Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Dipartimento di Fisica e Astronomia
Corso di Laurea Magistrale in Fisica

Statistical and network-based methods for the analysis of chromatin accessibility maps in single cells

Relatore:
Prof. Daniel Remondini

Correlatore:
PD Dr. Karsten Rippe

Presentata da:
Daniele Tavernari

Anno Accademico 2015/2016

Acknowledgements

Alla mia famiglia, i miei genitori Patrizia e Roberto, mio fratello Riccardo, i miei nonni, i miei zii, i miei cugini, devo i più importanti insegnamenti, e a loro è dedicato ogni mio traguardo. Mio padre e mia madre, in particolare, come già ho scritto loro due anni fa, rappresentano il migliore esempio da seguire, l'ideale al quale tendere. Un grande ringraziamento va anche ad Elisa e a tutti i miei amici, in particolare gli Oltreuomini, i Trenari e i Kirschgartenstrasse, per il supporto e l'affetto che mai mi hanno fatto mancare.

Mai dimenticherò quanto sono fortunato ad essere circondato da queste persone.

Finally, a huge "Thank you!" goes to my great supervisors, Daniel Remondini and Karsten Rippe, whose excellent support and mentoring meant a lot to my academic and personal education, and to the whole Genome Organization and Function group, especially to my collaborators Lara and Philipp.

Sommario

In questo lavoro, metodi provenienti dalla Fisica, dalla Statistica e dalla Teoria dei Grafi sono stati impiegati per caratterizzare ed analizzare profili di apertura e accessibilità della cromatina ottenuti con la tecnica ATAC-seq in singole cellule, nella fattispecie linfociti B provenienti da tre pazienti affetti da Leucemia Linfocitica Cronica.

Una pipeline bioinformatica è stata sviluppata per processare i dati di sequencing ed ottenere le posizioni accessibili del genoma per ciascuna cellula. La quantità di regioni aperte e la loro distribuzione spaziale lungo il DNA sono state caratterizzate. Infine, l'apertura simultanea nelle stesse singole cellule di regioni regolatrici è stata impiegata come metrica per valutare relazioni funzionali, e in questo modo grafi tra enhancer e promoter sono stati costruiti e le loro proprietà sono state analizzate.

La distribuzione spaziale lungo il genoma di regioni aperte consecutive ricapitola proprietà strutturali come gli array di nucleosomi e le strutture a loop della cromatina. Inoltre, i profili di accessibilità delle regioni regolatrici sono significativamente conservati nelle singole cellule. I network tra enhancer e promoter forniscono un modo per caratterizzare la rilevanza di ciascuna regione regolatrice in termini di centralità. Le statistiche sulla connettività tra enhancer e promoter confermano il modello di relazione uno-a-uno come il più frequente, in cui un promoter è regolato dall'enhancer ad esso più vicino. Infine, anche il funzionamento dei superenhancer è stato indagato.

In conclusione, ATAC-seq si rivela un'efficace tecnica per indagare l'apertura del DNA in singole cellule, i cui profili di accessibilità ricapitolano caratteristiche strutturali e funzionali della cromatina. Al fine di indagare i meccanismi della malattia, il panorama di accessibilità dei lifociti tumorali può essere confrontato con quello di cellule sane e cellule trattate con farmaci epigenetici.

Abstract

Introduction Physical and mathematical modeling find a wide range of application in many fields of natural sciences.

In Genomics and Epigenetics, the advent of Next Generation Sequencing technologies has paved the way for the development of a variety of high-throughput experimental assays. Based on DNA or RNA sequencing, these techniques are capable of generating a huge amount of data at many levels and for many features and biological mechanisms, such as whole genome sequence, gene expression, protein binding, DNA methylation, chromatin conformation and accessibility.

As a consequence, the bottleneck of genomic research is gradually shifting to the data analysis. In this framework, a current challenge is to develop physical, statistical and computational methods to characterize such huge amount of data, remove biases and extract meaningful features and patterns which can give insights into biological mechanisms. In cancer research, a deeper understanding at the system level of such mechanisms in healthy and tumor cells is essential to develop and enhance therapeutic approaches in Personalized Medicine.

Methods In this work, methods from Physics, Statistics and Graph Theory have been employed to characterize and analyze chromatin accessibility profiles obtained with ATAC-seq technique in single cells. The cells investigated were B-lymphocytes coming from three patients affected by Chronic Lymphocytic Leukemia.

A customized bioinformatic pipeline has been developed to process raw sequencing data and obtain accessible loci in each single cell. The distribution of the number of open regions as well as their spatial distribution along the genome have been characterized. Finally, the occurrence of accessible loci in regulatory regions such as promoters and enhancers have been investigated; their co-occurrence in the same single cells have been employed as a metric for functional linkages, and in this way enhancer-promoter networks have been constructed and their properties could be analyzed.

Results The variability of the number of open regions reveals a remarkable heterogeneity across the single cells.

The distribution along the genome of accessible loci recapitulates known structural properties such as nucleosome arrays and chromatin loops.

Despite the heterogeneity mentioned above, the accessibility of regulatory regions is highly conserved across the single cells. Moreover, cells coming from the same patients show more similar patterns than inter-patient sets of cells.

Enhancers-promoters networks provided a tool to explore the relevance of each regulatory element with centrality metrics. Statistics of enhancer-promoter connectivity confirmed that the most frequent linkage model is the one-to-one, in which one promoter is regulated by its closest enhancer. However, exceptions are not rare.

Superenhancers functioning was also investigated, and peaks of localized accessible chromatin within them have been found.

Discussion ATAC-seq technique is a valuable tool to investigate open chromatin in single cells. Accessibility profiles recapitulate structural features of chromatin and can be employed to assess functional relationships between regulatory elements. In order to investigate the mechanisms of the disease, the accessibility landscape of CLL B-lymphocytes can thus be compared with healthy controls and tumor cells treated with epigenetic drugs.

Contents

List of Figures

List of Acronyms

3C Chromosome Conformation Capture

ATAC-seq Assay for Transposase-Accessible Chromatin followed by high-throughput sequencing

ChIA-PET Chromatin Interaction Analysis by Paired-End Tag Sequencing

ChIP-seq Chromatin immunoprecipitation followed by high-throughput sequencing

CLL Chronic Lymphocytic Leukemia

DFT Discrete Fourier Transform

DNase-seq Deoxyribonuclease I reaction followed by high-throughput sequencing

FFT Fast Fourier Transform

gof Goodness of fit

hg19 Human Genome 19

Hi-C High-throughput Chromosome Conformation Capture

HMM Hidden Markov Model

IGV Integrative Genome Viewer

IPR Inverse Participation Ratio

K-S Kolmogorov-Smirnov

MNase-seq Micrococcal Nuclease reaction followed by high-throughput sequencing

MACS Model-Based Analysis of ChIP-seq

NGS Next Generation Sequencing

PE Paired-End (sequencing)

PCR Polymerase Chain Reaction

PWM Position Weight Matrix

PDF Probability Density Function

PMF Probability Mass Function

SE Single-End (sequencing)

scATAC-seq single cell ATAC-seq

SNP Single Nucleotide Polymorphism

TF Transcription Factor

TSS Transcription Start Site

WGBS Whole Genome Bisulfite Sequencing

Introduction

Chronic Lymphocytic Leukemia (CLL) is the most common leukemia among adults in the western world, with a share of 40% of all types [1].

CLL is a malignancy of B-lymphocytes, which grow in an uncontrolled manner and accumulate in the blood, bone marrow and other lymphoid tissues. Despite the homogeneous morphological phenotype, the clinical outcome of CLL is variable. CLL is also known to be an epigenetic disease [2], and one of the main alterations that it causes is the deregulation of enhancers.

As a consequence, a complete and accurate characterization of the epigenome of CLL B-lymphocytes is essential to understand the mechanisms of the disease and how to treat it in the most precise and effective way. This task has to be accomplished at the various levels of epigenetic marks: DNA methylation, histone modifications, protein binding, chromatin conformation and accessibility.

At least three kinds of comparisons in terms of epigenetic profiles are possible: tumor cells versus healthy controls, tumor cells belonging to different patients and tumor cells before and after treatment with epigenetic drugs.

In this framework, this work is a result of the analysis of open chromatin profiles assayed at the single cell level in CLL B-lymphocytes with ATAC-seq technique. These experiments have been performed in the research group Genome Organization and Function, headed by PD Dr. Karsten Rippe and member of the Bioquant and the German Cancer Research Center, in Heidelberg.

The first chapter provides a generic introduction to the biological features and mechanisms involved, such as chromatin architecture, epigenetic marks and regulation of gene expression; in the second chapter, various methods and techniques for data analysis are introduced, from fields such as mathematics, bioinformatics and statistics; the third chapter deals with the analysis of many statistical properties of open chromatin profiles; in the fourth chapter the construction and the analysis of enhancer-promoter networks is reported; the fifth and final chapter contains the conclusions and the future directions that might follow this work.

Chapter 1

Structure and regulation of the human genome

1.1 Overview

In molecular biology, the genetic material that carries the information used for growth, development, functioning and reproduction of a living organism is referred to as *genome*. The genome consists of Deoxyribonucleic acid or DNA, a long polymer made of repeated units called nucleotides. Each nucleotide is composed of a sugar, called deoxyribose, a phosphate group and a nitrogenous base, which can be either an adenine (A), a thymine (T), a cytosine (C) or a guanine (G). In the human genome, the frequency of C or G ("GC-content") is 40%. The nucleotides are joined to one another by covalent bonds between the sugar of one nucleotide and the phosphate of the next. The DNA polymer thus formed is bound to another polymer via hydrogen bonds between the nitrogenous bases, which are paired in a specific manner (A with T, C with G). The resulting pair of complementary DNA molecules is structured as a double helix, with a pitch of 3.4 nm and a radius of 1 nm [3]. The information is therefore encoded in the sequence of base pairs (bp) along the double helix.

Another nucleic acid implicated in various biological roles is the RNA, which is also a polymer of nucleotides but differs from the DNA for the sugar (the ribose instead of the deoxyribose), for one of the four types of nitrogenous bases (the thymine is replaced by the uracil (U)) and for the structure, which has a single filament instead of the double helix.

The flow of genetic information is summarized by the *central dogma of molecular biology* [4]: the DNA is transcribed into RNA, and the RNA is then translated into proteins. In this context, it is possible to define a *gene*

as a region or locus of the genome that undergoes transcription into messenger RNA (mRNA). Not all the genes give rise to protein synthesis, since some of them get transcribed into RNA that is not translated ("non-coding RNA" or ncRNA) but has regulatory or structural functions. The number of protein-coding genes in the human genome is roughly 20000, but the number of translated proteins is much higher. This increase is achieved through a process termed *alternative splicing*: the transcribed RNA is divided into subregions called introns and exons, and only the sequence resulting from a subsampling of the exons gets translated. Different subsamples of the exons coming from the same gene give rise to different proteins [5].

1.2 Chromatin and spatial organization of the genome

The human genome contains approximately $3.2 \cdot 10^9$ base pairs and undergoes many levels of spatial folding and compaction [6]. To this end, proteins such as histones and transcription factors (TFs) interact with the DNA molecule in order to form a large macromolecular complex termed chromatin [5]. The primary functions of chromatin are therefore to package the DNA into a smaller volume to fit in the cell nucleus, to allow DNA replication, transcription, recombination and repair and to control gene expression 1.1.

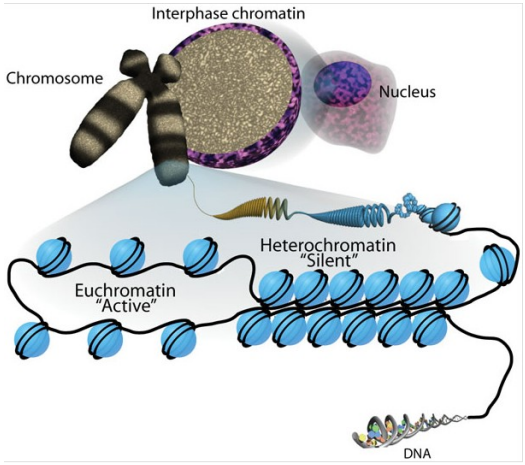


Figure 1.1: The chromatin complex (licence: Wikimedia Commons)

There are three main levels of chromatin organization [5]:

1. the DNA wraps around histone proteins in 147-bp coils and the resulting *nucleosomes* arrange in a "beads on a string" structure;

2. multiple histones cluster into arrays of 30 nm termed *fibers*

3. fibers further package into metaphase chromosomes during mitosis and meiosis

The densely packed form of chromatin is called heterochromatin and is transcriptionally inactive, while in the lightly packed euchromatin the DNA dynamically unfolds from the histones to allow spatially and temporally regulated gene expression 1.2.

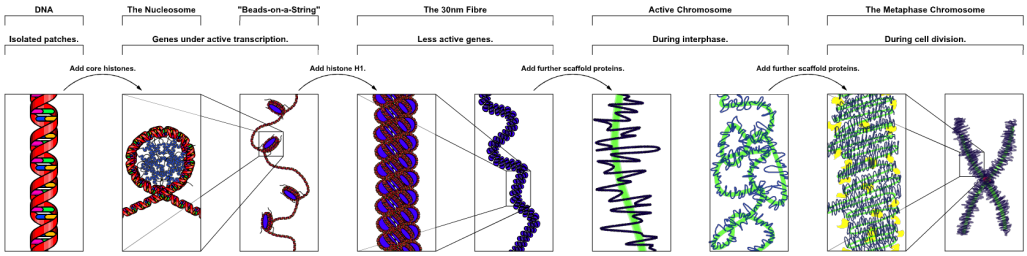


Figure 1.2: The levels of organization of chromatin (licence: Wikimedia Commons)

Recent studies [7] [8] have shown intermediate levels of spatial organization. The genome is compartmentalized into Topologically Associating Domains (TADs) of a few Megabases: within a TAD, which is usually delimited by binding sites of the CTCF transcription factor (known as an "insulator" protein), the genomic loci are in closer proximity than between different TADs. On a smaller scale ($\sim$ 100 kbp), chromatin "loops" bring distal genomic elements into spatial contact; consecutive loops form "rosette"-like structures with many sequences clustered as a "hub".

1.3 Regulation of gene expression

1.3.1 Regulatory elements

The regulation of gene expression includes the series of mechanisms by which cells activate, increase, decrease or cease the production of specific gene products. Gene regulation is involved in cell type development and differentiation, as well as responses to environmental stimuli and may be the main driver

of phenotypic divergence in evolution. Non coding DNA is largely responsible for it through interaction with the transcriptional regulatory machinery, which consists of proteins and RNA molecules [5].

Regulatory elements can be classified by their position relative to the genes they control:

i *cis*-regulatory elements are found in the vicinity (up to a few Megabases) of the genes they regulate, and typically function as binding sites for one or more transcription factors;

ii *trans*-regulatory elements may regulate the expression of distant genes, even in an allele-independent way; they are the DNA sequences that encode the transcription factors which compose the regulatory machinery.

As an additional subdivision, *cis*-regulatory elements can be classified as *promoters* or *enhancers*. Promoters are sequences found around the Transcription Start Sites (TSSs) of genes, and are therefore defined as TSSs +/- 500-1000 bp. The role of promoters is to toggle the start of transcription by recruiting RNA polymerase if they are bound by TFs. On the other hand, enhancers are sequences that can be up to a few Megabases upstream or downstream of the TSS. The binding of transcription factors on them can either intensify or repress transcription of their target genes.

Enhancers can have a variable size, ranging from 100-200 bp to several kbp. Bigger enhancers, namely the ones that are longer than roughly 18 kbp, are termed *superenhancers*. It is still unclear whether they function as a single element or as an array of independent smaller enhancers close to each other.

1.3.2 Epigenetic marks

Since all the cells of an organism share the same genome, gene expression regulation requires inputs outside of the DNA. *Epigenetics* is the field that studies how gene expression is controlled by modifications, spatial folding and accessibility of the chromatin complex. The epigenome is only partially inheritable, it is strongly influenced by environmental stimuli and can be altered in disease conditions. The most important epigenetic marks and mechanisms are listed below [5].

- *DNA methylation* is the process by which methyl groups ($-CH_3$) are attached to DNA, thus modifying its function. In mammalian cells, DNA methylation occurs at cytosine residues of CpG dinucleotides, at the 5-carbon (5mC). This modification is set and maintained by

a set of enzymes, the DNA methyltransferases (DNMTs). CpG sites are present throughout the genome but they are preferentially found at gene rich loci. DNA methylation in promoter regions can lead to stable gene silencing either by directly interfering with the binding of TFs or by inducing specific repressing chromatin states.

- *Histone modification* is a post-translational modification that alters histones in a nucleosome. Each nucleosome is composed by two copies of four core histones (H2A, H2B, H3 and H4), each of which can undergo enzymatic additions of acetyl ($COCH_3$) or methyl groups. Histone modifications are related to promoter and enhancer activity: active promoters show a high level of H3K4me3 (tri-methylation of lysine 4 in histone H3), while active enhancer are characterized by H3K4me1 and H3K27ac (acetylation of lysine 27 in histone H3).

- *Chromatin remodeling* refers to the rearrangement of chromatin from a densely packaged state to a transcriptionally accessible one, allowing TFs to bind regulatory sequences and control gene expression. Chromatin remodeling is achieved either through histone modifications or with ATP-dependent complexes.

- *Non-coding RNA* such as miRNA, siRNA, piRNA and lncRNA are also known to regulate gene expression at the transcriptional and post transcriptional level.

1.4 Experimental assays

Recent advances in molecular biology, both on the technological side and on the understanding of principles, have led to the development of novel experimental assays capable of generating huge amounts of data with standardized and automated protocols.

Many of these methods rely on Next Generation Sequencing techniques, which are capable of generating Gigabytes to Terabytes of genomic data in a limited amount of time. As a consequence, in parallel to the optimization of the experimental techniques, computational and statistical approaches are required for extensive data analysis, and these methods will be explored in more detail in the next chapter.

DNA sequencing One of the primary sources of data in Genomics and Epigenetics is constituted by the DNA sequences. DNA sequencing is the

process of experimentally determining the order of nucleotides within a DNA molecule.

Over the last few decades, the yield (in terms of base pairs sequenced per run) of sequencing machines has dramatically increased, outpacing the Moore's law for Big Data with massively parallel Next Generation Sequencing (NGS) technology [9]. Conversely, the cost per Gigabase has fallen: the 3 billion dollars required for the first sequenced human genome in 2001 have plummeted to less than 1000 dollars in 2014. These advantages allowed the implementation of population-scale sequencing and laid the foundations for personalized genomic medicine and sequencing-aided clinical decision making.

The main step in NGS is the incorporation, catalyzed by DNA polymerase, of fluorescently labeled deoxyribonucleotide triphosphates (dNTPs) into DNA template strands during sequential cycles of DNA synthesis. Instead of sequencing a single DNA fragment at a time, this process is extended across millions of fragments in a massively parallel fashion. The NGS experimental workflow includes 3 steps [9]:

1. *Library preparation.* The DNA is randomly fragmented and tagged with sequencing adapters. The resulting fragments are amplified by Polymerase Chain Reaction (PCR) cycles, which produce identical copies of the DNA templates.

2. *Cluster generation.* The library is loaded into a flow cell, where fragments are captured by oligonucleotides bound to the surface and complementary to the library adapters. DNA is amplified again with a process termed clonal bridge amplification, and clusters of identical fragments are thus formed.

3. *Sequencing.* In a process called "sequencing by synthesis", dNTPs bind specifically to the subsequent nucleotides, are excited with a light source and the emitted color defines the base pair. For each emission, the intensity of the color in a given cluster defines the quality of the call. This process is performed simultaneously for millions of clusters in a massively parallel fashion.

The final results are therefore the sequencing reads of each fragment, which come with a fixed length.

Sequencing can be performed in two ways:

- "Single-End" (SE): for each fragment, only one end is sequenced, up to the read length;

- "Paired-End" (PE): both ends of each fragment are sequenced in opposite directions. in this way, the amount of data is ideally twice as much as for Single-End sequencing. Additionally, Paired-End reads allow to determine the length of the original DNA fragments and leverage the accuracy of subsequent bioinformatic steps, such as aligning to a reference genome.

