Annotation of the human genome through the unsupervised analysis of high-dimensional genomic data

Jean-Baptiste Morlot

SORBONNE UNIVERSITÉS - PARIS

École Doctorale Physique en Île de France

THÈSE

pour obtenir le titre de

Docteur en Sciences, spécialité

PHYSIQUE

Présentée par

Jean-Baptiste MORLOT

Annotation of the human genome through unsupervised analysis of high-dimensional genomic data

Thèse dirigée par Julien MOZZICONACCI

préparée au Laboratoire de Physique Théorique de la Matière Condensée

soutenue le 12 decembre 2017

devant le jury composé de :

Rapporteurs :	Cedric VAILLANT
	Marc-Thorsten HUTT
Examinateurs :	Alessandra CARBONE
Directeur de thèse:	Julien MOZZICONACCI
Invités :	Annick LESNE

Research is to see what everybody else has seen and to think what nobody else has thought.

Albert Szent-Györgyi

Remerciements

Je tiens tout d'abord à remercier mon responsable, Julien Mozziconacci, pour m'avoir accompagné quotidiennement pendant ces trois années et avoir partagé ses conseils pertienents sur mon travail. Je remercie Annick Lesne pour avoir partagé son point de vue précis et rigoureux sur mon travail ainsi que sa vision de la recherche. Je remercie mes rapporteurs Cédric Vaillant, Marc-Thorsten Hutt et mon examinateur Alessandra Carbone d'avoir accepté de lire et de corriger mon travail.
Je remercie Charles Antoine pour ses conseils avisés autant dans la vie professionnelle que personnelle. Je remercie mes parents pour m'avoir soutenu pendant toutes mes études et ma soeur Anne pour m'avoir aidé à relire ce manuscrit. Je remercie Julia Antunes pour sa gentillesse et sa prévenance qui m'ont accompagné durant ces trois années difficiles. Et enfin, je remercie toute ma famille et tous mes amis pour m'avoir soutenu pendant toutes ces années.

Abstract

The human body has more than 200 different cell types each containing an identical copy of the genome but expressing a different set of genes. The control of gene expression is ensured by a set of regulatory mechanisms acting at different scales of time and space. Several diseases are caused by a disturbance of this system, notably some cancers, and many therapeutic applications, such as regenerative medicine, rely on understanding the mechanisms of gene regulation.

This thesis proposes, in a first part, an annotation algorithm (GABI) to identify recurrent patterns in the high-throughput sequencing data. The particularity of this algorithm is to take into account the variability observed in experimental replicates by optimizing the rate of false positive and false negative, increasing significantly the annotation reliability compared to the state of the art. The annotation provides simplified and robust information from a large dataset. Applied to a database of regulators activity in hematopoiesis, we propose original results, in agreement with previous studies.

The second part of this work focuses on the 3D organization of the genome, intimately linked to gene expression. This structure is now accessible thanks to 3D reconstruction algorithm from contact data between chromosomes. We offer improvements to the currently most efficient algorithm of the domain, ShRec3D, allowing to adjust the reconstruction according to the user needs.

Résumé

Le corps humain compte plus de 200 types cellulaires différents possèdant une copie identique du génome mais exprimant un ensemble différent de gènes. Le contrôle de l'expression des gènes est assuré par un ensemble de mécanismes de régulation agissant à différentes échelles de temps et d'espace. Plusieurs maladies ont pour cause un dérèglement de ce système, notablement les certains cancers, et de nombreuses applications thérapeutiques, comme la médecine régénérative, reposent sur la compréhension des mécanismes de la régulation géniques.

Ce travail de thèse propose, dans une première partie, un algorithme d'annotation (GABI) pour identifier les motifs récurrents dans les données de séquençage haut-débit. La particularité de cet algorithme est de prendre en compte la variabilité observée dans les réplicats des expériences en optimisant le taux de faux positif et de faux négatif, augmentant significativement la fiabilité de l'annotation par rapport à l'état de l'art. L'annotation fournit une information simplifiée et robuste à partir d'un grand ensemble de donnée. Appliqué à une base de donnée sur l'activité des régulateurs dans l'hématopoieïse, nous proposons des résultats originaux, en accord avec de précédentes études.

La deuxième partie de ce travail s'intéresse à l'organisation 3D du génome, intimement lié à l'expression génique. Elle est accessible grâce à des algorithme de reconstruction 3D à partir de donnée de contact entre chromosomes. Nous proposons des améliorations à l'algorithme le plus performant du domaine actuellement, ShRec3D, en permettant d'ajuster la reconstruction en fonction des besoins de l'utilisateur.

Contents

Chapter 1

Introduction

1.1 Gene regulation in human cells

1.1.1 Genes

Each cell of an organism possess in its nucleus a unique poly-nucleotide sequence of four nucleotides, adenine(A), guanine(G), cytosine(C) and thymine (T) [75] called the deoxyribonucleic acid or DNA Fig1.1, organized in a double helix structure, forming chromosomes.

Two categories can be identified within the DNA sequence: *coding sequences* and *non coding sequences* Fig1.1. *Coding sequences*, are sequences that can be transcribed by an enzyme, the RNA polymerase (RNA Pol II), which create a copy of the DNA strand into mRNA (messanger RiboNucleic Acid). In the following, we will refer to these sequences, which code for proteins as *genes*, albeit some alternative definition can be found. The mRNA produced by the RNA Pol II migrates outside the nucleus, where the ribosomes, a complex cellular machinery, will assemble amino-acids, corresponding to the RNA sequence, into a protein Fig1.1.

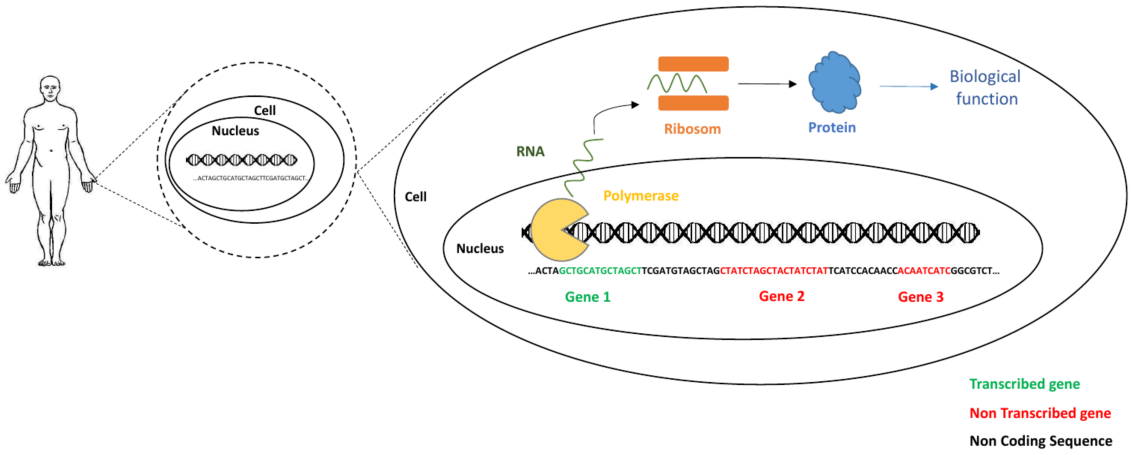

Figure 1.1 – Genes expression in human cells. Each nucleus posses DNA, which coding sequences can be transcribed by RNA Pol II. The RNA produced by the transcription of genes is transformed outside the nucleus by the ribosomes into proteins which fulfill many different biological functions

Proteins have many different functional roles in cells. For example, they can build the structural blocks forming the cell, transmit external signals into the cell or, and we will see this particular aspect in more details, control the genes expression.

1.1.2 Cell Identity

Among the 20.000 genes present in the human DNA, only a fraction is expressed in a cell type Fig1.2, encoding for a specific combination of proteins. For instance red blood cells produce abundant quantities of hemoglobin whereas muscle cells produce muscle fiber proteins. The regulation of gene transcription is orchestrated by an ensemble of regulators which ensures the precise coordination of genes expression during cell development and cells differentiation.

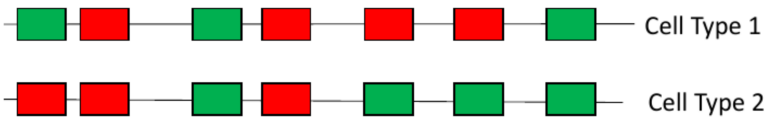

Figure 1.2 – Cell identity. Cell identity is defined by the combination of genes that are expressed (green) and repressed (red)

1.1.3 Regulators

There are two family of regulators: transcription factors (TFs) which binds on regulatory genomic sites, also called *cis-regulatory elements (CRE)*, and modulate the activity of nearby genes see Fig1.3(a), and biochemical marks that binds on the DNA molecule or on specific proteins, the *histones*, changing the activity of the genes nearby. We estimate that the number of CRE is six times larger than the number of genes, thus close to 120.000.

1.1.3.1 Transcription factors

TF proteins recognize specific small DNA sequences of about 10 bp and modify the expression of nearby genes Fig1.3(a). *Activators*, or *enhancers*, recruit the RNA Pol II in order to initiate genes transcription whereas *inhibitors* prevent this recruitment. *Insulators* are TF that create loops between an enhancer and the gene transcription starting site (TSS) Fig1.4. With the help of the cohesin and a mediator the enhancer forms a complex at the TSS which recruit the RNA Pol II Fig1.4. Insulators have a double function: they can block the action of a TF on its target gene when located between them Fig1.3(a) (right), or enable the action of TF by forming a loop Fig 1.3(a)(right). It has been reported that a mis-positioning of insulators may imply disease such as acute myeloid leukemia [37].

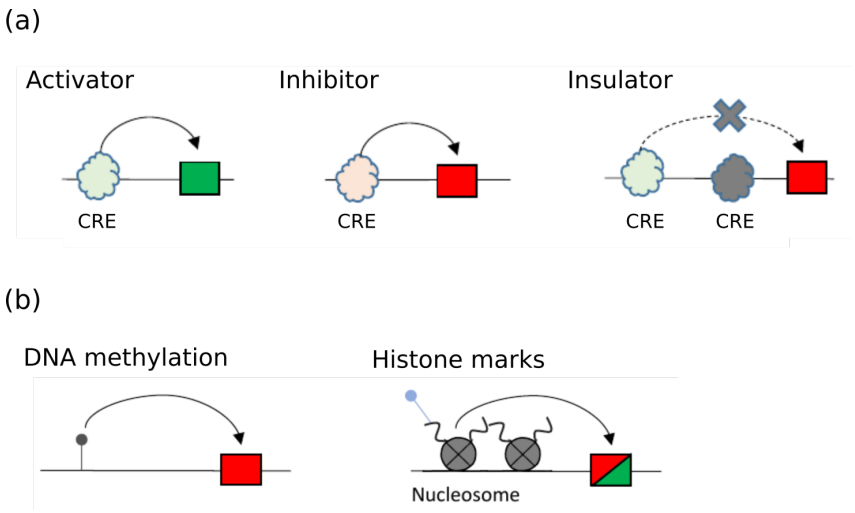

Figure 1.3 – The regulators can be separated in two categories: TF and biochemical marks. (a) TF are composed of *activators* (left), that activate nearby genes, *inhibitors* (middle), that inhibit nearby genes and *insulators* (right) that block or enable the influence of a TF on a gene. The color represents the transcribed (green) or inhibited (red) state of the gene. (b) Biochemical marks are composed of two kind of marks: marks that bind on the DNA, such as DNA methylation, and marks that binds on histones proteins, the histones marks

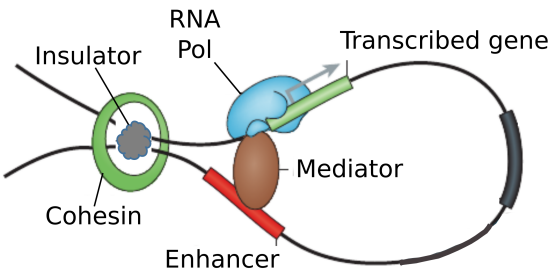

Figure 1.4 – Loops induced by the clipping of an insulator, enabling the recruitment of the RNA Pol II at the TSS (image adapted from [65])

1.1.3.2 Biochemical marks

Biochemical marks represent a second family of gene regulators. They bind either on DNA, like the *DNA methylation* or on proteins, the *histones*, composed of a protein core around which the DNA is wrapped at regular intervals forming nucleosomes. In this case the marks are called *histones marks*. These marks have an influence on gene transcription by acting on the nucleosomes condensation Fig1.5. In order to bind on the DNA, TFs need to remove nucleosomes at the CRE. This is only possible when the DNA is not highly condensed around nucleosomes Fig1.5.

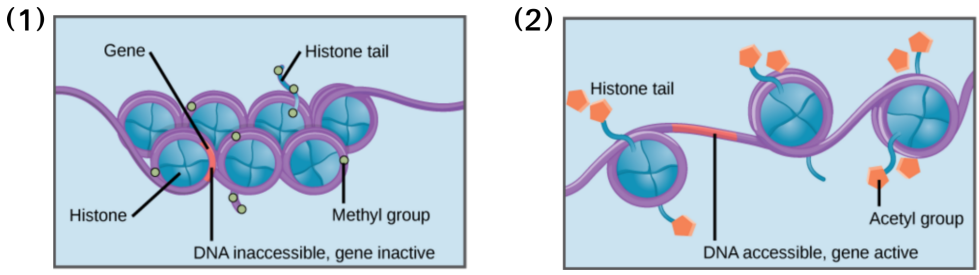

Figure 1.5 – The nucleosomes condensation changes the accessibility of the DNA. (1) The DNA is inaccessible to proteins. (2) The DNA is accessible to proteins. Figure adapted from [64]

DNA methylation DNA methylation is the addition of methyl group to the 5th carbon of cytosines, which condensed the DNA and repress the genes nearby. This process is mainly used during cell differentiation to constraint the different cellular pathways (from stem cell to mature tissues) [68], by definitively repressing some genes. A well known example is the inactivation of the chromosome X [9].

Histones marks Histones composing nucleosomes posses tails on which marks can bind and combine. Their actions are richer than DNA methylation because they can act either as gene activator or gene inhibitor by changing the level of compaction of the DNA [70] [71] and by regulating the type of proteins that can bind on the DNA (see appendix for further details).

1.1.4 The regulatory network

We have seen that cell types identity, defined by a specific pattern of gene activity, is controlled by an ensemble of regulators that act locally on the genes nearby. If the protein generated from the expression of a gene is a TF, then it will act as a regulator on one or many genes Fig1.6(a). Consequently, cell types identity emerges from the complex entanglement between regulators and genes activity (see Fig1.6(b-c) for a small example). The ensemble of causal relationships between the genes and regulators activity can be modeled by a network, namely a *regulatory network*.

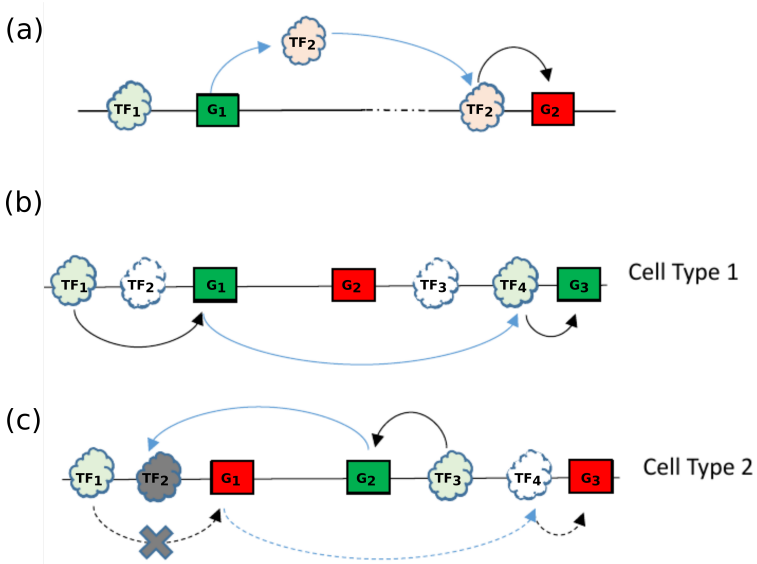

Figure 1.6 – Cell types identity emerge from the complex entanglement between regulators and genes activity. For simplicity, this figures represent only the regulation via TF. A more complete picture would also include the regulation via biochemical marks, also generated by gene transcription. (a)The gene G_1 is activated by the transcription factor TF_1. G_1 codes for a protein which cellular function is to act as TF TF_2 for another gene G_2. (b) The combination of genes emerge from a specific regulators-genes activity: TF_1 activate the G_1 which codes for TF_4 activating G_3. G_2 has no activators, thus remain inactive. (c) G_3 activates the gene G_2 which codes for TF_2 that insulate the activation of TF_1 over G_1. Therefore TF_4 is not produced and G_3 not expressed, forming another cell type

1.1.5 Motivations

1.1.5.1 Long-term motivation and practical applications

The long-term goal of this work is the establishment of the complete regulatory network, in order *to predict, control and understand the genes expression*. Therapeutic application are important because many diseases, such as cancers or immune diseases, takes roots in a dysfunction of the regulatory machine. A model would allow to understand, from observed consequence, the cause of the regulatory problem. Furthermore, other important field of application is the regenerative medicine which uses cellular reprogramming. It consists in changing the cell identity of a patient healthy cells, e.g. skin cells, to cure another cell type from a defective organ.

1.1.5.2 Challenges

To achieve this goal, we need to face multiple challenges. First, we need to discover all types of regulators and their regulatory mechanisms. Second, we need to identify these regulators and coding sequences genome-wide. And third, we need to establish the causal relationships between the regulators and coding sequences activity.

1.1.5.3 Purpose of this work

Our work has used the actual knowledge of regulators and regulatory mechanisms to focus on the genome wide identification of regulators and their involvement on these regulatory mechanisms.

More precisely, we have focused on the identification and the discovery of CRE functional role on the cell identity. This process is called *DNA sequence annotation* or *genome annotation*. We developed a tool, GABI (Genome wide Association with Bayesian Inference), which discovers the different activation patterns in NGS experiments and we applied it to CRE activation in different cell types. We will show that these regulatory patterns, also called *states*, have a deep biological meaning and that the number of patterns is several order less than the total number of CRE (about 120.000), which significantly reduce the number of element to study. We will also show that annotations provided by GABI has a higher reliability than the state of the art.

1.2 The datasets

1.2.1 Next Generation Sequencing

DNA sequencing technique was developed in 1977 by Sanger et al [60], which identify genomic features, such as binding sites of a given protein, accessible DNA regions or histones marks position. It is also used to recover the DNA sequence of an organism. In 2007, a new version of this technique has emerged called *Next Generation Sequencing* (NGS) [46]. We present the principle of this technique through the description of the DNAse-seq.

DNAse-seq DNase-seq (DNase I hypersensitive sites sequencing)[17] is an experiment providing a measure of DNA accessibility *which identify CRE* [68] [66]. It uses an enzyme, the DNase 1 [79], which slices nucleosomes free DNA regions, accessible to proteins Fig1.7(1). The sliced DNA fragments, called *reads*, are then amplified and sequenced[1] Fig1.7(2). These reads are aligned on a reference genome corresponding to a consensus genome sequence of an organism Fig1.7(3). And eventually, all the reads are summed in order to build a histogram-like signal Fig1.7(4), namely a *NGS profile* of the experiment.

There exist a large variety of NGS experiments, which differs by the measured genomic feature. For example, ChiP-seq uses chromatin immuno-precipitation (ChIP) to capture genomic regions interacting with a target protein. Another NGS technique maps reads from different part of the genome which are close in 3D space. This is the case of the Hi-C map, which will be presented in chapter 5.

[1] Their sequence is determined

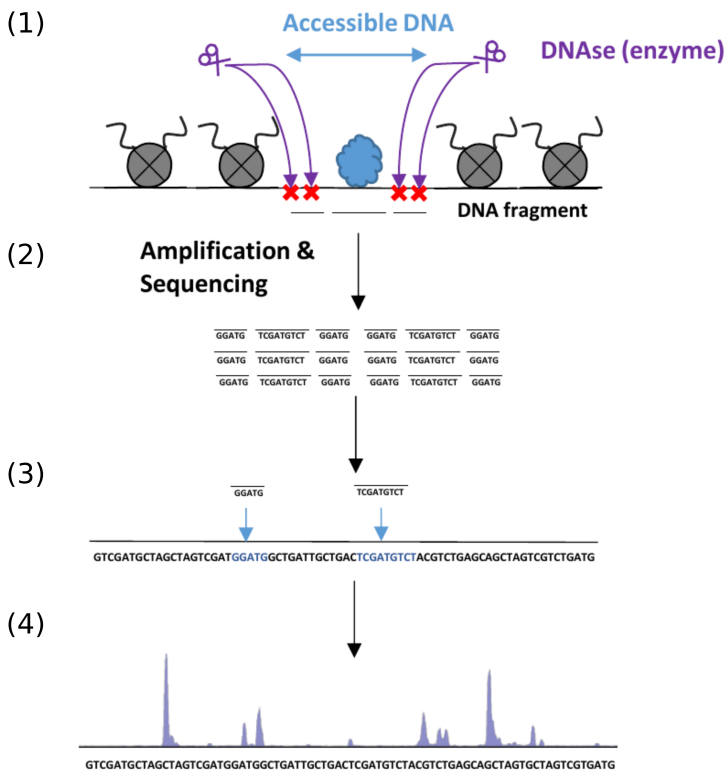

Figure 1.7 – DNAse-seq principle.(1) Generation of DNA fragments (reads). (2) Amplification and sequencing of these reads. (3) Alignment of the reads on a reference genome. (4) Sum of the reads

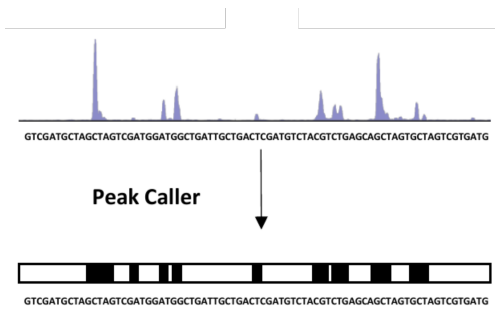

Figure 1.8 – The peak caller is a category of algorithm that detect peaks on NGS profiles. It sets a value of 1 to peaks, and 0 otherwise

Interpretation of NGS signal In order to increase the signal over noise ratio, most of NGS experiments are realized on a ensemble of cells. Consequently, in one dimensional NGS signal, the summation of the reads lets appear peaks of reads enriched genomic regions. They indicate the presence of the genomic feature, such as a CRE in the case of DNAse-seq. The height of these peaks depends on many parameters [47], such as the enzyme affinity to DNA sequence or the homogeneity of the pool of cells used for the experiment. This feature is complex to interpret, therefore, we have focused on the NGS peaks only. These peaks can be detected using a category

of algorithm, namely *peak callers* Fig1.8. It produces a binary signal which peaks are set to one and the rest to zero. In appendix "ZINB peak caller", we describe our own peak caller which improves two aspects of the most used peak caller, MACS [82]: The null model and the ability to process NGS samples formatted in arrays.

In the following, experiments resulting from an ensemble of cells will be called *pooled NGS profiles*.

.

1.2.2 ROADMAP dataset(DNAse-seq)

We have used in this work the DNAse-seq dataset produced by the Roadmap consortium [40], developed in more than 50 tissues and cell types from adult, fetal and induced stem cells Fig1.9. The dataset also posses 29 histones marks in these cell types, not used it in this work. These profiles are realized on an ensemble of cells, therefore correspond to pooled DNAse-seq profiles.

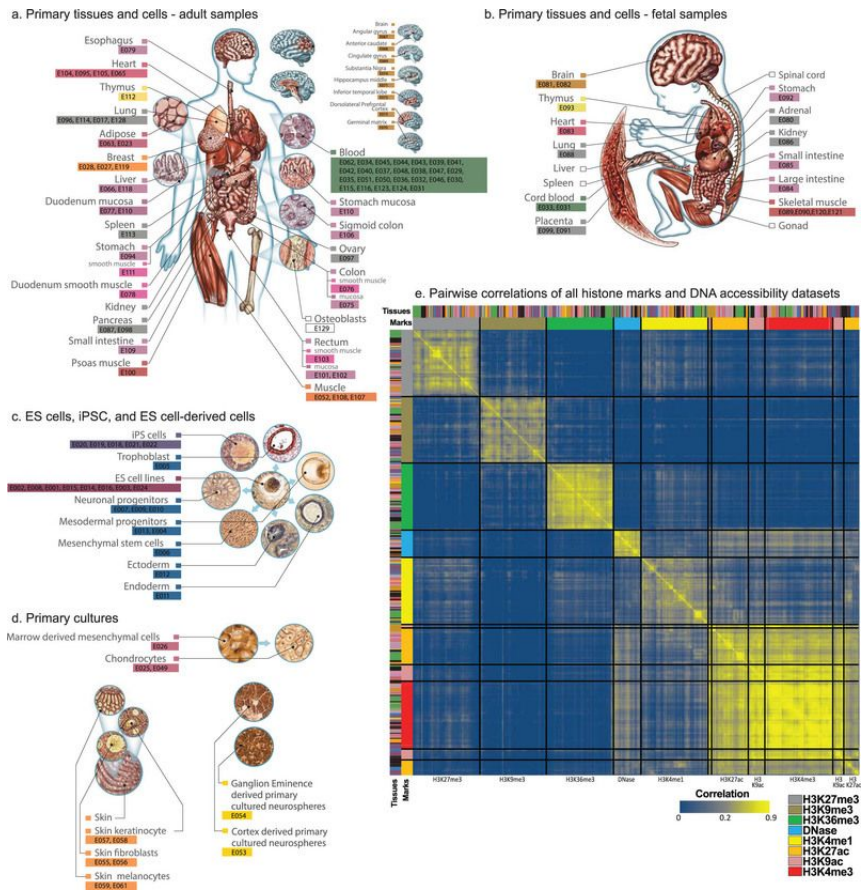

Figure 1.9 – (a-d) Samples available in the Roadmap dataset. (e) Correlation between the different NGS profiles, histones marks and the DNAse (figure from [40])

1.2.3 Single cell ATAC-seq: Hematopoiesis dataset

ATAC-seq ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) is a technique used in molecular biology to study chromatin accessibility. It was first described in 2013 [12] and it aims to identify accessible DNA regions, equivalent to DNase I hypersensitive sites. The key part of the ATAC-seq procedure is the action of the transposase Tn5 on the genomic DNA of the sample. Transposases are enzymes catalyzing the movement of transposons to other parts in the genome. While naturally occurring transposases have a low level of activity, ATAC-seq employs a mutated hyperactive transposase. The high activity allows for highly efficient cutting of opened DNA and simultaneous ligation of specific sequences, called adapters. Adapter-ligated DNA fragments are then isolated, amplified by PCR and used for next generation sequencing.

Pooled ATAC-seq Pooled ATAC-seq has many advantages over its concurrent, the pooled DNASe-seq: it needs only 500 to 50 000 cells compared to 50 million for the DNAse-seq and the protocol needs only 2 steps as opposed to 44 steps for the DNAse-seq [66]. The size of the cell bulk imply to sample more tissues, thus decrease the experiment homogeneity and therefore its quality.

