using simultaneous nonnegative matrix factorization. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*.

[Barneda-Zahonero and Parra, 2012] Barneda-Zahonero, B. and Parra, M. (2012). Histone deacetylases and cancer.

[Bartolomei and Ferguson-Smith, 2011] Bartolomei, M. S. and Ferguson-Smith, A. C. (2011). Mammalian genomic imprinting. *Cold Spring Harbor Perspectives in Biology*.

[Berdasco and Esteller, 2010] Berdasco, M. and Esteller, M. (2010). Aberrant Epigenetic Landscape in Cancer: How Cellular Identity Goes Awry.

[Bernstein et al., 2010] Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., Farnham, P. J., Hirst, M., Lander, E. S., Mikkelsen, T. S., and Thomson, J. A. (2010). The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology*.

[Bernstein and Allis, 2005] Bernstein, E. and Allis, C. D. (2005). RNA meets chromatin.

[Bertin et al., 2010] Bertin, N., Badeau, R., and Vincent, E. (2010). Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription. *IEEE Transactions on Audio, Speech and Language Processing*.

[Bone et al., 1994] Bone, J. R., Lavender, J., Richman, R., Palmer, M. J., Turner, B. M., and Kuroda, M. I. (1994). Acetylated histone H4 on the male X chromosome is associated with dosage compensation in Drosophila. *Genes and Development*.

[Borges and Martienssen, 2015] Borges, F. and Martienssen, R. A. (2015). The expanding world of small RNAs in plants.

[Brunet et al., 2004] Brunet, J.-P., Tamayo, P., Golub, T. R., and Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences of the United States of America*.

[Chen, 2008] Chen, T. (2008). Re-evaluation of HeLa, HeLa S3, and HEp-2 karyotypes. *Cytogenetic and Genome Research*.

[Cieślik and Bekiranov, 2014] Cieślik, M. and Bekiranov, S. (2014). Combinatorial epigenetic patterns as quantitative predictors of chromatin biology. *BMC Genomics*, 15(1).

[Deaton and Bird, 2011] Deaton, A. M. and Bird, A. (2011). CpG islands and the regulation of transcription. *Genes and Development*.

[Devarajan, 2008] Devarajan, K. (2008). Nonnegative matrix factorization: An analytical and interpretive tool in computational biology.

[Dodd et al., 2007] Dodd, I. B., Micheelsen, M. A., Sneppen, K., and Thon, G. (2007). Theoretical Analysis of Epigenetic Cell Memory by Nucleosome Modification. *Cell*.

[Douglas, 1973] Douglas, J. (1973). HeLa. *Nature*.

[Elgin, 1996] Elgin, S. C. (1996). Heterochromatin and gene regulation in Drosophila. *Current Opinion in Genetics and Development*.

[ENCODE Project Consortium et al., 2007] ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J. A., Dutta, A., and et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*.

[Ernst and Kellis, 2010] Ernst, J. and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome . Overview of Supplementary Materials : Supplementary Notes. *Nature Biotechnology*.

[Ernst and Kellis, 2017] Ernst, J. and Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. *Nature Protocols*.

[Feingold et al., 2004] Feingold, E. A., Good, P. J., Guyer, M. S., and et al. (2004). The ENCODE (ENCyclopedia of DNA Elements) Project.

[Frigyesi and Höglund, 2008] Frigyesi, A. and Höglund, M. (2008). Non-negative matrix factorization for the analysis of complex gene expression data: identification of clinically relevant tumor subtypes. *Cancer informatics*.

[Gandolfi and Tramontano, 2017] Gandolfi, F. and Tramontano, A. (2017). A computational approach for the functional classification of the epigenome. *Epigenetics and Chromatin*, 10(1):1–24.

[Gaujoux and Seoighe, 2010] Gaujoux, R. and Seoighe, C. (2010). A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*.

[Goldberg et al., 2007] Goldberg, A. D., Allis, C. D., and Bernstein, E. (2007). Epigenetics: A Landscape Takes Shape.

[Goll and Bestor, 2005] Goll, M. G. and Bestor, T. H. (2005). Eukaryotic cytosine methyltransferases. *Annual review of biochemistry*.

[Grewal and Moazed, 2003] Grewal, S. I. and Moazed, D. (2003). Heterochromatin and epigenetic control of gene expression.

[Hamilton and Baulcombe, 1999] Hamilton, A. J. and Baulcombe, D. C. (1999). A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science*.

[Heintzman and Ren, 2009] Heintzman, N. D. and Ren, B. (2009). Finding distal regulatory elements in the human genome.

[Hoffman et al., 2012] Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A., and Noble, W. S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature methods*.

[Hon et al., 2008] Hon, G., Ren, B., and Wang, W. (2008). ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. *PLoS computational biology*.

[Hotchkiss, 1948] Hotchkiss, R. D. (1948). The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *The Journal of biological chemistry*.

[Jeppesen and Turner, 1993] Jeppesen, P. and Turner, B. M. (1993). The inactive X chromosome in female mammals is distinguished by a lack of histone H4 acetylation, a cytogenetic marker for gene expression. *Cell*.