In both cases, if the sequencing read length is bigger than the size of the fragment, part of the adapter is also sequenced. This effect is known as *adapter contamination* and it requires a trimming step in the bioinformatic pipeline in order to get rid of the portion of adapters reported in the sequencing data.

DNA sequencing is important not only for whole genome reconstructions or variant calling, but also for experimental assays of epigenetic marks.

A recent development in the field is single cell sequencing, which provides a higher resolution of cellular heterogeneity. Single cell sequencing requires cells isolation, which can be achieved with automated microfluidics platforms: bulk of cells are divided into wells of small volumes (nL-pL) with microtubules; the DNA content of each single cell is then amplified with PCR and barcoded before sequencing, in order to retrieve to which cell the sequencing reads belong.

RNA-seq RNA-seq is a recently developed approach to transcriptome profiling that uses deep-sequencing technologies [10]. The transcriptome is the complete set of transcripts, namely all species of RNA, and their quantity, for a specific developmental stage or physiological condition of a cell. RNA-seq provides a powerful tool to investigate many aspects of gene expression: the transcriptional structure of genes in terms of their start sites and 5 and 3 ends; splicing patterns and other post-transcriptional modifications; the changes of the expression levels of each transcript during development and under different conditions.

RNA-seq is realized by high-throughput sequencing of the so called *complimentary DNA* (cDNA), namely the DNA sequence corresponding to the RNA molecule. There are essentially three general steps to prepare a cDNA library for RNA sequencing:

1. *RNA isolation.* RNA is isolated from tissue and mixed with deoxyribonuclease (DNase), an enzyme capable of digesting DNA. In this way, the amount of genomic DNA is reduced.

2. *RNA selection.* Depending on the type of experiment, the isolated RNA can either be kept as it is, filtered to remove ribosomal RNA (which

constitutes over 90% of RNA in a cell) or selected to bind specific DNA sequences.

3. *cDNA synthesis.* the RNA is reverse transcribed to cDNA. Reverse transcription results in loss of strandedness, which can be avoided with chemical labeling. Fragmentation and size selection are performed to purify sequences that are of appropriate length for the sequencing machine. The final cDNA fragments are then barcoded and sequenced.

Transcript levels for each gene are then determined with various data analysis techniques.

ChIP-seq Chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq)is a method to identify mammalian DNA sequences bound by transcription factors and other proteins *in vivo* and in a genome-wide fashion [11]. The process can be summarized in four steps:

1. DNA and associated proteins are cross-linked with formaldehyde fixation, thus strengthening their bond;
2. the DNA-protein complexes are sheared into 500 bp fragments by sonication or enzymatic digestion, using a micrococcal nuclease;
3. DNA fragments are selectively immunoprecipitated from the cell with the use of an antibody specific to the protein of interest;
4. the resulting fragments are purified by reversing the cross-linking of the DNA-protein complex, usually through heating;
5. the final DNA fragments are tagged with adapters and sequenced;
6. a subsequent computational analysis allows to determine the DNA sequences enriched against the background, which represent the ones associated with the protein of interest.

Enriched DNA sequences resulting from ChIP-seq experiments can be analyzed to look for characteristic patterns of nucleotides referred to as "motifs". For a protein of interest, the motif represents a sequence to which the protein is more likely to bind. Promoters and enhancers are thus enriched for motifs associated to transcription factors.

ChIP-seq can also be used to map histone modifications across the genome. A series of ChIP-seq experiments targeting H3K4me1, H3K27ac and H3K4me3 can provide the profiles necessary to detect the epigenetic patterns of regulatory sequences. ChIP-seq is therefore a powerful probe for detecting active enhancers and promoters genome-wide.

MNase-seq MNase-seq is a method for determining the positions of nucleosomes across the genome [12].

Micrococcal nuclease (MNase) is an enzyme capable of digesting nucleic acids. However, the nucleosome structure protects the DNA wrapped around the histone complex, making it resistant to the enzymatic digestion. As a consequence, the addition of MNase to the DNA leaves undigested footprints of about 150 bp, which is the length of the DNA wrapped around the histone octamers. Sequencing each end of these fragments and mapping them onto a reference genome is thus sufficient to find nucleosome positioning genome-wide.

DNase-seq DNase-seq is a method to find open chromatin regions in the genome [13].

It takes advantage of an enzyme, Deoxyribonuclease I (DNase I), which selectively digests nucleosome-depleted regions, whereas condensed and wrapped DNA is more resistant.

Open regions of chromatin are therefore preferentially and more frequently fragmented; as a consequence, the mapping of these fragments onto a reference genome allows to find the positions of open chromatin regions. Given the fact that some condensed regions can also be fragmented by chance, peak calling approaches performed on all the aligned reads give rise to the so called hypersensitive sites (HS), which are the regions most reliably and frequently open in the ensemble of cells considered.

A disadvantage of DNase-seq is that it usually requires a high quantity of input material (several millions of cells), on the one hand for the experimental assay itself and on the other hand to find the hypersensitive sites. However, a recent modification of the experimental technique has managed to apply DNase-seq to single cells [14].

ATAC-seq Assay for Transposase-Accessible Chromatin followed by high-throughput sequencing (ATAC-seq) is a novel technique developed by Buenrostro *et al* [15] to determine open regions of chromatin.

ATAC-seq takes advantage of a hyperactive prokaryotic enzyme, the Tn5 transposase, which is loaded *in vitro* to the cells. Transposases are a class of enzymes which catalyze the insertion of short DNA sequences (the transposons) into a genome. These transposons can be part of the same genome or external nucleotide sequences: in this case, the Tn5 transposase is loaded with sequencing adapters; as a consequence, its addition allows to simultaneously fragment and tag the DNA with adapters. This process is called "tagmentation" [16] and its principle is also employed in standard DNA se-

quencing.

In ATAC-seq, the Tn5 integrates only into regions of accessible chromatin, and therefore the resulting fragments come from nucleosome-depleted loci. After a PCR amplification step, the fragments can be sequenced to map open regions of the genome (Figure 1.3).

ATAC-seq profiles consistently overlap with DNase-seq hypersensitive sites and in some cases can also be used to identify transcription factor binding sites without ChIP-seq: CTCF, for example, has been shown to exhibit a peculiar pattern of ATAC-seq signal [15].

The main advantage with respect to DNase-seq is that ATAC-seq is feasible with low input material, and therefore it can be reliably applied to single cells (scATAC-seq) without particular restrictions [17].

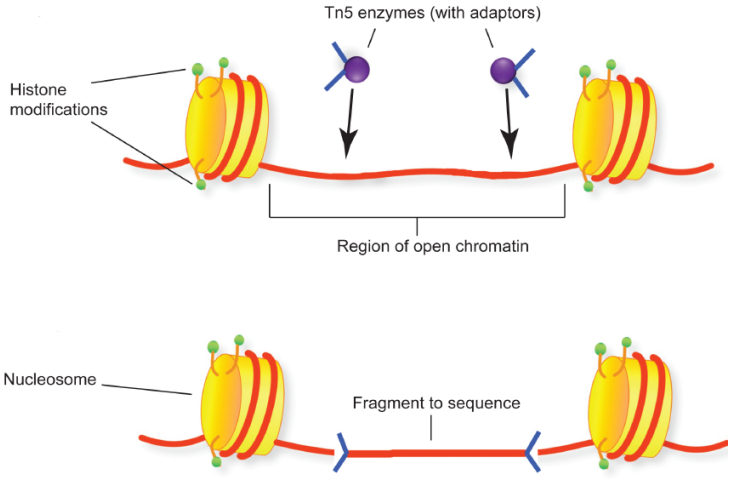


Figure 1.3: ATAC-seq technique (licence: Wikimedia Commons)

WGBS Whole Genome Bisulfite Sequencing (WGBS) is a technique that allows to map the pattern of DNA methylation across the genome.

Treatment of DNA with bisulfite (HSO_3^-) converts cytosine residues to uracil, but leaves 5-methylcytosines unaffected. Thus, the specific changes introduced in the DNA sequence by the bisulfite treatment depend on the methylation status of individual cytosine residues, yielding information about the methylation status of a segment of DNA at the single nucleotide resolution [18].

DNA methylation was previously investigated in a targeted manner for relatively small genomic regions; with the advent of WGBS, genome-wide mapping has become feasible.

3C, Hi-C Chromosome Conformation Capture techniques are used to analyze the spatial organization of chromatin in cell nuclei. They quantify the number of interactions between genomic loci that are nearby in 3D space, but may be separated by many nucleotides in the linear genome. First, the cell genomes are cross-linked, "freezing" existing interactions between genomic regions. The genome is then cut into fragments. Next, random ligation is performed in order to quantify the proximity of fragments, because fragments are more likely to be ligated to nearby mates.

The first method, 3C, was developed in 2002 by Dekker *et al.* [19]. It is capable of quantifying experimentally the interactions between a single pair of genomic loci, for example two candidate regulatory elements. As a consequence, it was a low-throughput technique. On the other hand, the recently developed Hi-C technique [20] uses high-throughput paired end sequencing, which retrieves short sequences from each end of each ligated fragment. As a consequence, all possible pairwise interactions can be found.

ChIA-PET Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) is a technique that incorporates chromatin immunoprecipitation, chromatin proximity ligation, Paired-End tags and high-throughput sequencing to determine long-range chromatin interactions genome-wide [21].

The ChIA-PET method combines ChIP-based methods and Chromosome Conformation Capture to extend the capabilities of both approaches. The issues of non-specific interaction noise found in ChIP-seq are solved by sonicating the ChIP fragments in order to separate random attachments from specific interaction complexes. Chromatin proximity ligation based on 3C and Paired-End sequencing allows then to map anchor regions of the same protein on the DNA strand, thus directly revealing three dimensional folding structures of chromatin such as loops and rosettes.

1.5 Enhancer-promoter targeting

Active regulatory elements need to be accessible and nucleosome-free in order to harbor transcription factors. As a consequence, they should overlap DNase I hypersensitive sites or ATAC-seq peaks. Moreover, as mentioned before active enhancers have flanking regions with nucleosomes carrying H3K4me1 and H3K27ac, with low levels of H3K4me3; conversely, active promoters are marked by high levels of H3K4me3 and low H3K4me1, and can also be defined as regions surrounding Transcription Start Sites. TSS can be mapped with transcription (RNA-seq) data. As a consequence, active regulatory sequences can be defined as regions with specific patterns of histone modifications which

overlap open chromatin loci [5]. Enhancer elements have also been shown to have low levels of DNA methylation [22].

1.5.1 The relevance of enhancers in diseases

Experimental assays have revealed a high degree of spatiotemporal cell type specificity of enhancers, whereas promoters appear to be more conserved across time points and developmental lineages [23]. It is thought that enhancers bound by critical lineage-specifying transcription factors help to establish the precise order of expression of both protein-coding genes and non-coding RNAs. As a consequence, changes in enhancer sequences, activity and targeting can be associated with diseases [23]. Single Nucleotide Polymorphisms (SNPs) can impact the binding of transcription factors, thus altering gene expression [22]. SNPs, somatic mutations and chromosomal rearrangements that relocate enhancers have been found to drive diseases such as Burkitts lymphoma, acute myeloid leukemia and also non-hematopoietic cancers [23]. Chronic lymphocytic leukemia (CLL) has also been indicated as an epigenetic disease, since it is associated with a deacetylation of H3K27, that causes enhancers to inactivate.

Mutated or variant enhancers, when associated with a particular disease, may become potential therapeutic targets for epigenetic drugs [23]. Importantly, the cell type specificity of enhancers may enable more precise therapy in the framework of personalized medicine, which is becoming a feasible option as the costs of sequencing experiments decrease.

1.5.2 Targeting models

As reported before, enhancers can be located at long distances upstream or downstream of target promoters/genes. Multiple models have been proposed to explain enhancer-promoter targeting. The two most common models are "scanning or tracking", in which TF-containing protein complexes bind at an enhancer and diffuse along the genome to search for a target promoter, and "looping", in which an enhancer is brought into direct contact to its target promoter by chromatin loops [23].

The "scanning" model implies that an enhancer should regulate exclusively the nearest active promoter. However, experimental evidences suggest that long-range interactions in which an enhancer bypasses multiple promoters to regulate a more distally located gene are also possible, therefore this model is not sufficient to explain the phenomenon. Moreover, recent studies on multiple nuclear architecture have provided further evidence in support of the "looping" model [23].

Chromatin loops occurs through protein-protein interactions mediated by transcription factors bound at the two ends of the loop. Structural proteins such as CTCF and proteins of the cohesin complex intervene in the formation and stabilization of the loop (Figure 1.4).

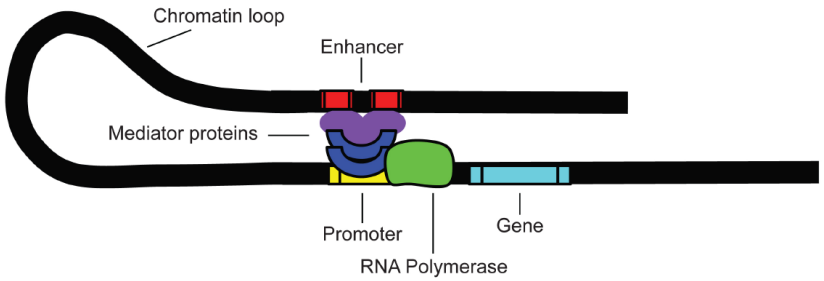


Figure 1.4: The structure of a chromatin loop (licence: Wikimedia Commons)

At both the global level and within a given cell type, the number of enhancers is bigger than the expressed genes. Therefore, enhancer-promoter interactions are not limited to one-to-one relationships: rather, an enhancer can contact multiple promoters and vice versa. Enhancer specificity suggests that a gene can be regulated by different enhancers in different cell types. As a consequence, enhancer targets have to be identified in a cell type specific manner.

1.5.3 Assaying enhancer-promoter linkages

There are two possible classes of methods to assay enhancer-promoter targeting:

1. methods based on physical interactions;

2. methods based on gene expression association.

The first class requires the use of chromosome conformation capture techniques. In Hi-C experiments, it has been found that only a small subset of chromatin interactions can be associated with enhancer-promoter linkages. As a consequence, Hi-C alone is not sufficient for the purpose. One way to increase resolution is to focus only on a subset of interactions of interest, using methods such as the previously described ChIA-PET. This method can be used to map all interactions at a subset of enhancers bound by a specific TF (for example CTCF) or showing RNA polymerase [23].

Chromatin compartmentalization arising from Hi-C experiments have suggested that enhancers are restricted to regulating promoters within specified chromatin boundaries, namely within TADs. CTCF is known to act as an insulator and can be found at TAD boundaries as well as in chromatin loops. However, CTCF binding sites are also located inside a loop. These bypassed CTCFs could be involved in enhancer-promoter interactions of other cell types, but this mechanism is unclear. Moreover, spatial contact does not necessarily imply a functional linkage, therefore methods based on physical interactions alone are not sufficient to describe enhancer-promoter targeting, although they provide valuable insights.

For the second class of methods, epigenetic techniques allow to correlate enhancer activity with target gene expression across different cell types and conditions. Activity can be assessed through assays for DNA accessibility, histone modifications or DNA methylation levels, while gene expression is measured with RNA-seq. These correlation-based analyses provide approaches to predict individual putative enhancer-gene linkages on a genome-wide scale, although some of these could be indirect.

1.5.4 Co-occurrence of open chromatin as a targeting model

In this work, an additional method based on histone modifications and open chromatin maps assayed at the single cell level is proposed.

As reported before, one of the main issues for some experimental techniques is the high number of input cells required. The inability to distinguish between individual cells is particularly detrimental when the goal is to investigate enhancers activity, since as reported before they exhibit temporal and cell cycle stage specificity. When large groups of cells are assayed together, this heterogeneity is masked and the predictions for enhancer-promoter linkages are less accurate.

Fortunately, ATAC-seq technique overcomes this limitation and allows single cell assays. As a consequence, with enhancers and promoters annotated via histone modification patterns, it is possible to evaluate which regulatory regions are accessible together in each single cell. If co-occurrence of open chromatin in two regulatory regions is recurrent across many single cells, then the two elements are likely to have a functional relationship. This feature can be robustly assessed with pairwise correlation scores.

In this way, putative chromatin networks between enhancers and promoters can be constructed and their change in disease conditions can be investigated, thus giving additional insights into epigenetic alterations and

possible therapeutic strategies to address them.

Chapter 2

Statistical and computational methods for genomic data analysis

2.1 Overview

Novel high-throughput experimental techniques such as Next Generation Sequencing yield a considerable amount of data, which allows the investigation of biological principles in a more systematic and robust way. To this end, data analysis requires tailored statistical as well as computational methods to be developed.

Here some of these techniques are introduced, with a focus on the ones that have been employed for the experimental data analyzed in this work.

2.2 Bioinformatic processing

Raw sequencing data is basically constituted by the sequence of nucleotides and the quality scores. Developing algorithm and pipelines to extract meaningful information is a challenge in Genomics and Epigenetics, since the enormous amount of data requires fine tuning and optimization to keep the computational times and costs low. Moreover, each experimental technique can give rise to biases in the data which have to be addressed throughout the whole downstream analysis.

The bioinformatic pipeline is specific for the type of experimental data available and the purpose of the analysis. However, some basic steps are performed most of the times, such as quality control, alignment to a reference genome, peak calling and visualization.

Quality control Data coming from NGS machines is usually in FASTQ format, which is a flat text file with four lines for each read. The first line contains meta information about the instrument name, the run ID, the location on the flowcell, the index sequence and whether it is Single-End or Paired-End. The second line is the raw read sequence, with an "N" where it was not possible to determine the base. The third line includes an optional description. The fourth line contains the Phred quality score for each nucleotide in hexadecimal ascii code, which is defined as

$$Q = -10 \cdot log_{10}(P) \tag{2.1}$$

where P is the probability that the call was incorrect. As a consequence, scores above 30 are considered high, since the probability of sequencing error is below 10^{-3}. The quality usually decreases towards the end of the read, since random trimmings of the fragments in the cluster accumulate, making the signal more heterogeneous. However, this is usually a problem only for long reads.

Tools such as FASTQC [24] provide summaries for quality assessments.

Alignment to a reference genome The most fundamental step is to align the raw reads to a reference genome assembly, in order to find which positions and which chromosomes they were fragmented from. Reference genomes have been constructed over the years from a large amount of sequencing data, and the assembly have been performed by concatenating partially overlapping reads. A version of the human genome often used as a reference is Human Genome 19 (hg19).

The most popular tools are Bowtie and Bowtie2 [25], which are fast and memory-efficient implementations of string alignment algorithms. It is possible to tune many parameters, such as the maximum number of mismatches allowed, how to handle low quality base calls and whether or not multiple aligning positions for a single read must be reported. For Paired-End reads, a maximum allowed length of the fragment can also be set. The alignment process is randomized in order to avoid mapping bias.

As reported before, it is possible that if the read length is bigger than the size of the fragment, the adapter at the opposite end (or part of it) can also be sequenced and therefore the read can not be mapped (adapter contamination). To solve this problem, substrings of the adapter can be trimmed from the raw data after the first mapping, and the resulting shorter reads can be remapped.

Alignments are reported in SAM format, with all the information on the mapping results and the positioning of each read (chromosome, start and

end points, "+" or "-" strand). With Samtools [26] it is possible to convert SAM files into a compressed binary version, namely the BAM format. A minimal and easy to read file format termed BED is also available; BED files can be generated with Bedtools [27] and contain only the information on the location and strand of the mapped reads.

From the alignment files, a metric termed *coverage* can be determined: the coverage describes the average number of reads that align to or "cover" known reference bases.

Peak calling As reported before, many experimental techniques generate reads which map genome-wide, but that can be enriched in certain genomic loci of interest. To extract these regions, a *peak calling* approach is required. Peak calling is equivalent to finding the genomic positions whose coverage is greater than the background in a statistically significant way.