Single cell ATAC-seq In 2015, ATAC-seq on individual cells has been performed by Buenrostro & al [13], namely *single cell ATAC-seq* or *scATAC-seq*, which provides an insight of cell-to-cell variation and different stages of cell type cycle. It solves the major issues of pooled experiments, the *homogeneity of the NGS samples*. Indeed, in pooled experiments, all extracted cells are not strictly in the same cellular states because some genes and regulator are transiently activated during the cell cycles. Moreover, extracted cells come from different cell types. For example, an extracted piece of lung, would contain blood vessels and muscles.

In this work, we have used a single cell dataset from hematopoeisis with more than 2.700 NGS profiles, including stem cells, hematopoietic-progenitor and differentiated blood cells, produced by Buenrostro & al [14]. Hematopoeisis is often used as reference model to study cell differentiation.

1.2.4 Datasets organization

Each NGS experiment gives after peak calling a binary vector. We have chosen to represent these vector at a 500 bp resolution, which represents 2.5 time size of a nucleosome (200bp x 2.5 = 500 bp).

All these vectors are concatenated in a matrix, Fig1.10, such that each line represents an experiment (or a profile) and each column a genomic site of 500 bp. At this resolution, each profile has 5.6 million dimensions, corresponding to the size of all the 22 human chromosomes (Y and M chromosomes removed) stacked together. The replicates are ordered together and Fig1.11 represents a sample of the formatted DNAse-seq database.

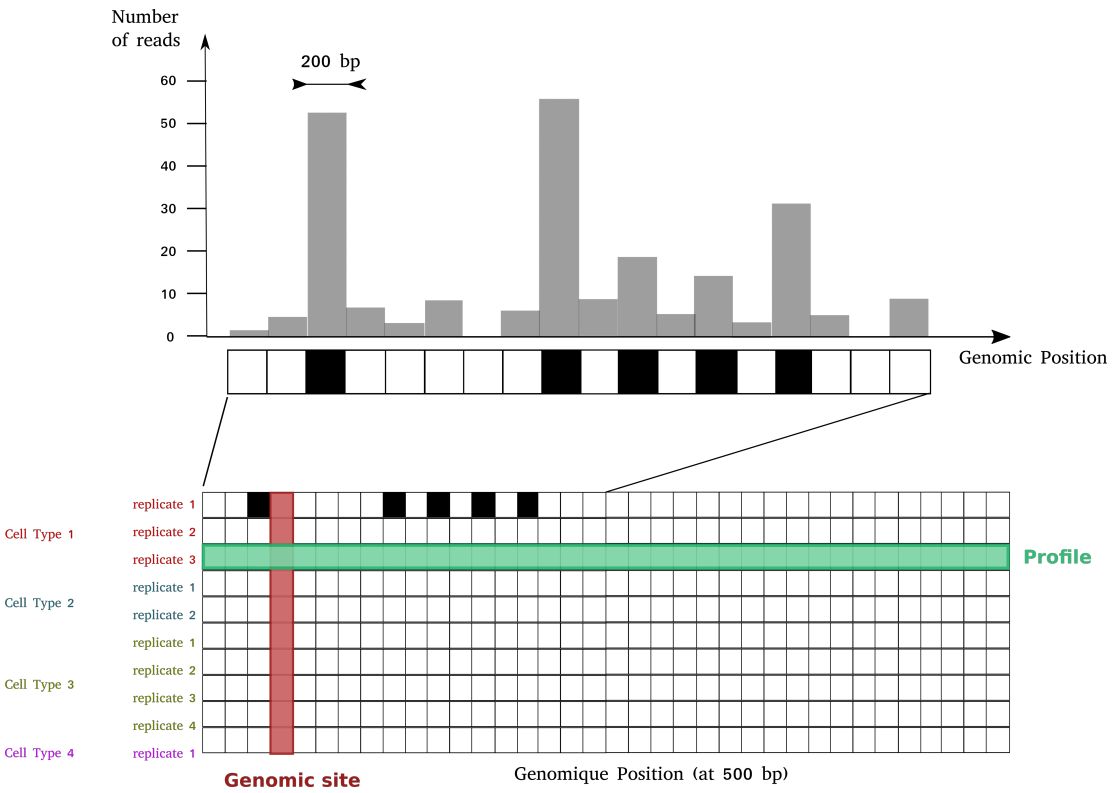

Figure 1.10 – The NGS profiles are stack in a matrix at 500 bp resolution

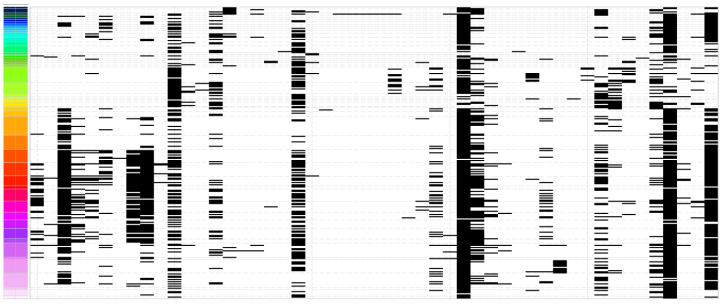

Figure 1.11 – Sample of the DNAse-seq database. Each line represent a profile, each color a different cell type and columns represent genomic sites at 500 bp

1.3 Plan of this manuscript

In chapter 2, we expose the problem of the reproducibility of Next Generation Sequencing experiments and develop a method to ensure the samples quality. In the chapter 3, we present our pattern identification algorithm GABI and its application, in chapter 4, on the single cell hematopoietic dataset which allow us to recover known results but also to discover original ones. In chapter 5 we present a different project aiming at reconstructing the 3D structure of chromosomes from their contacts.

Chapter 2

Quality of NGS experiments

2.1 Reproducibility in NGS experiments

The relevance of biological results based upon NGS experiments depends on the reliability of the NGS profiles. To increase this reliability, experiments are repeated multiple times, namely *replicates*, and a consensual profile for each experiment is deduced.

However, this is only possible if the replicates are "similar enough": if one or many samples has been disrupted, e.g. by a mistake in the experimental protocol or by a swap in the cell types labels, it may create artifacts and skew the biological conclusions .

We assert that a NGS profile is reliable if it is more similar to its replicates than to other experiments in the dataset. However, similarity among replicates is not obvious when comparing only some genomic site Fig2.1 due to the high variability of these samples. This is why we need a genome-wide comparison of profiles using *distance measures*.

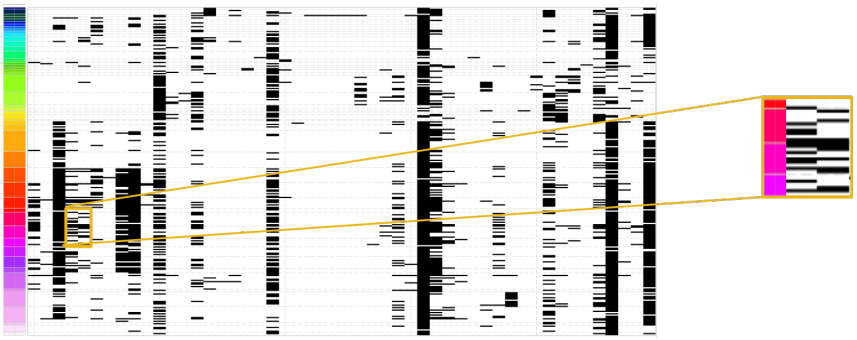

Figure 2.1 – An illustration of the variability between similar and different cell types on the DNAse-seq dataset. The yellow box represents three different cell types and two genomic sites. This variability shows the importance to produce replicates to ensure the reliability of an observation

2.2 The choice of a distance measure

There exists a large variety of distance measures and its choice is important to measure the quality of the profiles. A measure which cluster experimental replicates and separate different experiments Fig2.2(top right) leads to the identification of samples quality, whereas a mixed replicate clustering Fig2.2(down right), prevent the detection of anomalous samples.

There is three techniques, that can be combined, used to compute the distance between profiles[1]. First, the *feature extraction*, using for example the principal component analysis (PCA), which combined the features of the profiles, here corresponding to the genomic sites, in order to produce new features reducing the redundancy among features. Second, the *distance metric* corresponding to a function which takes a pair of profiles (N-dimensional vectors) and returns a scalar. In most studies, distance metrics are selected among the Pearson correlation, Euclidean distance, Spearman correlation for raw profiles or Jaccard distance for binary profiles. Third, the *kernel function* which are symmetric and positive semi-definite functions that has been wildly used in supervised learning [51]. We define a distance measure as the triplet combination of feature extraction, distance metric and kernel function. An extensive analysis of all distances measures doesn't exist so far to our knowledge. Consequently, we have developed a test to evaluate and find the optimal distance measures given a datasets.

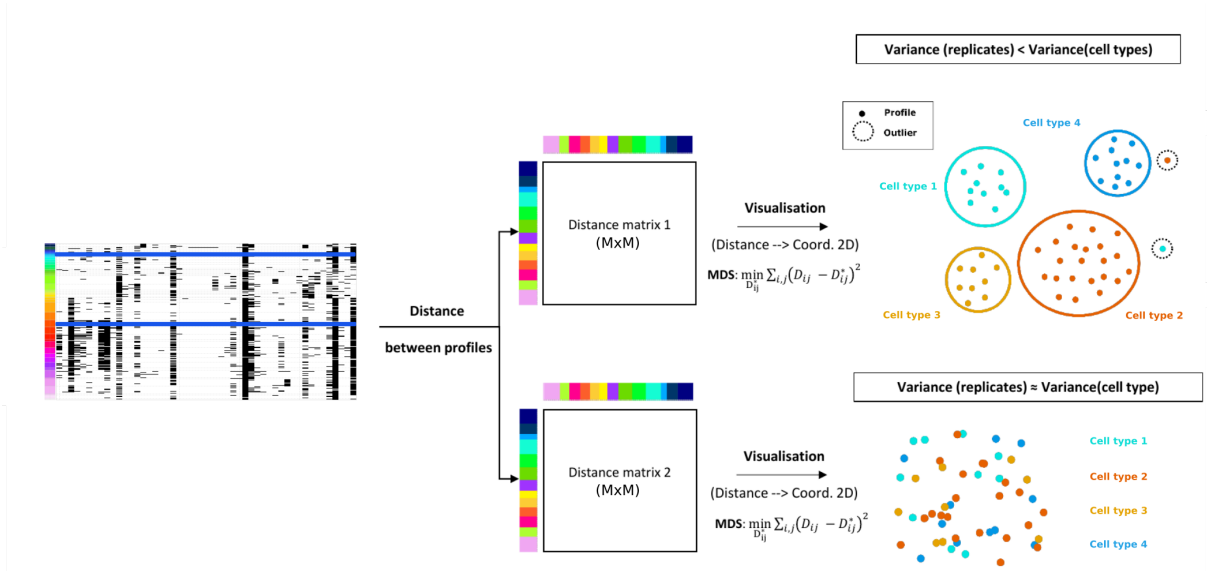

Figure 2.2 – Relevance of the choice of a distance measure. From the profile matrix, two different distance matrices are computed between the M profiles (in blue). The distance between profiles can be visualized using a Dimensionality Reduction (DR) algorithm such as the Multi-Dimensional Scaling (MDS) which minimizes the squared difference between the original distance and the 2D distance. Here each point represents an NGS profile. Because the difference between replicates should be less than the difference between cell types, profiles are expected to form *clusters*. In the case of the distance matrix 2, this property is not respected, which imply that this distance is less suitable than the distance 1 to represent this dataset. Our approach is to use *clustering indexes* to measure the goodness of the clusters, knowing the labels of the profiles

[1] Considered as vectors

To evaluate the quality of a distance measure, we have used indexes that measure the quality of replicates clusters, namely *clustering indexes*. These indexes are commonly used when the labels are not known for a given distance measure in order to find the best clustering algorithm. In our case the labels are given (the replicates), and we search the distance measure that optimize the clustering indexes value, ie produces the best clusters of replicates.

2.3 Clustering Indexes

We have selected two indexes to compare the different distance measures: Silhouette and connectivity. These indexes show reliable and robust values compared to other tested indexes[2]

Silhouette index Consider an ensemble of N points $x_{i\in[1,N]}$, grouped in C clusters $c_k \in C$. The Silhouette index [59] is the normalized difference of the mean distance between a point x_i and points within the same cluster, $a(x_i)$, and the smallest mean distance of x_i to points in other clusters, $b(x_i)$ Fig2.3. It is defined as:

$$S(C) = \frac{1}{N} \sum_{c_k \in C} \sum_{x_i \in c_k} S_i(C) \qquad S_i(C) = \frac{b(x_i) - a(x_i)}{max(a(x_i), b(x_i))}$$

with:

$$a(x_i) = \frac{1}{|c_k|} \sum_{\substack{x_j \in c_k \\ x_i \in c_k}} D(x_i, x_j)$$

$$b(x_i) = min_{c_l \in C \setminus c_k} \frac{1}{|c_l|} \sum_{\substack{x_j \in c_l \\ x_i \in c_k}} D(x_i, x_j)$$

The Silhouette value ranges from -1 to +1. A high $S_i(C)$ value indicates that i is closer to its clusters than the closest cluster, and a low value that the point i is closer to another cluster than its own. The total Silhouette value is the mean over all samples.

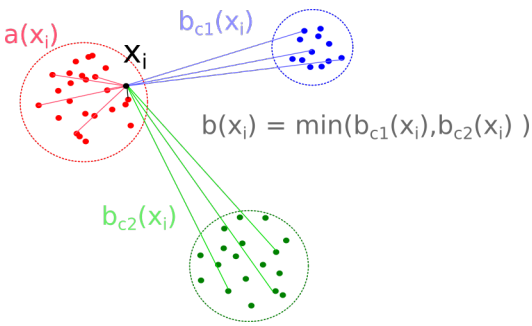

Figure 2.3 – An illustration of the Silhouette index on three clusters.

[2]Such as Calinski-Harabaz [15] which may diverge in some cases.

Connectivity index The connectivity index is defined by the number of nearest neighbor N_{neigh} chosen (= 5 in our test) . For a given point, the connectivity index is the sum over all its neighbors of the inverse of their rank[3] if they don't belong to the same cluster and zeros otherwise. The index is average over all the points value. For N observation points, K clusters, and M nearest neighbor the connectivity is defined as:

$$Conn = \frac{1}{N}\sum_{i=1}^{N} Conn_i$$

with:

$$Conn_i = \sum_{\substack{m=1 \\ c(x_i)\neq c(x_m)}}^{M} \frac{1}{rank_{x_i}(x_m)}$$

with $c(x_i)$ the cluster of the point x_i, and $rank_{x_i}(x_m)$ the rank of the point x_m from x_i.

The score range from $[0, \sum_{i=1}^{M} 1/m]$ and the lower the score the more separate are the clusters.

2.4 Distance measures description

2.4.1 Feature extraction

Feature extraction is an operation that create new features from original ones in order to reduce the redundancy and the noise among features. In this work, it is performed using the principal component analysis (PCA). PCA is an orthogonal linear transformation that convert the data features such that the first feature (or first component) maximizes the data variance, the second feature maximizes the remaining variance, and so on [54]. The PCA can be computed by singular value decomposition of the covariance matrix [1], which eigenvectors correspond to the principal components. We have defined the number of features by comparing the eigenvalue spectrum of the original data covariance matrix to a random model. We propose a random model based on the reshuffling instances (ones for binary and reads for raw datasets) conserving the row and column probability such that for a NGS profile matrix m indexed by row i and column j, the probability the null model has the value c at (i,j) is

$$P_{i,j}(m_{i,j} = c) = (P_i \,.\, P_j)^c$$

$$P_i = \frac{\sum_j m_{i,j}}{\sum_{i,j} m_{i,j}} \qquad P_j = \frac{\sum_i m_{i,j}}{\sum_{i,j} m_{i,j}}$$

with P_i and P_j the probability of the row i and column j respectively. The PCA on the random matrix is repeated multiple times and the number of components is selected such that the corresponding eigenvalues are above a z-score of 2.32 from eigenvalues of the random models. For the distance measures, we have tested either with and without PCA.

[3]The rank of the nearest neighbor correspond to k if the point is the k-nearest neighbor

2.4.2 Distance metrics

Distance metrics are implemented from different sources in order to make this study as exhaustive as possible. Let's consider two input N dimensional vectors (u, v) that can have either raw values $(u, v) \in \mathbf{N}^N$ or binary values $(u, v) \in \{0, 1\}^N$. There are two families of distance metrics: distances metrics that can be applied on both raw / binary profiles and distances metrics that can only be applied on binary profiles. These distances are defined in table 2.1 and table 2.2.

Distance	Expression
Pearson distance (1)	$1 - \frac{(u-\mu_u).(v-\mu_v)}{\sigma(u).\sigma(v)}$
Euclidean distance	$\sqrt{\|\|u-v\|\|_2}$
L2 distance	$\|\|u-v\|\|_2$
Mutual information	$\sum_i P_U(u_i).\log \frac{P_V(v_i)}{P_U(u_i)}$
Spearman distance (2)	$1 - \frac{(S(u)-\mu_u).(S(v)-\mu_v)}{\sigma(u).\sigma(v)}$
Chebyshev distance	$max_i \|\|u_i - v_i\|\|_1$
Canberra distance	$\sum_i \frac{\|u_i-v_i\|}{\|u_i\|+\|v_i\|}$
Bray Curtis distance	$\frac{\sum \|u_i-v_i\|}{\sum \|u_i+v_i\|}$
Cosine distance	$1 - \frac{u.v}{\|\|u\|\|_2.\|\|v\|\|_2}$
Seuclidean distance	$\sqrt{\sum_i (u_i - v_i)^2/v_i}$
Cityblock distance	$\sum_i \|u_i - v_i\|_1$

Table 2.1 – Distance metrics that can be applied on both raw and boolean profiles.
(1) μ the mean and σ the standard deviation (2) with $S(.)$ a sorting function

Distance	Expression
Jaccard distance	$\frac{C_{10}+C_{01}}{C_{11}+C_{01}+C_{10}}$
Yule distance	$\frac{2C_{10}C_{01}}{C_{11}.C_{00}+C_{01}.C_{10}}$
Dice distance	$\frac{C_{10}+C_{01}}{2C_{11}+C_{01}+C_{10}}$
Rogers Tanimoto distance	$\frac{2(C_{10}+C_{01})}{C_{11}+C_{00}+2(C_{10}+C_{01})}$
Sokal-Sneath distance	$\frac{2(C_{10}+C_{01})}{C_{11}+2(C_{10}+C_{01})}$
Sokal-Michener distance	$\frac{2(C_{10}+C_{01})}{2(C_{10}+C_{01})+(C_{11}+C_{00})}$
Russell-Rao distance	$\frac{N-C_{11}}{N}$
Hamming distance	$\frac{C_{10}+C_{01}}{N}$

Table 2.2 – Distance metrics that can be applied on binary profiles only. C_{00},C_{10},C_{01},C_{11} represents respectively the number of corresponding zeros, ones-zeros, zeros-ones and ones between u and v

2.4.3 Kernels

In order to expand the number of possibilities, we have also implemented variations of these distance metrics by applying *kernels functions* on top of them. Kernels functions are symmetric and positive semi-definite functions that has been wildly used in supervised learning [51]. One of the most used kernel is the *Radial Basis Function* [53] defined as the exponential of the L2 distance $K(x,y) = e^{-\gamma||x-y||^2}$ with x and y two vectors and γ a scalar parameter. Our extensions, rely on substituting the L2 distance by another distance metric. We assess three different kernels: Exponential kernel[4], sigmoid kernels and cosine kernels defined in table 2.3.

Kernel	Expression
Exponential (RBF) (1)	$K(x,y) = e^{-\gamma_e D(x,y)}$
Sigmoid (2)	$K(x,y) = 1 - tanh(-\gamma_s * D(x,y) + c)$
Cosine	$K(x,y) = 1 - (\sum_k D(x_k,y)D(x,y_k))/\|\|x\|\|_2\|\|y\|\|_2$

Table 2.3 – Kernels. (1) with γ_e a parameter (2) with γ_s and c parameters

Parameters selection Usually kernels parameters are selected using a grid search exploration. Here, in order to reduce the computing time, the parameters are computed according to the distance matrices: the γ_e parameter of the exponential kernel is computed as $\gamma_e = 1/\sum_k \sum_{i,j \in c_k, i>j} D(i,j)$. The intuition is that the standard deviation of the kernel should represent the mean size of the clusters. The γ_s of the sigmoid kernel is defined as $\gamma_s = \sum_k \sum_{i \in c_{k'\setminus k}, i>j} D(i,j)$ and the c is set to zero. The intuition is that the γ_s parameter represents the mean inter-clusters distance.

[4]The extension of radial basis function to any distance metric

2.5 Results

We have preformed our test on a raw and binary NGS dataset from pooled experiments and single cell experiments in order to identify the optimal distance in each cases. We have used the raw DNAse-seq profiles (without peak caller) as raw dataset, the binary DNAse-seq profiles (with peak caller) as binary dataset and the single cell ATAC-seq profiles as binary single cell experiment.

The figure Fig2.4 represents the twenty best scores of distance measures applied to the raw DNAse-seq dataset, computed by minimizing the sum of the Silhouette and connectivity indexes ranks. We identify two distance metrics that perform best on this dataset, Pearson and cosine, closely followed by Bray-Curtis and Spearman. PCA feature selection didn't improve the distance measure on this dataset and exponential kernel improved slightly the distance measure for all the aforementioned distance metrics.

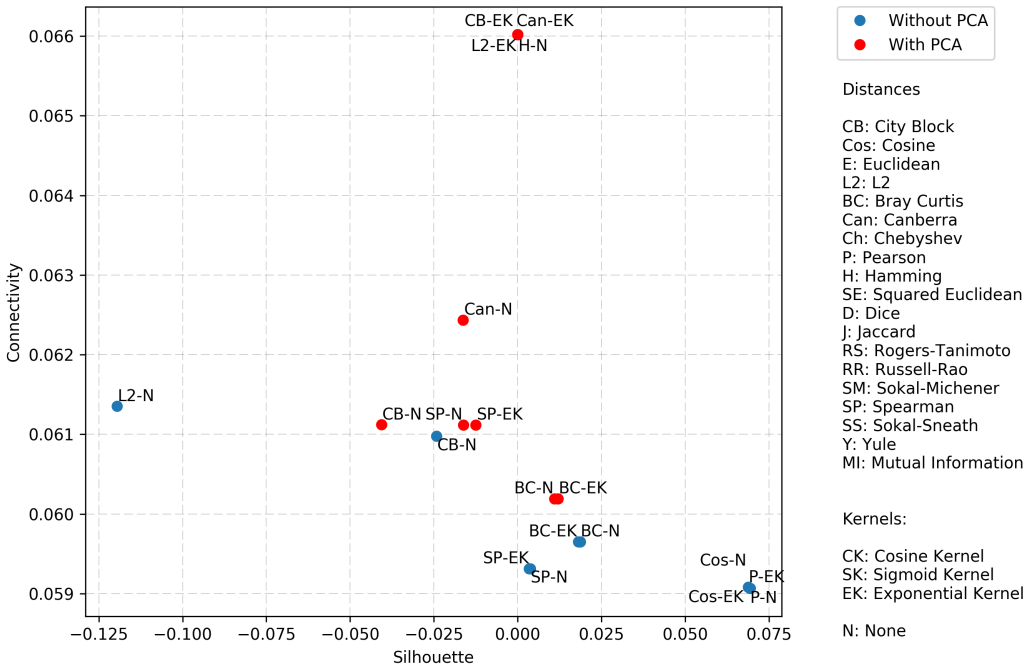

Figure 2.4 – Internal indexes applied to distances on the raw DNAse dataset. For Silhouette, the higher the score the better the replicates clustering while for connectivity, the lower the score the better the replicates clustering.

The figure Fig2.5 represents the twenty best Silhouette and connectivity scores of distance measures applied to the binary DNAse-seq dataset. We observe that optimal distance measures are Yule, cosine, Pearson, with and without PCA, and Spearman. These measures are again slightly increased with the exponential kernel. Connectivity measure the mixture of clusters while Silhouette measure the relative distance. Therefore, Yule distance allows to produce more

separated clusters than the other metrics. This is quite unexpected since Yule distance never appears in studies using binary NGS profiles, as opposed to Jaccard distance.

Comparing the indexes score of the best raw and binarized distance measures, we observe that binary distances achieve a better score. This indicates that binarization of NGS profiles with a peak caller is a good pre-processing for NGS profiles and confirms that peaks height is less relevant due to potential biases [47] than the peaks presence.

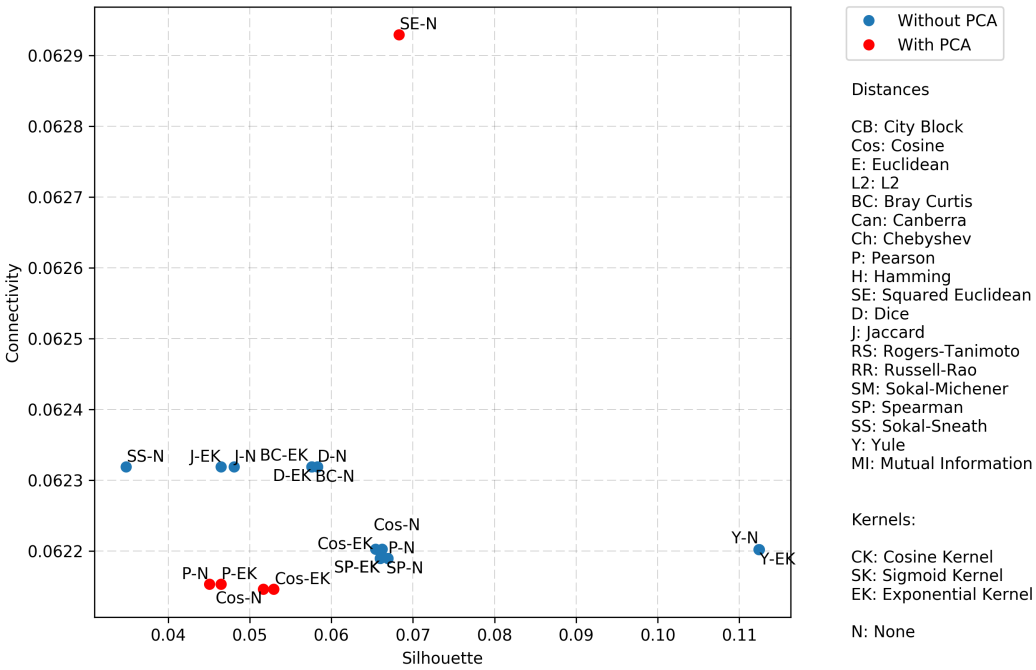

Figure 2.5 – Internal indexes applied to distances on the binary DNAse dataset. For Silhouette, the higher the score the better the replicates clustering while for connectivity, the lower the score the better the replicates clustering.

Eventually, Fig2.6 represents the twenty best Silhouette and the connectivity scores of distance measures applied to the binary single cell ATAC-seq dataset. Since the profiles are binary, Yule metric has been an expected candidate. But the inherent high false negative rate of single cell profiles leads Spearman with PCA feature extraction as the optimal distance measure. Contrarily to previous cases, exponential kernel doesn't improve the distance measure.