[Jessberger et al., 2007] Jessberger, S., Nakashima, K., Clemenson, G. D., Mejia, E., Mathews, E., Ure, K., Ogawa, S., Sinton, C. M., Gage, F. H., and Hsieh, J. (2007). Epigenetic Modulation of Seizure-Induced Neurogenesis and Cognitive Decline. *Journal of Neuroscience*, 27(22):5967–5975.

[Johnson et al., 2003] Johnson, J. M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P. M., Armour, C. D., Santos, R., Schadt, E. E., Stoughton, R., and Shoemaker, D. D. (2003). Genome-Wide Survey of Human Alternative Pre-mRNA Splicing with Exon Junction Microarrays. *Science*.

[Jones, 2012] Jones, P. A. (2012). Functions of DNA methylation: Islands, start sites, gene bodies and beyond.

[Jones and Taylor, 1980] Jones, P. A. and Taylor, S. M. (1980). Cellular differentiation, cytidine analogs and DNA methylation. *Cell*.

[Kayne et al., 1988] Kayne, P. S., Kim, U. J., Han, M., Mullen, J. R., Yoshizaki, F., and Grunstein, M. (1988). Extremely conserved histone H4 N terminus is dispensable for growth but essential for repressing the silent mating loci in yeast. *Cell*.

[Kiefer, 2007] Kiefer, J. C. (2007). Epigenetics in development.

[Kim and Tidor, 2003] Kim, P. M. and Tidor, B. (2003). Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Research*.

[Lee and Seung, 2001] Lee, D. and Seung, S. (2001). Algorithms for Non-negative Matrix Factorization. *Advances in Neural Information Processing Systems 13*, 1(1):556–562.

[Lee and Seung, 1999] Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*.

[Liu et al., 2004] Liu, L., Lai, S., Andrews, L. G., and Tollefsbol, T. O. (2004). Genetic and epigenetic modulation of telomerase activity in development and disease.

[Mammana and Chung, 2015] Mammana, A. and Chung, H. R. (2015). Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. *Genome Biology*.

[Mewes et al., 2002] Mewes, H. W., Frishman, D., Güldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Münsterkötter, M., Rudd, S., and Weil, B. (2002). MIPS: a database for genomes and protein sequences. *Nucleic acids research*.

[Mochizuki et al., 2002] Mochizuki, K., Fine, N. A., Fujisawa, T., and Gorovsky, M. A. (2002). Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in Tetrahymena. *Cell*.

[Partridge et al., 2002] Partridge, J. F., Scott, K. S., Bannister, A. J., Kouzarides, T., and Allshire, R. C. (2002). cis-acting DNA from fission yeast centromeres mediates histone H3 methylation and recruitment of silencing factors and cohesin to an ectopic site. *Current Biology*.

[Pascual-Montano et al., 2006] Pascual-Montano, A., Carazo, J. M., Kochi, K., Lehmann, D., and Pascual-Marqui, R. D. (2006). Nonsmooth nonnegative matrix factorization (nsNMF). *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Ramakrishna et al., 2012] Ramakrishna, M., Hinton, J., Alexandrov, L., and et al. (2012). Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell*.

[Rén et al., 2017] Rén, B., Pueyo, L., Zhu, G. B., Debes, J., and Duchêne, G. (2017). Non-negative Matrix Factorization: Robust Extraction of Extended Structures. *The Astrophysical Journal*, 852(2):104.

[Schatz et al., 1964] Schatz, G., Haslbrunner, E., and Tuppy, H. (1964). Deoxyribonucleic acid associated with yeast mitochondria. *Biochemical and Biophysical Research Communications*.

[Sutherland and Costa, 2003] Sutherland, J. E. and Costa, M. (2003). Epigenetics and the environment. In *Annals of the New York Academy of Sciences*.

[Ucar et al., 2011] Ucar, D., Hu, Q., and Tan, K. (2011). Combinatorial chromatin modification patterns in the human genome revealed by subspace clustering. *Nucleic Acids Research*.

[Waddington, 1942] Waddington, C. H. (1942). The epigenotype. *Endeavour*, 1(1):18–20.

[Wang and Cooper, 2007] Wang, G. S. and Cooper, T. A. (2007). Splicing in disease: Disruption of the splicing code and the decoding machinery.

[Weber et al., 2005] Weber, M., Davies, J. J., Wittig, D., Oakeley, E. J., Haase, M., Lam, W. L., and Schübeler, D. (2005). Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature Genetics*, 37(8):853–862.

[Wray et al., 2007] Wray, N. R., Goddard, M. E., and Visscher, P. M. (2007). Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Research*, 17(10):1520–1528.

[Xie et al., 2013] Xie, W., Schultz, M. D., Lister, R., and et al. (2013). Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell*.

[Yang et al., 2007] Yang, Z., Yuan, Z., and Laaksonen, J. (2007). Projective Non-Negative Matrix Factorization with Applications To Facial Image Processing. *International Journal of Pattern Recognitionand Artificial Intelligence*, 21(8):1353–1362.