Model-Based Analysis of ChIP-seq (MACS) [28] is a powerful software to identify statistically significant enriched genomic regions in ChIP-seq, DNase-seq and ATAC-seq data. MACS computes a background model as a random Poisson distribution with local lambda, which means that each enrichment is assessed locally. In this way, if the coverage is not uniform along the genome, peaks found in regions of low coverage are also detected.

The program yields the peak summits, their extension, the p-value and the fold enrichment against the background. The user can thus choose a cutoff and only keep higher peaks.

Visualization Integrative Genome Viewer (IGV) [29] is one of the most common tools to visually inspect mapped reads and called peaks. It allows to load multiple tracks from a variety of formats and the indexing of big files makes the browsing fast and memory-efficient.

Tracks loaded can be explored at any scale, from single nucleotide up to whole genome or whole chromosome. In this way, coverage of genomic loci of interest can be easily browsed, and the results of many experiments can be overlaid and inspected. For example, ChIP-seq alignment files or called peaks can be overlaid for different histone modifications: enhancer regions will show a good overlap between H3K4me1 and H3K27ac tracks, as shown in Figure 2.1.

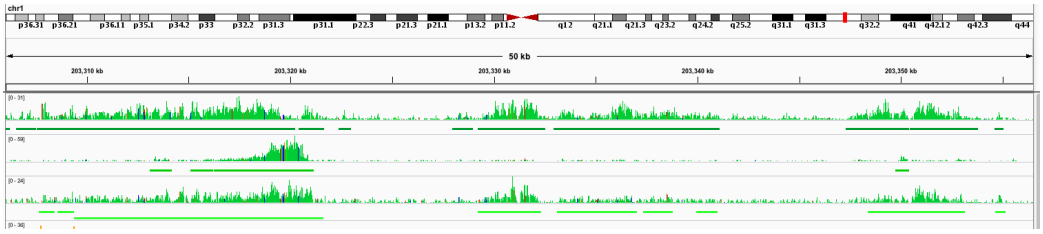


Figure 2.1: IGV snapshot for H3K4me1 (upper track), H3K4me3 (middle track) and H3K27ac (lower track)

2.3 Statistical methods

2.3.1 Probability distributions

In Statistics, a probability distribution is a function which describes a random phenomenon in terms of the probabilities of events [30].

A discrete probability distribution, also known as *probability mass function* (PMF), assigns probabilities p to an enumerable set of N different events characterized by an ordered, real variate x_i, such that

$$\sum_{i=1}^{N} p(x_i) = 1 \tag{2.2}$$

On the other hand, in a continuous probability distribution the probability P to find the random variable x in the interval $[x_1, x_2]$ is given by

$$P(x_1 < x < x_2) = \int_{x_1}^{x_2} f(x)dx \tag{2.3}$$

where $f(x)$ is known as *probability density function* (PDF). A discrete probability distribution can be described as continuous if one takes into account a set of Dirac delta functions.

Distributions have characteristic quantities associated to them that can be derived as *moments*. The n-th moment of $f(x)$ and $p(x)$ respectively is

$$\mu_n = E(x^n) = \int_{-\infty}^{\infty} x^n f(x)dx \tag{2.4}$$

and

$$\mu_n = E(x^n) = \sum_{k=1}^{\infty} x_k^n p(x_k) \tag{2.5}$$

The first moment is the mean μ; the second moment is the variance σ^2, which is related to the width of the distribution; the third moment is the skewness,

which is a measure of asymmetry; the fourth moment is termed kurtosis and it is related to the shape of the tails of the distribution.

The Gaussian distribution The Gaussian or Normal distribution with mean μ and variance σ^2 is defined as

$$N(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.6}$$

Many distributions arising from natural phenomena are well approximated by the Normal distribution. The reason is the effect of the *central limit theorem*, according to which the mean value of a large number N of independent random variables, obeying the same distribution with variance σ_0^2, approaches a Normal distribution with variance $\sigma = \sigma_0^2/N$.

The central limit theorem applies also when the individual variates follow different distributions provided that the variances are of the same order of magnitude.

The Negative Binomial distribution The Negative Binomial is a discrete probability distribution of the number of successes k in a sequence of independent and identically distributed Bernoulli trials (a random experiment with two possible outcomes, "success" with probability p or "failure" with probability $1-p$) before a specified number of failures r occurs.

The probability mass function of the Negative Binomial is

$$NB(k|r,p) = \binom{k+r-1}{k} \cdot (1-p)^r p^k \tag{2.7}$$

Its mean and its variance are respectively

$$\mu = \frac{pr}{1-p} \qquad \sigma^2 = \frac{pr}{(1-p)^2} \tag{2.8}$$

The Negative Binomial can be used as an alternative to the Poisson distribution to model count data. It is especially useful for discrete data over an unbounded positive range whose sample variance exceeds the sample mean. In such cases, the observations are overdispersed with respect to a Poisson distribution, for which the mean is equal to the variance, thus making the Poisson distribution not an appropriate model. Since the negative binomial distribution has one more parameter than the Poisson, this second parameter can be used to adjust the variance independently of the mean. This effect rises from the fact that the data is affected by an unobserved heterogeneity.

The Negative Binomial has been successfully employed in gene expression experiments [31]. One of the goal of RNA-seq data analysis is to find genes that are differentially expressed across groups of samples, and the statistical approaches have to account for small replicate numbers, discreteness, large dynamic range and the presence of outliers. In this context, read counts can be modeled as following a negative binomial distribution.

This approach can be extended for other comparative high-throughput sequencing assays, including ChIP-seq, chromosome conformation capture and chromatin accessibility experiments.

The Beta distribution The Beta distribution is a continuous probability distribution defined on the interval $[0, 1]$ and parametrized by two positive shape parameters, α and β.

The probability density function is

$$Beta(x|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{2.9}$$

where the normalization coefficient

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{2.10}$$

is the Beta special function.

The beta distribution has been applied to model the behavior of random variables limited to intervals of finite length in a wide variety of disciplines.

Especially, the beta distribution has been shown to provide a good model for distances between random points. Srinivasa and Haenggi [32] have proven that the distribution of the distance to the n-th neighbor from an arbitrarily chosen point in a population of independently and uniformly distributed points in a set W of arbitrary shape is a beta distribution.

Mixture models A mixture model is a probabilistic model for representing the presence of subpopulations within an overall population. It corresponds to the mixture distribution of two or more probability distributions describing the individual subpopulations.

For example, a Gaussian Mixture Model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions (in general multivariate) with unknown parameters [33].

The model is defined as a weighted sum of M Gaussian density components given by

$$p(x|\lambda) = \sum_{i=1}^{M} \omega_i g(x|\mu_i, \Sigma_i) \tag{2.11}$$

where x is a D-dimensional data vector, ω_i, $i = 1, \ldots, M$ are the mixture weights and each Gaussian component is defined as

$$g(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \; exp\left\{-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\right\} \tag{2.12}$$

with mean vector μ_i and covariance matrix Σ_i. The mixture weights are constrained to

$$\sum_{i=1}^{M} \omega_i = 1 \tag{2.13}$$

In Genomics, Mixture Models are applied when subpopulations or bimodal effects are expected in the data. For example, they have been recently employed in gene expression experiments with transcripts coming from two species [34].

Inverse participation ratio The inverse participation ratio (IPR) is a metric which can be used to identify the flatness or, conversely, the sharpness of an empirical distribution. It is defined as

$$IPR = \frac{1}{\sum_i^N p_i^2} \tag{2.14}$$

where p_i is the probability for value (or bin, in a histogram representation) i. If the distribution is sharp, showing one or a few peaks well above a background, the IPR has a small value, tending towards 1, which is obtained for a given k such that

$$p_{(i=k)} = 1 \qquad p_{(i\neq k)} = 0 \tag{2.15}$$

On the other hand, if the distribution is flat the IPR tends towards the total number of values or bins N, which is the case for

$$p_i = \frac{1}{N} \qquad \forall i \in \{1, ..., N\} \tag{2.16}$$

Inverse participation ratio is frequently applied in various fields of Physics as a measure of localization [35] [36] [37].

2.3.2 Correlation coefficients

Two random variables or two sets of data have an *association* if they are statistically related. As a particular case of association, *correlation* implies a dependence, linear or monotony. Formally, *dependence* refers to any situation in which random variables do not satisfy a condition of probabilistic independence, therefore it does not necessarily imply causality.

Correlation assessments are applied extensively in many fields of biological and natural sciences.

Pearson correlation coefficient Pearson product-moment correlation coefficient is a measure of the linear dependence between two variables obtained by dividing their covariance by the product of their standard deviations.

The population correlation coefficient $r_{X,Y}$ between two random variables X and Y with expected values μ_X and μ_Y and standard deviations σ_X and σ_Y is defined as

$$r_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{2.17}$$

The sample correlation coefficient, on the other hand, is obtained by substituting the expected values with finite sums over the n terms:

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{2.18}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of n variates.

As a consequence, $r_{xy} = 1$ represents total positive correlation, $r_{xy} = -1$ total negative correlation or anticorrelation and $r_{xy} = 0$ no correlation.

Statistical significance of the Pearson correlation coefficient can be assessed with permutation tests, bootstrap or with a statistical test.

In the null case of no correlation, the statistic

$$t = r \cdot \sqrt{\frac{n-2}{1-r^2}} \tag{2.19}$$

is distributed like a Student's t-distribution with $\nu = n-2$ degrees of freedom [38]. As a consequence, the significance is higher for high numbers of terms n and great values of correlation.

Spearman correlation coefficient Spearman's rank correlation (ρ) is a nonparametric measure of statistical dependence between the ranking of two variables, i.e. the relative positioning of the observations within them. It is

calculated as the Pearson correlation coefficient between the ranked variables $Rank_X, Rank_Y$:

$$\rho = \frac{cov(Rank_X, Rank_Y)}{\sigma_{Rank_X}\sigma_{Rank_Y}} \tag{2.20}$$

While Pearson's correlation assesses only linear relationships, Spearman's correlation is suitable also for monotonic relationships. Significance can be calculated similarly as for the Pearson correlation coefficient.

The Phi coefficient The Phi coefficient ϕ is a measure of association between binary variables, which is equivalent to the Pearson's correlation coefficient between them. Given the $2x2$ contingency table

	Vector-2	
Vector-1	1	0
1	a	b
0	c	d

Table 2.1

The Phi coefficient is defined as

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \tag{2.21}$$

ϕ is related to a χ^2 statistic with $(2-1)x(2-1) = 1$ degree of freedom as

$$\chi^2 = n \cdot \phi^2 \tag{2.22}$$

and this equation allows to evaluate significance. Similarly to other correlation metrics, ϕ is significant if its value or the number of observations n are high.

2.3.3 Stochastic processes

A stochastic process is a probability model used to describe phenomena that evolve over time or space. As opposed to a deterministic process, it represents evolution of a system described by a variable whose change is subject to a random variation.

A stochastic process can be formalized as the joint probability density function of a collection of random variables $X(t)$ representing the evolution in time (or, depending on the system, space) from the initial state to state

at the time point t. This evolution is not deterministic, therefore it can't be described with a differential equation. In the most extreme cases, for two different time points t_1 and t_2, the variables $X(t_1)$ and $X(t_2)$ can even be stochastically independent, although processes of interest usually show correlation between the states.

Stochastic processes can be classified as

- Markovian or non-Markovian, if each state depends uniquely on its immediately previous one or not;
- with discrete or continuous time points
- with discrete or continuous state values

Stochastic processes can be described with stochastic differential equations, in which fluctuating terms with certain statistical properties are superimposed to a deterministic equation. They are used to model real systems for which the randomness and stochasticity can not be neglected.

Random walk The random walk is a well known stochastic process which describes a succession of random steps. It is an example of Markov chain with discrete time points and states.

In one dimension, it represents the movement of a particle in one of the two directions, with probability p and $1-p$. As a consequence, the state at time t_i only depends on the state at time t_{i-1}.

Formally, given a succession of independent random variables $Z_1, Z_2, ...$ which can take either the value $+1$ or -1 with probabilities p and $1-p$ respectively, the random walk S_n can be defined as the series

$$S_n = \sum_{j=1}^{n} Z_j \tag{2.23}$$

which corresponds to the position after n steps of the walk. The mean value is

$$E(S_n) = \sum_{j=1}^{n} E(Z_j) = 0 \tag{2.24}$$

whereas the expected translation distance $E(|S_n|)$ after n steps is of the order of $\sqrt{n}$.

The random walk has been used as a model for both DNA sequences and biological networks for disease gene identification [39] [40].

Hidden Markov model A hidden Markov model (HMM) is a statistical model in which the system is assumed to be a Markov process with unobserved states.

In simpler Markov models the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a hidden Markov model, the state is not directly visible, but the output, dependent on the state, is visible. Each state has a probability distribution over the possible outputs. Therefore, the sequence of outputs generated by an HMM gives some information about the sequence of states.

Formally, the observed output sequence can be denoted as $X = x_1x_2...x_L$ and the underlying state sequence as $Y = y_1y_2...y_L$, where y_n is the underlying state of the n-th observation x_n. Each output takes on a finite number of possible values from the set of observations $O = \{O_1, O_2, ..., O_N\}$ and each state takes one of the values from the set of states $S = \{1, 2, ..., M\}$.

As a Markov chain, the state j only depends on state i occurring at the immediately previous time point. The probability for making the transition from i to j is termed *transition probability* $t(i, j)$. The *initial state probability* is defined as $\pi(i) = P\{y_1 = i\{$, for all $i \in S\}$. Finally, the *emission probability* that the n-th observation $x_n = x$ only depends on the underlying state y_n, hence $P\{x_n = x|y_n = 1\{= e(x|i)$ for all possible observation, states and time points.

The three probability measures completely specify an HMM, and this set of parameters can be denoted as Θ.

Therefore, the probability that the HMM will generate the observation sequence $X = x_1x_2...x_L$ with the underlying state sequence $Y = y_1y_2...y_L$ can be computed as

$$P\{X, Y|\Theta\{= P\{X|Y, \Theta\{P\{Y|\Theta\{ \tag{2.25}$$

where

$$P\{X|Y, \Theta\{= e(x_1|y_1)e(x_2|y_2)...e(x_L|y_L) \tag{2.26}$$

$$P\{Y|\Theta\{= \pi(y_1)t(y_1, y_2)t(y_2, y_3)...t(y_{L-1}, y_L) \tag{2.27}$$

The model can be learned given the output sequences by finding the best set of transition and emission probabilities with parameter estimation algorithms such as maximum likelihood estimators.

Hidden Markov models frequently find applications in Genomics [41]. An useful implementation is the automatic chromatin state discovery and characterization realized by the ChromHMM software [42].

As reported before, histone modification marks assayed with ChIP-seq experiments are characteristic of chromatin states with different functional

roles, such as promoters and enhancers. ChromHMM is based on a multivariate Hidden Markov Model that allows to learn such hidden chromatin states directly from histone marks, modeled as outputs, by estimating the transition and emission probability parameters. As input, it receives a list of aligned reads for each chromatin mark, which are automatically converted into presence or absence calls for each mark across the genome, based on a Poisson background distribution. ChromHMM then outputs both the learned chromatin state model parameters and the state assignments for each genomic position, thus providing an efficient way to automatically extract enhancer and promoter regions from ChIP-seq experiments.

2.4 Fourier analysis

In mathematics, Fourier analysis is the study of the way general functions can be represented by sums of simpler trigonometric functions, and it finds applications in many fields of science, engineering and computation. The Fourier transform can be an efficient computational tool for accomplishing certain common manipulations of data or, in other cases, it is itself of intrinsic interest [38]. In signal processing, the decomposition into periodic functions provides insights into characteristic frequencies and their contribution to the signal.

For a one dimensional continuous function $f(x)$ of a real variable x, the Fourier transform $F(u)$ is a linear operation defined as

$$F(u) = \int_{-\infty}^{\infty} f(x) e^{-i2\pi ux} dx \tag{2.28}$$

Similarly, the inverse transform expresses the original function in terms of its Fourier transform as

$$f(x) = \int_{-\infty}^{\infty} F(u) e^{i2\pi ux} dx \tag{2.29}$$

where x is the variable of the *spatial* or *temporal domain* and u of the *frequency domain*.

If, on the other hand, the domain is discrete, such as for a set of M sampled data points, the Discrete Fourier Transform (DFT) is defined as a Fourier series:

$$F(u) = \sum_{x=0}^{M-1} f(x) e^{-i2\pi ux/M} \tag{2.30}$$

and the inverse DFT is

$$f(x) = \sum_{x=0}^{M-1} F(u) e^{i2\pi ux/M} \tag{2.31}$$

In this way, the frequency components of a periodic or semi-periodic signal can be investigated.

An efficient implementation of the Discrete Fourier Transform is the Fast Fourier Transform (FFT), which reduces the computational costs from $\mathcal{O}(M^2)$ to $\mathcal{O}(M \log M)$.

Genomes show periodicities at many levels, from DNA sequences to chromatin organization. As a consequence, Fourier analysis is a powerful tool in Genomics, and it is applied frequently. For example, differential analysis of genomic sequences belonging to genes or to non coding regions have pointed to a different Fourier spectrum capable of predicting gene positioning [43].

2.5 Graph models

For structured data, graphs provide an intuitive yet effective model which is frequently adopted in many fields of science.

In Genomics, graphs are frequently applied to model biological interactions at many levels.

For example, Gene Regulatory Networks is a class of graphs defined by collections of molecular regulators that interact with each other and with other substances in the cell to govern the gene expression levels of mRNA and proteins. These regulators can be DNA, RNA or proteins, and the interactions can be direct or mediated by RNAs or proteins. The study of such networks can provide insight into the underlying biological principles and allow *in silico* predictions and experiments, such as how gene expression is affected if one or more nodes of the network are disrupted.

More frequently, graphs have been applied to chromatin conformation data. Hi-C experiments provide chromosome-wise and genome-wide maps of interactions between genomic loci, which can be modeled as networks. In this way, the structure and folding principles of chromatin can be investigated by analyzing the properties of the graphs.

2.5.1 Basic definitions

A graph or network, denoted as $G = (V, E)$ consists of a finite and non-empty set of vertices or nodes V and a set of edges E, which are 2-element subsets of V, namely node pairs [44].

Graphs can be *undirected* if the edges are symmetric or *directed* if they rather have an orientation, from one node to the other. If the edges represent relationships of variable intensity they are associated with weights, and in this case the graph is termed *weighted*. An undirected graph is *connected* if for each pair of vertices there exist a path composed by one or more edges that connect them. A *bipartite* graph is a graph whose vertex set can be partitioned into two subsets V_1 and V_2 such that the edges connect only nodes belonging to V_1 to nodes belonging to V_2.

The *connectivity degree* of a node is the number of nodes it is connected to or, for weighted graphs, the sum of the weights of the edges it shares.

2.5.2 Matrix representations

A natural and useful way to represent graphs is in matrix form.

Adjacency matrix A non weighted graph with N nodes can be represented with a square $N \times N$ adjacency matrix A where for each pair of nodes i, j $A_{i,j} = 1$ if i and j are connected and $A_{i,j} = 0$ otherwise. If the graph is weighted, $A_{i,j} = w_{i,j}$, namely the value of the edge between i and j.

The adjacency matrix is symmetric for undirected graphs, and the main diagonal contains only $0's$ if the graph does not have loops (nodes connected to themselves).

Link matrix The link matrix is a $L \times 2$ matrix where L is the number of edges in the graph. Each row contains the first and the second node connected by each edge. It provides an efficient representation for sparse graphs, i.e. for graphs with a small number of edges which would give rise to a sparse adjacency matrix.

Laplacian matrix The Laplacian matrix is defined as $L = D - A$, where A is the adjacency matrix of the graph and D is the diagonal matrix containing each node's connectivity degree. The Laplacian matrix is an useful representation for certain types of network analysis, for example for finding the disconnected components and subgraphs.

2.5.3 Centrality metrics

Analyzing a graph often involves assessing the relevance of its nodes. This can be accomplished through various centrality measures.

Degree centrality The simplest metric for the centrality of a vertex j is its connectivity degree, also termed degree centrality C_D. As reported before, it is calculated as the number of nodes connected to the given vertex or the sum of the edge weights shared by it, in case of a weighted graph:

$$C_D(j) = \sum_{k|(j,k)\in E} w_{j,k} \tag{2.32}$$

For a directed graph, degree centrality is divided into *indegree* and *outdegree*, taking into account the direction of the edges connected to the given vertex.