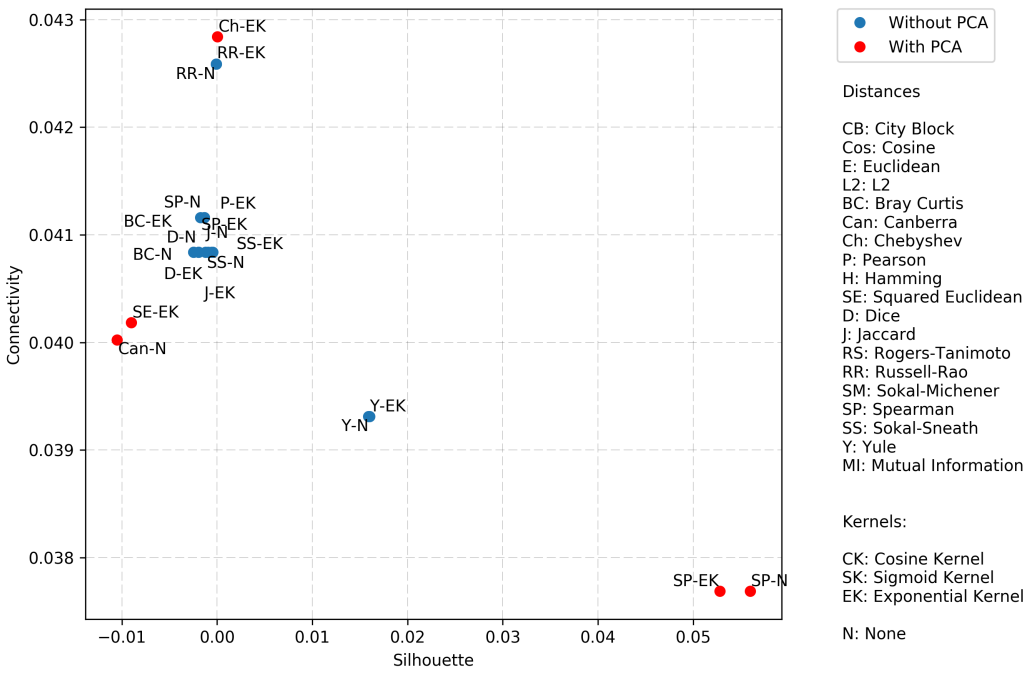

Figure 2.6 – Internal indexes applied to distances on the single cell ATAC-seq dataset. For Silhouette, the higher the score the better the replicates clustering while for connectivity, the lower the score the better the replicates clustering.

2.6 Identification of reliable NGS profiles

We assert that a NGS profile is reliable if it is more similar to its replicates than to other experiments in the dataset. With the optimal distance measure presented above we can now score the profiles membership to their clusters by comparing the intra-cluster distance and the inter-cluster distance. One way of doing it is by comparing the *Silhouette index for each profiles* (at a fixed distance measure), $S_i(C)$, to the Silhouette index of a null model, defined by reshuffling the cluster affiliation C_R (labels) multiple times.

$$S_i(C) = \frac{1}{N} \sum_{c_k \in C} \frac{b(x_i, c_k) - a(x_i, c_k)}{max(a(x_i, c_k), b(x_i, c_k))}$$

The distance to the null model is expressed via a Z-score such that:

$$SZ_i = \frac{S_i(C) - \overline{S_i(C_R)}}{\sigma(S_i(C_R))}$$

with $\overline{S_i(C_R)}$, $\sigma(S_i(C_R))$ the mean and the standard deviation of the Silhouette index over the reshuffled tests and SZ_i the Z-score of the Silhouette score of the profile i. Profiles with a Z-score

below a threshold T, $SZ_i < T$, are considered as outliers and thus removed. The choice of the threshold depends on the dataset: if the clusters correspond to well separated cell types (eg: muscle, lung, brain), we advice a threshold of 2.32 (associated in this case to a p-value of 10^{-3}) or 1.65 (associated in this case to a p-value of 10^{-2}). In the case of similar cell types (eg: muscle of back, muscle of arm, skeleton muscle) in the dataset, we advice to choose a threshold of 0 in order to remove only far outliers.

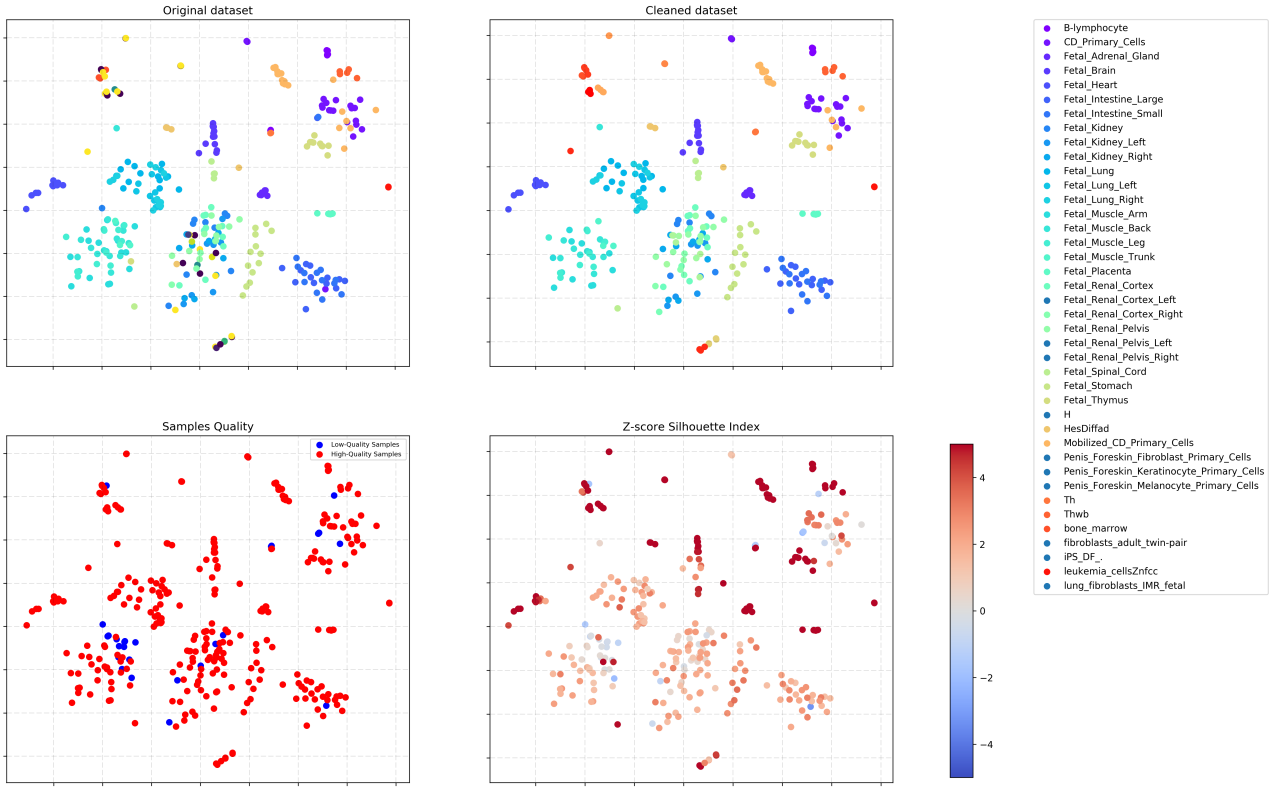

Figure 2.7 – NGS Quality procedure applied to raw DNAse-seq dataset. The scatter plot is performed using the t-SNE dimensionality reduction algorithm [45].(Original Dataset) The dataset with all the profiles. (Cleaned Dataset) The dataset without outliers. Here the threshold is set to zero (Samples Quality) The outliers are displayed in blue and the rest in red (Z-score Silhouette Index) Z-score of the profiles

We have performed this NGS profile selection pipeline on the three datasets. In Fig2.7, Fig2.8 and Fig2.9 we represent four images representing the NGS profiles using their respective optimal distance for each and displayed in two dimensions using the t-SNE algorithm [45]. It constructs a probability distribution of all pairs of points in the high dimensional space and optimizes the position of these points in the low dimensional space by minimizing the Kullback-Liebler divergence [39] between the probability distribution in the high and low dimensional space. The parameters of the t-SNE are optimized by maximizing the Silhouette score of the two dimensional representation. Among the four images, we display the original dataset (top left), the samples Silhouette Z-score (bottom right) the samples below and upper the threshold (below left) and the dataset removed from samples below the threshold (upper right). For the three datasets, the threshold

has been set to zero to remove only far outliers.

To underline the statement on the importance of the distance measure choice to detect low quality profiles, we have performed in appendix the same quality procedure on the single cell ATAC-seq dataset using a distance measure not adapted to these profiles.

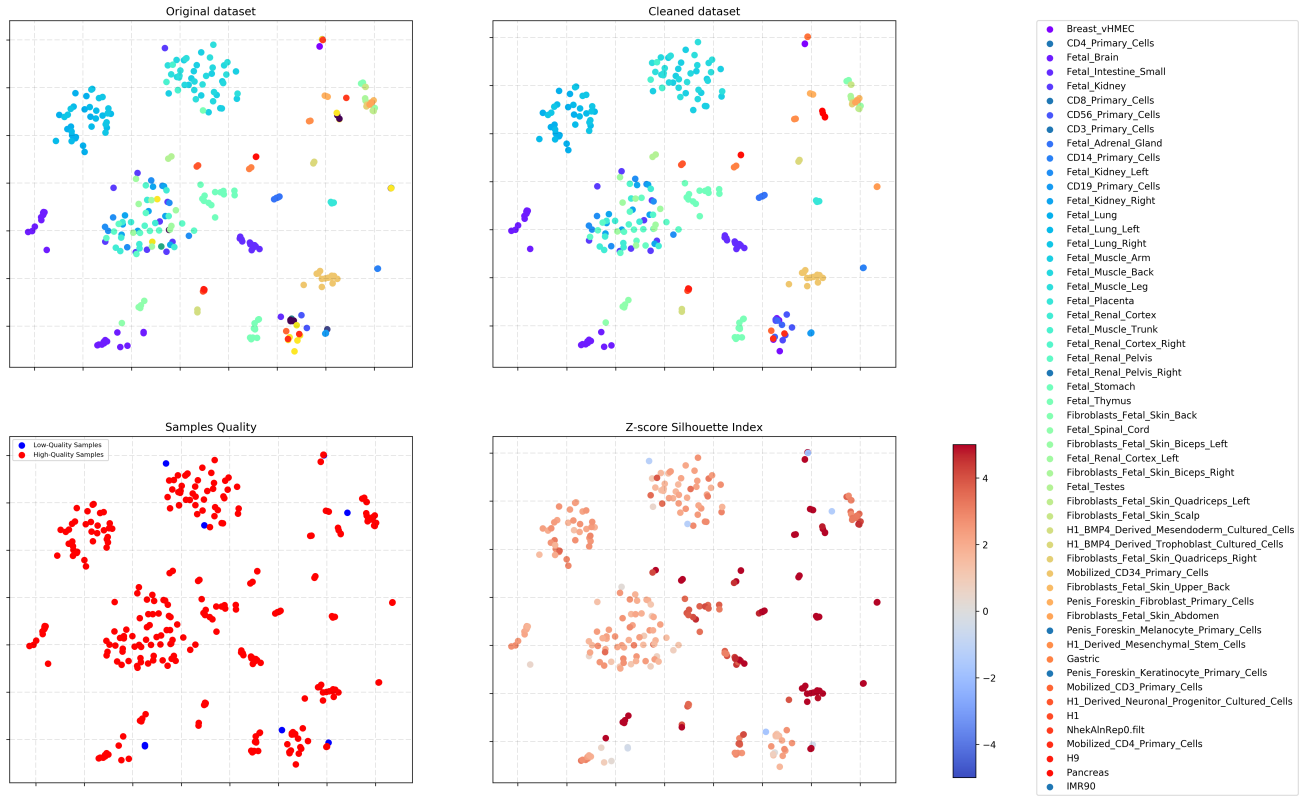

Figure 2.8 – NGS Quality procedure applied to binary DNAse-seq dataset. The scatter plot is realized using the t-SNE dimensionality reduction algorithm [45]. (Original Dataset) The dataset with all the profiles. (Cleaned Dataset) The dataset without outliers. Here the threshold is set to zero. (Samples Quality) The outliers are displayed in blue and the rest in red (Z-score Silhouette Index) Z-score of the profiles

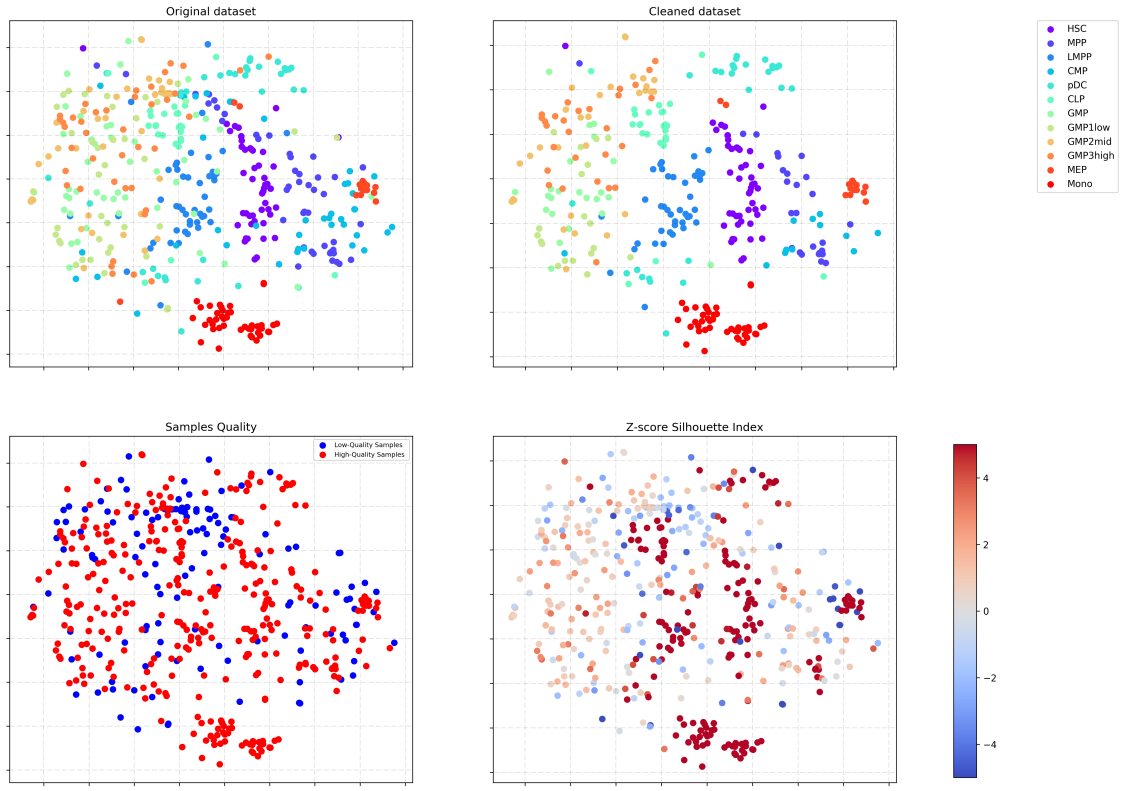

Figure 2.9 – NGS Quality procedure applied to binary single cell ATAC-seq dataset. The scatter plot is realized using the t-SNE dimensionality reduction algorithm [45]. (Original Dataset) The dataset with all the profiles. (Cleaned Dataset) The dataset without outliers. Here the threshold is set to zero. (Samples Quality) The outliers are displayed in blue and the rest in red (Z-score Silhouette Index) Z-score of the profiles

2.7 Conclusion

We have proposed a procedure to identify the optimal distance measure for a NGS dataset involving a feature extraction, a distance metric and a kernel function. We have identified that Spearman, Bray-Curtis, Pearson and Cosine distances, without PCA, with and without exponential kernels perform best on raw (discrete) NGS profiles.

For NGS profiles binarized by a peak caller, Yule distance, with and without exponential kernel, outperforms the other distance measures. Pearson, Cosine and Spearman distances with PCA, with and without exponential kernels, have a slightly better connectivity index but lower Silhouette score. Interestingly , Yule distance is never used, to our knowledge, in NGS studies, preferring Jaccard, Euclidean or Pearson for instance.

For single cell NGS profiles, we show that PCA feature extraction combined with Spearman distance, with an without exponential kernel, outperformed all the other distance measures.

We have also defined a protocol to select high quality NGS profiles based on the optimal distance measure. We provide a user-friendly package in python to test the distance measures and the profiles quality at `https://github.com/jbmorlot/NGSQualityTester` .

Chapter 3

Automatic genome annotation

3.1 The concept

Automatic annotation of the genome is the association of a biological function to DNA locii sharing common genomic features. It relies on two categories of algorithms. The first, called *supervised*, estimates the patterns of genomic features based on classified genomic sites. For example, we can discover the recurrent patterns associated to enhancers, given two ensembles of genomic sites classified as enhancers and not enhancers. These algorithms are optimized by minimizing the error between inferred classes and known classes 3.1(a). The second, called *unsupervised*, estimates the recurrent patterns without external data, by minimizing the difference between the original data and the data reconstructed from the patterns Fig3.1(b). Many different cost function can be designed to measure this difference depending on the chosen model (eg Max Log-Likelihood, Mean Square Error, Kullback-Liebler divergence).

Genome annotation is a wide field since different parts of the genome are annotated using very different biological experiments depending on the studied biological process[80] [67] [5] [32]. In this work, we focus on *unsupervised automatic annotation from NGS experiments* in order to discover new structures in the data and not be constrained on the availability of external data.

3.2 State of the art of human genome automatic annotations from NGS experiments

The two most important human genome automatic annotations from NGS experiments lately are: The ENCODE and Roadmap Project Consortiums.

The ENCODE Consortium [18] have analyzed 1,640 profiles resulting from 7 experiments (RNA-seq, CAGE, RNA-PET, ChIP-seq, DNAse-seq, FAIRE-seq, RRBS) on 147 different cell types, from which they have annotated 80% of the human genome with a biochemical function. They have used the two annotations strategies: Supervised annotation to identify putative enhancer and unsupervised annotation to discover combinations of histone modification, using a hidden Markov model *ChromHMM*[25]. Histone modifications combinations assign a biological function to genomic sites, such as promoters, enhancers or heterochromatine. References and further explanations are provided in appendix *Histone modifications and references*.

The Roadmap Epigenomics Consortium [40] have provided 2,805 genome-wide datasets in

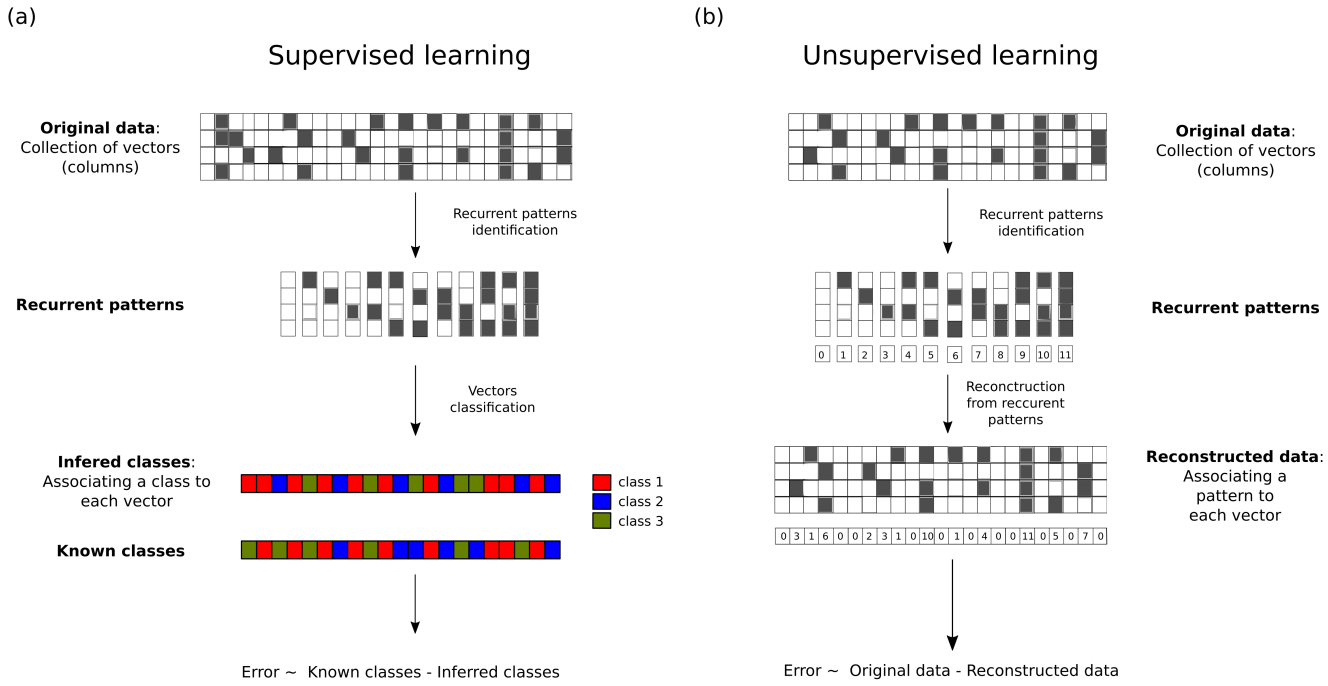

Figure 3.1 – Principle of supervised and unsupervised learning. (a) In supervised learning, classes of the data are known. The algorithm infers recurrent patterns by minimizing the error between known classes and inferred classes. (b) In unsupervised learning algorithm, classes are not known. The algorithm infers the recurrent patterns by minimizing the error between the original data and the reconstructed data.

127 cell types, including 1,821 histone modifications data sets, 360 DNA accessibility data sets, 277 DNA methylation datasets, and 166 RNA-seq data sets. Missing profiles among the experiments are imputed [27] resulting in 28 histones marks, DNAse-seq, RNA sequencing and methylation sequencing in 127 cell types. Again ChromHMM has been used and they have annotated the genome of the 127 cell types with a biochemical function.

3.3 Pitfall with the current annotation algorithms

Current annotation algorithms, such as ChromHMM, focus on the discovery of unconstrained patterns. However, if applied directly to an ensemble of NGS experiments with replicates, such as Fig3.3(a), the discovered patterns won't share a common value among the replicates Fig3.2(b). To circumvent this problem, the usual protocol is to *merge the replicates* first and then discover patterns among the cell types which reduces the information for the annotation. Moreover, there exists different techniques to merge the replicates and no studies, to our knowledge, has tested the impact of these approach to the final result.

To increase the reliability by removing the merging step, we have chosen to develop our own annotation algorithm *GABI* which discovers patterns among all NGS profiles by constraining the value among the replicates to be exactly identical Fig3.2(a).

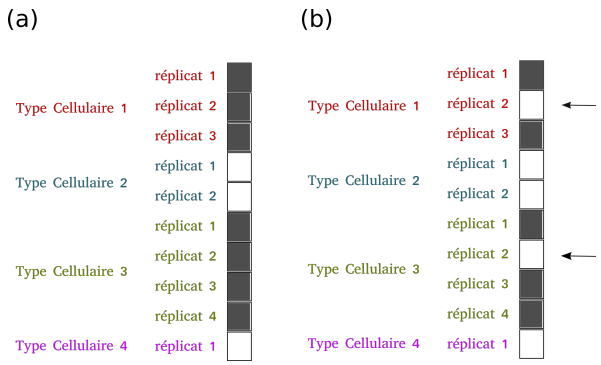

Figure 3.2 – Actual annotation algorithm pitfalls. (a) Expected pattern (b) With an usual annotation tool, the value of the replicates is not constrained to be the same (see arrows)

3.4 Genome-wide Annotation with Bayesian Inference (GABI)

GABI is an unsupervised genome annotation based on a probabilistic model which discovers recurrent patterns in *binary* NGS profiles under the constrain that replicates are exactly identical. The optimization of such algorithms is performed using the *Expectation-Maximization (EM) algorithm* which maximizes the likelihood of a probability distribution. In GABI, this distribution is defined as the similarity between the genomic sites, representing all genomic positions, Fig3.3(a), and an ensemble of vectors, namely the *states of reference* Fig3.3(b), which represents all possible patterns. This distribution is parametrized by the False Negatives (FP), the False Positives (FP) of each profiles and the probability of a reference state τ Fig3.3(c).

The probability distribution is written such that:

$$P(\hat{i}=i,\hat{k}=k,\theta)=\frac{1}{R}\prod_{d=1}^{D} a_{11,d}^{X_{id}Z_{kd}}\ a_{10,d}^{X_{id}(1-Z_{kd})}\ a_{01,d}^{(1-X_{id})Z_{kd}}\ a_{00,d}^{(1-X_{id})(1-Z_{kd})}\ \tau_k$$

with $P(\hat{i}=i,\hat{k}=k,\theta)$ the probability of a reference state k and the genomic site i, D the number of profiles,

$$P(\hat{i}=i)=\prod_{d=1}^{D} P(\hat{X}_{id}=X_{id})\ \text{ with }\ X_{id}=\{0,1\}$$

the probability of the genomic site i defined as the product of the probability of the D NGS profiles $d\in[1,D]$ to be equals to 0 or 1.

$$P(\hat{k}=k)=\prod_{d=1}^{D} P(\hat{Z}_{kd}=Z_{kd})\ \text{ with }\ Z_{kd}=\{0,1\}$$

the probability of reference state k. $a_{00,d},a_{01,d},a_{10,d},a_{11,d}$ are respectively as the True Positive (TP), False Negative (FN), False Positive (FP), True Negative (TN) rates of the profile d, τ_k is the

probability of the state k and R is a normalization factor, defined as:

$$
\begin{aligned}
a_{11,d} &= P(\hat{X}_{id} = 1 | \hat{Z}_{kd} = 1, \theta) \\
a_{10,d} &= P(\hat{X}_{id} = 1 | \hat{Z}_{kd} = 0, \theta) \\
a_{01,d} &= P(\hat{X}_{id} = 0 | \hat{Z}_{kd} = 1, \theta) \\
a_{00,d} &= P(\hat{X}_{id} = 0 | \hat{Z}_{kd} = 0, \theta) \\
\tau_k &= P(\hat{k} = k | \theta)
\end{aligned}
$$

An application of this distribution on a small example is provided Fig3.4. The complete development of the calculations with the EM algorithm is provided in the appendix *GABI calculation and algorithm description*.

