Determinants of Nucleosome Organisation and Transcription Regulation by Histone Marks

Jeremie Becker

Department of Statistics

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary term 2012

To my family.

Acknowledgements

First of all, I would like to thank my supervisors Chris Holmes and John Hancock for giving me this opportunity and providing many stimulating discussions and guidance throughout my DPhil. Not only did they offer me the perfect conditions for my DPhil, they also opened me to the field of Bayesian statistics and encouraged me to develop initiatives.

Second, I would like to thank Chris Yau and George for their precious advises in many analyses as well as their help in turning intuitions into formal statistical form. I am also extremely grateful to Chris for sharing his experience on how to manage a DPhil.

I also must thank all the people in the Holmes group, the Mammalian Genetics Unit at Harwell and the Department of Statistics for the interesting discussions and making these places friendly and positive environments. More particularly, I would like to thank my office mates over the years, Alex, Avi, Chris, Eleanor, Eleni, James, Jo, Julia, Peter, Quin, Rahul, Shahzia, Tristan as well as Maddy for countless conversations, cups of tea/coffee as well as various rubber-band/water-gun/bouncing-ball related distractions.

I am very grateful to Elisabeth who supported and motivated me everyday of my DPhil, who visited me on a monthly basis and greatly contributed to make me discover England outside Oxford.

I would like to thank my family who have also been very supportive and who tried to take interest in my project even when the explanations were confused.

I would also like to thank my housemates and my friends Habib, Mathieu, Klaus, Aziz, Frederic, Sylvestre, Cecile for making the last five years fun and memorable.

I gratefully acknowledge the MRC and EPSRC for funding my position.

Abstract

Epigenetics is the study of heritable changes in gene expression that do not involve changes in the underlying DNA sequence. Epigenetic pathways appeared in eukaryotes where multicellular organisms differentiate into different cell types, associated with different phenotypes. The differentiation process is achieved by modifications of the chromatin structure which, by altering the access of *trans*-factors to the DNA, result in gene activation and repression. Epigenetic mechanisms are therefore viewed as an "extra" layer of information that modulate the genetic information in time and space, necessary for the development and the response to environment stimuli. Although the recent development of high-throughput sequencing technologies has provided an unprecedented insight into epigenetic pathways, the mechanisms controlling the chromatin dynamic as well as their downstream effects on cellular processes are far from being fully understood.

In this thesis, our focus will be restricted to mechanisms acting on the nucleosome level. The first chapter will present the factors known to influence nucleosome positioning as well as the challenges related to the measurement of the nucleosome architecture. The second chapter will introduce a statistical approach, NucleoFinder, capable of iden-

tifying nucleosomes consistently positioned in a population of cells. In chapter three, we will make use of NucleoFinder to investigate the importance of *cis* and *trans*-factors on the nucleosome architecture in human and show that, despite variation across functional regions, *cis*-factors have a very modest influence on nucleosome positioning. Finally, in chapter four, we will aim to identify clusters of histone modifications specific to functionally distinct regions, characterize their function and their association with gene expression level.

Contents

Chapter 1

Introduction

1.1 Introduction

1.1.1 Epigenetics

Transcription, translation and post-translational modifications represent the transfer of the genetic information from the DNA to the mRNA and protein levels. In multicellular organisms, although the genomic sequence is identical across cells, cell types exhibit important variations in their profile of gene expression, resulting in different cell phenotypes and functions. Many of these patterns characteristic of differentiated cells arise during development whereby a single fertilized egg cell changes into many cell types (neurons, muscle cells, etc.) by activating some genes while inhibiting others. After differentiation, apart from possible modifications triggered by the environment, cells' gene expression pattern is maintained over divisions. Therefore, in addition to inheriting genetic information, the cell also receives information during mitosis that is not encoded in the genomic se-

quence and is known as epigenetic information (Gibney and Nolan, 2010).

A recent review in Science stated that "the epigenetic information resides in self-propagating molecular signatures that provide a memory of previously experienced stimuli, without irreversible changes in the genetic information" (Bonasio et al., 2010). In other words, a molecular signal can be considered as epigenetic if it is heritable (possible transmission to the progeny), self-sustaining (there exists pathways responsible for reproducing the molecular signal after DNA replication), and reversible. These epigenetic signals have commonly been classified into two categories, (i) *trans*-acting mechanisms which, as indicated by their name, are mediated by molecules other than those forming the chromatin. *Trans*-acting mechanisms influence transcriptional states through feedback loops, transcription factors (particularly in simple organisms such as single-cell eukaryotes) and different classes of non-coding RNA (sRNAs : Bourchis and Voinnet 2010, lncRNAs : Saxena and Carninci 2011). (ii) *Cis*-acting mechanisms, on the other hand, are physically associated with the locus on which they act. They include DNA methylation, histone post-translational modifications, nucleosome positioning and higher-order chromatin structure (Bonasio et al., 2010).