The distribution of the degree centrality provides information about properties of the graph such as scaling and global connectivity.

Betweenness centrality The betweenness centrality C_B quantifies the number of times a node j acts as a bridge along the shortest path between two other nodes. It can be defined as

$$C_B(j) = \sum_{s\neq j\neq t} \frac{\sigma_{st}(j)}{\sigma_{st-tot}} \tag{2.33}$$

where σ_{st-tot} is the total number of shortest paths from node s to node t and $\sigma_{st}(j)$ is the number of those paths that pass through j.

The betweenness centrality of a node scales with the number of pairs of nodes (excluded the given node itself) as implied by the summation indices. As a consequence, the metric can be rescaled by dividing by $(N-1)(N-2)/2$ such that $C_B \in [0,1]$. This is particularly advisable when comparing betweenness centralities across different graphs.

Salience centrality Another metric based on shortest paths is salience centrality [45]. For a given node j, the collection of shortest paths to all other nodes defines the shortest-path tree $T(j)$. $T(j)$ can be represented as a $N \times N$ matrix with $1's$ in the positions associated with the links belonging to at least one of the shortest paths and $0's$ elsewhere. Salience can be defined as the average shortest path tree

$$S = \frac{1}{N}\sum_k T(k) \tag{2.34}$$

and salient centrality C_S as the sum of the salience of each link shared by the given node j.

Chapter 3

Statistical properties of open chromatin profiles and chromatin loops

3.1 Chromatin accessibility in single cells - state of the art

Following very recent technological development, single cell high-throughput experiments are gaining popularity in Genomics. The ability to assay single cells allows to investigate biological mechanisms with an unprecedented accuracy, since the issues concerning the effects of smoothing introduced by the heterogeneous ensembles of cells are thus overcome. Most notably, single cell sequencing gives insights into cell-cell heterogeneity, thus providing a new framework to characterize cell state in a more precise way.

While single-cell RNA-sequencing is now mature and frequently adopted, development of assays for chromatin accessibility in single cells has been realized only in the last two years.

3.1.1 scDNase-seq

Jin *et al.* [14] managed to overcome the limitation of the high input material requirement of DNase-seq technique with scDNase-seq. Single cells were separated by flow cytometry. To prevent loss of the extremely small amount of DNase I hypersensitive DNA released by DNase I digestion of single cells, a large amount of circular plasmid DNA as carrier DNA was added. Before sequencing, the fragments were amplified by PCR with a method capable of increasing the amplification of genomic DNA versus carrier DNA.

The technique was applied to tumor samples coming from patients affected by thyroid cancer as well as to normal cells.

Since DNase-seq experiments on ensembles of cells were also performed, the DNase I hypersensitive sites (DHSs) could be compared. Each single cell showed an occupancy of 35-59% of the ensemble DHSs.

By integrating single cells RNA-seq experiments, it could be found that the single-cell DHSs predict enhancers that regulate cell specific gene expression programs and the cell-to-cell variations of DHS are predictive of gene expression.

Finally, a comparison of the DHSs in tumor versus normal samples detected thousands of tumour-specific DHSs, many of which were associated with promoters and enhancers critically involved in cancer development.

3.1.2 scATAC-seq

Developed by Buenrostro *et al.* [17], scATAC-seq is a fast and relatively cheap technique to assay single cell open chromatin profiles. The experimental procedure, described previously, is less complicated and only require an efficient way to isolate single cells.

Buenrostro *et al.* realized the isolation with a programmable microfluidics platform, which divided the cells into wells that could be observed under an optical microscope and hosted the complete reaction procedure. In this way, single cells from K562 chronic myelogenous leukaemia and GM12878 lymphoblastoid cell lines were investigated.

Single cells accessibility profiles overlapped known DHSs, and cell-cell heterogeneity was explored.

Single cell ATAC-seq was also performed by Cusanovich *et al.* [46] with a combinatorial indexing approach and without physically dividing the cells. First, populations of cell nuclei were molecularly barcoded in each of many wells. Then intact nuclei were pooled, diluted, and redistributed to a second set of wells, where a second barcode was introduced. Because the overwhelming majority of nuclei pass through a unique combination of wells, they are compartmentalized by the unique barcode combination that they receive.

3.2 Investigation of primary leukemia B-lymphocytes with scATAC-seq

3.2.1 Chronic Lymphocytic Leukemia - overview

Chronic Lymphocytic Leukemia (CLL) is the most prevalent type of leukemia in the western world, accounting for $\sim 40\%$ of all adult leukemias [1]. Its incidence increases with age and reflects a combination of environmental and genetic factors.

CLL is a malignancy of mature clonal B-lymphocytes that grow in an uncontrolled manner and accumulate in the blood, bone marrow and other lymphoid tissues for extensive periods of time. Despite the homogeneous morphological and immunological phenotype, the clinical outcome of CLL is variable, reflecting the existence of two main disease subtypes, defined by the mutational status of the variable region of the immunoglobulin (Ig) genes. Mutated Ig chain (IGHV-M) is associated with a good prognosis, whereas CLL with unmutated IGHV (IGHV-UM) is known to be more aggressive [1] [2]. The current standard therapies for CLL involve combinations of chemotherapy and drugs impacting B-cell receptor (BCR) signaling, such as *ibrutinib* [1].

CLL is also known to be an epigenetic disease, since it shows subtype-specific epigenome signatures [2] and remarkably elevated levels of histone deacetylase (HDAC) isoenzymes [47]. As reported before, histone acetylation is a modification associated with enhancer activity; its depletion can therefore significantly affect gene expression regulation, even though it is not fully understood how these processes are involved in CLL clinical development. Nevertheless, there are epigenetic drugs known as HDAC inhibitors (HDACi) capable of counteracting histone deacetylaytion, for example *panobinostat* [48]. These drugs could provide effective and less aggressive therapeutic alternatives to chemotherapy, treating the disease in a targeted and precise way.

As a consequence, in order to more deeply understand the mechanisms of the disease and to implement the new approaches mentioned, detailed characterizations of the epigenetic profiles and heterogeneity of CLL are needed.

3.2.2 ATAC-seq experiments

In this work, raw data coming from scATAC-seq assays performed on B-lymphocytes of CLL patients are analyzed. ATAC-seq experiments have

been performed[1] in the laboratories of Genome Organization and Function research group of the Bioquant and the German Cancer Research Center (DKFZ), in Heidelberg, as part of a project aiming at characterizing the disease with various techniques such as ChIP-seq, RNA-seq and ATAC-seq, both in ensembles of cells and at the single cell level.

For scATAC-seq, cells from three patients (who will be referenced as *B01*, *B02* and *B03*) have been investigated. Both Single-End and Paired-End sequencing have been performed, with a read length of 51 bp. The quantities of single cells for each experiment are summarized in Table 3.1.

Patient	# Single-End	# Paired-End
B01	89	48 + 74
B02	38	-
B03	65 + 85	70 + 47

Table 3.1: Composition of data set. "+" denotes the presence of two biological replicates.

3.2.3 Bioinformatic processing of raw data

Raw data coming from the sequencing facilities were in FASTQ format. Sequencing quality has been inspected with FastQC tool [24] and have turned out to be above 30. The high quality was also due to the short read length.

Single-End reads for each single cell have been aligned to a a reference human genome (hg19) with Bowtie [25], allowing for a maximum of 2 mismatches. Reads not mapping uniquely (i.e., to multiple locations of the reference genome) were discarded. Paired-End reads were aligned with Bowtie2 [49], which is known to perform better than Bowtie for Paired-End data, allowing fragments up to 2000 bp to align. Bowtie2 has a more advanced strategy to score aligned reads than Bowtie: since discarding non-uniquely mapping reads with mismatches can be too conservative and therefore it can lead to throw away a lot of data, Bowtie2 assigns a score to each alignment based on how "unique" it is. As a result, alignments with a low score can be filtered out afterwards. In this case, a threshold of 10 has been chosen. A significant number of reads presented adapter contamination, containing the whole Tn5 transposase adapter or a substring of it. These reads

[1] by Dr. Jan-Philipp Mallm, Lara Klett and Sabrina Schumacher, all members of Genome Organization and Function research group headed by PD DR. Karsten Rippe

were trimmed with a custom Python [50] script, taking advantage of Biopython [51] library functions, and then aligned again.

The resulting .sam files were converted to .bam with SAMtools [26], sorted and indexed with IGVtools for visualization on IGV Genome Browser [29]. A second conversion to .bed format was performed and reads mapping to mitochondrial DNA were removed.

For each read, the true insertion position of the Tn5 transposase was calculated as the read start plus 4 bp for the "+" strand and minus 4 bp for the "-" strand. The 29-bp binding region was then determined by extending the insertion position by 14 bp on both sides with a Perl [52] script. Afterwards, exact duplicates (i.e. regions with the same start and end positions) were collapsed to one entry in each single-cell .bed file, since they mostly constituted PCR amplification artifacts. In this way, however, insertions in homologous chromosomes that might have by chance occurred in the same genomic positions were lost, but this distortion is marginal.

After this processing, unusual peaks of mapped reads in certain genomic regions were found. This phenomenon was clearly an artifact since many ($> 10 - 20$) reads occurred at a distance of 4-6 bp, and this can't be explained biologically, since a single cell can contain a maximum of 4 copies of the genome (when the cell is in a replicative state). It turned out that these regions overlapped the High Mappability Islands thus defined by the ENCODE [22] consensus blacklist of hg19[2]. These blacklisted regions are located within highly repetitive portions of the genome but nevertheless a lot of sequencing reads from various experiments usually manage to pass the uniquely mapping filter, thus generating a biased high signal. All of the reads falling within them were therefore removed.

In many cases more than 4 insertions still occurred within a 51-bp window on the same strand orientation. This effect was probably caused by random trimmings of the fragments during the PCR amplification. A second collapsing operation was therefore performed on the data.

These pre-processing steps had to be performed in such an accurate way because given the low amount of reads, even relatively small artifacts could have a high impact on the downstream analysis.

3.2.4 Tn5 sequence specificity

As mentioned before, ChIP-seq data have shown that transcription factors are more likely to bind specific patterns of nucleotides known as motifs. In

[2] `http://www.broadinstitute.org/~anshul/projects/encode/rawdata/blacklists/hg19-blacklist-README.pdf`

order to verify whether or not the Tn5 transposase have a similar sequence preference, all the 29-bp insertion regions from all the single cells have been aggregated and the actual sequences have been extracted from hg19 with *getFasta* function of BEDtools [27]. A Position Weight Matrix (PWM), a 4×29 matrix with relative frequencies of each of the 4 nucleotides along the insertion region, has thus been calculated with Biopython functions [51] and is represented in Figure 3.1.

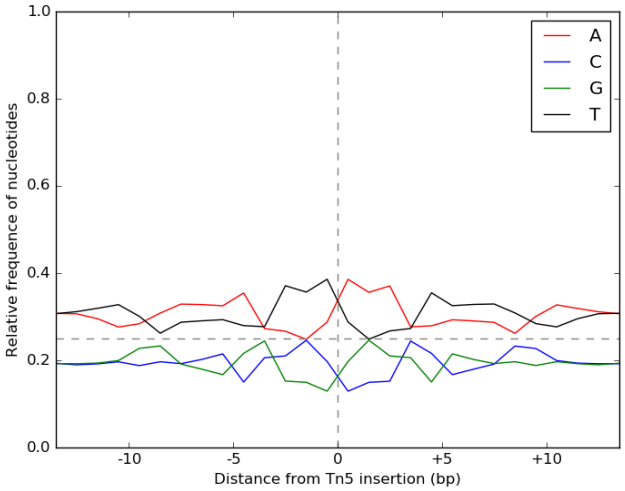


Figure 3.1: Tn5 transposase sequence specificity

As expected, the plots of complementary nucleotides (A-T, C-G) are symmetric with respect to the exact insertion position since the sequencing of the "+" or the "-" strand is equally probable. It can be noticed that the frequencies are closer to their average values in the human genome (with a GC content of $\sim 40\%$) at the boundaries of the binding region. On the other hand, there are fluctuations around the exact insertion position, even though there is no clear predominance and no frequency exceeds 0.4. The PWM calculated here is consistent with the one reported by Buenrostro *et al* [15]. In conclusion, a binding specificity is present but it is not too prominent.

3.3 Statistical properties and characterization

3.3.1 Single-End data

The downstream data analysis has been performed with MATLAB [53]. After all the filtering operations, the number of unique insertions in a single cell ranges from a minimum of 2588 to a maximum of 252438. The histogram distribution with 60 bins of the total number of insertions is shown in Figure 3.2 with various fitting functions. Particularly, as previously mentioned the Negative Binomial has been used in literature to model count data [31].

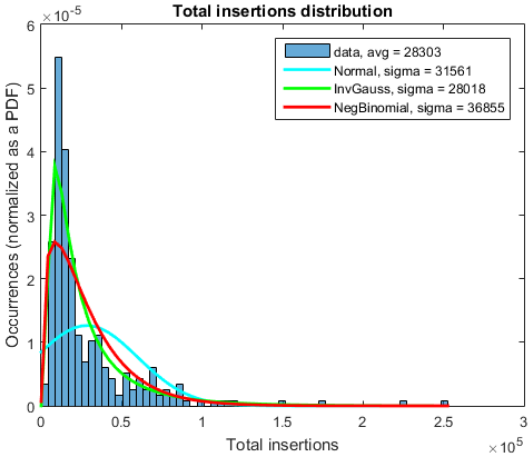


Figure 3.2: Histogram of total insertions across cells

Despite having followed the same experimental procedure, the average number of insertions differs patient-wise, with patient B02 yielding about twice as much as patient B01. The reason for this heterogeneity of the insertion efficiency is not clear, but the right tail values are hypothesized to represent chromatin that completely unfolded, causing the Tn5 transposase to insert anywhere. As a consequence, cells with a number of insertions 3 standard deviations (calculated with the Negative Binomial fit) above the average were discarded as well as cells with less than 0.2 times the average. With these thresholds the number of single cells decreases from 277 to 265.

3.3.2 Paired-End data

A similar analysis performed on Paired-End data showed only one clear outlier, which had a number of insertions more than 10 times above the average. Given the fact that with Paired-End sequencing both the ends of each DNA

fragment are detected, Paired-End data show a significantly higher average number of insertions than Single-End (after filtering, 37830 versus 26440), thus enabling a more robust analysis.

The composition of the data set after the filtering operations is summarized in Table 3.2.

Patient	# Single-End	# Paired-End
B01	79	48 + 74
B02	38	-
B03	63 + 85	70 + 46

Table 3.2: Filtered data set

Paired-End data retain information not only on the positions of accessible chromatin but also on DNA fragments. Fragment lengths have been calculated with a Perl script, and their distribution is shown in Figure 3.3.

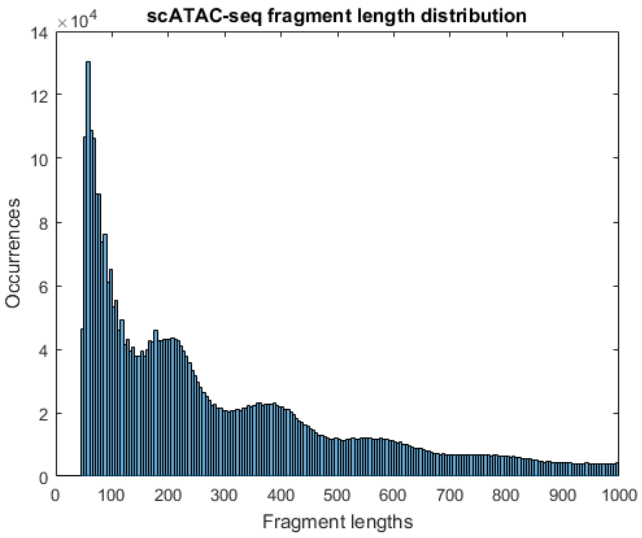


Figure 3.3: Distribution of fragment lengths

It can be noticed that the distribution has a remarkable periodicity of ~ 200 bp. The peaks corresponding to integer multiples of 200 bp represent fragments whose two ends are located at opposite sides of an array of one or more nucleosomes, since DNA is wrapped around them for ~ 147 bp. As a consequence, fragments whose length is significantly different from a multiple of 200 bp are less likely to occur, and therefore less frequent.

A zoom-in of the fragment length distribution over the small lengths is reported in Figure 3.4.

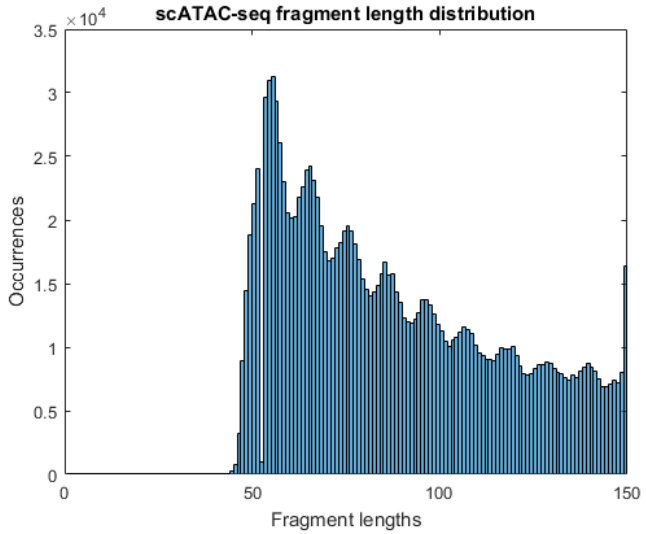


Figure 3.4: Distribution of fragment lengths - zoom-in

Another type of periodicity can be seen here, with peaks every ~ 10 bp. This feature reflects the pitch of the DNA double helix, which has a similar size.

To assess these periodicities in a more accurate way, it is possible to perform a Fourier analysis. The counts of the fragment length distribution can be interpreted as a semi-periodic signal with spatial periods. A Fast Fourier Transform of this signal, shown in Figure 3.5, yields a decaying Fourier spectrum, since the signal itself is not purely periodic. The most intense component is the zero frequency, also referred to as "continuous component", but two peaks are also present: the first one is located around a frequency of 0.006 bp^{-1}, which gives a period of 167 bp, whereas the second one corresponds to a frequency of 0.0941 bp^{-1} and consequently a period of around 10.6 bp. These values are therefore even closer to the theoretical ones for the nucleosome arrays and the DNA helical pitch respectively.

Fitting the Fourier spectrum with a power law (Figure 3.6) shows a behavior of $S(f) \sim f^{-0.82}$. This exponent is consistent with the model of "pink noise" or "$1/f$ noise", for which the low frequency components are dominant, as in this case.

Both the nucleosome and the helical periodicities have also been reported in other studies [16] [15] [17].