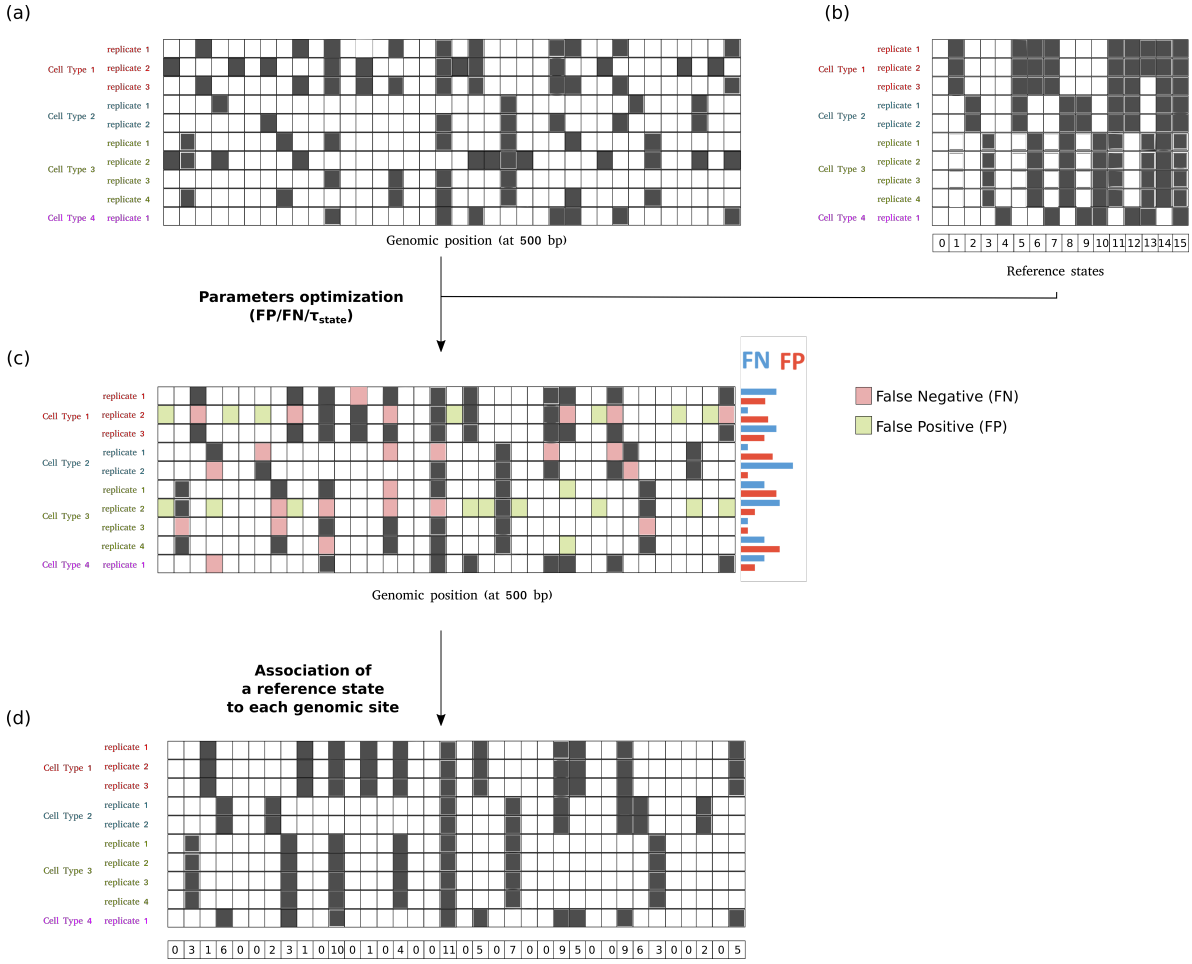

Figure 3.3 – Principle of Genome annotation. (a) Schematic representation of an usual NGS profiles matrix (b) Reference states corresponding to all the possible combinations among the cell types with the same value among replicates. (c) Parameter optimization reveals FN, FP rates and the probability τ_{state} of a reference state. (d) Reconstructed matrix by association of the reference states to each genomic sites

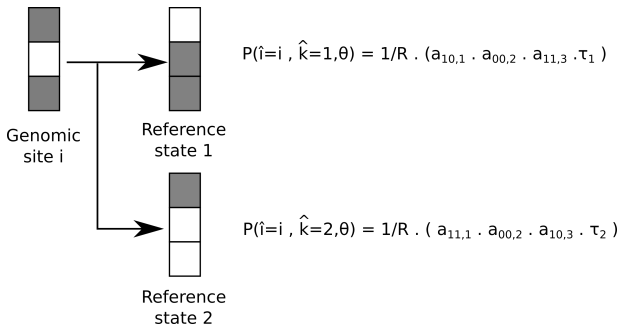

Figure 3.4 – The probability of reference states k and a genomic site i is adapted through the FP, FN rates of each profile and the states probability τ_k

3.5 Comparison with the state of the art methods

We have compared GABI to the states of the art of annotation, which consists in merging the replicates and possibly uses an annotation tool. We have explored four different merging techniques, commonly used in many studies. *Sum merging*, which sums the replicates reads over the replicates before to apply a peak caller. *Uniform merging*, which sums the reads of the replicates and down-samples them in order to have the same number of reads for all merged profiles. The procedure is then followed by the application of a peak caller. And *Proportional merging* which applies the peak caller first on all profiles before merging the replicates, assigning for each genomic sites a non zero value if the number of replicates with a non zero value is larger than the number of replicates with a zero value. We have compared GABI to the most used annotation algorithm, the hidden Markov models and we used ChromHMM [25] as benchmark.

We have performed a synthetic dataset, in order to test the performance of these techniques. The synthetic matrix comes from a binary matrix, representing a noise free signal, i.e. in which replicates are identical. We have token five cell types of 30 replicates each. Every genomic site is set to a random combination among the $2^5 = 32$ possibilities following the distribution of states of five DNAse-seq cell types. The matrix is then modified using a False Positive and False Negative rate identical for each profile. We have evaluated the methods by computing the area under the precision-recall curve (AUC) [10].

We define the probability to have an observation for the reconstructed matrix at a genomic site i for GABI as $P_{GABI}(i,d) = \sum_{k=1}^{K} P(\hat{k} = k|\hat{i} = i, \theta)\, Z_{kd}$ and for the HMM as $P_{HMM}(i,d) = \sum_{k=1}^{K_{HMM}} P(\hat{k} = k|\hat{i} = i, \theta)\, E_{kd}$ with E the K_{HMM} x D emission matrix corresponding to the K_{HMM} different optimized patterns of the HMM [81].

Unlike GABI, HMM optimizes the patterns and the number of states is defined before the optimization. We have defined the number of states in two different ways: either by fixing it to $2^5 = 32$, corresponding to all the possible combinations, which should be equivalent to GABI, either by minimizing the Bayesian information criterion (BIC) criterion [61] [84]. This criterion is commonly used with probabilistic model to select the optimal number of parameters, along with the Akaike information criterion (AIC) [3][1].

[1] It often doesn't lead to a minimum because it is less constrained than BIC

3.5.1 Results

Table 3.1 presents the results of the different annotation procedures on a synthetic dataset of five cell types and 150 profiles. We have performed a simulated matrix disturbed with different (FN,FP). Bold values represent the top performers for each FN-FP.

We observe that GABI with no merge outperforms all the other annotation procedures especially in the case of high (FN,FP). Pooled DNAse-seq correspond to $(FN, FP) = (0.01, 0.3)$ in average, which correspond to an AUC of 1.0 and single cell ATAC-seq, $(FN, FP) = (0.01, 0.9)$ in average, which represents an AUC of 0.93 with GABI.

Not merging replicates and defining reference states with identical value among replicates allows to score the sample by their (FN,FP) and to more efficiently recover, for each genomic site, the associated state. The Table 3.1 shows that after a merging operation, the annotation, using GABI, using HMM with 32 states or using HMM with BIC, don't increase significantly the AUC and is even slightly lower in average than the merge alone.

Unlike GABI, HMMs optimize the states and a state transition matrix. Transition matrix of HMM increases slightly the optimization of successive genomic sites most likely by inferring the same state, since the transition from a state to itself are more probable. Handling transitions is time consuming compared to the potential gain, therefore in GABI, we have decided to focus on independent genomic sites. All possible combinations evolves as $2^{N_{CT}}$ with N_{CT} the number of cell types. When this number is large, and we will show in chapter 4 on hematopoiesis, the number of states covering most of the genomic sites is a small subset of all possible combinations. Furthermore, unlike the HMMs, taking all states allows to discover rare combinations and avoid the optimization of the number of states which needs to launch multiple times the HMM in order to optimize a parameter such as BIC.

GABI				
(FN , FP)	Merge Uniform	Merge Sum	Merge Prop	No Merge
(0.1 , 0.01)	0.732189	0.796823	0.999995	**1.000000**
(0.3 , 0.01)	0.998388	**1.000000**	0.694963	**1.000000**
(0.5 , 0.01)	0.973686	0.799805	0.741249	**1.000000**
(0.7 , 0.01)	0.801364	0.932318	0.395107	**0.999909**
(0.9 , 0.01)	0.456105	0.421016	0.000000	**0.938357**
(0.1 , 0.1)	0.718551	0.922829	**1.000000**	**1.000000**
(0.3 , 0.1)	0.704560	0.939165	0.995550	**1.000000**
(0.5 , 0.1)	0.698318	0.949744	0.759204	**0.999755**
(0.7 , 0.1)	0.512336	0.827230	0.373358	**0.962490**
(0.1 , 0.3)	0.304047	0.221072	0.962585	**1.000000**
(0.3 , 0.3)	0.382416	0.286358	0.935671	**0.998661**
(0.5 , 0.3)	0.294445	0.255401	0.686937	**0.901771**

Merge only			
(FN , FP)	Merge Uniform	Merge Sum	Merge Prop
(0.1 , 0.01)	0.999872	1.000000	**1.000000**
(0.3 , 0.01)	0.998265	1.000000	0.992804
(0.5 , 0.01)	0.966284	0.999272	0.753350
(0.7 , 0.01)	0.810993	0.934845	0.575678
(0.9 , 0.01)	0.588915	0.580188	0.000000
(0.1 , 0.1)	0.898596	0.975558	**1.000000**
(0.3 , 0.1)	0.892679	0.979086	0.992085
(0.5 , 0.1)	0.868799	0.976877	0.756513
(0.7 , 0.1)	0.669710	0.852179	0.574683
(0.1 , 0.3)	0.607109	0.593999	0.980094
(0.3 , 0.3)	0.604118	0.591697	0.972883
(0.5 , 0.3)	0.585676	0.593660	0.712568

HMM (32 states)			
(FN , FP)	Merge Uniform	Merge Sum	Merge Prop
(0.1 , 0.01)	0.995474	0.995325	0.995325
(0.3 , 0.01)	0.993287	0.994787	0.989211
(0.5 , 0.01)	0.967301	0.994739	0.751071
(0.7 , 0.01)	0.823715	0.943245	0.575678
(0.9 , 0.01)	0.559255	0.572448	0.000000
(0.1 , 0.1)	0.715494	0.919793	0.994566
(0.3 , 0.1)	0.721250	0.930621	0.989950
(0.5 , 0.1)	0.684582	0.924658	0.754425
(0.7 , 0.1)	0.518374	0.817632	0.574683
(0.1 , 0.3)	0.211866	0.171423	0.943976
(0.3 , 0.3)	0.203138	0.176628	0.916812
(0.5 , 0.3)	0.192547	0.176496	0.669103

HMM (BIC optimized)			
(FN , FP)	Merge Uniform	Merge Sum	Merge Prop
(0.1 , 0.01)	0.996400	0.996355	0.996355
(0.3 , 0.01)	0.995096	0.995894	0.992335
(0.5 , 0.01)	0.970686	0.997315	0.749209
(0.7 , 0.01)	0.820915	0.944901	0.440885
(0.9 , 0.01)	0.465967	0.554817	0.000000
(0.1 , 0.1)	0.722972	0.913078	0.996325
(0.3 , 0.1)	0.697447	0.934549	0.991746
(0.5 , 0.1)	0.676707	0.922686	0.751401
(0.7 , 0.1)	0.506039	0.811436	0.448965
(0.1 , 0.3)	0.386469	0.272341	0.937451
(0.3 , 0.3)	0.367177	0.272639	0.928915
(0.5 , 0.3)	0.290557	0.237094	0.675321

Table 3.1 – Comparison of the different annotation procedures on five cell types with 30 replicates each. The value correspond to the AUC precision-recall for different FN-FP. Bold values represent the top performers for each of these matrices. (GABI) GABI annotation is applied on each merging procedure and without merging. (Merge Only) No annotation procedure where applied on top of the merging procedure. (HMM 32 states) ChromHMM is given the maximal number of states in the dataset. (HMM BIC optimized) ChromHMM number of states is defined by minimizing the BIC.

3.6 Influence of the number of cell types and number of replicates

The precision and the recall depends on the number of cell types and replicates. In figure Fig3.5 we represent the AUC on three different simulated datasets with different cell types and replicates. The (FN,FP) of the three datasets correspond to usual cases: $(\mathrm{FN}, \mathrm{FP}) = (0.1, 0.01)$ and $(\mathrm{FN}, \mathrm{FP}) = (0.3, 0.01)$ correspond to common pooled NGS profiles and $(\mathrm{FN}, \mathrm{FP}) = (0.9, 0.01)$ to common single cell NGS profiles. For $\mathrm{FN} = 0.3$ the AUC of GABI is higher than 0.9 for more than two replicates, independently of the number of cell types. For $\mathrm{FN} = 0.5$, the AUC of GABI is higher than 0.9 for more than 5 replicates by cell types or with five cell types with at least 2 replicates. And eventually, in the case of $\mathrm{FN} = 0.9$, at least 30 replicates per cell types are required to get an AUC higher than 0.9. Whereas pooled datasets with 30 replicates is rare, it is more common in single cell datasets. In summary, GABI is well-adapted for analyzing all kind of NGS datasets.

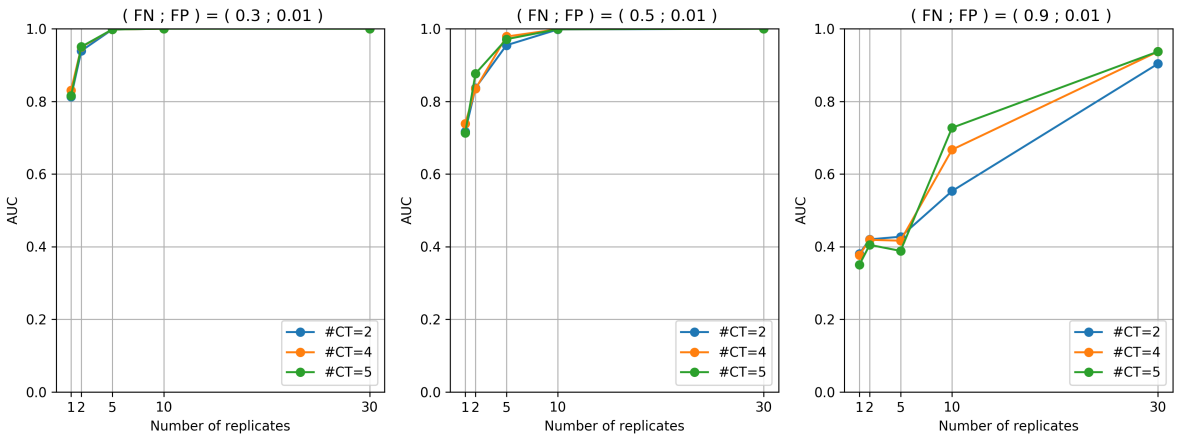

Figure 3.5 – AUC on three different simulated datasets with different cell types and replicates. The (FN,FP) of the three datasets correspond to usual cases: $(\mathrm{FN}, \mathrm{FP}) = (0.1, 0.01)$ and $(\mathrm{FN}, \mathrm{FP}) = (0.3, 0.01)$ correspond to common pooled NGS profiles and $(\mathrm{FN}, \mathrm{FP}) = (0.9, 0.01)$ to common single cell NGS profiles.

Annotation with multiple cell types is especially interesting when the number of replicates is low. We present in Fig3.6 a dataset of two cell types with different replicates each and we measure the AUC of one of the two. When increasing the number of replicates in cell type 1 from three to nine for three replicates in cell type 2, the AUC of cell type 2 increases from 0.45 to 0.7. This effect diminish when the number of replicates in the cell type 2 increases. The inferred combinations links the replicates among cell types, and therefore, cell types sharing multiple non null combinations will increase their reliabilities. More the cell types in a dataset are similar, more the fact to increase replicates of one cell types will be equivalent to increase replicates in all cell types.

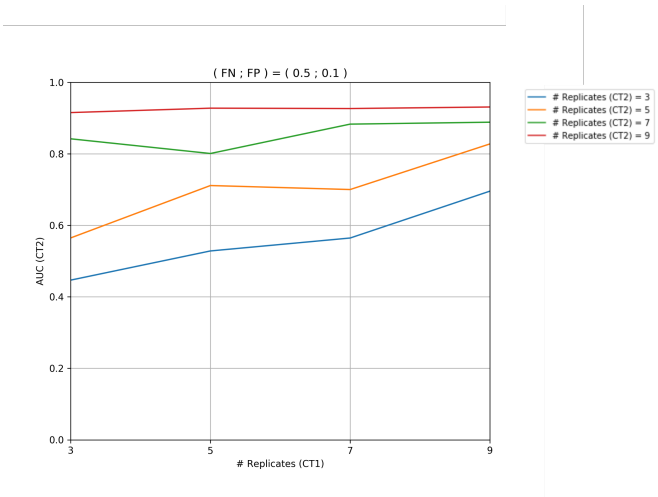

Figure 3.6 – Influence of the replicates on other cell types. The dataset posses two cell types. The AUC of cell type 2 is measured for different number of replicates in each cell type.

3.7 Inference of false positive and false negative

We have used the synthetic matrix defined above to compare the FN and FP rates inferred by GABI and the real ones Fig3.7. We observe that the real FP rate is well-recovered by GABI under 0.1 and with a larger variance at 0.3. The inferred FN rate also match to the real FN rate, with a larger variance for high values above 0.9. Therefore, GABI recovers well the (FP,FN) rates in the datasets, which is an indicator of profiles quality. This information can be used to study the origins of these differences among replicates.

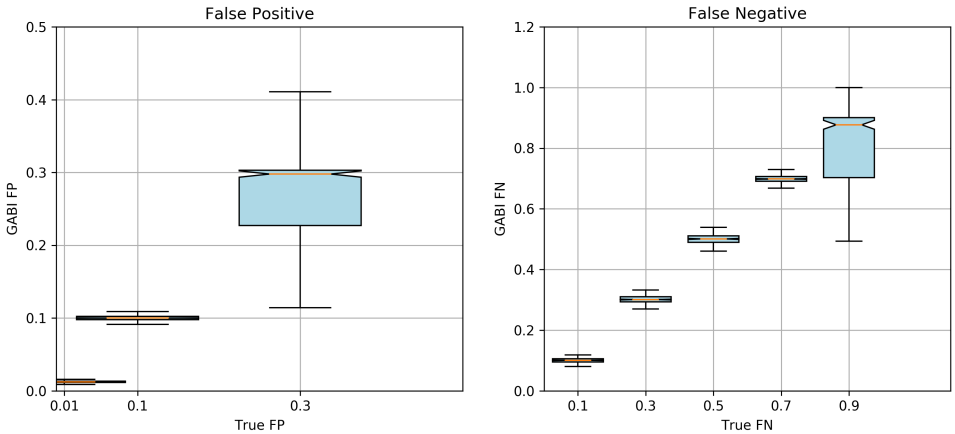

Figure 3.7 – Inference of the False Negative (FN) rates and False Positive(FP) rates by GABI. The simulated datasets mixes FP and FN ranging from $FP \in [0.01, 0.3]$ and $FN \in [0.1, 0.9]$ respectively. The boxes represent the upper and lower quartile for each true (FN,FP), the orange bar, represents the median and the whiskers show the range of the data

3.8 Conclusion

We have developed an unsupervised annotation method on binary NGS profiles, GABI, that uses the information of all the replicates by inferring a FP, FN rate for each profile.

Any NGS experiment is subject to artifacts that could come from the biological diversity of the sample, an issue in the experimental protocol or an error in the computational treatment of the profiles. Annotation simplifies and increases the reliability of the biological information contained in a dataset. Genome annotation on NGS profiles is an extension of differential analysis, which studies the differences between only two ensembles of profiles and interprets their differences. GABI has shown to outperform other procedures, especially in hard cases such as high FN and FP rates, which makes it suitable for *single cell* datasets. We provide a user-friendly package in python to use GABI at `https://github.com/jbmorlot/GABI` .

Chapter 4

Hematopoiesis analysis

4.1 Introduction to the hematopoiesis

The hematopoiesis is the study of blood cells genesis. It is one of the most studied system for cell differentiation models due to the presence of hematopoietic stem cells (HSC), the precursors of the hematopoietic lineage, in bone marrow. During hematopoiesis, HSC either self-renew either differentiates into blood cell types through successive stages of lineage commitment. This process has become a prototype of multi-lineage diversification from stem cells [69] [24] [41] [16] [62]. The prevailing model of hematopoiesis Fig4.1 predicts a first differentiation step of HSC toward Multi-Potent Progenitors (MPP) before splitting into the *myeloid* and the *lymphohoid* lineages. This model lies on the identification of the Common Lymphoid Progenitors (CLP) [35] and the Common Myeloid Progenitor (CMP) [4]. However, recent single cell in vivo experiments [55] highlight that this representation is still incomplete and misunderstood.

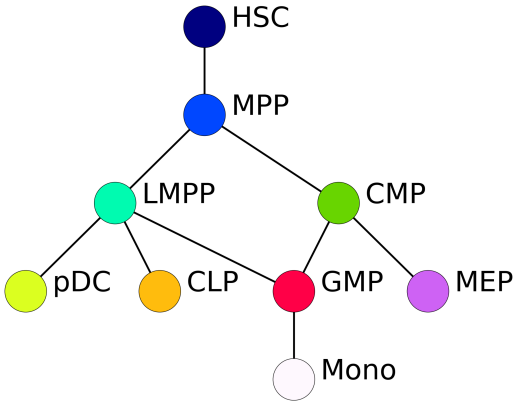

Figure 4.1 – Hematopoiesis development pathways. **HSC**: Hematopoietic Stem Cells. **MPP**: MultiPotent Progenitor. **LMPP**: Lymphoid-primed MultiPotent Progenitor. **CMP**:Common Myeloid Progenitor. **CLP**: Common Lymphoid Progenitor. **pDC**: plasmacytoid Dendritic Cell. **GMP**: Granulocyte-Macrophage Progenitors. **MEP**: Megakaryocyte Erythroid Progenitor. **Mono**: Monocyte.

4.2 Annotation

4.2.1 NGS quality selection

We have performed a quality test on NGS profiles, keeping the top 100 profiles per cell type which respect a Silhouette z-score of at least 1.65 (equivalent to a p-value of 10^{-2}). We have used the optimal distance measure defined in chapter 2: PCA feature extraction with Spearman distance. The table 4.1 presents the cell types mean z-score value, which is an indicator of profiles similarity within clusters, and the number of samples per cell types.

Name	Number of samples	Mean z-score Silhouette
HSC	100	7.97534057337
MPP	24	3.29652016787
LMPP	100	7.80727017549
CMP	100	6.02394166809
pDC	73	14.4182383982
CLP	100	26.007825668
GMP	32	2.62027218729
MEP	100	11.0472076293
Mono	84	28.1673353084

Table 4.1 – Mean z-score value among the profiles and number of samples within each clusters (cell types)

We have used the powerful t-distributed Stochastic Neighbor Embedding (t-SNE) [45] DR algorithm to represent the filtered single cell profiles Fig4.2. We observe that some profiles are spitted in more than one cluster, some even mixed with other cell types. This is the case, for example, of MPP-LMPP, GMP-CLP, CLP-pDC or CMP-pDC. This result might indicate different populations of the same cell types or new developmental stages.

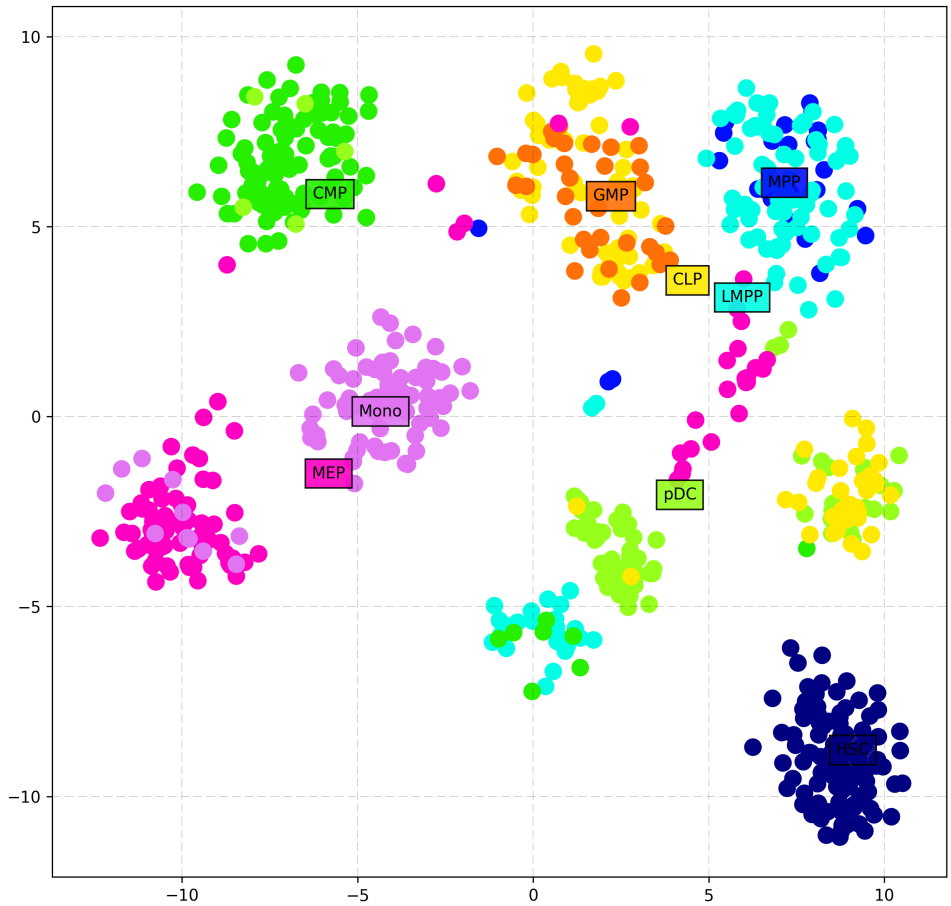

Figure 4.2 – t-SNE of the filtered single cell profiles. Each color represent a different cell type

4.2.2 Annotation

We have applied GABI on the filtered profiles. Among the $2^9 = 512$ possible states, only 121 states can account for 90% of the non null genomic sites Fig4.3. Consequently the analysis of hematopoeisis with this dataset is reduced to the study of these 121 combinations.