It remains unclear whether all these mechanisms determining transcriptional states are truly epigenetic. The most studied of them, DNA methylation, fulfills all three criteria : (i) *in vitro* methylated DNA remains methylated after several generations (Wigler et al., 1981), (ii) the division of a double stranded DNA sequence decorated with methylation marks leads to two hemi-methylated double strands that can be restored to fully methylated thanks to maintenance

methyltransferases (Goll and Bestor, 2005), (iii) the DNA demethylase plays an antagonist role to methyltransferase. By contrast, it is less clear whether histone modifications (and many noncoding RNAs) are also epigenetic since mechanisms of self-perpetuating and inheritance have not been identified for most of them. Propagation mechanisms have been proposed that explain how to re-establish histone modifications on daughter strands from parental histones carrying such modifications. However it is not known yet whether and how parental nucleosomes are re-used after DNA replication.

In this thesis we will restrain our focus to *cis*-acting epigenetic mechanisms, with a particular emphasis to (i) nucleosome positioning and (ii) histone modifications that will be at the heart of this work. These epigenetic factors, embedded in the chromatin, have been shown to play a fundamental role in the regulation of many cellular processes, including DNA-protein interactions, gene expression, suppression of transposable element mobility, cellular differentiation, X-chromosome inactivation and genomic imprinting (Zhou et al., 2011). In this section we introduce the epigenetic mechanisms acting at the different levels of the chromatin, then, after a brief summary of epigenetic modifications associated with human diseases, we will lay out the global structure of the thesis.

1.1.2 The Epigenetic Information Occurs at Different Level of the Chromatin Structure

At the root of the chromatin structure is the nucleosome which appears as "beads on a string" in electron micrograph and consists of approximately 150 base pairs

coiled 1.65 times around a histone octamer core comprising two H2A/H2B and H3/H4 heterodimers (figure 1.1) . The "string" also called linker varies from 10 bp to 80 bp across species, cell types and chromatin regions (Luger et al., 1997)). Nucleosomes can further be folded into higher order organisations to eventually form the chromosome. These structures include the 30 nm fiber whereby histone H1 binds to both nucleosome and linker DNA, and the 100 nm fiber (Felsenfeld and Groudine, 2003). Chromatin has commonly been viewed as either in an active or repressed form called euchromatin and heterochromatin. The former, characteristic of active transcribed regions, is lightly packed (up to the 30 nm fiber compaction) and allows regulatory proteins and transcription complexes to bind to the DNA. The latter, tightly packed, is associated with gene silencing and the protection of chromosome structures such as centromeres and telomeres (Elgin, 1996). Among the superimposed layers that constitute the chromatin, epigenetic modifications are present at three different levels (see figure 1.1) :

- the DNA level, where methylation of cytosines leads to transcriptional repression;
- the nucleosome level, where the nucleosome position, the use of histone variants (e.g. H2A.Z) and post-translational modifications of histones can induce or repress transcription;
- higher level where the chromatin itself interacts with other components present in the nucleus such as small interfering RNA or the nuclear lamina (structure made up of protein filaments located in inner face of the nuclear envelope) (Guelen et al., 2008).