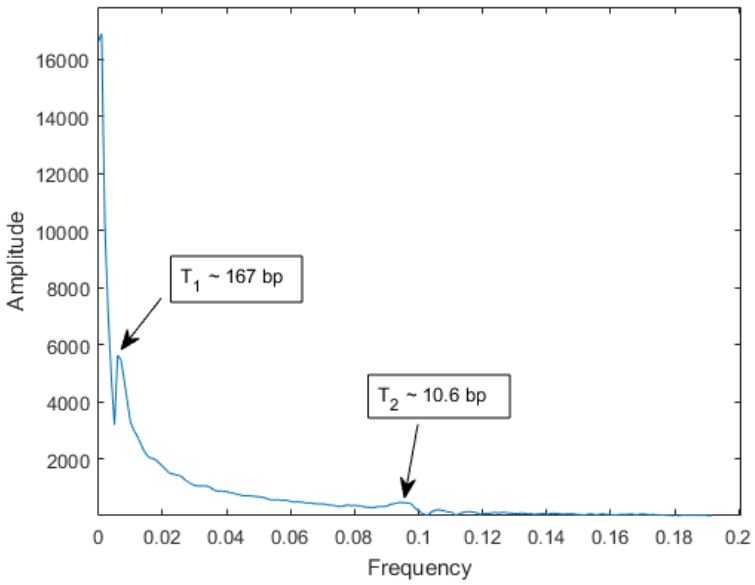


Figure 3.5: Fourier spectrum for the fragment length distribution

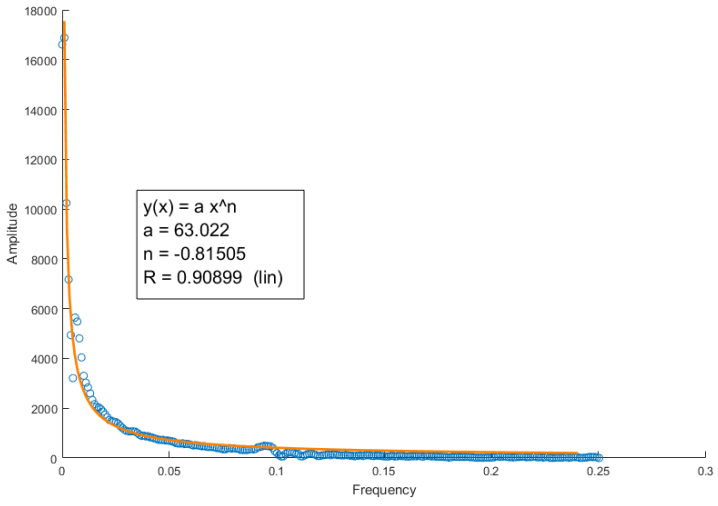


Figure 3.6: Power law fitting of the Fourier spectrum

3.4 Characteristic distances and chromatin loops

In Genomics, distances between features along the DNA can reveal statistical properties related to biological features, and as reported before stochastic processes such as random walks and HMM provide good model for such features.

For example, Paci *et al.* [54] investigated the distances between each dinucleotide along the genome for many species. A significant difference was found between mammals and other species for the distribution of the CG dinucleotide distances, pointing to the fact that the different functional role of DNA methylation in mammals is reflected by its peculiar statistical properties.

3.4.1 Distributions of insertion distances in scATAC-seq data

Since Tn5 transposase insertions tag only open and uncondensed regions of the DNA, it can be hypothesized that the its spatial distribution along the genome is capable of revealing structural properties of chromatin.

It has already been shown by the fragment length distribution in Paired-End data that nucleosomes protect the DNA, thus making it inaccessible; in order to investigate if this feature can be seen also for higher-order structures at bigger length scales, it is possible to calculate the distances between consecutive insertions of the transposase and observe the shape of the distribution they give rise to.

As shown in Figure 3.7 for Single-End data, this distribution is clearly dominated by small distances, probably since a given open region could host many close by insertions. As a consequence, a distribution of the distances in logarithmic scale (with a base of 10), plotted in Figure 3.8, can be more informative.

It can be noticed that the latter is remarkably bimodal.

The first peak, located around an exponent of 3, shows the small oscillations characteristic of the nucleosome periodicity, the first of which is, as expected, close to 200 bp.

On the other hand, the second peak is very pronounced and is located around an exponent of 5. This peak can be hypothesized to represent a characteristic order of magnitude for the length of chromatin loops. As previously mentioned, loops have actually a length of about 10^5 bp. Moreover, they are known to bring distal regulatory elements into close spatial proximity, and a necessary condition for their activity is their accessibility. As a consequence, it is reasonable that the two ends of a loop are simultaneously open.

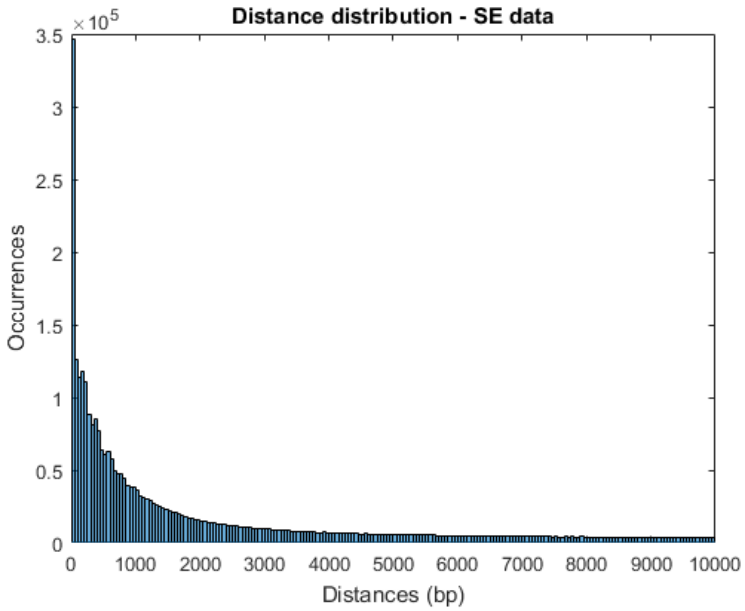


Figure 3.7: Distance distribution up to 10000 bp

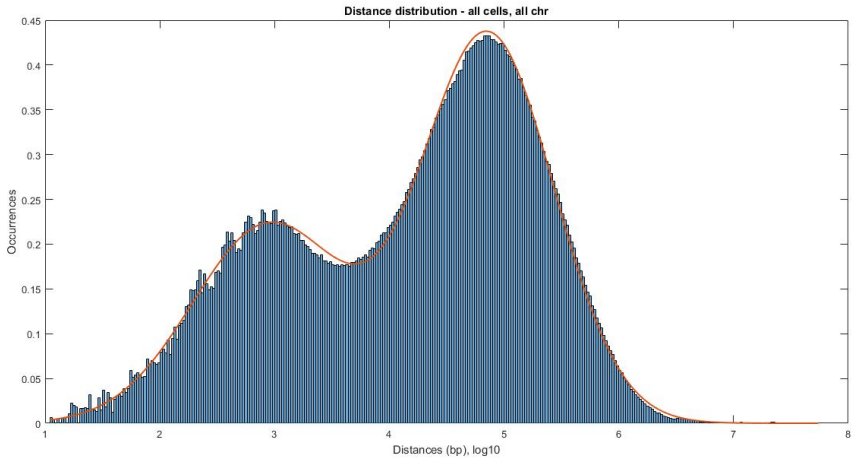


Figure 3.8: Distribution of logarithmic distances - SE data

In order to quantify the peaks, a Gaussian Mixture Model has been fitted to the data with MATLAB *gmdistribution.fit*, and the resulting function has been overlaid to the plot as a red line.

Bayesian parameter estimation with Expectation-Maximization algorithm yields the parameters

$$\lambda = \{\omega_i, \mu_i, \Sigma_i\} \quad i = 1, \dots, M \tag{3.1}$$

In this case, the model is fitted on 1-dimensional data (the logarithmic distances) for two Gaussians

$$p(x|\omega_1, \mu_1, \sigma_1, \omega_2, \mu_2, \sigma_2) = \omega_1 \frac{1}{\sqrt{2\pi}\sigma_1} e^{\left(\frac{x-\mu_1}{2\sigma_1^2}\right)} + \omega_2 \frac{1}{\sqrt{2\pi}\sigma_2} e^{\left(\frac{x-\mu_2}{2\sigma_2^2}\right)} \tag{3.2}$$

The resulting parameter values are reported in Table 3.3.

	ω	μ	σ
Peak 1	0.37	2.9572	0.4482
Peak 2	0.63	4.8581	0.3322

Table 3.3

A similar trend can be noticed in Paired-End data (Figure 3.9a). However, in this case the oscillations caused by the nucleosome periodicity are more pronounced in the first peak.

The irregular shape does not make the distribution for Paired-End data a good candidate for a Gaussian Mixture Model fit. As a consequence, a standard Gaussian fit can be performed on the distribution truncated around the second peak. This fitting and the following ones have been realized with R package *fitdistrplus* [55] [56]. In Figure 3.9 the fit and the P-P plot of empirical versus estimated probability are shown.

The parameter values and the goodness of fit calculated with a Kolmogorov-Smirnov test (K-S gof) are reported in 3.4.

μ	σ	K-S gof
5.1294	0.5468	0.024

Table 3.4

The second peak of Paired-End data is therefore only slightly greater than the corresponding one in Single-End data.

In order to validate this feature of chromatin accessibility maps, Paired-End data originated from another scATAC-seq experiment on cells from K562 cell line (Buenrostro *et al.*, 2015 [17]) have been downloaded and analyzed in the same way.

The distribution of logarithmic distances is very similar to the one presented before for CLL samples (Figure 3.10a.

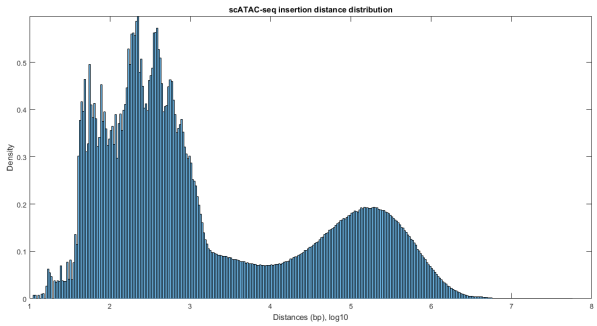
The results of the Gaussian fit of the second peak (Figure 3.10) are reported in Table 3.5.

μ	σ	K-S gof
4.9654	0.5375	0.036

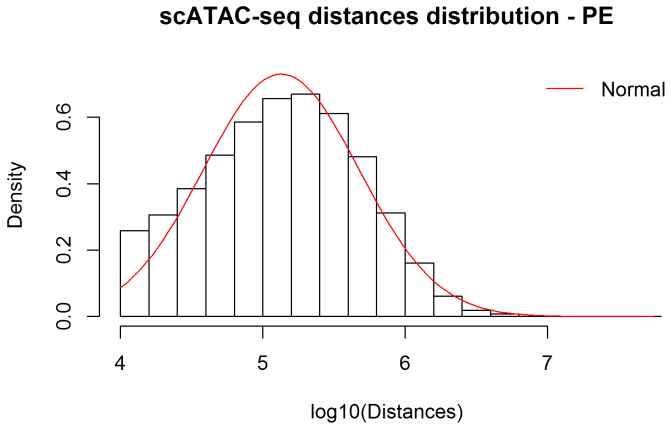
Table 3.5

From the three datasets, the characteristic exponent of the second peak can therefore be estimated by their average value μ_{ATAC} and its standard deviation as $\sigma_{ATAC} = \sum_{i=1}^{3} \sigma_i^2/3$:

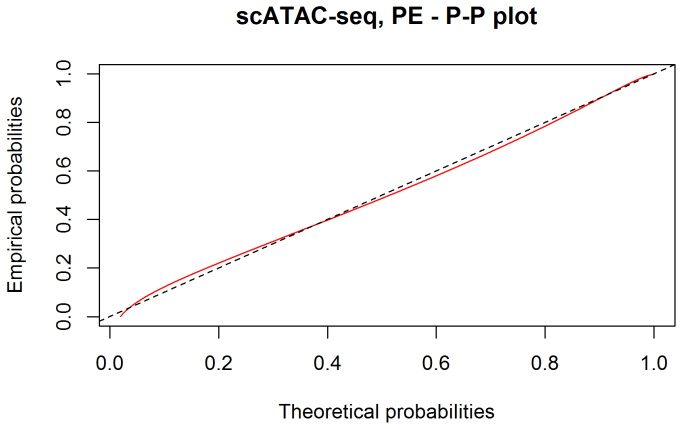
$$\mu_{ATAC} = 4.98 \quad \sigma_{ATAC} = 0.28 \tag{3.3}$$



(a) Distribution of logarithmic distances - PE data

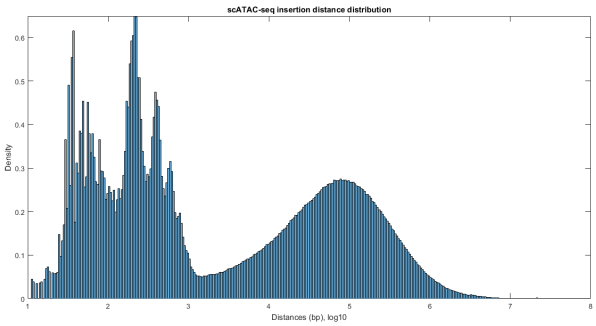


(b) Gaussian fit of the second peak

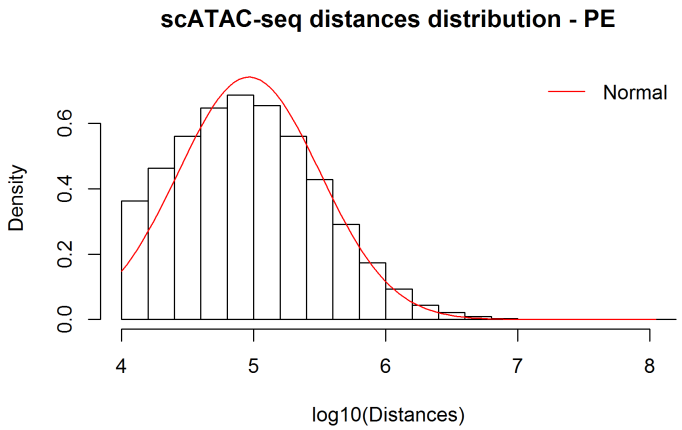


(c) P-P plot of the Gaussian fit

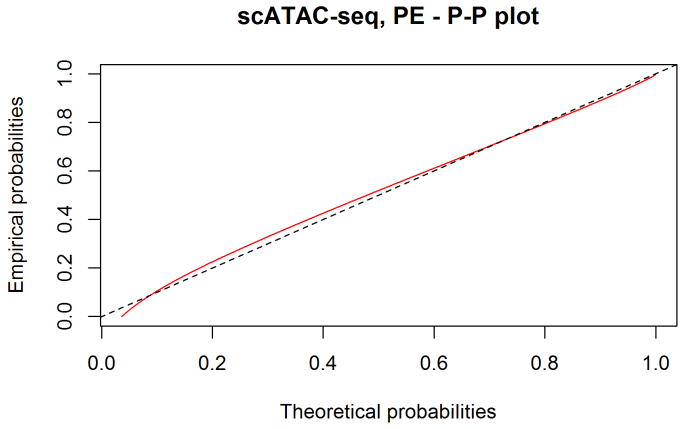
Figure 3.9: Paired-End scATAC-seq distances



(a) Distribution of logarithmic distances - PE data



(b) Gaussian fit of the second peak



(c) P-P plot of the Gaussian fit

Figure 3.10: Paired-End scATAC-seq distances, Buenrostro *et al.*

3.4.2 Distributions of anchor region distances in ChIA-PET data

In order to validate this characteristic length, ChIA-PET data for proteins involved in chromatin loops have been downloaded from GEO database [57] and analyzed. Experiments have been performed on a variety of cell lines (Table 3.6) in many labs [58] [59] [60] [61] [62] [63].

Protein	Cell line
CTCF	GM12878, HeLa, K562, MCF-7
ESR1	MCF-7
POLR2A	HCT116, HeLa-S3, MCF-7, NB4
RAD21	GM12878
RNAPII	GM12878, HeLa, HUVEC, K562, MCF-7
SMC1	T-ALL
H3K4me1	K562
H3K4me3	K562
H3K27ac	K562

Table 3.6

Proteins analyzed have been selected according to their known structural or functional role:

- CTCF and ESR1 are transcription factors; CTCF is known to act as an insulator between TADs or chromatin loops;
- POLR2A and RNAPII are RNA polymerases involved in transcription;
- RAD21 and SMC1 are proteins belonging to the cohesin complex, which is known to tight together the two ends of chromatin loops;
- H3K4me1, H3K4me3 and H3K27ac are histone modifications associated, as mentioned before, with the activity of enhancers and promoters.

The distributions of anchor distances for all the proteins taken into account but RNAPII (in Figure 3.11) show a peak around 10^5 bp. For POLR2A there is a peak also around 10^8 bp. Histograms and Gaussian fits are shown in Figure 3.12.

The parameters resulting from the fits are summarized in Table 3.7.

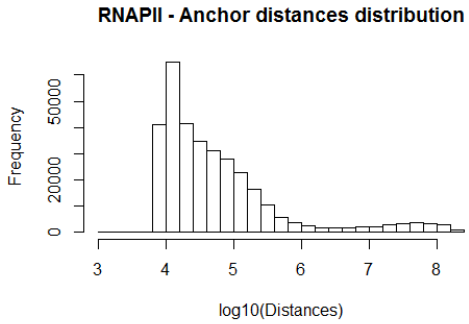


Figure 3.11: RNAPII

Protein	μ	σ	K-S gof
CTCF	5.1371	0.5302	0.02
ESR1	4.867	0.612	0.06
RAD21	5.174	0.408	0.02
SMC1	5.124	0.464	0.04
H3K4me1	4.981	0.290	0.06
H3K4me3	4.979	0.400	0.08
H3K27ac	4.945	0.288	0.06
POLR2A	4.739	0.556	0.07

Table 3.7

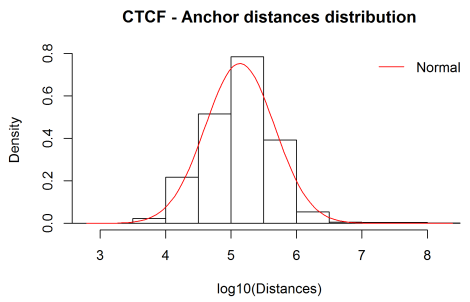
The average value of the exponent for ChIA-PET data and its standard deviation result

$$\mu_{ChIA-PET} = 4.99 \quad \sigma_{ChIA-PET} = 0.16 \tag{3.4}$$

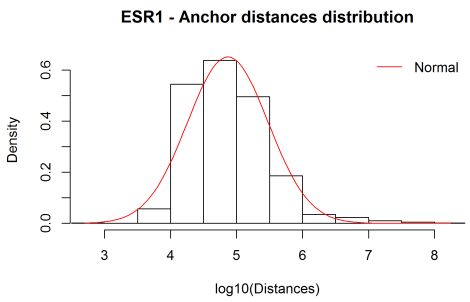
As a consequence, μ_{ATAC} and $\mu_{ChIA-PET}$ turn out to be very close to each other,

$$\mu_{ATAC} = 4.98 \pm 0.28 \approx \mu_{ChIA-PET} = 4.99 \pm 0.16 \tag{3.5}$$

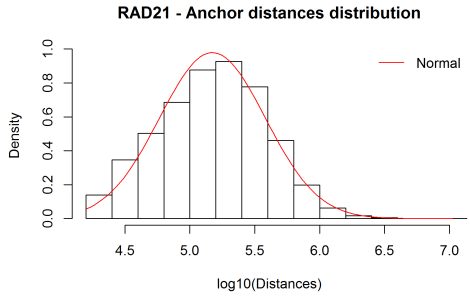
thus suggesting that ATAC-seq insertions occur often at the two ends of a chromatin loop, whose characteristic length is around 100 kbp.