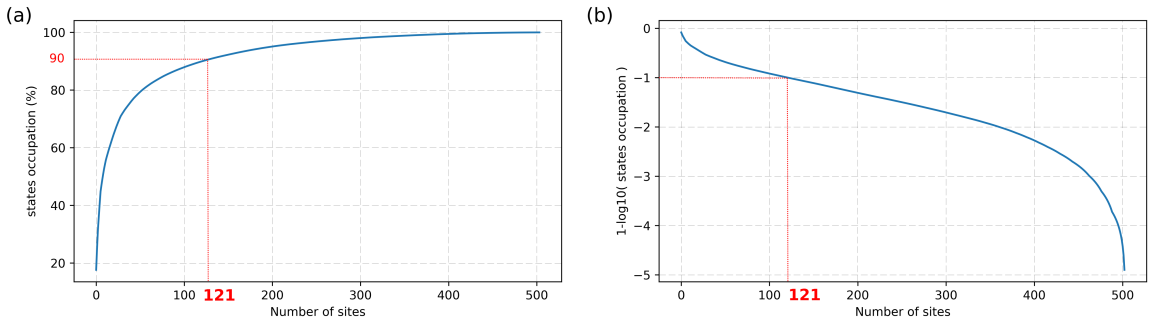

Figure 4.3 – Percentage of occupied genomic sites by summing the contribution of the most present states. (a)Percentage of occupied genomic sites (b) 1-log10 of this percentage

As explained in Chapter 3, an annotation tool gives the states and their position along the genome. Replacing the states at each genomic position gives the *inferred matrix*. To asses the quality of the annotation, besides the convergence of the model's likelihood, we represent the mean value of the genomic sites associated to each states Fig4.4. We observe that the annotation added more features than removed, considering that in this case the FN rate of the profiles is high ((FN, FP) $(0.85, 0.01)$). This is expected for single cell profiles because they have a low number of reads compared to profiles produced from a set of cells. We notice that the FN and FP rates of the profiles are not homogeneous, which shows potential experimental biases to further explore. Eventually, we represent a sample of the original matrix and the inferred matrix Fig4.5 and we observe that GABI succeeds in discovering redundant states in the original profiles matrix, despite a high FN rate. Based on the results on our synthetic datasets, we can assert that the AUC of this annotation is higher than 0.9.

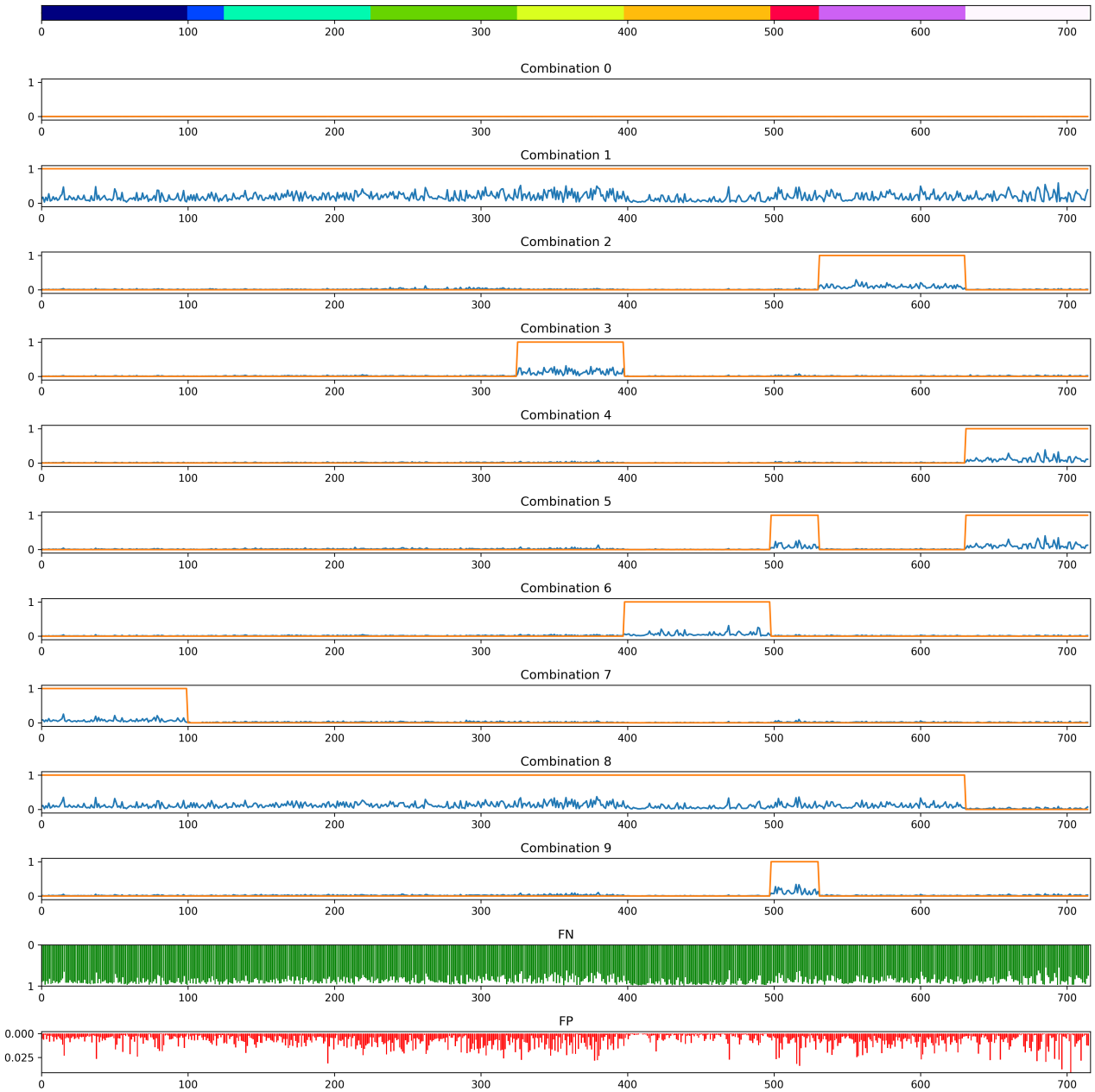

Figure 4.4 – Comparison of the reconstruction versus the original matrix. Plots represent the ten most present states ordered from top to bottom. For each plots, the blue curve corresponds to the mean of the genomic sites associated to one state, and the orange curve to the corresponding state inferred (the x axis represent the different profiles ordered by cell types). Below, the FN and FP rates for each profile are presented. These rates are GABI model parameters (see appendix *GABI calculation and algorithm description* for more details)

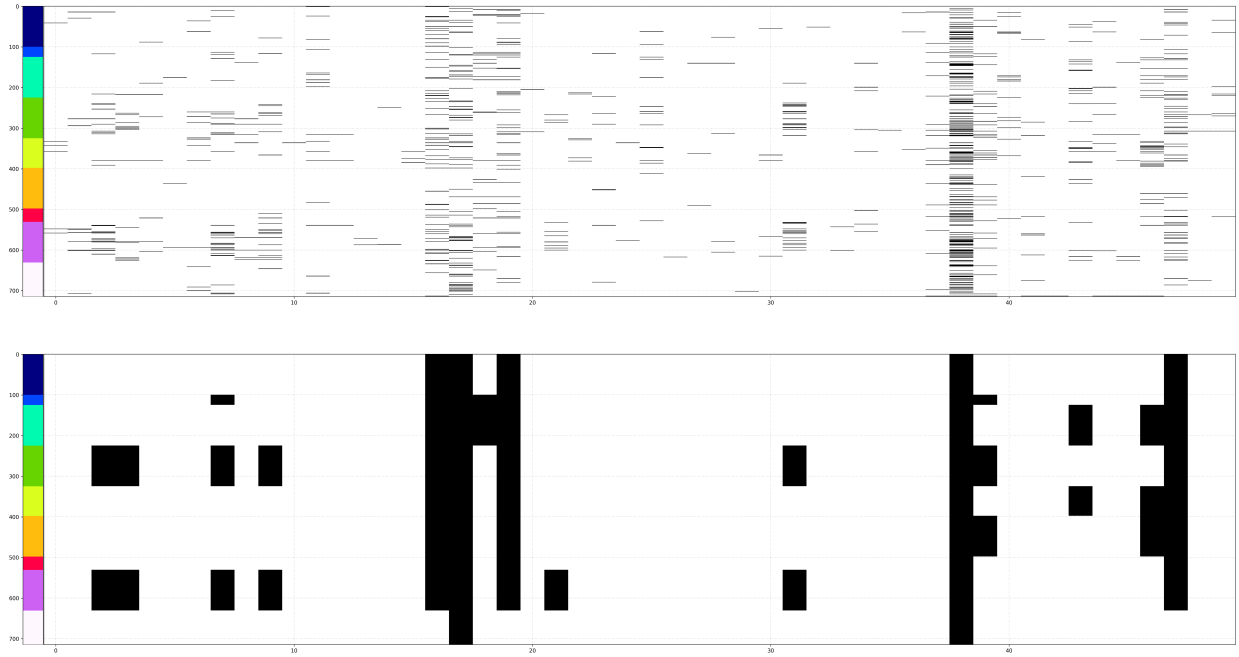

Figure 4.5 – Visualization of the inferred matrix compared to the original (top) Original profiles (bottom) Inferred profiles

4.3 Interpretation of the annotation

4.3.1 Interpretation of the states

Annotation provides a reliable and a simple information from a large and complex dataset by summarized in some combinations. The study of cell types combinations on the ATAC-seq dataset reveals informations about cellular differentiation of blood cells. The figure Fig4.6 represents nine of the most present states, mapped on the hematopoietic tree. Interestingly most of the non null combinations, *representing* 78% *of the genomic sites, are in agreement with the hematopoietic tree* (result with a p-value of 10^{-5} compared to a random states conserving the number of instance per state). These states could indicate potential developmental triggers, conserved or vanishing regulators during blood cells differentiation. To go further, the genomic sites associated to these states need to be compared with other genomic data to develop a model of the underlying mechanisms.

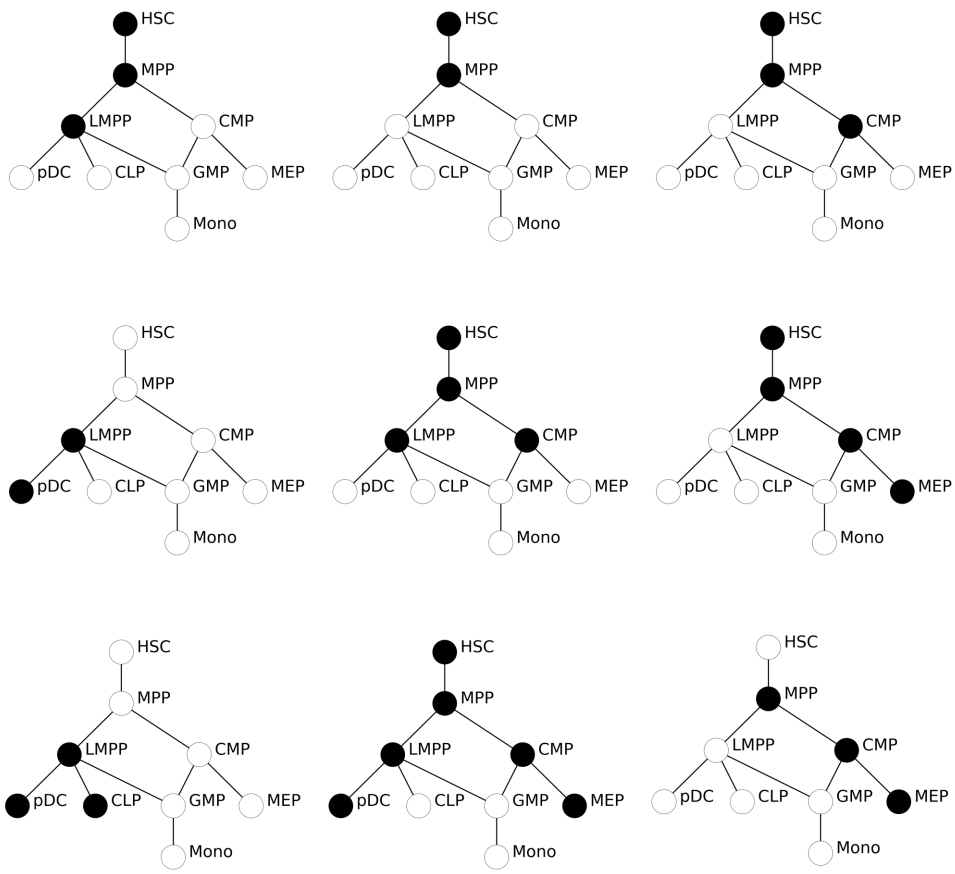

Figure 4.6 – Sample of combinations from the top most present states.

Some states, don' t follow the hematopoietic tree. For instance Fig4.7 shows three states

which shortcut MPP. This result is consistent with recent studies [2] [78] describing that MPP cellular state is can be avoided when HSC cells are set in specific conditions. The study of genomic site associated with these states might highlight some processes responsible of this effect.

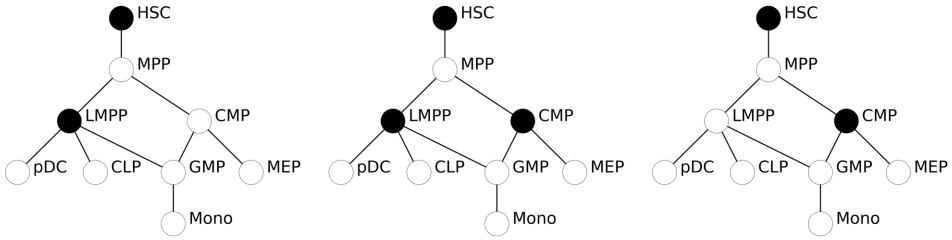

Figure 4.7 – Sample of combinations which shortcut the MPP cell type

4.3.2 Transcription factors analysis

We have compared the matrix inferred by GABI to one of the most used genomic feature to study cell regulation: the motif analysis of transcription factors binding sites. We have measured the enrichment of TFs in different cell types, using the motif scanner FIMO [29] applied on the HOCOMOCO v10 database [38], containing 601 TFs motifs. This enrichment is computed as a z-score to a null model, equivalent to the model presented in 2.4.1 (further details are provided in appendix *TF enrichment*). Then, we have mapped this score on the hematopoietic tree Fig4.8.

We recover results such that NANOG and SOX13 are involved in stem cells regulation [49] [77] and that PAX5 encodes the B-cell lineage specific activator protein. It is expressed at early, but not late stages of B-cell differentiation, corresponding to precursor of CLP [11]. PITX2 is known to have a role in hematopoiesis [33] and according to Fig4.8, its role is limited to the HSC-CMP-GMP-Mono transition.

We also observe that PAX5 and PITX2 are under-expressed in MPP whereas they are over-expressed in HSC and following cell types in the tree. This could indicate that these two TF are involved in the MPP shortcut.

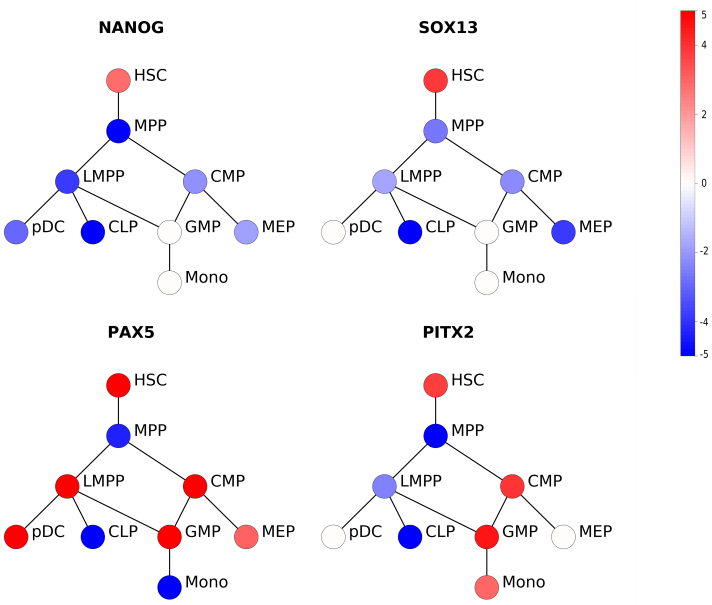

Figure 4.8 – Transcription factor enrichment in hematopoietic cell types computed using a z-score. Z-scores between -1.65 and 1.65 are set to zero (corresponding to a p-value of 10^{-2}). This figure is a sample of the HOCOMOCO TF database

4.3.3 Differential analysis

During cell development, regulators and genes are either kept or changed in order to change the cell identity. The figure Fig4.9 represents the proportion of sites kept during the development of the pDC,CLP,Mono and MEP. We observe that the number of accessible sites kept from HSC decreases with the maturity of cells, implying a differentiation mechanism based on locii closing. This trend is consistent with previous results on polled DNAse-seq on three stages of hematopoiesis [69] but we observe a higher conservation percentage on mature cells than this previous study.

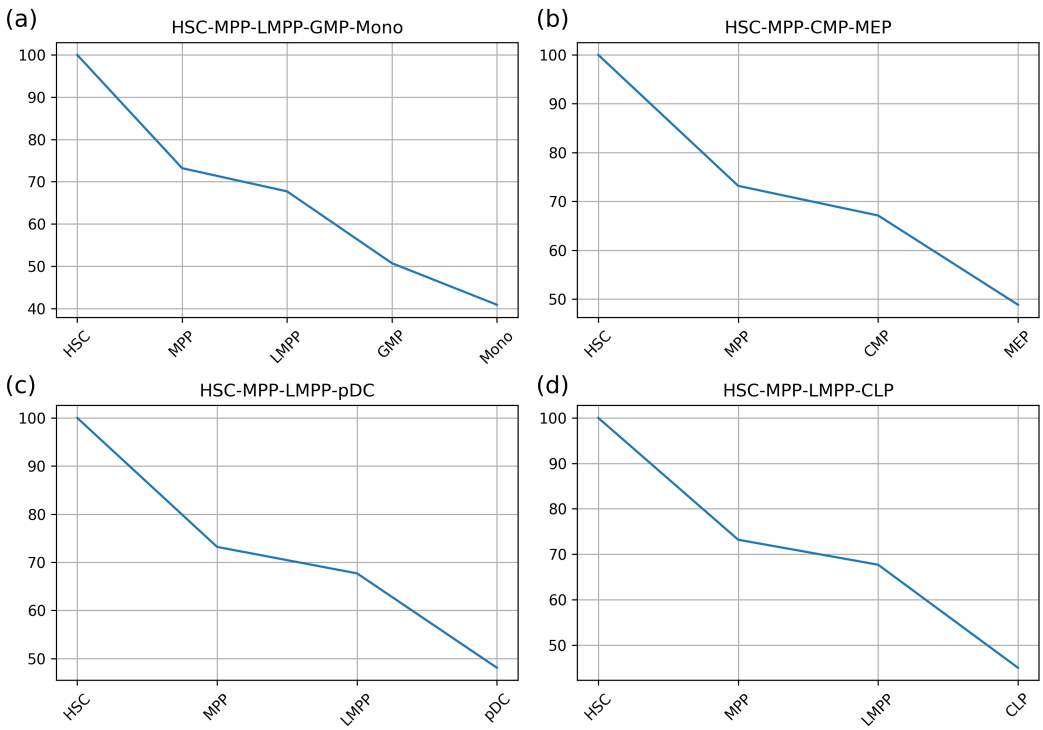

Figure 4.9 – Percentage of accessible sites kept from the HSC (a) HSC-MPP-LMPP-GMP-Mono transitions (b) HSC-MPP-CMP-MEP transitions (c) HSC-MPP-LMPP-pDC transitions (d) HSC-MPP-LMPP-CLP transitions

We have studied cell types transitions by representing in Fig4.10 the proportion of sites kept and changed for all pairs of cell types. We observe that the number of sites changed and kept during transitions is heterogeneous, implying transitions with more important modifications of the cellular landscape, eg HSC-MPP (25% of sites changed), than others, eg CMP-MEP (less that 10% of sites changes).

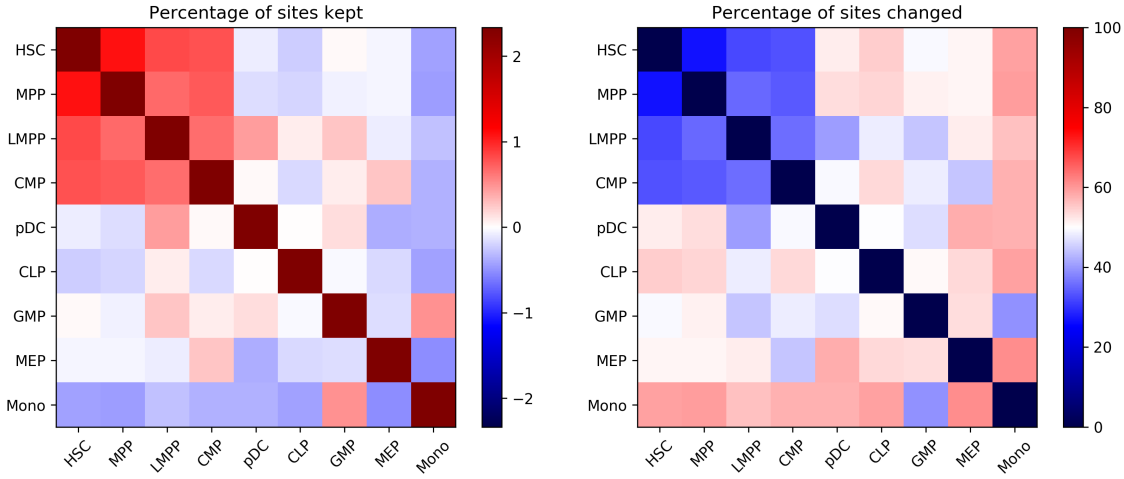

Figure 4.10 – Differential analysis between the profiles; (a) Percentage of genomic sites kept in the transition (b) Percentage of genomic sites changed in the transition. The summation of the two matrices equals 100

4.4 Conclusion

We have applied the entire pipeline describe in this thesis, from sample selection to GABI in order to produce a reliable annotation, on the hematopoiesis single cell ATAC-seq dataset. We highlight 121 combinations representing the entire dataset which potentially describe developmental triggers, conserved and vanishing regulators. Further studies are necessary to adapt these result to a therapeutic application but these results show the potential of our method to analyze large and noisy NGS datasets.

Chapter 5

Analysis of the 3D genome organization with network concepts

5.1 Introduction

As mentioned in the introduction, the gene regulation is encoded by CRE which can either act as activator, inhibitor or insulator. Activators and inhibitors respectively activate or repress the target genes. Insulators define gene domain boundaries. Enhancers and inhibitors are supposed to act only on genes within these domains.

At larger scales, the genome is separated in two compartments: the *euchromatine*, a region dense in active genes and active regulators, and *heterochromatine* a region dense in repressive biochemical marks with no active genes. Therefore, short and long range 3D structures are closely related to gene expression. In this chapter, we present a tool to reconstruct the 3D organization of the genome from chromosomal contacts, using tool from graph theory. This works as been published [50].

5.1.1 Experimental data

The Chromosome conformation technique *HiC* is a method that identify nearby genomic loci in the 3D space[1] [22, 44, 58]. Close genomic sites are aggregated forming a contact. All contacts between regions are reported in a matrix, namely the HiC matrix Fig5.1. In contrast with the single cell , this experiment is performed on an ensemble of cells, averaging the chromosomal contacts.

[1]Which may be separated by many nucleotides in the linear genome.

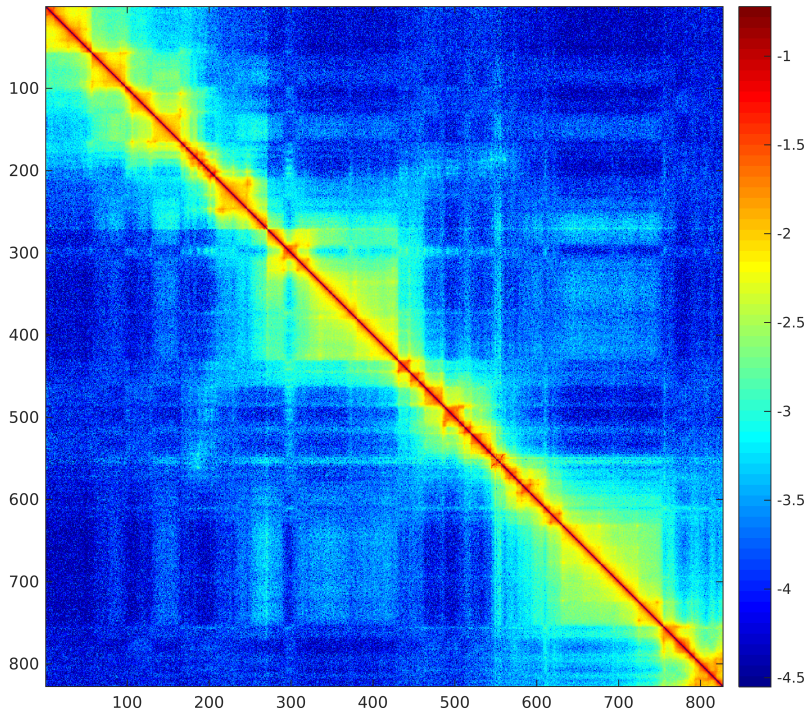

Figure 5.1 – HiC matrix sample representing a 10Mb-fragment of human chromosome 1 at a resolution of 10kb from[58]. The color code represents the contact frequency ($-\log_{10}$ units).

In this work, we consider this contact matrix as the adjacency matrix of an undirected network, called *contact network*. This approach allows to use tools from graph theory in order to compute the distance between genomic loci as a distance on a graph.

In the first part of this chapter, we will study the use of the contact network representation to compute the graph distance between any pair of genomic sites, including those displaying no (or very few) contact(s). The contact network representation has been exploited in [43] to derive a fast reconstruction algorithm, named ShRec3D for Shortest-path 3D Reconstruction. This reconstruction algorithm uses graph distance to impute missing distances between pairs of genomic regions before using a MDS for reconstructing the 3D genome structure.

We propose in the second part an extension of this reconstruction algorithm, involving a tunable graph distance and two dimensionality reduction (DR) algorithms. According to experiments using fluorescence in-situ hybridization (FISH) data, which evidenced a power-law correlation between contact frequencies and measured distances [44], we explore the relationships between the contact frequencies, the graph distances, and the distances within the reconstructed 3D structures. We study the transformations achieved by the different steps of the algorithm and benchmark its possible variants. As a result, we identify two parameters which allows to tune the final reconstruction according the user needs.

5.2 3D Reconstruction algorithm

5.2.1 Contact network normalization

There are basically two types of networks: the networks with multiple edges and the networks with weighted edges. An illustrative example of the former is the binary adjacency matrix defined by a threshold on the edges values(see example Fig5.2). In this work, we used the weighted edges' network because it uses the number of reads as a feature, under the condition that the weights are correctly normalized in order to remove biological biases. The SCN [19] iteratively divide the columns and the lines of the contact map by their sum. The result is that the probability for a given genomic region i to have a contact is the same for all region j. Formally, if we consider C_{ij} a contact at the coordinates (i,j), $P(I = i|J = j)$ the probability of having a contact between the line i and column j given the column j, and N the number of genomic regions:

$$SCN(C_{ij}) = P(I = i|J = j) = P(J = j|I = i)$$

with:

$$\sum_{i=1}^{N} P(I = i|J = j) = 1, \forall (i, j) \in [1, N]^2$$

which is also equivalent to:

$$P(I = i) = P(J = j), \forall (i, j) \in [1, N]^2$$

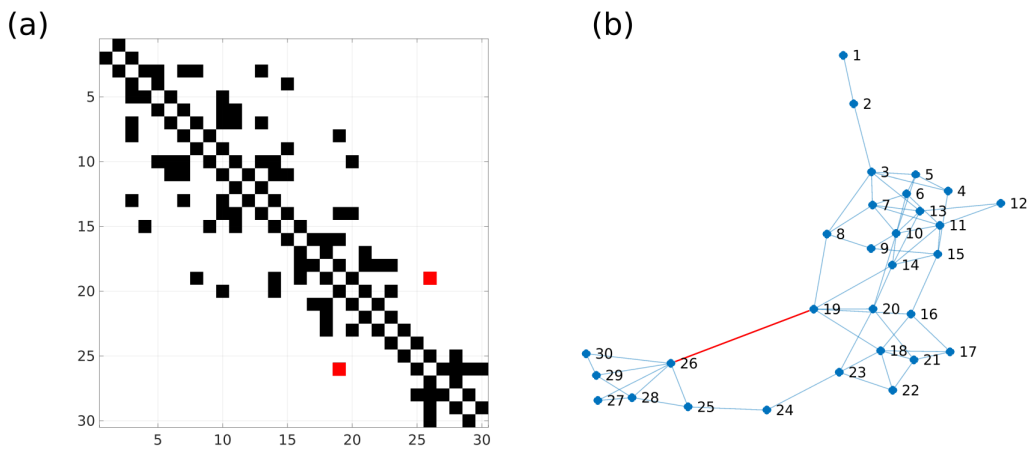

Figure 5.2 – A simple example of a (a) binary contact map and (b) its associated network. The number on the nodes refer to their linear ordering in the contact map

5.2.2 The challenges

The first issue is the large size of genomic contact maps, which requires fast reconstruction algorithms. Methods for the reconstruction of the native structure of the proteins from its contact map, e.g. by targeted growth [73], are limited to a few hundred of elements at the very most,

hence do not apply to the large HiC contact maps. The same issue apply to standard reconstruction methods which are based on iterative optimization of the reconstructed structure over the observed datas [63].

The second issue lies in the fact that part of all contact are detected, therefore the reconstruction algorithm must first infer the missing data corresponding to these non reported contacts.

5.2.3 ShRec3D

To tackle these issues, we used Shortest path Reconstruction 3D (ShRec3D) algorithm [43]. The different steps of the algorithm are Fig5.3(a):

1. **Convert contact matrix to distance matrix.** The standard method to compute a distance from conformational capture data is to consider the inverse of the contact frequency. Let F_{ij} be the contact frequency between the DNA regions (i,j), then the associated distance L_{ij} is defined as $L_{ij} = 1/F_{ij}$ [28].

2. **Complete missing data with shortest path.** Considering the distance matrix as a network, we can compute the shortest path distance D between all paires of genomic regions Fig5.3(b). The Floyd–Warshall algorithm [74] is commonly used for this purpose. It consists in computing for each triplet (i,j,k) if the distance between (i,j) is shorter than the distance between (i,k) and (k,j), $D_{ij} = min_k(D_{ij}, D_{ik} + D_{kj})$, with D initialized as equal to L.

3. **Convert the distance matrix into coordinates with the MDS** The MDS is a technique developed to reduce the dimension of the coordinates of points which conserves as much as possible the distance between the points [72] (see appendix MDS). There exists two ways to optimize the MDS, either iteratively with gradient descend, either algebraically, which is simple to implement and fast to compute. The spectrum of the Gram matrix reflects up to what point the matrix D is close to the Euclidean distance matrix of a single 3D structure, by analyzing the gap between the highest three eigenvalues and the rest of the spectrum Fig5.3(c-d). In the case of a real 3D structure, only the first three eigenvalues are non zeros.

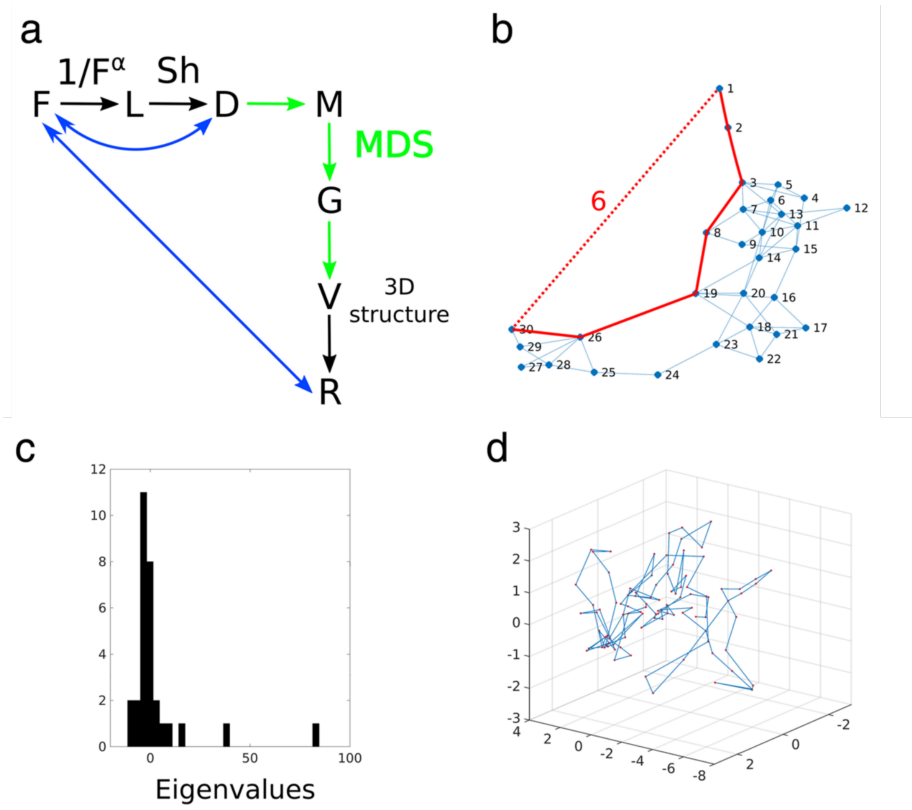

Figure 5.3 – Principle of ShRec3D reconstruction algorithm (a) Summary the notation used in the papers and the pipeline of ShRec3D. The green arrows represent the MDS steps and the blue arrows the comparisons between the distance matrices that we have investigated to assess the impact of the factor α on the different steps of the algorithm (b) An example of contact network converted in distance network and application of the shortest path. (c) The Gram spectrum from this distance matrix.(d) The corresponding 3D reconstruction.

5.2.4 An improvement of Shrec3D based on fluorescence in-situ hybridization (FISH) experiments

FISH protocol associates fluorescent tags to a few specific genomic sites. It allows in a population of fixed cells the accurate measurement of the spatial distances between these sites. Compared with the HiC, we can link the number of contact and the in-vivo distance between genomic sites. On the contrary of HiC, the FISH protocols tags only few samples but it provides a totally independent constraint on the 3D reconstruction from the HiC maps. A negative correlation has been observed for the sites tagged by FISH between their distance d_{ij} (average over numerous single cells) and the number C_{ij} of Hi-C reads, or equivalently the contact frequency F_{ij} [44] Fig5.4. A simple fit of this result is through a power-law such that $L_{ij} \sim F_{ij}^{-\alpha_{FISH}}$ with $\alpha_{FISH} \approx 0.227$. We add this extension to the ShRec3D algorithm and investigate the influence of the value of α on the properties of the shortest-path distance matrix D and the matrix reconstructed from the 3D coordinates R with two cases $\alpha = 0.2$ (the rounded value of the exponent observed experimentally in the above-described situation) and $\alpha = 1$ (the value adopted in the original algorithm). Moreover, we compared the MDS used previously, also called *classical MDS* (cMDS)

with a MDS variant, the Sammon mapping (see appendix). By construction of its cost function, Sammon mapping is more precise for the reconstruct of the small distances, as opposed to the classical MDS which is more precise for the large distance. In the following, we will first study the effect of the the shortest path and the dimensionality reduction on the value of α, before exploring the relationship between the reconstruction error and α. This studies are performed on Hi-C data from human cells in lymphoblastoid (GM12878) as in [44] but at higher resolution [58].

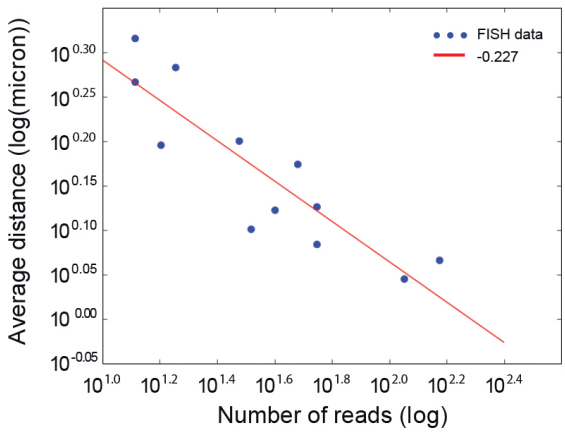

Figure 5.4 – Contacts recorded in FISH experiment in log-log scale and the power-law fit

5.2.5 The effect of the shortest path

By definition, the shortest-path distance D_{ij} is always smaller or equal to the edge length L_{ij}, this is why we can see on Fig5.5(b) that the slope α_{sh} of the fit of $logD_{ij} = -\alpha_{sh} \times logF_{ij}$ is smaller or equal than α given by $L_{ij} = 1/F_{ij}^{\alpha}$. The dependence of α_{sh} as a function of α is shown on Fig.5.5(d) and a crossover is observed at a value $\alpha \approx 0.2$. It is expected that D does not rely on low contact frequencies, associated with long edges in the contact network. Fig5.5(b) shows that the difference between D and L is indeed more marked for smaller contact frequencies, i.e. larger distances. However, when α decrease, the distances tends to uniformize and small contact frequency are less and less changed by the shortest path. This is illustrated Fig5.5(c) by the evolution with α of the proportion of rewirering, i.e. the proportion of edges $L_{ij} > D_{ij}$. The choice of α impacts the use of low contacts for the final reconstruction. For high values of α, only high contact value, thus contact near the diagonal will contribute to the reconstruction. The other contacts will be inferred based on the shortest path distance of high contact values. On the other hand, for small α, low contact values will contribute to the reconstruction which produce a 3D structure with more information but also with more noise.

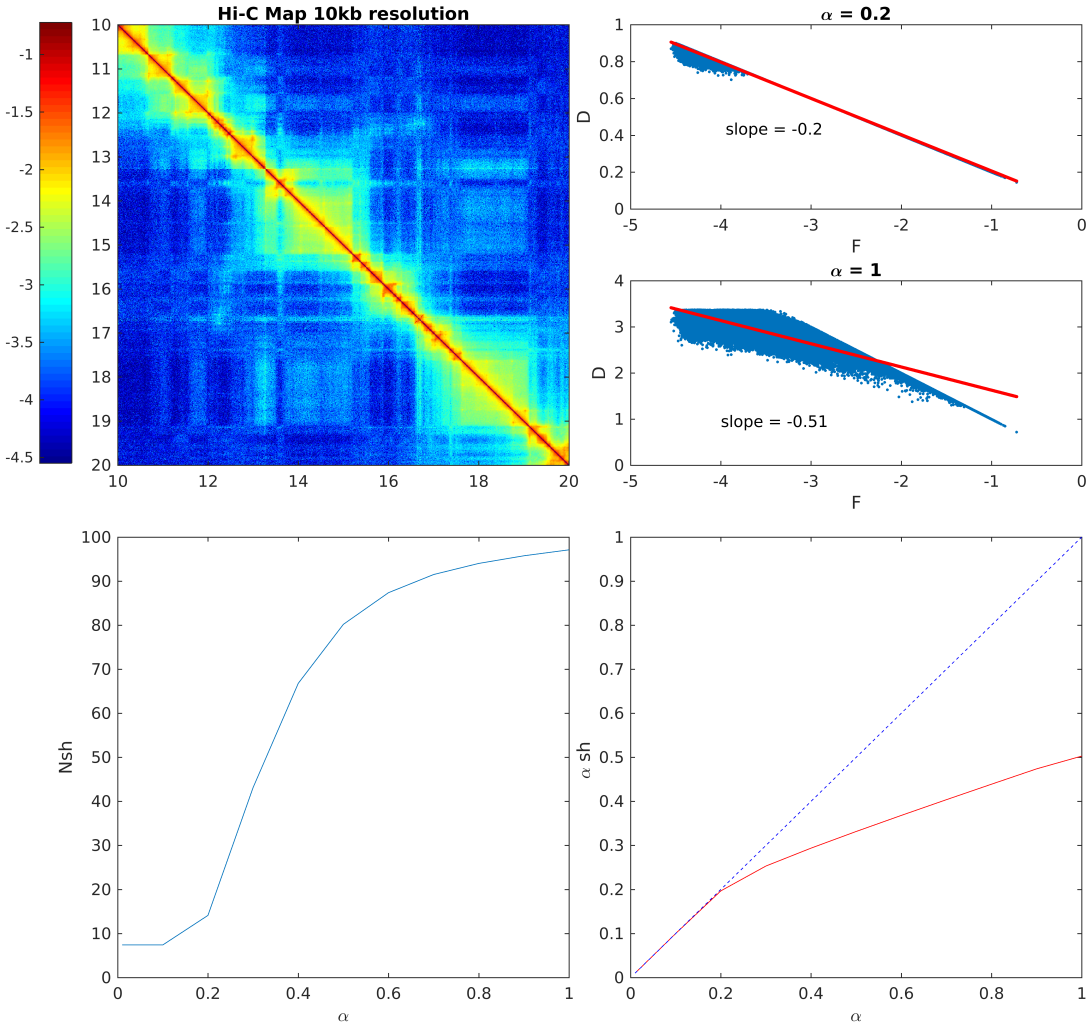

Figure 5.5 – Analysis of the weighted graph distance. (**a**) Hi-C contact map (**b**) Log-log scatter plot of the shortest-path distances D_{ij} with respect to the contact frequencies F_{ij}, for two values $\alpha = 0.2$ (top) and $\alpha = 1$ (bottom) of the exponent α involved in the prescription of the edge length. The upper boundary of the cloud of points is a line of slope $-\alpha$, corresponding to the pairs of sites for which the direct edge (i, j) of length L_{ij} is the shortest path. Minus the slope of the red line gives the exponent α_{Sh} of the best power-law fit $D_{ij} \sim F_{ij}^{-\alpha_{Sh}}$. (**c**) Increase of the percentage N_{Sh} of pairs of sites for which the direct connection (i, j) is not the shortest path, when α increases. (**d**) Exponent α_{Sh} as a function of α; the dashed blue line indicates the diagonal $\alpha_{Sh} = \alpha$

5.2.6 The effect of the dimensionality reduction

We further explored the relationship between the reconstructed distances R and the contact frequencies F (long blue arrow in Fig.5.3(a)) as a function of α. We moreover compared two DR techniques, corresponding to different optimization criteria hence different approximations. Classical MDS correspond to the minimization of $\sum_{i,j}(D_{ij} - R_{ij})^2$. The strength of this method is the fast calculation of the 3D structure by computing the three first eigenvectors of the metric matrix M (as explained above). Its weakness is the low constraint on small distances, since minimizing the error is achieved mainly by controlling the large distances. This dominance of large distances can be corrected by considering the relative error $\sum_{i,j}(D_{ij} - R_{ij})^2/D_{ij}$ [83], leading to the so-called *Sammon mapping* (see Appendix). Importantly, DR implementations essentially requires a complete distance matrix. Otherwise when applied to L the reconstruction is highly unstable, due to the treatment of infinite or abnormal components of L. As shown in Fig5.6, we observe a correlation between the reconstructed distances R and the contact frequencies F, which can be summarized by a powerlaw with exponent α^*, depending on the value of α and the cMDS implementation. Note that we do not claim that these power-laws have a deep meaning, reflecting e.g. some self-similar or fractal structure of the chromosomes; the range of the fit is not large enough to make such a claim. These power-laws are used as the simplest way to quantitatively describe the correlation between F and distances matrices L, D and R. The comparison of the exponent α^* with α_{Sh} (Fig5.6(c)) and α (Fig5.6(d)) provides a global quantification of the effect on the distances of the dimensionality reduction step and the integrated algorithm, respectively. A local quantification will be implemented in the next section.

The value of α initially taken in the expression of edge lengths L is not recovered in the relationship between the reconstructed distance and the contact frequencies, with exponent α^*. Part of the difference between the two exponents comes from the shortest-path computation, Fig5.6 (d), and part from the MDS dimensional reduction, Fig5.6 (c). This latter figure shows that Sammon mapping has a smaller impact on the exponent α^* than classical MDS. Using Fig5.6d, it is possible to choose a value of α to get the desired correlation behavior in the reconstructed structure, with some limitations. Noticeably, the effect of MDS on α^* is weaker at larger α and the value $\alpha_{FISH} = 0.227$ is observed at the lower boundary of the accessible range for α^*.

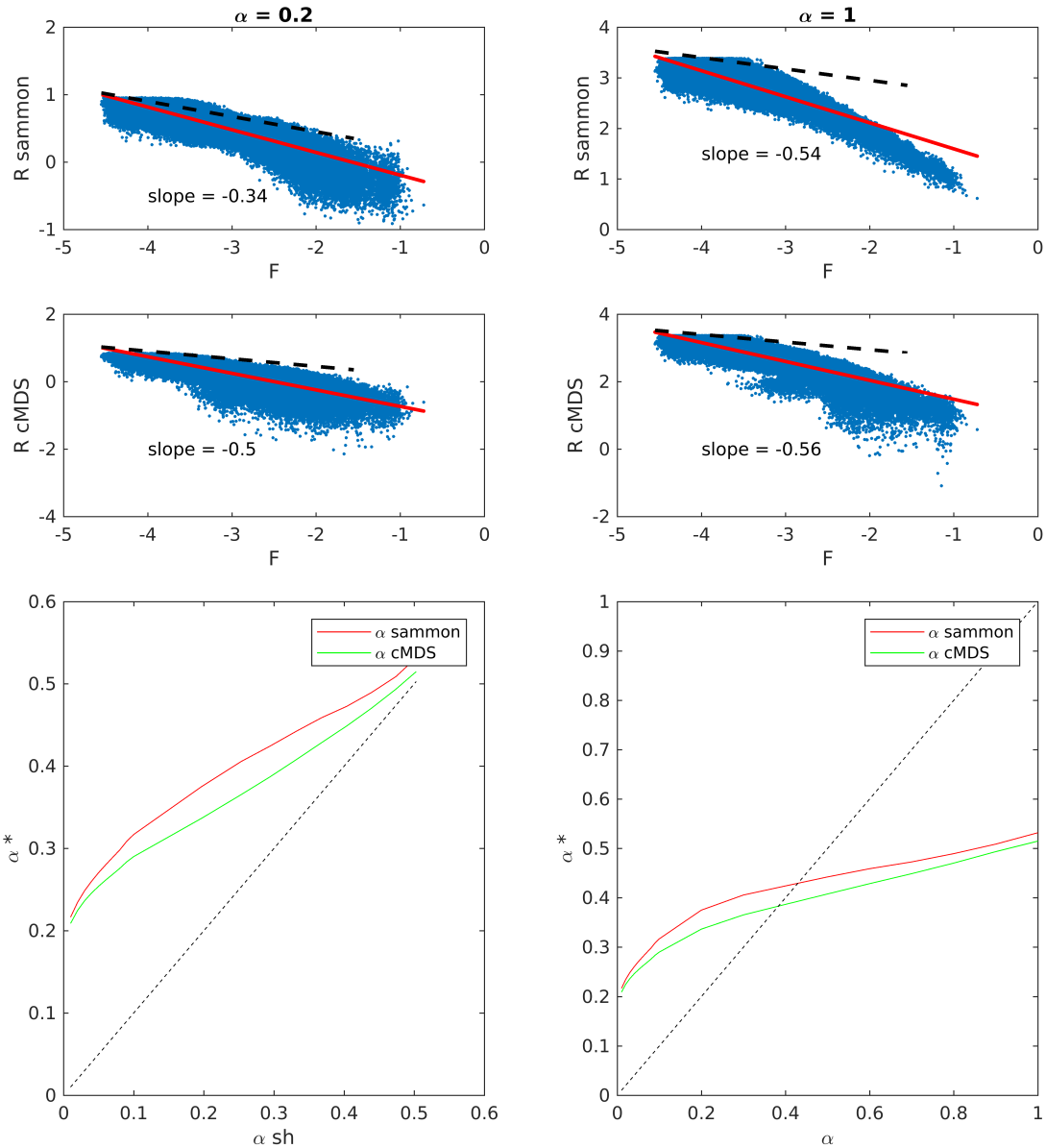

Figure 5.6 – Joint influence of the exponent α and DR implementation. (**a, b**) Log-log scatter plot of the reconstructed distances R with respect to the contact frequencies F for two values $\alpha = 0.2$ (a) and $\alpha = 1$ (b) of the edge-length exponent α, and two RD implementations: Sammon mapping (Sammon, top) and classical MDS (cMDS, bottom). Minus the slope of the red line gives the exponent α^* of the best power-law fit $R_{ij} \sim F_{ij}^{-\alpha^*}$. As a guide for the eyes, the dashed black lines, with the same starting point as the red lines, represent the line with slope $-0.227 = -\alpha_{FISH}$. (**c**) Exponent α^* as a function of α_{Sh} for cMDS (green line) and Sammon (red line); the dashed blue diagonal corresponds to $\alpha^* = \alpha_{Sh}$. (**d**) Exponent α^* as a function of α for cMDS (green line) and Sammon (red line); the dashed blue diagonal corresponds to $\alpha^* = \alpha$

5.2.7 Reconstruction Error

We computed the component-wise relative error $|D_{ij} - R_{ij}| / D_{ij}$ Fig5.7 (a-b) and its mean over small ranges Fig5.8 in order to analyze quantitatively the action of the DR algorithms on the different scales. We observe as expected, and it is more significant for $\alpha = 1$, that the small distances are better reconstructed for the Sammon mapping and the larger distance are better reconstructed with the classical MDS.

It is also apparent that the Sammon mapping reproduces small structures (e.g. small loops) while global shape is more clearly represented by the cMDS Fig5.7 (d). For small values of α Fig5.7 (c) the structure is more fuzzy and compact, because more mid-range contact were taken into account before the shortest path, representing closer the results of imaging experiments. For larger values of α Fig5.7 (d), the reconstructed 3D structure is more extended, representing the skeleton of the overall shape, which is specially suitable for 3D genome browsers.

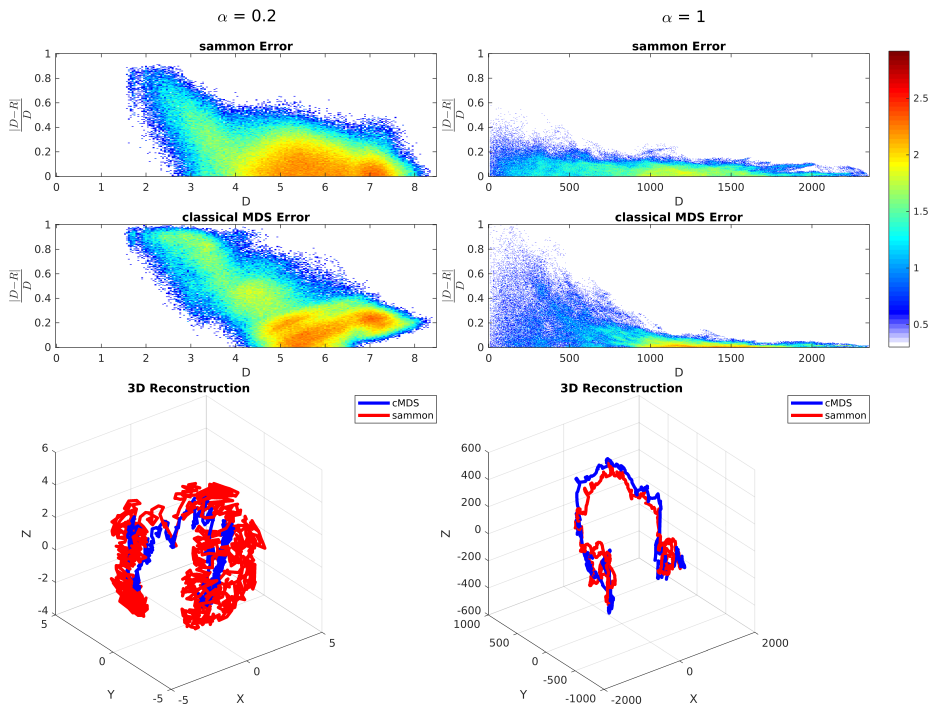

Figure 5.7 – Comparison of classical MDS and Sammon mapping reconstruction error. (**a, b**) Action of the MDS step at various scales of the 3D structure, analyzed quantitatively by computing for each pair of sites (i, j) the relative difference $|D_{ij} - R_{ij}| / D_{ij}$ between the shortest-path distances D_{ij} and the reconstructed distances R_{ij}. This relative difference is represented component-wise as a scatter plot with respect to the distances D_{ij} for two values $\alpha = 0.2$ and $\alpha = 1$ of the edge-length exponent α, for both Sammon mapping (Sammon, top) and classical MDS (cMDS, bottom). The color scale is related to the density of points in the scatter plot (increasing density from blue to red). (**c, d**) 3D structures obtained for $\alpha = 0.2$ and $\alpha = 1$ with classical MDS (blue) and Sammon mapping (red). A comparison between panels c and d would require a suitable 3D alignment, see Fig5.9 below