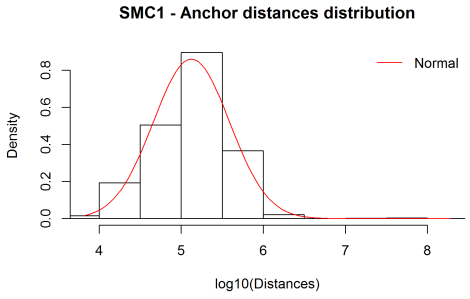
(a) CTCF



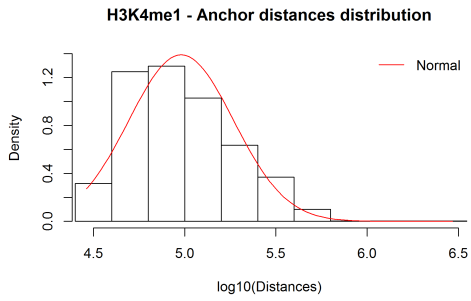
(b) ESR1



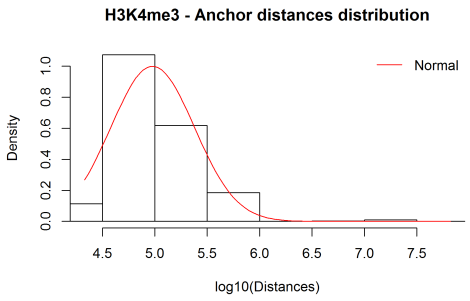
(c) RAD21



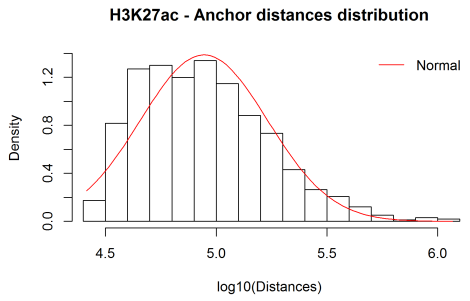
(d) SMC1



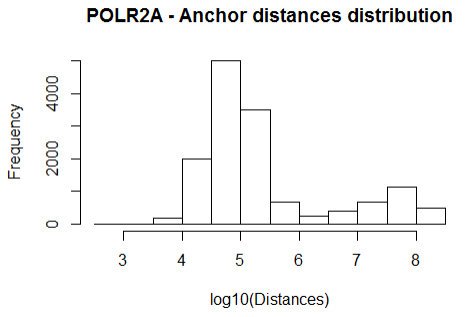
(e) H3K4me1



(f) H3K4me3



(g) H3K27ac



(h) POLR2A

Figure 3.12: ChIA-PET distances

3.4.3 Null model for insertion distances

In order to investigate which is the distribution of distances under the null hypothesis of uniformly distributed insertions, a null model has been created.

For each single cell, the same number of insertions as for the scATAC-seq experiments has been spread randomly over a segment of unitary length. The distances between consecutive insertions have been calculated and aggregated for all the cells. The resulting distribution in both linear and logarithmic scale has been then investigated.

In Figure 3.13 and 3.14 the null models overlaid to the original scATAC-seq distributions are shown for Single-End and Paired-End data respectively. It can be noticed that the null model has a good correspondence to the second peak of the experimental data in both cases.

As reported before, the beta distribution provides a good model for random variables limited to intervals of finite length. In this case, the distances are limited by the finite length of the human genome, around $3.2 \cdot 10^9$ bp. As a consequence the null model can be fitted with a beta distribution: with the domain in logarithmic scale $y = \log_1 0(x)$, the beta distribution becomes

$$Beta(y|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} 10^{y\alpha} \cdot (1 - 10^y)^{\beta - 1} \cdot ln(10) \tag{3.6}$$

For Single-End data, a fit with this function on the null model yielded the parameters

$$\alpha = 0.684 \pm 0.016 \quad \beta = (2.59 \pm 8) \cdot 10^4 \tag{3.7}$$

whereas for Paired-End data they are

$$\alpha = 0.811 \pm 0.008 \quad \beta = (9.71 \pm 0.12) \cdot 10^3 \tag{3.8}$$

Therefore, the second peak of scATAC-seq data, although consistent with chromatin loop lengths, can be explained also with a random process.

In conclusion, the analysis of the distances between scATAC-seq insertions results in a bimodal distribution. The first peak, for short range distances, shows the periodicity characteristic of nucleosome arrays, which are more pronounced in Paired-End data since the second end of each DNA fragment is always hit. The second peak at long range distances is on the one hand consistent with a characteristic length for chromatin loops assayed with ChIA-pet, and on the other hand can be modeled with a beta distribution derived from random "cuts" on a segment of finite length. This points to the fact that a share of scATAC-seq insertions occur with the maximum entropy along the DNA, and since this subset is consistent with chromatin loops it might be possible that these features occur with the maximum entropy as well in the genome.

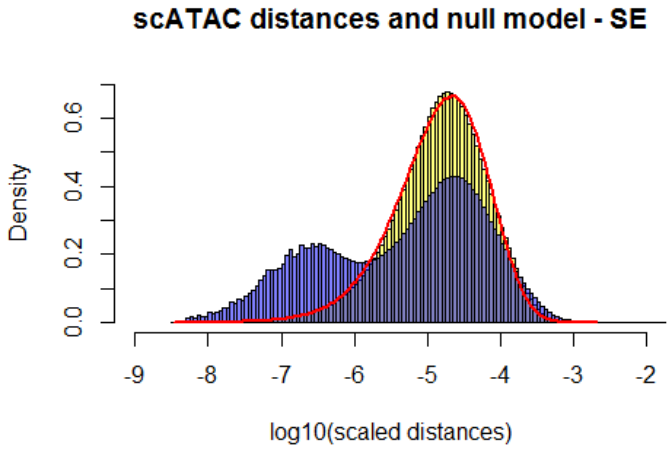


Figure 3.13: scATAC distance distribution and null model - SE data

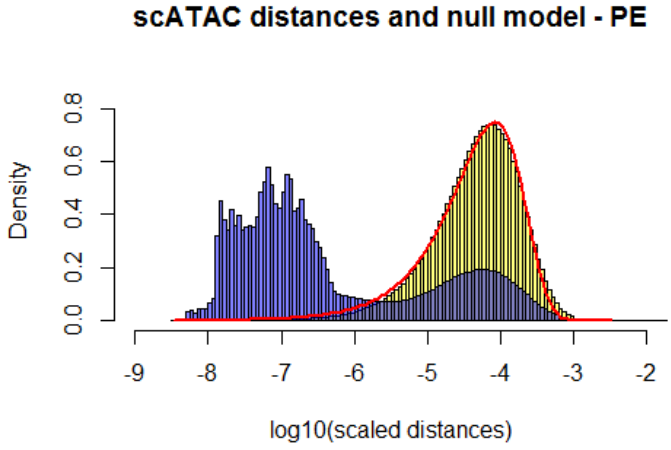


Figure 3.14: scATAC distance distribution and null model - PE data

Chapter 4

Enhancer-Promoter correlation networks

4.1 Regulatory elements in CLL B-lymphocytes

As reported before, regulatory elements such as promoters and enhancers are associated with characteristic histone modification profiles. Promoter sequences carry a high level of H3K4me3, whereas enhancers are associated with high levels of H3K4me1 and H3K27ac marks.

As a consequence, the combination of different ChIP-seq experiments for various histone marks and a computational analysis of the profiles provide a way to predict the location and extension of regulatory sequences in a given cell type.

ChIP-seq experiments and analyses have been performed[1] on Chronic Lymphocytic Leukemia B-lymphocytes for target proteins including H3K4me1, H3K4me3 and H3K27ac in the laboratories of the Computational Oncology group, which affiliates to the Theoretical Bioinformatics division at the German Cancer Research Center (DKFZ), Heidelberg.

The segmentation of promoter and enhancer regions have been realized with ChromHMM software described before. ChromHMM yielded a total of 238820 enhancer-like and 62065 promoter-like regions in the genome. An additional promoter list can be obtain from the annotated 23143 Transcription Start Sites (TSSs) of the RefSeq database [64]: as reported before, promoter regions can be defined as TSSs +/- 1000 bp.

The number of regulatory regions for each chromosome correlates with the chromosome length ($r = 0.67 - 0.78$, depending on the region set), as can be noticed in Figure 4.1 for RefSeq promoters. In some cases chromosomes show

[1] by Dr. Naveed Ishaque, PostDoc in the mentioned group

a higher density of enhancers or promoters: chromosome 19, for example, is gene-rich and therefore it shows an above average abundance of regulatory regions with respect to its size.

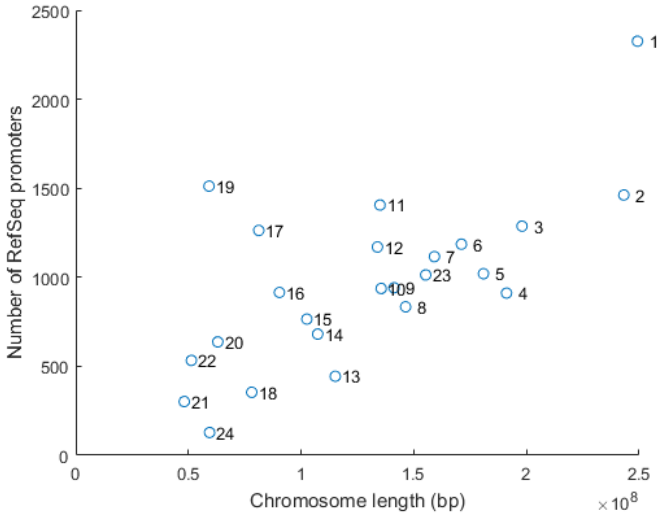


Figure 4.1: Number of RefSeq promoters vs chromosome lengths

4.2 Accessibility of regulatory elements in single cells

4.2.1 Overlap between open chromatin and regulatory elements

As previously reported, active regulatory sequences require to be open and accessible in order to harbor transcription factors and regulate gene expression. As a consequence, an overlap between them and ATAC-seq insertion profiles is expected.

With the available enhancer and promoter lists, this overlap can be computed as the average share of scATAC-seq insertions occurring within them across the single cells. The percentages of open regulatory regions can also be computed, in two ways: the average number of open enhancers/promoters in a single cell and the percentage of open enhancers/promoters for the aggregated data set, obtained by merging all the insertion positions for all the

single cells. These quantities can be defined as follows, both for Single-End and Paired-End data:

$$S_{sc} = \langle \frac{\#\ insertions\ within\ regulatory\ elements}{\#\ total\ insertions} \rangle_{average\ over\ all\ single\ cells} \tag{4.1}$$

$$Ec_{sc} = \langle \frac{\#\ open\ enhancers}{\#\ total\ enhancers} \rangle_{average\ over\ all\ single\ cells} \tag{4.2}$$

$$Ec_{aggr} = \{ \frac{\#\ open\ enhancers}{\#\ total\ enhancers} \}_{aggregated\ data\ set} \tag{4.3}$$

The last two definitions can be naturally extended also to ChromHMM promoters Pc and RefSeq promoters Pr.

The results are reported in Table 4.1 and Table 4.2.

	ChromHMM enhancers	ChromHMM promoters	RefSeq promoters
$S_{sc}, Single-End$	18.2%	20.0%	14.9%
$S_{sc}, Paired-End$	17.0%	17.6%	13.2%

Table 4.1

	Single-End	Paired-End
Ec_{sc}	1.60%	1.17%
Ec_{aggr}	73.5%	70.8%
Pc_{sc}	5.28%	4.10%
Pc_{aggr}	79.9%	76.9%
Pr_{sc}	10.5%	8.14%
Pr_{aggr}	94.2%	91.0%

Table 4.2

As a consequence, 30-38% of scATAC-seq insertions occur within active regulatory regions. 5-10% of total promoters and only 1-2% of total enhancers are open, on average, in a single cell. On the other hand, if one pools all the single cells in an aggregated data set, 77-94% of promoters and 71-74% of enhancers are represented.

In conclusion, the accessibility of enhancers and promoters shows a remarkable heterogeneity between each single cell, probably due to different cell states.

4.2.2 Genome-wide accessibility matrices

Enhancer and promoter lists can be merged together and a binary genome-wide matrix of accessibility M can be constructed. This matrix has a number of rows equal to the number of single cells and a number of columns equal to the number of regulatory sequences, enhancers or promoters. Each entry $M_{i,j}$ has a "1" if the region j is open in cell i, i.e. if at least one scATAC-seq insertion is present, otherwise $M_{i,j} = 0$. The binarization comes from the fact that the accessibility of a region is a binary phenomenon, and the fact that one region can have more than one insertion is probably simply due to the stochasticity of the events.

The two alternative definitions of promoters reported above give rise to two different merged enhancer-promoter lists, one with ChromHMM enhancers and ChromHMM promoters and the other with ChromHMM enhancers and RefSeq promoters. As a consequence, four accessibility matrices can therefore be constructed, given the two different types of experimental data (Single-End and Paired-End).

In Figure 4.2 the accessibility matrix for ChromHMM enhancers-RefSeq promoters and Single-End data is shown.

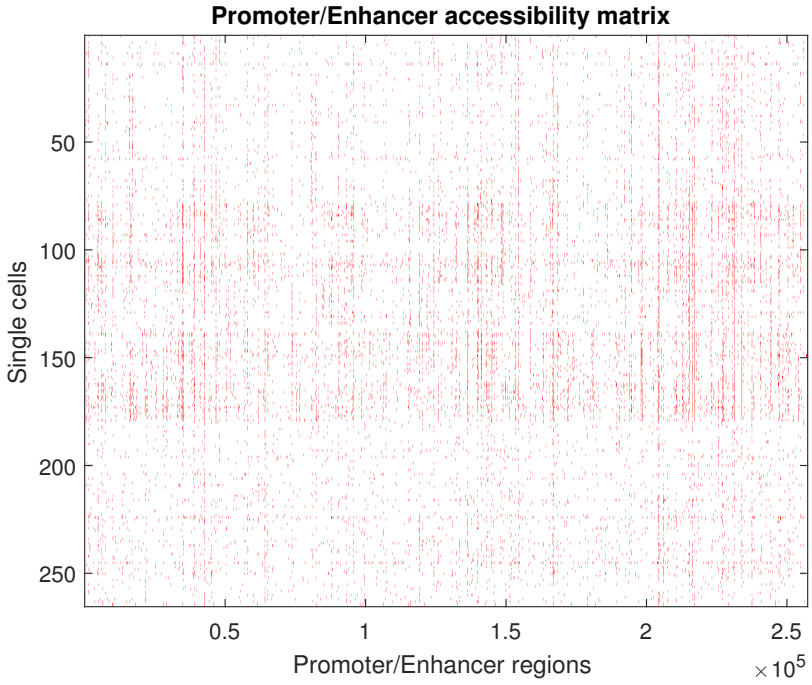


Figure 4.2: Accessibility matrix

It is possible to notice that some of the regulatory regions are open in most of the cells, yielding vertical stripes, while others are blank. The cells with highest total insertion efficiency also show the highest number of open promoters/enhancers.

4.2.3 Accessibility patterns across patients

In order to evaluate patient-patient heterogeneity one might compare the open chromatin patterns for promoter regions across cells of the same patient (*intra*-patient) or belonging to different subjects (*inter*-patient). This analysis has been performed on Single-End profiles since data for three patients were present (versus two in Paired-End data).

A promoter pattern $P_{patient}$ can be defined as a 1D vector whose elements are the ratios of open regions for each promoter across the cells of the given patient s:

$$P_s = \left(\frac{\sum_{i \in s} m(i,1)}{N_s}, \dots, \frac{\sum_{i \in s} m(i,k)}{N_s}, \dots, \frac{\sum_{i \in s} m(i,N_{promoters})}{N_s} \right) \quad (4.4)$$

where N_s is the number of cells belonging to patient s, and $k = 1, \dots, N_{promoters}$ are the indices of promoter regions in the accessibility matrix. A comparison between the promoter open chromatin patterns of different patients $s1$ and $s2$ can be performed by calculating Pearson's and Spearman's sample correlation coefficients between them:

$$r_{Pearson}(P_{s1}, P_{s2}) = \frac{cov(P_{s1}, P_{s2})}{\sigma_{P_{s1}} \sigma_{P_{s2}}} \quad (4.5)$$

$$r_{Spearman}(P_{s1}, P_{s2}) = \frac{cov(rank(P_{s1}), rank(P_{s2}))}{\sigma_{rank(P_{s1})} \sigma_{rank(P_{s2})}} \quad (4.6)$$

Since the distributions of values of these promoter patterns are skewed towards the left, they can't be assumed as Gaussian, and therefore rank correlation metrics are more suitable. Standard deviations can be calculated by repeating the evaluations on 10 subsampled data sets. Tables 4.3 and 4.4 report the *inter*-patient correlation scores thus calculated, taking into account both the promoter definitions reported above (RefSeq and chromHMM respectively).

As reported before, significance can be assessed via a t-test with the statistic $t = r\sqrt{\frac{N-2}{1-r^2}}$ with $N-2$ degrees of freedom for $r_{Pearson}$, and with exact permutation tests for $r_{Spearman}$. In all cases the p-value was negligible.

The resulting correlation values suggest a consistency between the patterns of open chromatin in promoter regions of the different patients. As

Coefficient	RefSeq B01-B02	RefSeq B01-B03	RefSeq B02-B03
$r_{Pearson}$	0.67 ± 0.02	0.71 ± 0.03	0.70 ± 0.03
$r_{Spearman}$	0.70 ± 0.03	0.68 ± 0.04	0.68 ± 0.03

Table 4.3

Coefficient	chromHMM B01-B02	chromHMM B01-B03	chromHMM B02-B03
$r_{Pearson}$	0.68 ± 0.02	0.70 ± 0.04	0.69 ± 0.03
$r_{Spearman}$	0.64 ± 0.03	0.64 ± 0.04	0.61 ± 0.02

Table 4.4

a further validation step, the calculations can be repeated on a randomly shuffled accessibility matrix for promoters. As expected, all the resulting correlation scores are very close to zero.

A similar analysis can be done to assess the *intra*-patient consistency. In this case, one can randomly sample in two groups of equal or similar size the single cells belonging to a given patient and then calculate the pairwise correlations of the resulting promoter patterns. To increase randomization and avoid getting a biased splitting, this procedure can be repeated many times, so that more reliable correlation scores result from averaging over the different subsamples. In this way, standard deviations can also be calculated. The results for the *intra*-patient consistency thus evaluated for 10 subsamples are reported in Tables 4.5 and 4.6

Coefficient	RefSeq B01	RefSeq B02	RefSeq B03
$r_{Pearson}$	0.905 ± 0.003	0.915 ± 0.007	0.947 ± 0.003
$r_{Spearman}$	0.831 ± 0.002	0.840 ± 0.007	0.866 ± 0.004

Table 4.5

The correlation scores are therefore significantly higher within the patients than between them, suggesting a clearer consistency between cells belonging to the same subject, as would be biologically expected. To further validate this result and avoid any possible bias due to the fact that the number of cells for *intra*-patient is lower than for *inter*-patient evaluation, it

Coefficient	chromHMM B01	chromHMM B02	chromHMM B03
$r_{Pearson}$	0.922 ± 0.004	0.928 ± 0.008	0.953 ± 0.003
$r_{Spearman}$	0.734 ± 0.003	0.741 ± 0.007	0.704 ± 0.003

Table 4.6

is possible to apply the randomized subsampling approach to *inter*-patient correlation assessments as well. It turns out that even for cell groups of the same size, *intra*-patient correlation scores are consistently higher than *inter*-patient ones.

4.3 Analysis of correlation networks

4.3.1 Co-occurrence of open chromatin across the single cells

The binary accessibility matrices as the one reported in Figure 4.2 show which regulatory regions are open in which single cells. As mentioned, each column reports the pattern of accessibility for an enhancer or a promoter.

As mentioned before, if an enhancer regulates a promoter, both of them need to be accessible in order to harbor transcription factors. As a consequence, it can be hypothesized that two regulatory elements showing the same accessibility pattern, i.e. being consistently open or closed in the same single cells, might have a functional relationship.

This relationship can be assessed by measuring the statistical association between the pair of columns of the accessibility matrix. Since they are binary vector, rank correlation is unfeasible and the most appropriate metric is the Phi coefficient, which in practical terms matches the Pearson correlation coefficient.

Co-occurrence of open chromatin can be assessed for each pair of columns, thus creating a symmetric square correlation matrix C where entry $C_{i,j}$ is the Phi coefficient between regulatory elements i and j.

However, since inter-chromosomal enhancer-promoter targeting are biologically unlikely, the correlation matrices can be constructed in a chromosome-wise fashion, by limiting the pairwise correlation assessment only to regulatory elements belonging to the same chromosome.

In this way, 22 (one for each chromosome, discarding X and Y) correlation matrices can be constructed for each enhancer-promoter list and each type

of experiment (Single-End and Paired-End). In order to explore differences between patients, patient-wise matrices can also be derived by limiting the correlation assessment to the subset of rows corresponding to cells of each patient.

Correlation matrices between close regulatory elements can also be derived as block matrices along the main diagonal of the chromosome-wise correlation matrices.