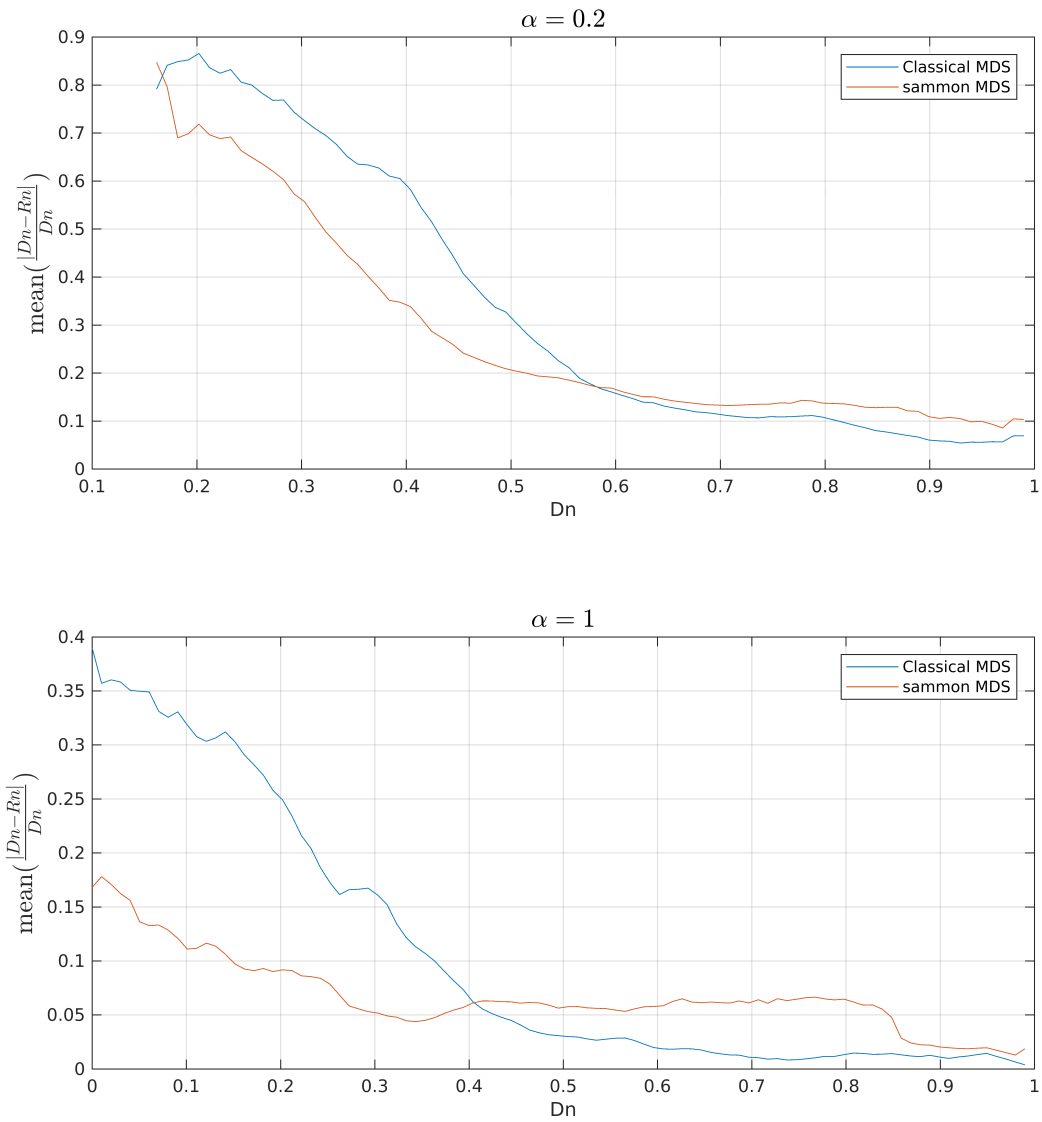

Figure 5.8 – Relative error according to the distance range. The figure offers an alternative representation of Fig5.7. It displays the average of the relative error $|D_{ij} - R_{ij}|/D_{ij}$ originating from the DR step as a function of the normalized distance range D_n, i.e. the distance divided by the maximal component of the matrix D for the cMDS and Sammon mapping. The average of the relative error is taken over pairs (i, j) corresponding to the same value of the normalized distance D_n. Classical MDS performs better at large scale, whereas Sammon mapping reproduce better the features associated with small distances. This difference is more marked for $\alpha = 1$ (bottom) than for $\alpha = 0.2$ (top)

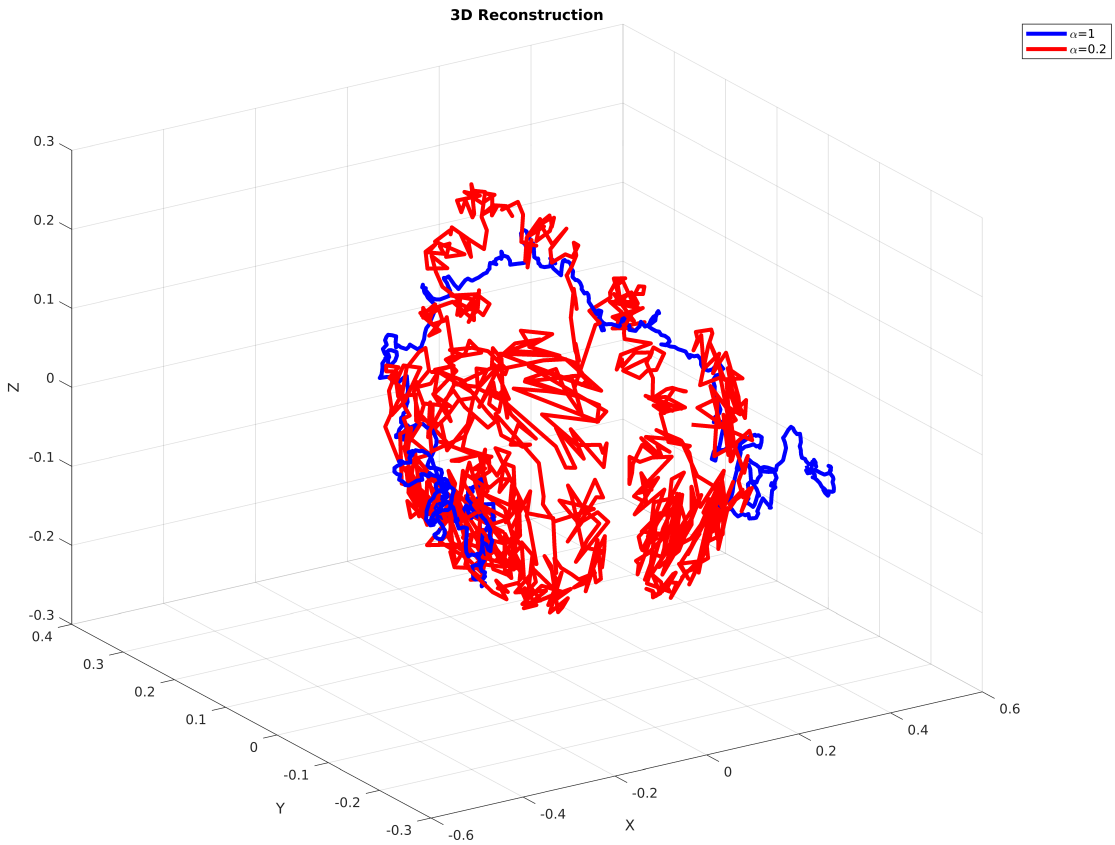

Figure 5.9 – 3D reconstruction with Sammon mapping and different values of the exponent α. Same as Fig5.7 (fragment of 10 Mb of human chromosome 1). While local features are best reconstructed with $\alpha = 0.2$ (blue) and overall shape with $\alpha = 1$ (red)

5.3 Conclusion

Reconstruction algorithm allows to get access to an unobservable data, the 3D structure, from observable contacts between the chromosomes. The most important steps in these algorithms are the imputation of the missing data and the 3D representation. In this project we explore different aspects of these steps with ShRec3D as a starting point. The extension presented here, allows to select the *scale of interest* (Global:cMDS,Local:Sammon) and the *level of information* from the contact matrix that will be used in the reconstruction (α) Eventually, we demonstrate the capacity of our extension by applying ShRec3D to the first human chromosom at a resolution of 10kb. We have used $\alpha = 1$ in order to lower the level of details and noise due to the large size of the reconstruction and used Sammon mapping in order to focus on short range distances.

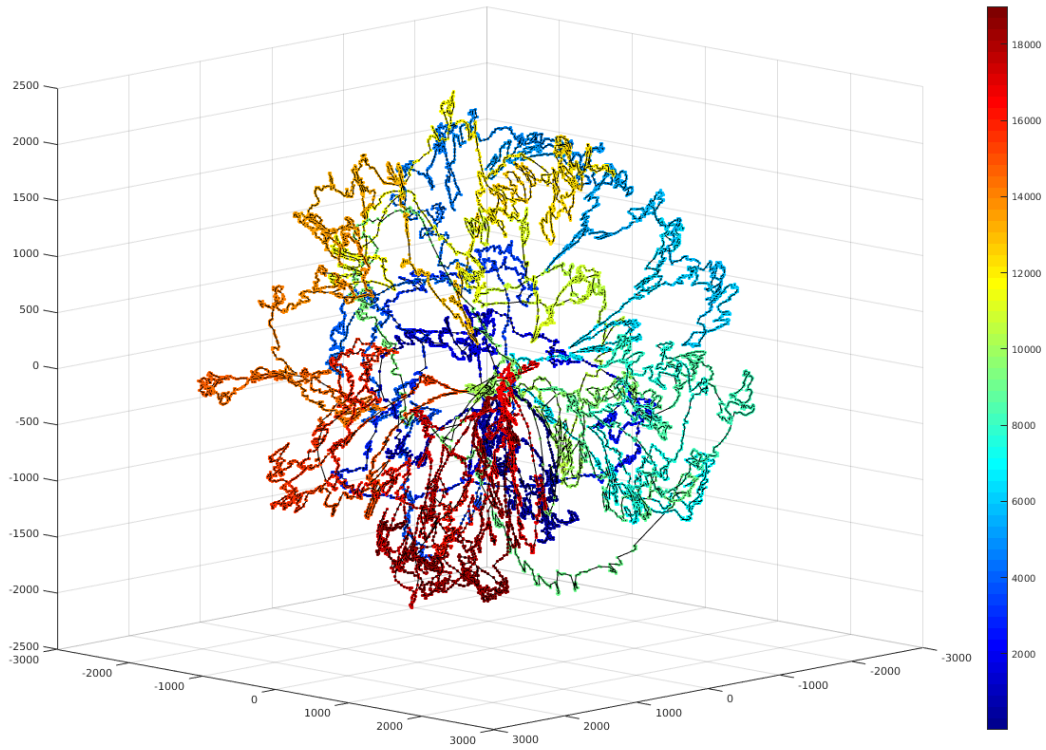

Figure 5.10 – 3D structure of the chromosome 1 at a resolution of 10kb (18,000 fragments), with Sammon mapping and $\alpha = 1$. The sites are colored according to their linear position along the genome in order to keep track of the 1D connectivity

Chapter 6

Appendix

6.1 Histone modifications and references

Many studies has studied the relationship between the histones modifications and their associated biological functions Table 6.1. Recently, the Roadmap consortium [40] provided an automatic annotation of the genome based on the redundant combination of histones marks in 127 cell types. Based on the annotation provided by ChromHMM, we reconstructed the histone marks combinations Fig6.1, which complete the table Table 6.1.

Functional Annotation	**Histone Marks**	**References**
Promoter	H3K4me3	[7],[34] [56]
Bivalent/Poised Promoter	H3K4me3/H3K27me3	[7]
Transcribed Gene Body	H3K36me3	[6]
Enhancer (Active & Repressed)	H3K4me1	[30]
Poised Developmental Enhancer	H3K4me1/H3K27me3	[20][57]
Active Enhancer	H3K4me1/H3K27ac	[20],[57],[30]
Polycomb Repressed	H3K27me3	[8],[42]
Heterochromatine	H3K9me3	[48]

Table 6.1 – Histones marks and their related functions

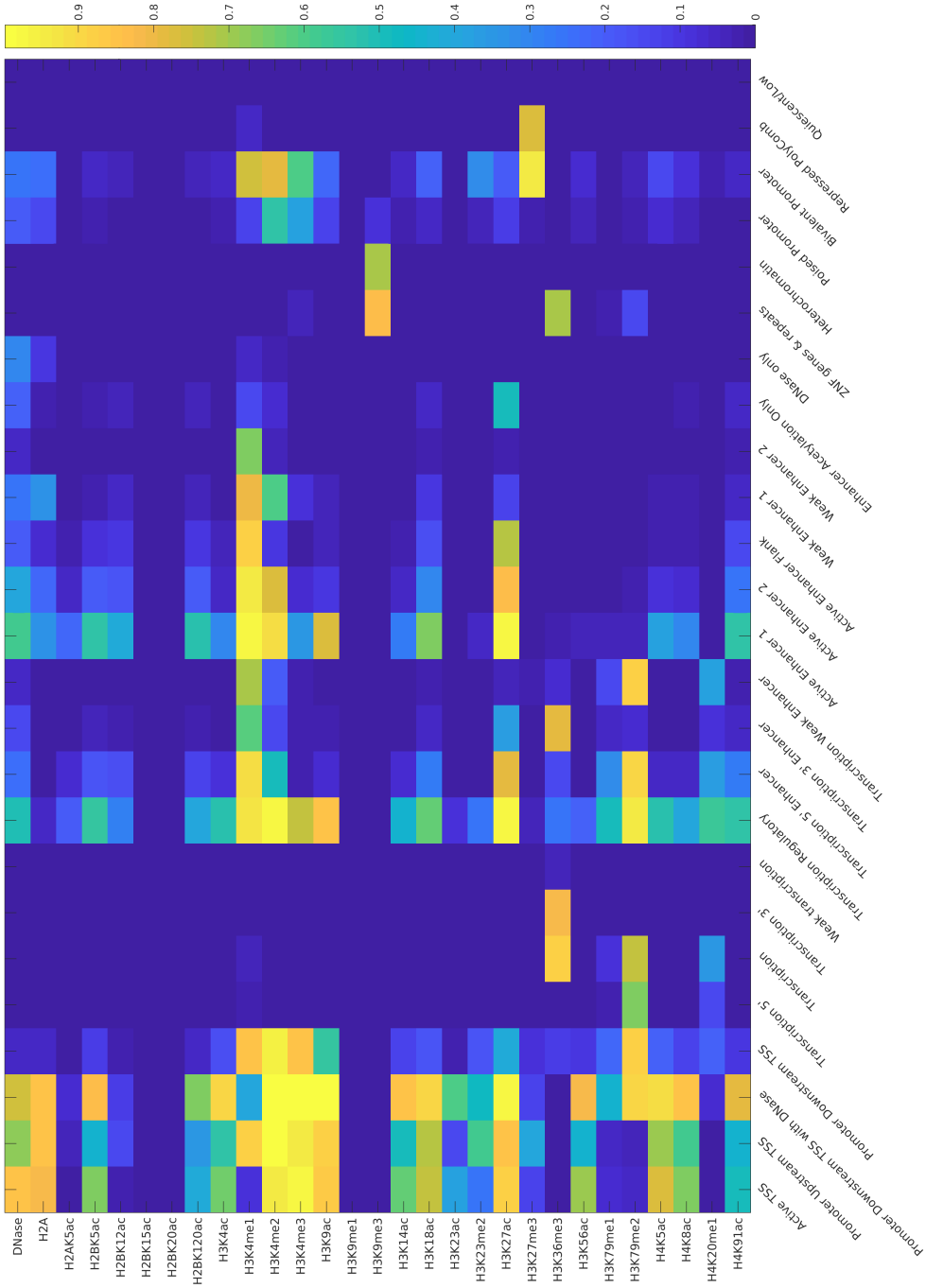

Figure 6.1 – The annotation has been realized with ChromHMM [26] on 29 histone marks in 127 cell types. The figure represents, for all combinations, the mean over all their associated genomic sites

6.2 Peak caller

NGS profiles are expected to show peaks. These peaks are defined as regions with an enrichment in reads compared to the background and represent the experimental genomic feature. Peak callers aims at detecting these peaks and therefore produce a binary signal with non zero value at peaks.

Dozens of different peak caller exist [36] and one of the best performer and commonly used is MACS [82]. The principle this algorithm is to fit the number of reads with a *Poisson law*, representing a null model of reads randomly distributed over the genome. Peaks in the profiles are DNA regions with a number of reads above a threshold, defined by a p-value, of a Poisson law[1]. In order to take into account the evolution of the background in NGS profiles [47] MACS first defines the potential peaks by fitting the entire genome before confirming their existence by fitting the read in small windows around the peaks.

6.2.1 ZINB peak caller

However, MACS algorithm has a pitfall: it needs a specific format as input which is available only after the alignment and unusually given. Aligning the reads can be tricky due to potential sequencing biases known by the experimentalist. Therefore, we developed our own peak caller which could be used on segmented NGS profiles vectors (see Fig1.10). We used the same principle than MACS but we improved the probability distribution, since it has been shown [21] that the zero inflated binomial negative distribution (ZINB) fit the NGS noise better than the Poisson distribution. This result has been obtained by comparing the likelihood of the Poisson and the ZINB law on a NGS control experiment Fig6.2. These control experiments are profiles generated with the same protocol but without the enzyme used to detect the genomic feature of interest. Therefore, these profiles represent the noise background of a NGS experiment.

We compare the results of our peak caller with MACS on a DNAse sample Fig 6.3, taking the same p-value of 10^{-5}. The figure Fig6.3 shows that 80% of the ZINB peak caller peaks are shared with MACS where most differences occurs at the peaks edges. Since Poisson law has a smaller variance than ZINB, it is normal to observe that MACS find more peaks for the same p-value. Therefore, our peak caller results are similar to MACS results.

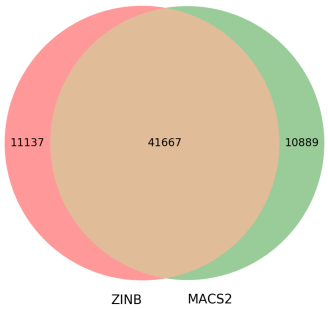

Figure 6.3 – Venn Diagram comparing the number of peaks similar and different. The fact that MACS has more peaks is coherent with the fact that its distribution is underestimating the threshold value compared to ZINB distribution

[1]MACS uses a p-value of 10^{-5}

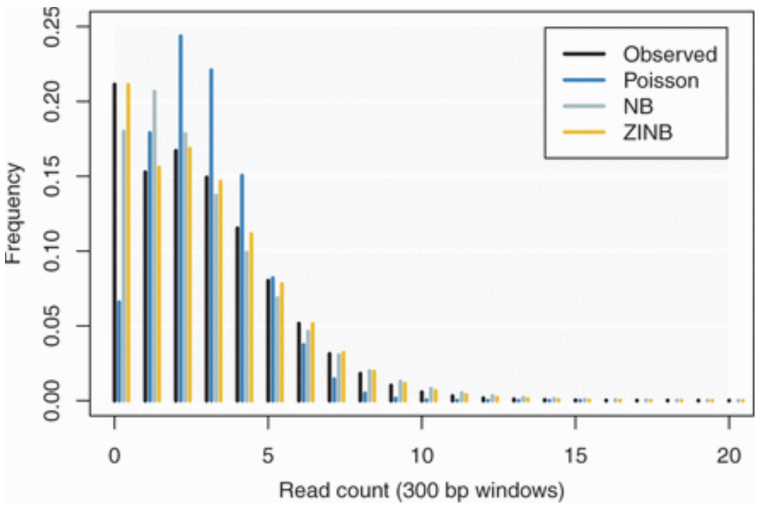

Figure 6.2 – Reads from a mock control dataset. The histogram of the read counts is shown in black. The colored histograms show the maximum likelihood fit of the Poisson, Negative Binomial (NB) and Zero-Inflated Negative Binomial (ZINB) distributions. The fit of the Poisson distribution is poor. The NB distribution gives a good fit at the tail, but not for windows with 0 and 1 read. The ZINB distribution gives a good fit over the whole range. Data from ENCODE file ENCFF000VEK (Figure from [21])

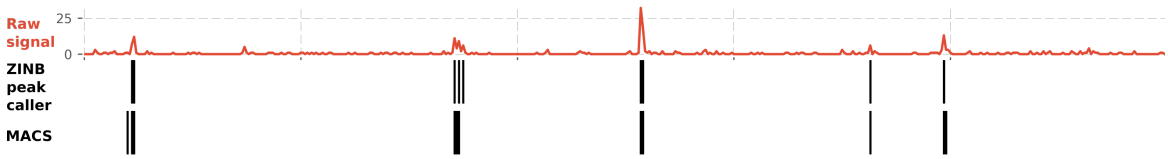

Figure 6.4 – Difference between MACS and ZINB peak caller on raw signal

For convenience, because MACS is not applicable on our formatted datasets, we used our ZINB peak caller for the binarization step.

6.2.2 Zero Inflated Negative Binomial (ZINB) distribution

The zero inflated negative binomial distribution extend the definition of the negative binomial to take into account unmappable regions. The probability to have y_i reads at a locus $i \in [1, N]$ is defined as

$$P(Y = y_i) = (\pi + (1-\pi)p_0^{\alpha})^{1-y_i} \text{ if } y_i = 0$$
$$P(Y = y_i) = (1-\pi)\frac{\Gamma(\alpha + y_i)}{\Gamma(\alpha)y_i!}p_0^{\alpha}p_1^{y_i} \text{ otherwise}$$

with π the zero-inflation or mixture parameter (proportional to the number of zeros in the profile), α the parameter dictating the distribution of the reads and p_0, p_1 the probabilities indicating the average number of reads in each windows k, linked by the relation $p_0 + p_1 = 1$

6.2.3 Algorithm description

The different steps of our peak caller are:

1. Compute the histogram of the entire profile, remove the 1% highest peaks[2] and normalize it in order to get a probability.

2. Fit the Mean Square Error (MSE) between the data distribution and the ZINB by using the Nelder-Mead method implementation of the Scipy library [31]. This methods uses the Simplex algorithm [52] and is robust for optimization for which the derivatives are not defined.

3. Repeat *Step 2* multiple times with different parameters initialization in order to get the optimal fit of the entire signal distribution.

4. Split the profile in windows of size WS and fit the data distribution in each using the entire signal distribution parameters as initialization.

5. Each fit gives the p-value of the data in its window. All the windows are concatenated and the profile is binarized by setting a threshold on the p-value.

[2]Which are not necessary for the fitting the noise, induce larger vectors to fit and thus is more computationally expensive

6.3 NGS Illumina Technique

One of the most used NGS technique is the Illumina sequencing which is based on *sequencing by synthesis*. It can be decomposed in the following steps:

First the experiment needs to generate reads, corresponding to the regions of interest. Then some oligo (or adapters) are attached to each part of the fragment allowing it to bind to a flowcell surface Fig6.5(1). Once binded, the reads are duplicated starting from the oligo attached to the cell before being removed, leaving a strongly attached copy Fig6.5(2). The copy then fold on the complementary oligo attached at step 1 and allows the polymerase to create a complementary copy of the read Fig6.5(3). This last operation is repeated multiple times in in order to increase the read signal Fig6.5(4). All the reverse reads are then removed Fig6.5(5). The sequence of the reads is then determined by building the complementary sequence with oligo elements that emits light when they bind to a sequence Fig6.5(6), allowing to identify the read sequence. Eventually the read is folded and duplicated and sequenced backward, to improve the signal quality Fig6.5(7). Reads are treated in parallel, allowing the sequencing of billions of reads for each experiment Fig6.5(8).

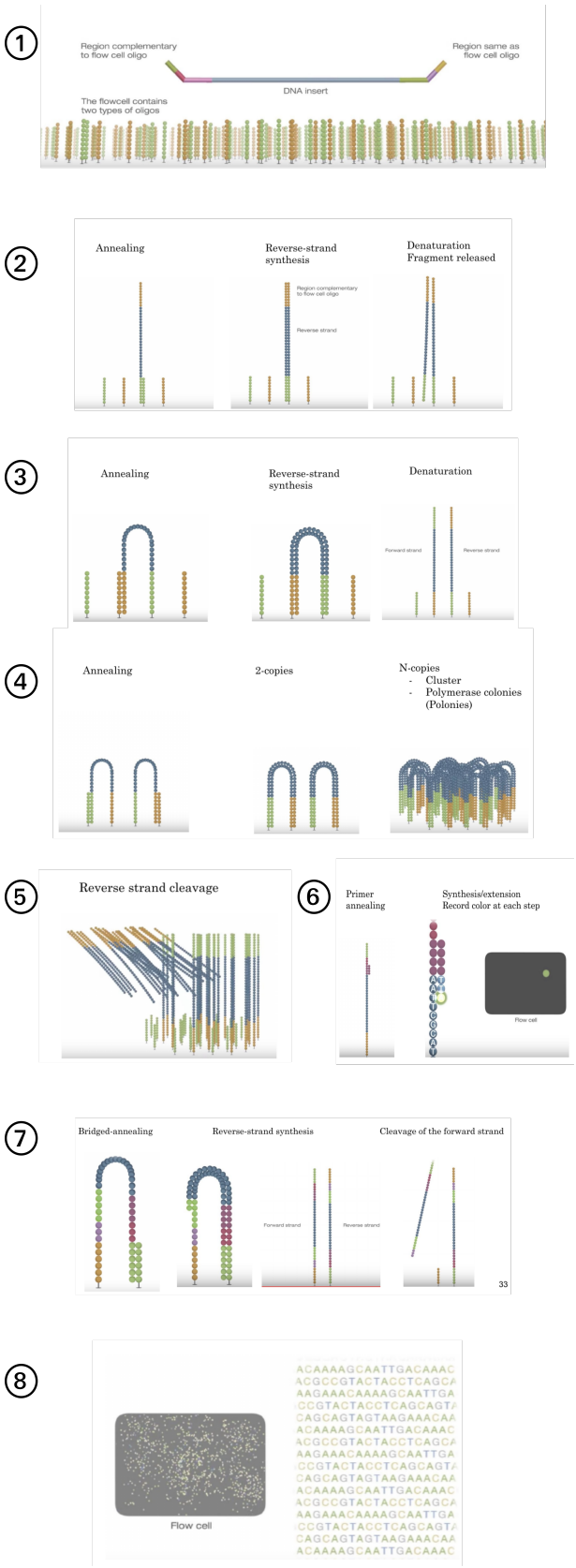

Figure 6.5 – The different steps to sequence with the Illumina technology *(source http://www.illumina.com/company/video-hub/HMyCqWhwB8E.html)*

6.4 Distance measure not adapted to dataset

To underline the statement of the importance of the distance measure choice to detect low quality profiles, we performed the same quality procedure on the single cell ATAC-seq dataset using a distance measure not adapted to these profiles, Roger-Tanimoto without PCA or kernels functions (resulting in a Silhouette score of -0.22 and a connectivity score of 0.043) Fig6.6. Because experimental replicates are mixed, almost all the samples are defined as outliers and thus removed.

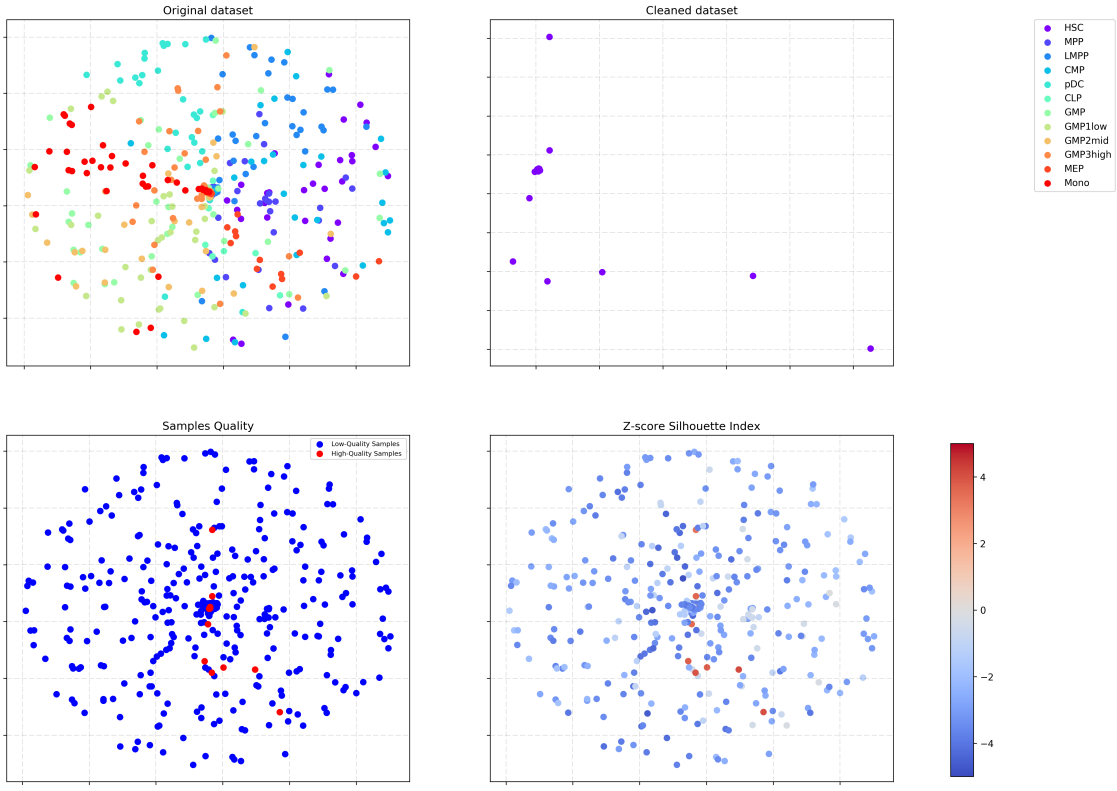

Figure 6.6 – NGS Quality procedure applied to binary single cell ATAC-seq using a distance measure not adapted to this dataset.