In Figure 4.3, two examples of correlation matrices are reported. The values of the Phi coefficients are represented with the heatmap shown on the right of each matrix. The main diagonal has $1's$, since it represent the correlation of regions with themselves. The blue rectangles overlaid to the matrices represent the boundaries of Topologically Associating Domains (TADs) found in literature [65] and derived from IMR90 cell line.

In order to avoid spurious correlations, each regulatory element was required to be accessible in at least 5% of the cells, otherwise it was removed.

Phi coefficient significance has been assessed with a χ^2 test, as reported before. Correlations whose p-value was below 0.05 were discarded and set to zero.

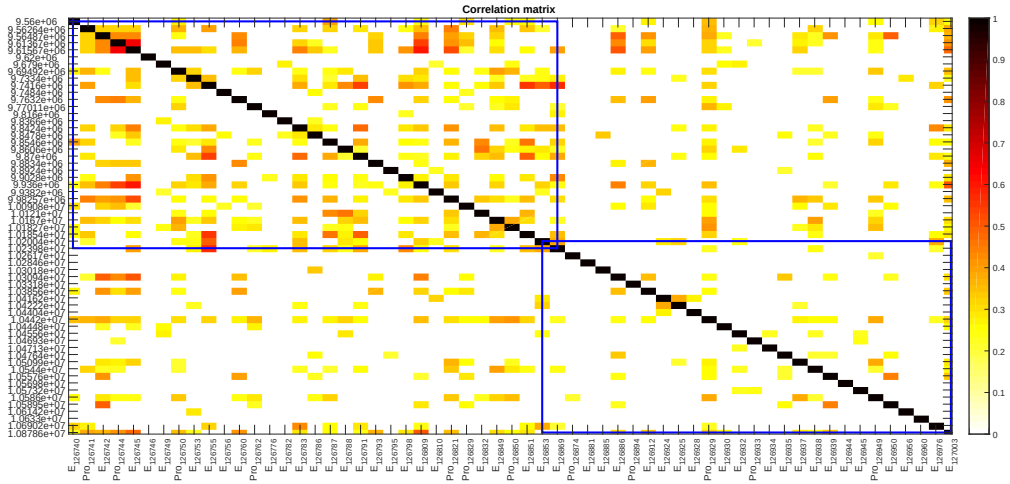
4.3.2 Correlation networks

Correlation matrices can be interpreted as adjacency matrices of weighted networks, where the nodes are constituted by enhancers and promoters and the edge weights are the correlation scores. If enhancer-enhancer and promoter-promoter correlations are ignored, the graphs are bipartite, only connecting enhancers to their target promoters.

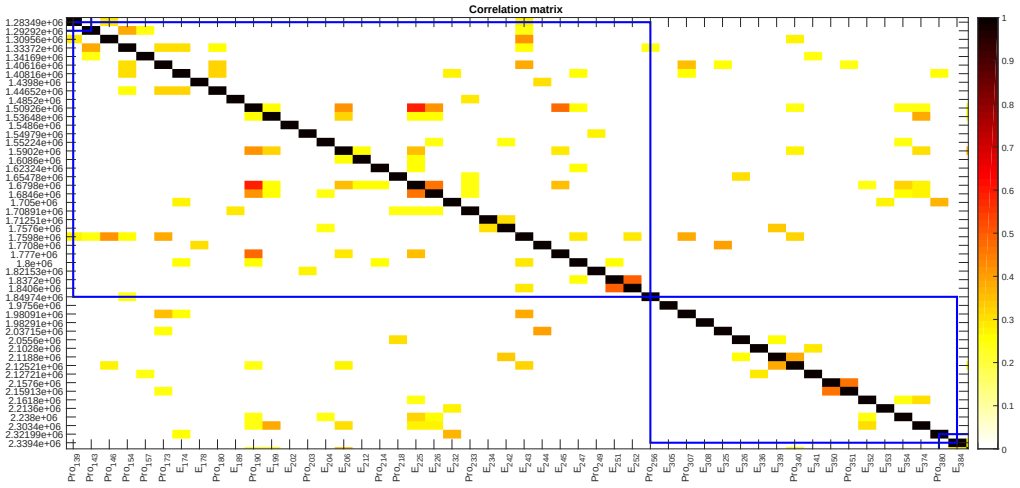
Submatrices extracted as blocks along the main diagonal of a correlation matrix can also give rise to (sub-)graphs. One example is shown in Figure 4.4.

Indeed, even though correlations can occur between any two elements along a chromosome, enhancers are known to regulate promoters up to a few Megabases upstream or downstream the DNA. As a consequence, the correlation scores that can be better explained from a biological point of view are the ones which are relatively close to the main diagonal.

In support to this point, the dependence of the correlation values versus the genomic distance between loci can be investigated. To this end, distances up to 10 Mbp have been sampled into 100 bins of 100 kbp each. The average correlation values related to each distance bin have thus been determined. This assessment has been performed for each patient, promoter list and type of experiment.



(a) Patient B03, chromosome 2, Single-End

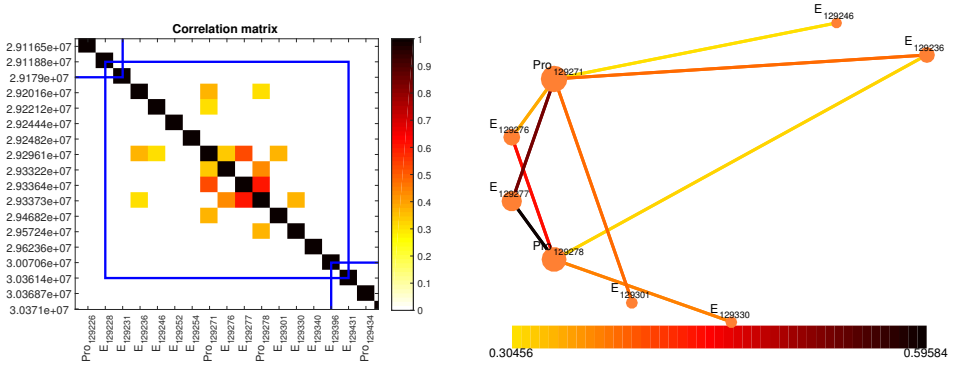


(b) Patient B01, chromosome 1, Paired-End

Figure 4.3: Examples of block correlation matrices

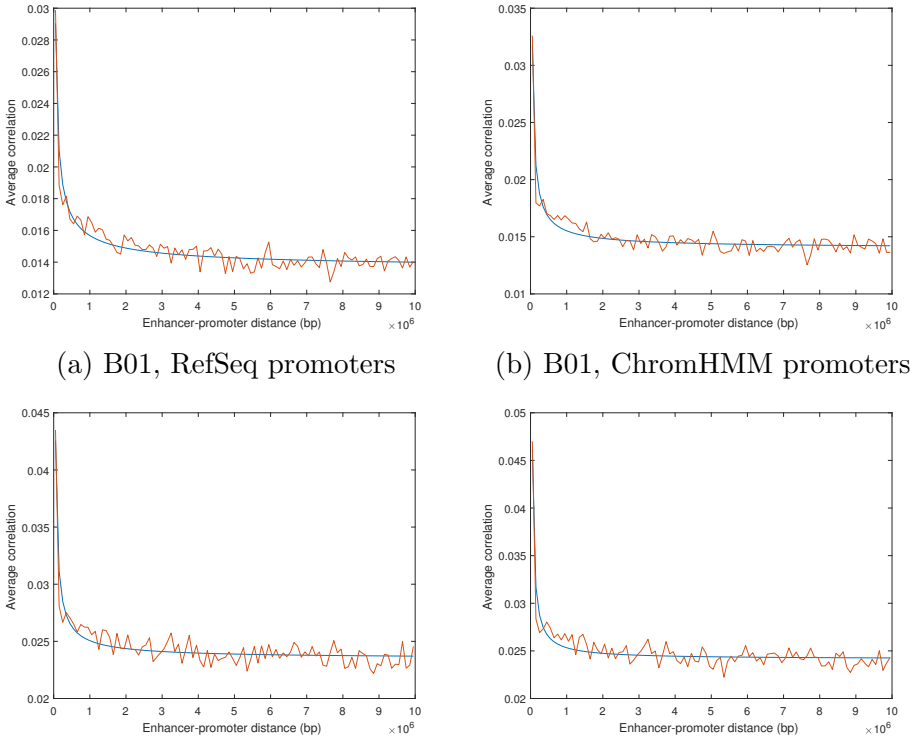
Figure 4.5 reports the average correlation versus genomic distance for Paired-End experiments. The data points have been fitted with a power law, whose exponents are reported in Table 4.7.

It can be noticed that the average correlation scores are higher for distances up to 2-3 Megabases and afterwards they become substantially uniform and independent of the genomic distance. The same trend is present by assessing distances up to 20 or 30 Mbp, thus it can be hypothesized that after 2-3 Mbp a background of spurious correlations is reached. This feature is in agreement with the fact that enhancer-promoter linkages occur more



(a) Correlation matrix (b) Enhancer-promoter network

Figure 4.4: Example of correlation network in patient B03, chromosome 2, Single-End data



(a) B01, RefSeq promoters (b) B01, ChromHMM promoters

(c) B03, RefSeq promoters (d) B03, ChromHMM promoters

Figure 4.5: Correlation versus genomic distance

frequently for elements closer than a few Megabases.

In this sense, these curves indirectly reflect the spatial folding of chro-

Promoter	Single-End			Paired-End	
	B01	B02	B03	B01	B03
RefSeq	-0.5 ± 0.2	-0.38 ± 0.12	-0.67 ± 0.15	-0.66 ± 0.07	-0.66 ± 0.12
HMM	-0.43 ± 0.2	-0.31 ± 0.13	-0.58 ± 0.17	-0.82 ± 0.09	-0.97 ± 0.12

Table 4.7: Power law exponents of average correlation decay over genomic distance

matin. It has been found [20] that for Hi-C data the probability of contact as a function of genomic distance decreases as a power law with exponent close to -1, thus suggesting a fractal globule-like folding of the DNA. In this case, the exponents are slightly below the one reported for Hi-C contact, and they are closer to -1 for Paired-End data, which should indeed reflect more short-range correlations. For this reason, the following analyses have been performed only on Paired-End data.

4.3.3 Centrality of enhancers and promoters

In order to quantify the relevance of enhancers and promoters, an approach based on their role as nodes of these correlation networks can be adopted. In fact, their centrality in the graphs can be investigated with different metrics.

Degree, betweenness and salience centrality scores have been calculated for each regulatory region. In order to retain only those interactions that are explainable biologically, the maximum distance allowed for a correlation was set to 1, 2 and 3 Mbp, as suggested by the correlation-distance curves reported before. As a consequence, the centrality of each node has been assessed by sliding a 2, 4 and 6 Mbp window over the main diagonal of the correlation matrices.

Figure 4.6 shows a matrix reporting the Spearman correlation between various promoter centrality vectors. The centrality vectors have been calculated for Paired-End data of patients B01 and B03, maximum distance 1, 2 and 3 Mbp and with the three metrics mentioned. As a consequence, a 18×18 symmetric correlation matrix is constructed.

It can be noticed that centrality vectors assessed for different maximum distances are strongly correlated. On the other hand, the correlation is very small for different patients. Finally, betweenness and salience centrality are highly correlated, whereas degree centrality is substantially different.

Among most central genes there are some transcription factors (FOSL1, ZNF763, ZNF554) and also genes involved in mechanisms characteristic of B-cells. For example, PTPRCAP, a regulator of T- and B-lymphocyte activa-

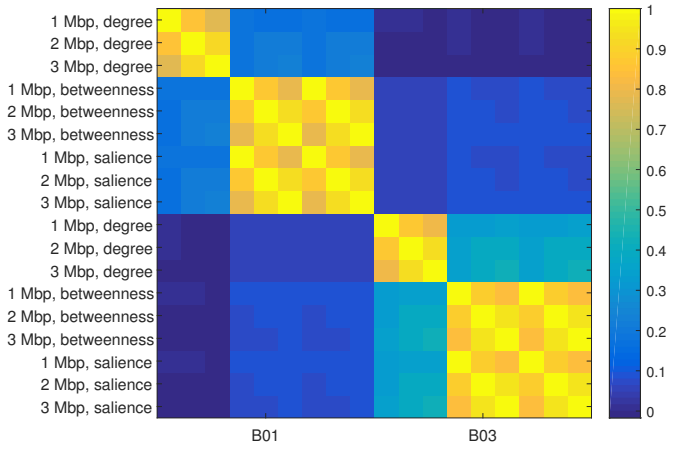


Figure 4.6: Correlation between centrality vectors

tion and ABHD16A, known to be involved in some aspects of immunity [66], rank among the ones with highest degree centrality.

4.3.4 Statistics of enhancer-promoter connectivity

Correlation networks can give insights into enhancer-promoter connectivity models. It is still unclear whether a gene is regulated by one or by multiple enhancers, and whether one enhancer targets only one or more promoters. In order to try to address these questions, it is possible to evaluate how many strong correlations each element is subjected to.

By limiting again the assessments to a maximum distance of 3 Mbp, for each promoter the number of enhancers correlating to it by more than a certain threshold has been counted. Figure 4.7 reports the histogram for patient B01, RefSeq promoters and threshold 0.3.

It can be noticed that the most frequent connection model is one promoter to one enhancer. This feature is even more pronounced if stricter thresholds are used. In fact, the probability of interaction with one or more enhancers shows an exponential decay, with shorter characteristic lengths (and thus steeper decays) for bigger thresholds. Figure 4.8 reports the exponential fits of curves obtained with thresholds of 0.3, 0.4 and 0.5.

The same results are obtained for patient B03 and ChromHMM promoter list.

A similar analysis can be repeated for enhancer regions, looking at how many promoters they regulate. Also in this case, the one enhancer to one promoter model is the most frequent. However, the exponential decays are

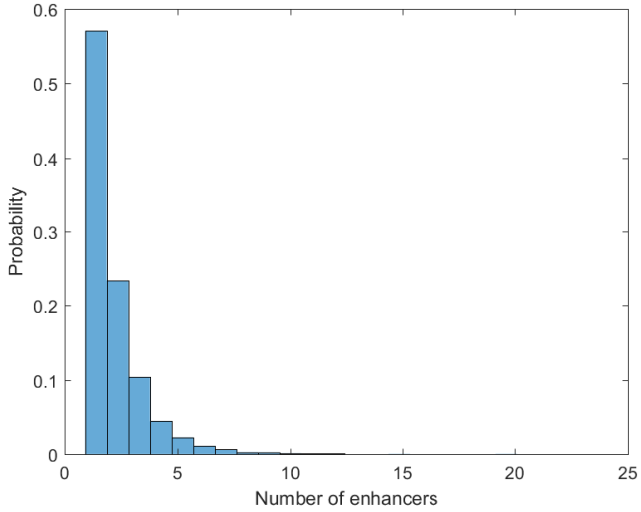
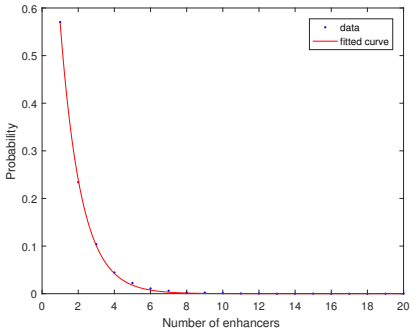
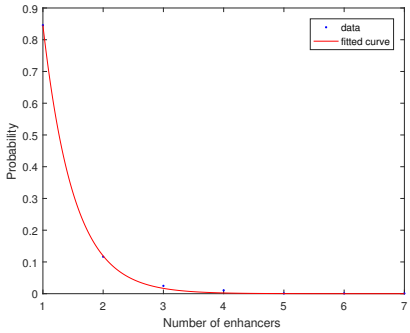


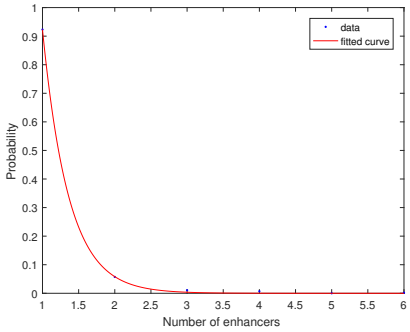
Figure 4.7: Histogram of the number of enhancers connected to one promoter



(a) Threshold = 0.3



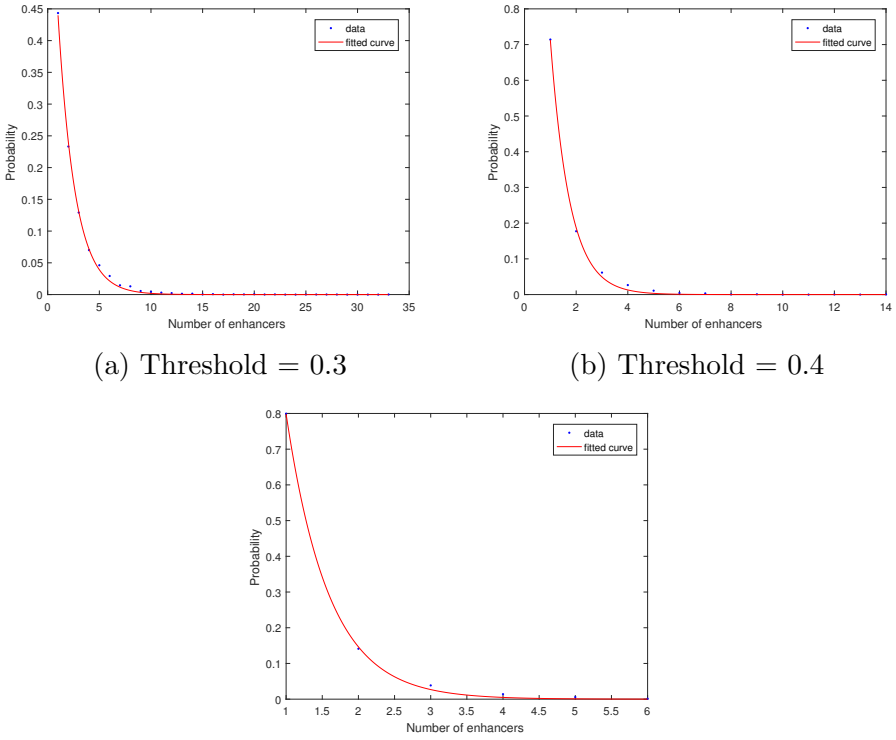
(b) Threshold = 0.4



(c) Threshold = 0.5

Figure 4.8: Exponential fits for promoters connection models

less steep, as reported in Figure 4.9.



(a) Threshold = 0.3

(b) Threshold = 0.4

(c) Threshold = 0.5

Figure 4.9: Exponential fits for enhancers connection models

A direct comparison between the exponential decays of enhancer and promoter connections is reported in Table 4.8. As a result, it happens more frequently that an enhancer targets more than one promoter than that one promoter is regulated by more than one enhancer.

Region type	Threshold 0.3	0.4	0.5
Promoter	-0.86 ± 0.02	-1.96 ± 0.11	-2.9 ± 0.5
Enhancer	-0.60 ± 0.01	-1.33 ± 0.07	-1.70 ± 0.15

Table 4.8: Decay constants for enhancer's and promoter's number of connections at different thresholds

Another analysis that can be performed is to check whether or not a promoter is regulated by its closest enhancer. To do so, the enhancer with

strongest correlation to each promoter can be determined. Figure 4.10 shows the distribution of the number of enhancers "skipped" by promoters. The first bin corresponds to zero skipped enhancers, meaning that only 35% of the times a promoter is regulated by its closest enhancer, and this number is consistent to the ones reported in literature [23].

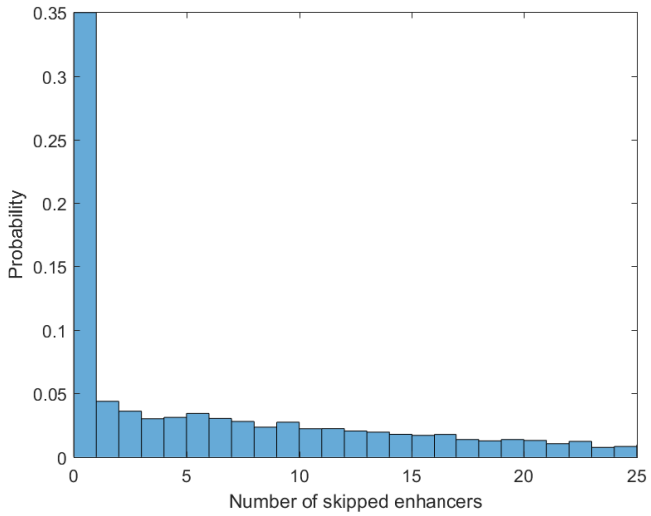


Figure 4.10: Histogram of number of skipped enhancers for each promoter

This feature is partially recapitulated by the distribution of the distances between promoters and enhancers with strongest connections (Figure 4.11). Around 31% of the times the distance is less than 200 kbp.

4.3.5 Investigation of superenhancers functioning

As mentioned before, a relevant question in gene regulation is about the superenhancer functioning. Superenhancers, simply defined as "large" (over 18 kbp long) enhancers, have been hypothesized to either function as a single, big, enhancer or as multiple nearby enhancers.

Insight into this question can be provided by single cell open chromatin profiles. The coverage of superenhancers, namely the positions of open chromatin for the aggregated single cells, can be investigated in order to possibly find peaks of accessible chromatin within them; in this case, superenhancers would be more likely composed by multiple, localized, nearby enhancers.

An example of scATAC-seq coverage in a superenhancer is shown in Figure 4.12. At least two, if not four peaks can be noticed, suggesting the presence of an array of enhancers.

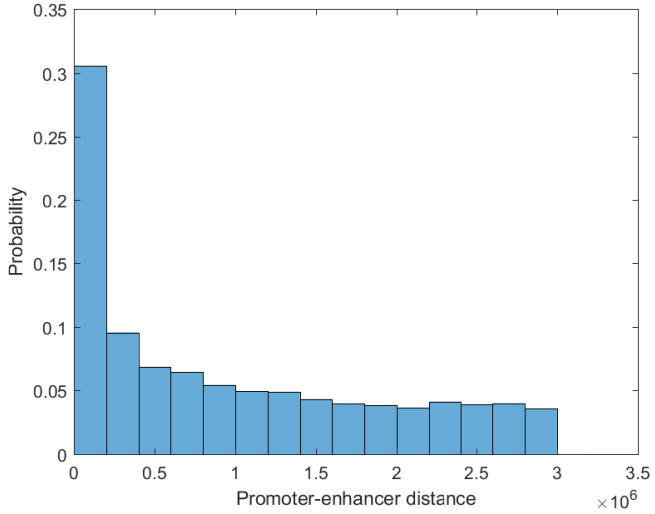


Figure 4.11: Histogram of distances between promoters and enhancers with strongest correlation

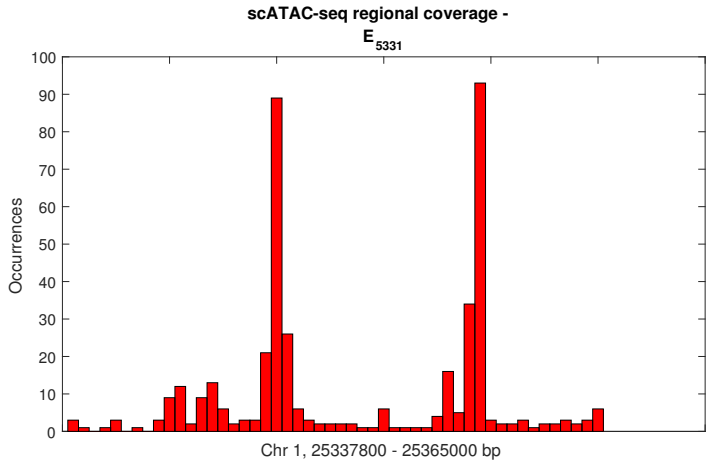


Figure 4.12: Example of open chromatin coverage for a superenhancer

In other cases, a more uniform coverage is revealed. In order to assess this in a systematic way, the inverse participation ratio reported before can be used as a metric for localization. IPR can be calculated on the binned coverage of each superenhancer, using the same number of bins set to 50. Following its definition, it ranges from 1 to 50, and values close to 1 indicate peaked distributions.

Figure 4.13 reports the distribution of values of inverse participation ratio.

Figure 4.14 shows the two superenhancers having the smallest and greatest IPR respectively. It can be noticed that the distribution is slightly skewed towards small values, corresponding to highly localized profiles. This would point to the fact that superenhancers rarely act as single blocks but rather as arrays of multiple smaller enhancers.

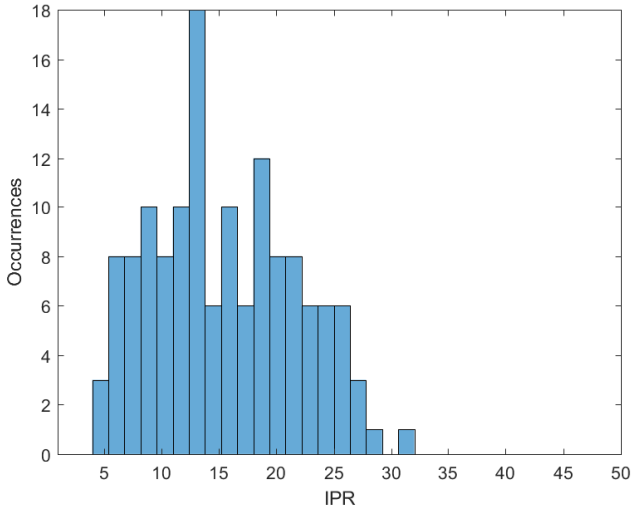
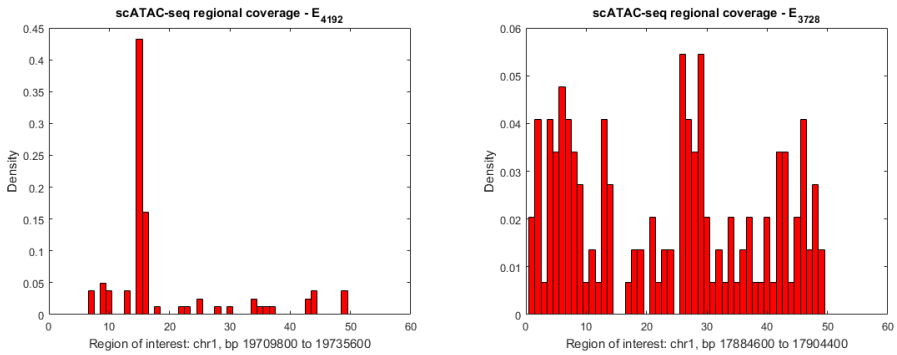


Figure 4.13: Histogram of inverse participation ratio values for superenhancers



(a) Smallest IPR (b) Greatest IPR

Figure 4.14: Superenhancers with extreme IPR

Chapter 5

Conclusions and future prospects

In the era of single cell Genomics, scATAC-seq has proven itself to be a valuable technique to investigate open chromatin in primary tumor cells with an unprecedented precision. Accessibility profiles in single cells could be derived for CLL B-lymphocytes and their analysis yielded information on chromatin structure and mechanisms of gene expression regulation. Importantly, they also provide the basis for comparisons with healthy cells and tumor cells treated with epigenetic drugs, which would leverage the possibilities of Personalized Medicine.

Given the fact that the experimental technique is very recent, the processing of raw sequencing data and their characterization presented a challenge since no standard pipeline was well established in literature. Biases introduced by PCR artifacts were removed and the sequence specificity of the Tn5 transposase was characterized.

Statistical analyses of open chromatin maps across the single cells recapitulated known structural properties of chromatin architecture. The distribution of fragment lengths has shown two types of periodicity, one which is characteristic of nucleosome arrays and the other representing the DNA helical pitch. Moreover, the distances between consecutive open regions were consistent with typical lengths of chromatin loops and could partially be explained with a null model of uniformly spread positions along the genome, which follows a beta distribution.

As expected, a considerable share of open chromatin regions overlapped regulatory sequences such as promoters and enhancers. The pattern of accessibility of promoters was substantially conserved when different subset of cells were compared. Remarkably, intra-patient consistency was significantly higher than inter-patient one.

The accessibility profiles of enhancers and promoters across the single cells provided a tool to assess functional relationships between regulatory elements. In this way, chromosome-wise correlation networks could be constructed and analyzed. Correlation values decreased as a power law with the genomic distance, in agreement with the fact that enhancers are known to regulate promoters whose distance is within a few Megabases. From the networks, the relevance of regulatory elements could be assessed with various centrality metrics.

The statistics of enhancer-promoter connectivity were in substantial agreement with known targeting models. The model according to which one promoter is regulated only by its closest enhancer was indeed the most frequent, but examples of one-to-many relationships have also been found. Moreover, in some cases promoters "skipped" enhancers in their neighborhood and were most strongly linked to distal ones.

Finally, scATAC-seq coverage of superenhancers gave also insights into their functioning. The frequent localization of open chromatin hotspots within them pointed to the "array-of-enhancers" model.

The features and analyses reported here are good candidates for comparisons between tumor samples and healthy controls. In particular, healthy cells should show a higher number of active and accessible enhancers, since CLL is associated with histone deacetylation. As a consequence, enhancer-promoter networks should change and promoters should in principle show, on average, higher values of centrality in healthy controls, especially for disease-associated genes.

In addition, comparisons with tumor cells treated with epigenetic drugs such as deacetylase inhibitor Panobinostat could give feedback on their therapeutic efficacy. In the best case scenario, treated cells should recover a healthy state and share similarities with healthy controls. If the response to drugs is variable, open chromatin profiles in single cells could help patient stratification.

In this framework, ATAC-seq is fast and affordable enough to be employed for clinical decision making, and therefore it is a good candidate to play a pivotal role in Personalized Medicine in the future.

Bibliography

[1] Giulia Fabbri and Riccardo Dalla-Favera. The molecular pathogenesis of chronic lymphocytic leukaemia. *Nature Reviews Cancer*, 16(3):145–162, 2016.

[2] André F Rendeiro, Christian Schmidl, Jonathan C Strefford, Renata Walewska, Zadie Davis, Matthias Farlik, David Oscier, and Christoph Bock. Chromatin accessibility maps of chronic lymphocytic leukaemia identify subtype-specific epigenome signatures and transcription regulatory networks. *Nature Communications*, 7, 2016.

[3] James D Watson, Francis HC Crick, et al. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

[4] Francis Crick et al. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[5] Rippe (ed.) et al. *Genome Organization And Function In The Cell Nucleus*. Wiley-VCH, 2011.

[6] Arthur Lesk. *Introduction to genomics*. Oxford University Press, 2012.

[7] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012.

[8] HHQ Heng, SA Krawetz, W Lu, S Bremer, G Liu, and CJ Ye. Redefining the chromatin loop domain. *Cytogenetic and Genome Research*, 93(3-4):155–161, 2001.

[9] Illumina. An introduction to next generation sequencing technology (and references). Technical report.

[10] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.

[11] Gordon Robertson, Martin Hirst, Matthew Bainbridge, Misha Bilenky, Yongjun Zhao, Thomas Zeng, Ghia Euskirchen, Bridget Bernier, Richard Varhol, Allen Delaney, et al. Genome-wide profiles of stat1 dna association using chromatin immunoprecipitation and massively parallel sequencing. *Nature methods*, 4(8):651–657, 2007.

[12] Kairong Cui and Keji Zhao. Genome-wide approaches to determining nucleosome occupancy in metazoans using mnase-seq. *Chromatin Remodeling: Methods and Protocols*, pages 413–419, 2012.

[13] 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.

[14] Wenfei Jin, Qingsong Tang, Mimi Wan, Kairong Cui, Yi Zhang, Gang Ren, Bing Ni, Jeffrey Sklar, Teresa M Przytycka, Richard Childs, et al. Genome-wide detection of dnase i hypersensitive sites in single cells and ffpe tissue samples. *Nature*, 2015.

[15] 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.

[16] Andrew Adey, Hilary G Morrison, Xu Xun, Jacob O Kitzman, Emily H Turner, Bethany Stackhouse, Alexandra P MacKenzie, Nicholas C Caruccio, Xiuqing Zhang, Jay Shendure, et al. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome biology*, 11(12):1, 2010.

[17] 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):486–490, 2015.

[18] Yuanyuan Li and Trygve O Tollefsbol. Dna methylation detection: bisulfite genomic sequencing analysis. *Epigenetics Protocols*, pages 11–21, 2011.

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

[20] 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.

[21] Melissa J Fullwood and Yijun Ruan. Chip-based methods for the identification of long-range chromatin interactions. *Journal of cellular biochemistry*, 107(1):30–39, 2009.

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

[23] Lijing Yao, Benjamin P Berman, and Peggy J Farnham. Demystifying the secret mission of enhancers: linking distal regulatory elements to target genes. *Critical reviews in biochemistry and molecular biology*, 50(6):550–573, 2015.

[24] Simon Andrews et al. Fastqc: A quality control tool for high throughput sequence data. *Reference Source*, 2010.

[25] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome biology*, 10(3):1, 2009.

[26] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, et al. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[27] Aaron R Quinlan and Ira M Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[28] 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):1, 2008.

[29] James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. Integrative genomics viewer. *Nature biotechnology*, 29(1):24–26, 2011.

[30] Gerhard Bohm and Günter Zech. *Introduction to statistics and data analysis for physicists*. DESY, 2010.

[31] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1, 2014.

[32] Sunil Srinivasa and Martin Haenggi. Distance distributions in finite uniformly random networks: Theory and applications. *IEEE Transactions on Vehicular Technology*, 59(2):940–949, 2010.

[33] D. Peel McLachlan, G. *Finite Mixture Models*. John Wiley & Sons, Inc., 2000.

[34] Yuhua Su, Lei Zhu, Alan Menius, and Jason Osborne. Mixture models for gene expression experiments with two species. *Human genomics*, 8(1):1, 2014.

[35] Franz Wegner. Inverse participation ratio in 2+ ε dimensions. *Zeitschrift für Physik B Condensed Matter*, 36(3):209–214, 1980.

[36] F Evers and AD Mirlin. Fluctuations of the inverse participation ratio at the anderson transition. *Physical review letters*, 84(16):3690, 2000.

[37] Yan V Fyodorov and Alexander D Mirlin. Analytical derivation of the scaling law for the inverse participation ratio in quasi-one-dimensional disordered systems. *Physical review letters*, 69(7):1093, 1992.

[38] William H Press. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[39] Feng-lan Bai, Ying-zhao Liu, and Tian-ming Wang. A representation of dna primary sequences by random walk. *Mathematical biosciences*, 209(1):282–291, 2007.

[40] Yongjin Li and Jinyan Li. Disease gene identification by random walk on multigraphs merging heterogeneous genomic and phenotype data. *BMC genomics*, 13(Suppl 7):S27, 2012.

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

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

[43] Shrish Tiwari, S Ramachandran, Alok Bhattacharya, Sudha Bhattacharya, and Ramakrishna Ramaswamy. Prediction of probable genes by fourier analysis of genomic sequences. *Computer applications in the biosciences: CABIOS*, 13(3):263–270, 1997.

[44] Richard J Trudeau. *Introduction to graph theory*. Courier Corporation, 2013.

[45] Daniel Grady, Christian Thiemann, and Dirk Brockmann. Robust classification of salient links in complex networks. *Nature communications*, 3:864, 2012.

[46] Darren A Cusanovich, Riza Daza, Andrew Adey, Hannah A Pliner, Lena Christiansen, Kevin L Gunderson, Frank J Steemers, Cole Trapnell, and Jay Shendure. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237):910–914, 2015.

[47] JC Wang, MI Kafeel, B Avezbakiyev, C Chen, Y Sun, C Rathnasabapathy, M Kalavar, Z He, J Burton, and S Lichter. Histone deacetylase in chronic lymphocytic leukemia. *Oncology*, 81(5-6):325–329, 2012.

[48] Claudia V Andreu-Vieyra and James R Berenson. The potential of panobinostat as a treatment option in patients with relapsed and refractory multiple myeloma. *Therapeutic advances in hematology*, 5(6):197–210, 2014.

[49] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359, 2012.

[50] Python software foundation. python language reference, version 2.7. `http://www.python.org`.

[51] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[52] The perl programming language. `https://www.perl.org/`.

[53] Matlab r2016a, the mathworks, inc., natick, massachusetts, united states.

[54] Giulia Paci, Giampaolo Cristadoro, Barbara Monti, Marco Lenci, Mirko Degli Esposti, Gastone C Castellani, and Daniel Remondini. Characterization of dna methylation as a function of biological complexity via dinucleotide inter-distances. *Phil. Trans. R. Soc. A*, 374(2063):20150227, 2016.

[55] R Core Team et al. R: A language and environment for statistical computing. 2013.

[56] Marie Laure Delignette-Muller, Christophe Dutang, et al. fitdistrplus: An r package for fitting distributions. *Journal of Statistical Software*, 64(4):1–34, 2015.

[57] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. Ncbi geo: archive for functional genomics data setsupdate. *Nucleic acids research*, 41(D1):D991–D995, 2013.

[58] Zhonghui Tang, Oscar Junhong Luo, Xingwang Li, Meizhen Zheng, Jacqueline Jufen Zhu, Przemyslaw Szalaj, Pawel Trzaskoma, Adriana Magalska, Jakub Wlodarczyk, Blazej Ruszczycki, et al. Ctcf-mediated human 3d genome architecture reveals chromatin topology for transcription. *Cell*, 163(7):1611–1627, 2015.

[59] Xiong Ji, Daniel B Dadon, Benjamin E Powell, Zi Peng Fan, Diego Borges-Rivera, Sigal Shachar, Abraham S Weintraub, Denes Hnisz, Gianluca Pegoraro, Tong Ihn Lee, et al. 3d chromosome regulatory landscape of human pluripotent cells. *Cell Stem Cell*, 18(2):262–275, 2016.

[60] Tatyana Kuznetsova, Shuang-Yin Wang, Nagesha A Rao, Amit Mandoli, Joost HA Martens, Nils Rother, Aafke Aartse, Laszlo Groh, Eva M Janssen-Megens, Guoliang Li, et al. Glucocorticoid receptor and nuclear factor kappa-b affect three-dimensional chromatin organization. *Genome biology*, 16(1):1, 2015.

[61] Fabian Grubert, Judith B Zaugg, Maya Kasowski, Oana Ursu, Damek V Spacek, Alicia R Martin, Peyton Greenside, Rohith Srivas, Doug H Phanstiel, Aleksandra Pekowska, et al. Genetic control of chromatin states in humans involves local and distal chromosomal interactions. *Cell*, 162(5):1051–1065, 2015.

[62] Denes Hnisz, Abraham S Weintraub, Daniel S Day, Anne-Laure Valton, Rasmus O Bak, Charles H Li, Johanna Goldmann, Bryan R Lajoie,

Zi Peng Fan, Alla A Sigova, et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, 351(6280):1454–1458, 2016.

[63] Nastaran Heidari, Douglas H Phanstiel, Chao He, Fabian Grubert, Fereshteh Jahanbani, Maya Kasowski, Michael Q Zhang, and Michael P Snyder. Genome-wide map of regulatory interactions in the human genome. *Genome research*, 24(12):1905–1917, 2014.

[64] Kim D Pruitt, Garth R Brown, Susan M Hiatt, Françoise Thibaud-Nissen, Alexander Astashyn, Olga Ermolaeva, Catherine M Farrell, Jennifer Hart, Melissa J Landrum, Kelly M McGarvey, et al. Refseq: an update on mammalian reference sequences. *Nucleic acids research*, 42(D1):D756–D763, 2014.

[65] Benjamin D Pope, Tyrone Ryba, Vishnu Dileep, Feng Yue, Weisheng Wu, Olgert Denas, Daniel L Vera, Yanli Wang, R Scott Hansen, Theresa K Canfield, et al. Topologically associating domains are stable units of replication-timing regulation. *Nature*, 515(7527):402–405, 2014.

[66] Michael Rebhan, Vered Chalifa-Caspi, Jaime Prilusky, and Doron Lancet. Genecards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*, 14(8):656–664, 1998.