6.5 GABI calculation and algorithm description

Consider $X = \{X_i\}_{i\in[1,N]}$, the matrix of NGS profiles, as collection of vectors corresponding to N genomic sites indexed by the random variable $i \in [1,N]$. We define

$$P(\hat{i} = i) = \prod_{d=1}^{D} P(\hat{X}_{id} = X_{id}) \text{ with } X_{id} = \{0,1\}$$

the probability of the genomic site i with D the number of NGS profiles. Consider $Z = \{Z_k\}_{k\in[1,K]}$ as collection of vectors corresponding to the reference states, indexed by the random variable $k \in [1,K]$ with K the number of states. We define

$$P(\hat{k} = k) = \prod_{d=1}^{D} P(\hat{Z}_{kd} = Z_{kd}) \text{ with } Z_{kd} = \{0,1\}$$

the probability of reference state k. To assign a reference state to each genomic site, we compute $P(\hat{k} = k|\hat{i} = i, \theta)$ using Bayes formula on $P(\hat{i} = i|\theta)$. This probability distribution is optimized using the maximum likelihood estimation (MLE), with θ an ensemble of parameters to optimized. The likelihood is defined such that:

$$\begin{aligned}
P(X|\theta) &= \prod_{i=1}^{N} P(\hat{i} = i|\theta) \\
&= \prod_{i=1}^{N} \sum_{k=1}^{k} P(\hat{i} = i, \hat{k} = k|\theta) \\
logP(X|\theta) &= \sum_{i=1}^{N} log\left(\sum_{k=1}^{k} P(\hat{i} = i, \hat{k} = k|\theta)\right) \\
&= \sum_{i=1}^{N} log\left(\sum_{k=1}^{k} R(\hat{k} = k|\theta)\frac{P(\hat{i} = i, \hat{k} = k|\theta)}{R(\hat{k} = k|\theta)}\right)
\end{aligned}$$

with $R(\hat{k} = k|\theta)$ the distribution of $\hat{k}$. The principle of the MLE, is to maximize $P(X|\theta)$ according to θ, thus it is the same as maximizing the lower bound of $logP(X|\theta)$. According to the Jensen inequality on concave function, $log(\sum_i \lambda_i x_i) \geq \sum_i \lambda_i log(x_i)$ with $\sum_i \lambda_i = 1$. Consequently:

$$\begin{aligned}
logP(X|\theta) &\geq \sum_{i=1}^{N}\sum_{k=1}^{k} R(\hat{k} = k|\theta) log\left(\frac{P(\hat{i} = i, \hat{k} = k|\theta)}{R(\hat{k} = k|\theta)}\right) \\
&= Q(\theta)
\end{aligned}$$

with $Q(\theta)$ the lower bound of the log likelihood. We see that if $R(\hat{k} = k|\theta) = P(\hat{k} = k|\hat{i} = i, \theta)$, then the lower bound is optimal, which means that $logP(X|\theta) = Q(\theta)$:

$$\begin{aligned}
Q(\theta) = &\sum_{i=1}^{N}\sum_{k=1}^{k} P(\hat{k} = k|\hat{i} = i, \theta) logP(\hat{i} = i, \hat{k} = k, \theta) \\
&- \sum_{i=1}^{N}\sum_{k=1}^{k} P(\hat{k} = k|\hat{i} = i, \theta) logP(\hat{k} = k|\hat{i} = i, \theta)
\end{aligned}$$

with the posteriori probability defined using Bayes rule:

$$P(\hat{k}=k|\hat{i}=i,\theta)=\frac{P(\hat{i}=i|\hat{k}=k,\theta)P(\hat{k}=k|\theta)}{\sum_{k=1}^{K}P(\hat{i}=i|\hat{k}=k,\theta)P(\hat{k}=k|\theta)}$$

Let's now define the core of this method, the a priori probability $P(\hat{i}=i|\hat{k}=k,\theta)$.

$$\begin{aligned}P(\hat{i}=i|\hat{k}=k,\theta)&=\prod_{d=1}^{D}P(\hat{X}_{id}=X_{id}|\hat{Z}_{kd}=Z_{kd},\theta)\\&=\prod_{d=1}^{D}P(\hat{X}_{id}=1|\hat{Z}_{kd}=1,\theta)^{X_{id}Z_{kd}}P(\hat{X}_{id}=1|\hat{Z}_{kd}=0,\theta)^{X_{id}(1-Z_{kd})}\\&\qquad P(\hat{X}_{id}=0|\hat{Z}_{kd}=1,\theta)^{(1-X_{id})Z_{kd}}P(\hat{X}_{id}=0|\hat{Z}_{kd}=0,\theta)^{(1-X_{id})(1-Z_{kd})}\end{aligned}$$

Now we parametrize the probability:

$$\begin{aligned}a_{11,d}&=P(\hat{X}_{id}=1|\hat{Z}_{kd}=1,\theta)\\a_{10,d}&=P(\hat{X}_{id}=1|\hat{Z}_{kd}=0,\theta)\\a_{01,d}&=P(\hat{X}_{id}=0|\hat{Z}_{kd}=1,\theta)\\a_{00,d}&=P(\hat{X}_{id}=0|\hat{Z}_{kd}=0,\theta)\\\tau_k&=P(\hat{k}=k|\theta)\end{aligned}$$

and we define $\theta=(a_{11,d},a_{10,d},a_{01,d},a_{00,d},\tau_k)\forall d\in[1,D]$. Eventually the lower bound is:

$$\begin{aligned}Q(\theta)=&\sum_{i=1}^{N}\sum_{k=1}^{k}P(\hat{k}=k|\hat{X}_i=X_i,\theta)\\&\sum_{d=1}^{D}(X_{id}Z_{kd})log(a_{11,d})+((1-X_{id})Z_{kd})log(a_{01,d})\\&\quad+(X_{id}(1-Z_{kd}))log(a_{10,d})+((1-X_{id})(1-Z_{kd}))log(a_{00,d})+\log(\tau_k)\\&-\sum_{i=1}^{N}\sum_{k=1}^{k}P(\hat{k}=k|\hat{X}_i=X_i,\theta)log\left(P(\hat{k}=k|\hat{X}_i=X_i,\theta)\right)\end{aligned}$$

Expectation Maximization Algorithm The Expectation Maximization algorithm (EM) [23] is an optimization technique which has been developed to find iteratively the maximum likelihood. It is mainly divided in two steps:

1. **Initialization**: Initialize randomly the parameters in θ and compute $P(\hat{i}=i|\hat{k}=k,\theta)$
2. **Expectation Step**: Compute the posteriori probability such such that:

$$P(\hat{k}=k|\hat{i}=i,\theta)=\frac{P(\hat{i}=i|\hat{k}=k,\theta)\tau_k}{\sum_{k=1}^{K}P(\hat{i}=i|\hat{k}=k,\theta)\tau_k}$$

3. **Maximization Step:** Compute $\frac{\partial Q}{\partial \theta} = 0$ for all parameters in $\theta = (a_{11}, a_{10}, a_{01}, a_{00}, \tau_k)$ and update them. It gives

$$
\begin{aligned}
a_{10,d} &= \frac{\sum_{i=1}^{N}\sum_{k=1}^{k} P(\hat{k}=k|\hat{i}=i,\theta)X_{id}(1-Z_{kd})}{\sum_{i=1}^{N}\sum_{k=1}^{k} P(\hat{k}=k|\hat{i}=i,\theta)(1-Z_{kd})} \\
a_{01,d} &= \frac{\sum_{i=1}^{N}\sum_{k=1}^{k} P(\hat{k}=k|\hat{i}=i,\theta)(1-X_{id})Z_{kd}}{\sum_{i=1}^{N}\sum_{k=1}^{k} P(\hat{k}=k|\hat{i}=i,\theta)Z_{kd}} \\
a_{11,d} &= 1 - a_{01,d} \\
a_{00,d} &= 1 - a_{10,d} \\
\tau_k &= P(\hat{k}=k,\theta)
\end{aligned}
$$

4. Repeat the Expectation and Maximization until $Q(\theta^{t+1}) - Q(\theta^t) < \epsilon$.

6.6 TF enrichment

The result of FIMO on all genomic sites is a matrix, TFPos, which rows and columns represent the TFs and the genomic position respectively, and the value the number of times FIMO has detected a TF at a given genomic position[3].

The enrichment is computed as the z-score to a null model, equivalent to the model presented in 2.4.1 : it is based on reshuffling instances of the original matrix conserving the row and column probability such that for a NGS profile matrix m indexed by row i and column j, the probability the null model has the value c at (i,j) is

$$P_{i,j}(m_{i,j} = c) = (P_i \,.\, P_j)^c$$

$$P_i = \frac{\sum_j m_{i,j}}{\sum_{i,j} m_{i,j}} \qquad\qquad P_j = \frac{\sum_i m_{i,j}}{\sum_{i,j} m_{i,j}}$$

with P_i and P_j the probability of the row i and column j respectively. Here, we randomize the inferred matrix resulting from GABI annotation, AnnMat, which rows and columns represents the cell types and the genomic position.

The z-score for the TF k and the cell type j is then computed such that:

$$Z_{CT_j,TF_k} = \frac{\text{TFMat}(j,k) - \mu_{\text{TFMatR}}(j,k)}{\sigma_{\text{TFMatR}}(j,k)}$$

$$\text{TFMat}(j,k) = \sum_{\text{genomic site i}} \text{TFPos}(k,i) \,.\, \text{AnnMat}(j,i)$$

$$\text{TFMatR}(j,k) = \sum_{\text{genomic site i}} \text{TFPos}(k,i) \,.\, \text{AnnMatR}(j,i)$$

with AnnMatR = null model(AnnMat) and μ_{TFMatR}, σ_{TFMatR} the mean and the standard deviation of TFMatR repeated multiple times (100 times in our case).

[3]Each sequence is 500 bp

6.7 Dimensionality reduction methods

6.7.1 Classical MDS (cMDS)

The algebraic version of classical MDS of from a distance matrix D is defined in three steps [76]:

1. **Compute the Gram matrix** The Gram matrix M, defined as:

$$M_{ij} = \frac{1}{2}\left[D_{0i}^2 + D_{0j}^2 - D_{ij}^2\right] \quad D_{0i}^2 = \frac{1}{N}\sum_{j=1}^{N} D_{ij}^2 - \frac{1}{N^2}\sum_{j=1}^{N}\sum_{k>j}^{N} D_{jk}^2$$

   can be computed using the double centering on the distance matrix such that:

$$D_{ij}^{(2)} = D_{ij}^2 \quad M = \mathbf{Id}_N - N^{-1}\mathbf{1}_N$$

   where $\mathbf{Id}_N$ is the the $N \times N$ identity matrix and $\mathbf{1}_N$ the $N \times N$ matrix with all components equal to 1 [76].

2. **Compute the coordinates of the data points** The coordinates are defined as the eigenvectors of the Gram matrix such that:

$$V_{\kappa,i} = \sqrt{\lambda_\kappa} \times E_\kappa(i), \qquad (\kappa = 1, 2, 3)$$

   for a 3D structure.

6.7.2 Sammon mapping

This method is based on the minimization of the relative stress

$$\epsilon = \frac{1}{\sum_{i<j} D_{ij}} \sum_{i<j} \frac{(D_{ij} - R_{ij})^2}{D_{ij}}$$

with (i, j) the line and column indexes of the matrix. In contrast with classical MDS, there is no longer an analytical solution relating D with the optimal coordinates. The minimization of the stress is achieved by iterative optimization. Noticeably, the procedure takes as a starting point the 3D structure provided by classical MDS, in order to reduce the nonconvex optimization problem to a local minimization problem and exploit the efficient dimensional reduction of the cMDS.

Bibliography

[1] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010. 16

[2] Jörgen Adolfsson, Robert Månsson, Natalija Buza-Vidas, Anne Hultquist, Karina Liuba, Christina T Jensen, David Bryder, Liping Yang, Ole-Johan Borge, Lina AM Thoren, et al. Identification of flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential: a revised road map for adult blood lineage commitment. *Cell*, 121(2):295–306, 2005. 42

[3] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974. 29

[4] Koichi Akashi, David Traver, Toshihiro Miyamoto, and Irving L Weissman. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature*, 404(6774):193–197, 2000. 35

[5] Robin Andersson, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, Xiaobei Zhao, Christian Schmidl, Takahiro Suzuki, et al. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455, 2014. 25

[6] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, 2007. 61

[7] Bradley E Bernstein, Michael Kamal, Kerstin Lindblad-Toh, Stefan Bekiranov, Dione K Bailey, Dana J Huebert, Scott McMahon, Elinor K Karlsson, Edward J Kulbokas, Thomas R Gingeras, et al. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell*, 120(2):169–181, 2005. 61

[8] Bradley E Bernstein, Tarjei S Mikkelsen, Xiaohui Xie, Michael Kamal, Dana J Huebert, James Cuff, Ben Fry, Alex Meissner, Marius Wernig, Kathrin Plath, et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125(2):315–326, 2006. 61

[9] Adrian Bird. Dna methylation patterns and epigenetic memory. *Genes & development*, 16(1):6–21, 2002. 4

[10] Kendrick Boyd, Kevin H Eng, and C David Page. Area under the precision-recall curve: Point estimates and confidence intervals. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 451–466. Springer, 2013. 29

[11] Cyril Broccardo, Nicole Dastugue, Marina Bousquet, Julien Familiades, Étienne Coyaud, Cathy Quelen, Pierre Brousset, and Éric Delabesse. Pax5, oncogène majeur des leucémies aiguës lymphoblastiques b. *Hématologie*, 13(6):391–394, 2007. 43

[12] Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature methods*, 10(12):1213–1218, 2013. 9

[13] Jason D Buenrostro, Beijing Wu, Ulrike M Litzenburger, Dave Ruff, Michael L Gonzales, Michael P Snyder, Howard Y Chang, and William J Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486, 2015. 9

[14] JD Buenrostro, MR Corces, CA Lareau, B Wu, AN Schep, MJ Aryee, R Majeti, HY Chang, and WJ Greenleaf. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell*, 173(6):1535, 2018. 9

[15] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974. 15

[16] Howard Cedar and Yehudit Bergman. Epigenetics of haematopoietic cell development. *Nature Reviews Immunology*, 11(7):478, 2011. 35

[17] Peter N Cockerill. Structure and function of active chromatin and dnase i hypersensitive sites. *The FEBS journal*, 278(13):2182–2210, 2011. 6

[18] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012. 25

[19] Axel Cournac, Hervé Marie-Nelly, Martial Marbouty, Romain Koszul, and Julien Mozziconacci. Normalization of a chromosomal contact map. *BMC genomics*, 13(1):436, 2012. 49

[20] Menno P Creyghton, Albert W Cheng, G Grant Welstead, Tristan Kooistra, Bryce W Carey, Eveline J Steine, Jacob Hanna, Michael A Lodato, Garrett M Frampton, Phillip A Sharp, et al. Histone h3k27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107(50):21931–21936, 2010. 61

[21] Pol Cuscó and Guillaume J Filion. Zerone: a chip-seq discretizer for multiple replicates with built-in quality control. *Bioinformatics*, 32(19):2896–2902, 2016. 63, 64, 86

[22] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *science*, 295(5558):1306–1311, 2002. 47

[23] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977. 70

[24] Connie J Eaves. Hematopoietic stem cells: concepts, definitions, and the new reality. *Blood*, 125(17):2605–2613, 2015. 35

[25] Jason Ernst and Manolis Kellis. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature biotechnology*, 28(8):817–825, 2010. 25, 29

[26] Jason Ernst and Manolis Kellis. Chromhmm: automating chromatin-state discovery and characterization. *Nature methods*, 9(3):215–216, 2012. 62, 85

[27] Jason Ernst and Manolis Kellis. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nature biotechnology*, 33(4):364–376, 2015. 26

[28] James Fraser, Mathieu Rousseau, Solomon Shenker, Maria A Ferraiuolo, Yoshihide Hayashizaki, Mathieu Blanchette, and Josée Dostie. Chromatin conformation signatures of cellular differentiation. *Genome biology*, 10(4):R37, 2009. 50

[29] Charles E Grant, Timothy L Bailey, and William Stafford Noble. Fimo: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011. 43

[30] Nathaniel D Heintzman, Rhona K Stuart, Gary Hon, Yutao Fu, Christina W Ching, R David Hawkins, Leah O Barrera, Sara Van Calcar, Chunxu Qu, Keith A Ching, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics*, 39(3):311, 2007. 61

[31] Eric Jones, Travis Oliphant, Pearu Peterson, et al. Scipy: Open source scientific tools for python. 2001–. 65

[32] Manolis Kellis, Barbara Wold, Michael P Snyder, Bradley E Bernstein, Anshul Kundaje, Georgi K Marinov, Lucas D Ward, Ewan Birney, Gregory E Crawford, Job Dekker, et al. Defining functional dna elements in the human genome. *Proceedings of the National Academy of Sciences*, 111(17):6131–6138, 2014. 25

[33] Aurélie Kieusseian, Jalila Chagraoui, Cécile Kerdudo, Philippe-Emmanuel Mangeot, Philip J Gage, Nicole Navarro, Brigitte Izac, Georges Uzan, Bernard G Forget, and Anne Dubart-Kupperschmitt. Expression of pitx2 in stromal cells is required for normal hematopoiesis. *Blood*, 107(2):492–500, 2006. 43

[34] Tae Hoon Kim, Leah O Barrera, Ming Zheng, Chunxu Qu, Michael A Singer, Todd A Richmond, Yingnian Wu, Roland D Green, and Bing Ren. A high-resolution map of active promoters in the human genome. *Nature*, 436(7052):876, 2005. 61

[35] Motonari Kondo, Irving L Weissman, and Koichi Akashi. Identification of clonogenic common lymphoid progenitors in mouse bone marrow. *Cell*, 91(5):661–672, 1997. 35

[36] Hashem Koohy, Thomas A Down, Mikhail Spivakov, and Tim Hubbard. A comparison of peak callers used for dnase-seq data. *PLoS One*, 9(5):e96303, 2014. 63

[37] Peter Hugo Lodewijk Krijger and Wouter De Laat. Regulation of disease-associated gene expression in the 3d genome. *Nature reviews molecular cell biology*, 17(12):771–782, 2016. 2

[38] Ivan V Kulakovskiy, Ilya E Vorontsov, Ivan S Yevshin, Anastasiia V Soboleva, Artem S Kasianov, Haitham Ashoor, Wail Ba-alawi, Vladimir B Bajic, Yulia A Medvedeva, Fedor A Kolpakov, et al. Hocomoco: expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic acids research*, 44(D1):D116–D125, 2016. 43

[39] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951. 22

[40] Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Pouya Kheradpour, Zhizhuo Zhang, Alireza Heravi-Moussavi, Yaping Liu, Viren Amin, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317, 2015. 8, 25, 61, 81

[41] Elisa Laurenti and Berthold Göttgens. From haematopoietic stem cells to complex differentiation landscapes. *Nature*, 553(7689):418, 2018. 35

[42] Tong Ihn Lee, Richard G Jenner, Laurie A Boyer, Matthew G Guenther, Stuart S Levine, Roshan M Kumar, Brett Chevalier, Sarah E Johnstone, Megan F Cole, Kyo-ichi Isono, et al. Control of developmental regulators by polycomb in human embryonic stem cells. *Cell*, 125(2):301–313, 2006. 61

[43] Annick Lesne, Julien Riposo, Paul Roger, Axel Cournac, and Julien Mozziconacci. 3d genome reconstruction from chromosomal contacts. *Nature methods*, 11(11):1141–1143, 2014. 48, 50

[44] Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293, 2009. 47, 48, 51, 52

[45] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008. 22, 23, 24, 36, 82

[46] Elaine R Mardis. The impact of next-generation sequencing technology on genetics. *Trends in genetics*, 24(3):133–141, 2008. 6

[47] Clifford A Meyer and X Shirley Liu. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews. Genetics*, 15(11):709, 2014. 7, 20, 63

[48] Tarjei S Mikkelsen, Manching Ku, David B Jaffe, Biju Issac, Erez Lieberman, Georgia Giannoukos, Pablo Alvarez, William Brockman, Tae-Kyung Kim, Richard P Koche, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553, 2007. 61

[49] Kaoru Mitsui, Yoshimi Tokuzawa, Hiroaki Itoh, Kohichi Segawa, Mirei Murakami, Kazutoshi Takahashi, Masayoshi Maruyama, Mitsuyo Maeda, and Shinya Yamanaka. The homeoprotein nanog is required for maintenance of pluripotency in mouse epiblast and es cells. *cell*, 113(5):631–642, 2003. 43

[50] Jean-Baptiste Morlot, Julien Mozziconacci, and Annick Lesne. Network concepts for analyzing 3d genome structure from chromosomal contact maps. *EPJ Nonlinear Biomedical Physics*, 4(1):2, 2016. 47

[51] Nasser M Nasrabadi. Pattern recognition and machine learning. *Journal of electronic imaging*, 16(4):049901, 2007. 14, 18

[52] John A Nelder and Roger Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965. 65

[53] Jooyoung Park and Irwin W Sandberg. Universal approximation using radial-basis-function networks. *Neural computation*, 3(2):246–257, 1991. 18

[54] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. 16

[55] Leïla Perié, Ken R Duffy, Lianne Kok, Rob J de Boer, and Ton N Schumacher. The branching point in erythro-myeloid differentiation. *Cell*, 163(7):1655–1662, 2015. 35

[56] Dmitry K Pokholok, Christopher T Harbison, Stuart Levine, Megan Cole, Nancy M Hannett, Tong Ihn Lee, George W Bell, Kimberly Walker, P Alex Rolfe, Elizabeth Herbolsheimer, et al. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell*, 122(4):517–527, 2005. 61

[57] Alvaro Rada-Iglesias, Ruchi Bajpai, Tomek Swigut, Samantha A Brugmann, Ryan A Flynn, and Joanna Wysocka. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, 470(7333):279, 2011. 61

[58] Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014. 47, 48, 52, 84

[59] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987. 15

[60] Frederick Sanger, Steven Nicklen, and Alan R Coulson. Dna sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12):5463–5467, 1977. 6

[61] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978. 29

[62] Jun Seita and Irving L Weissman. Hematopoietic stem cell: self-renewal versus differentiation. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 2(6):640–653, 2010. 35

[63] François Serra, Marco Di Stefano, Yannick G Spill, Yasmina Cuartero, Michael Goodstadt, Davide Baù, and Marc A Marti-Renom. Restraint-based three-dimensional modeling of genomes and genomic domains. *FEBS letters*, 589(20PartA):2987–2995, 2015. 50

[64] Norman E. Sharpless, NCI Staff, and NCI. Studies identify potential treatments for dipg brain tumors. 4, 81

[65] Daria Shlyueva, Gerald Stampfel, and Alexander Stark. Transcriptional enhancers from properties to genome-wide predictions. *Nature Reviews Genetics*, 15(4):272, 2014. 3, 81

[66] Lingyun Song and Gregory E Crawford. Dnase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*, 2010(2):pdb–prot5384, 2010. 6, 9

[67] Lincoln Stein. Genome annotation: from sequence to biology. *Nature reviews genetics*, 2(7):493, 2001. 25

[68] Andrew B Stergachis, Shane Neph, Alex Reynolds, Richard Humbert, Brady Miller, Sharon L Paige, Benjamin Vernot, Jeffrey B Cheng, Robert E Thurman, Richard Sandstrom, et al. Developmental fate and cellular maturity encoded in human regulatory dna landscapes. *Cell*, 154(4):888–903, 2013. 4, 6

[69] Andrew B Stergachis, Shane Neph, Alex Reynolds, Richard Humbert, Brady Miller, Sharon L Paige, Benjamin Vernot, Jeffrey B Cheng, Robert E Thurman, Richard Sandstrom, et al. Developmental fate and cellular maturity encoded in human regulatory dna landscapes. *Cell*, 154(4):888–903, 2013. 35, 44

[70] Minjia Tan, Hao Luo, Sangkyu Lee, Fulai Jin, Jeong Soo Yang, Emilie Montellier, Thierry Buchou, Zhongyi Cheng, Sophie Rousseaux, Nisha Rajagopal, et al. Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell*, 146(6):1016–1028, 2011. 4

[71] Zhixin Tian, Nikola Tolić, Rui Zhao, Ronald J Moore, Shawna M Hengel, Errol W Robinson, David L Stenoien, Si Wu, Richard D Smith, and Ljiljana Paša-Tolić. Enhanced top-down characterization of histone post-translational modifications. *Genome biology*, 13(10):R86, 2012. 4

[72] Warren S Torgerson. Theory and methods of scaling. 1958. 50

[73] Michele Vendruscolo, Edo Kussell, and Eytan Domany. Recovery of protein structure from contact maps. *Folding and Design*, 2(5):295–306, 1997. 49

[74] Stephen Warshall. A theorem on boolean matrices. *Journal of the ACM (JACM)*, 9(1):11–12, 1962. 50

[75] James D Watson and Francis HC Crick. Genetical implications of the structure of deoxyribonucleic acid. *Nature*, 171(4361):964–967, 1953. 1

[76] Florian Wickelmaier. An introduction to mds. *Sound Quality Research Unit, Aalborg University, Denmark*, 46, 2003. 72

[77] Megan Wilson and Peter Koopman. Matching sox: partner proteins and co-factors of the sox family of transcriptional regulators. *Current opinion in genetics & development*, 12(4):441–446, 2002. 43

[78] Carolien M Woolthuis and Christopher Y Park. Hematopoietic stem/progenitor cell commitment to the megakaryocyte lineage. *Blood*, pages blood–2015, 2016. 42

[79] Carl Wu, Paul M Bingham, Kenneth J Livak, Robert Holmgren, and Sarah CR Elgin. The chromatin structure of specific genes: I. evidence for higher order domains of defined dna sequence. *Cell*, 16(4):797–806, 1979. 6

[80] Mark Yandell and Daniel Ence. A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, 13(5):329, 2012. 25

[81] Byung-Jun Yoon. Hidden markov models and their applications in biological sequence analysis. *Current genomics*, 10(6):402–415, 2009. 29

[82] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, et al. Model-based analysis of chip-seq (macs). *Genome biology*, 9(9):R137, 2008. 8, 63

[83] ZhiZhuo Zhang, Guoliang Li, Kim-Chuan Toh, and Wing-Kin Sung. 3d chromosome modeling with semi-definite programming and hi-c data. *Journal of computational biology*, 20(11):831–846, 2013. 54

[84] Walter Zucchini, Iain L MacDonald, and Roland Langrock. *Hidden Markov models for time series: an introduction using R*. Chapman and Hall/CRC, 2016. 29

List of Figures

List of Tables