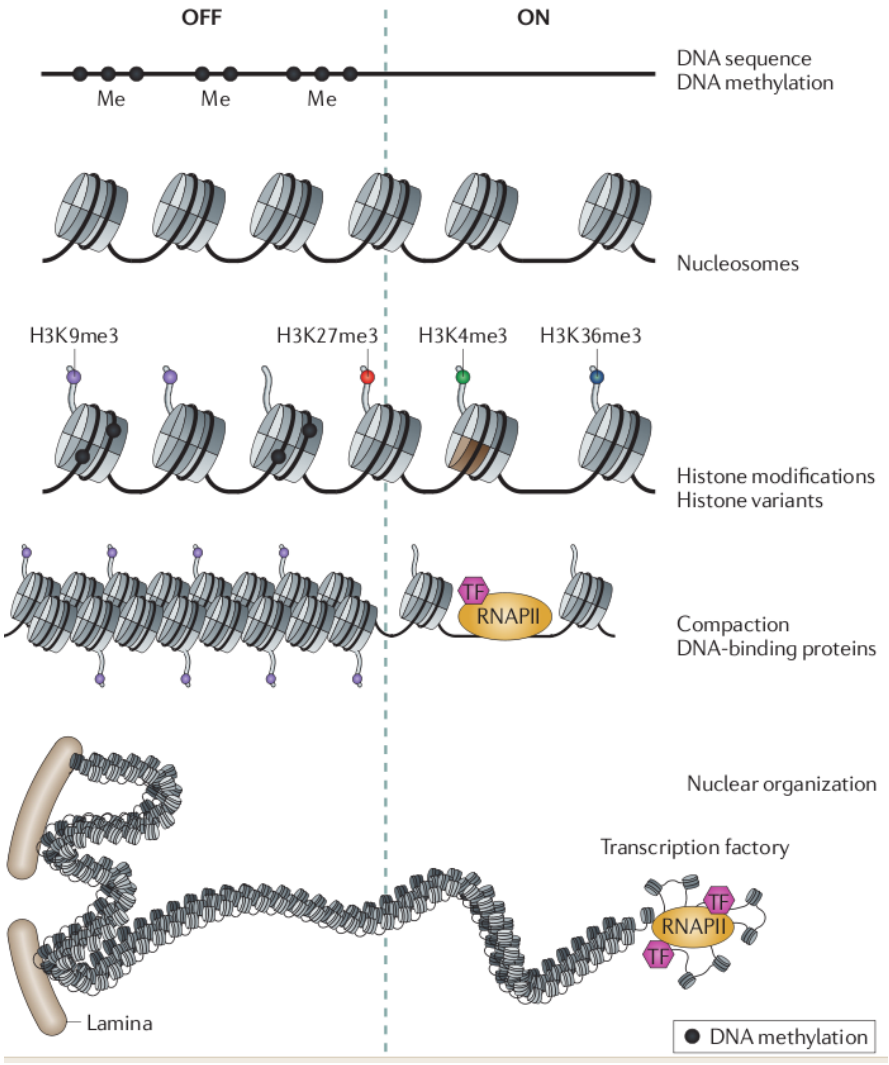

Figure 1.1: The epigenetic information is carried at various levels of the chromatin organisation, from DNA (top) to higher-order structure (bottom), figure taken from Zhou et al. (2011). The features displayed on the left and right sides are associated with inactive and active transcription states respectively. The first level displays DNA sequence, where cytosine nucleotides are methylated in repressed region; the second is the nucleosome level consisting of $\sim$ 150 bp DNA wrapped around a histone octamer; the third level shows post-translational modifications of the histone tails (coloured dots) along with histone variants such as H2A.Z. At the fourth level, differences of nucleosome compaction are visible in active and repressed chromatin regions. The last level illustrates possible higher-order organisation of the chromatin, including interactions with the nuclear lamina.

The chromatin structure at a given locus is defined by the totality of these features and is thought to affect transcription by both modifying the accessibility of DNA to transcription factors and recruiting transcriptional activators and repressors. In this work, our focus will be restricted to mechanisms acting on the nucleosome level, i.e. nucleosome positioning and post-translational modifications of histones.

1.1.2.1 Nucleosome positioning

Close observations of euchromatin have revealed a particular nucleosome organisation around genes loci. Despite variations across species, three features are commonly observed : (i) the $\sim$ 150 base pairs region upstream of the TSS is often depleted of nucleosome, (ii) nucleosomes show a regular spacing at the 5' end of genes and (iii) this regular nucleosome spacing, also called phasing, gradually decreases from the 5' to the 3' end of the coding region, as shown by the *in vivo* profile figure 1.2 (Bai and Morozov, 2010). The nucleosome depleted region (NDR) has a crucial role in the initiation of transcription by facilitating the transcription machinery to assemble onto the promoter region. In human, the first nucleosome downstream of the TSS (the 5' boundary of the NDR) has been demonstrated to be located $\sim$ 10 bp downstream of the TSS at inactive (paused) genes, whereas it is found $\sim$ 30 bp further downstream at active genes (Schones et al., 2008). In yeast, Shivaswamy et al. showed that gene activation and repression are mainly accompanied by changes (eviction, appearance or repositioning) of one or two nucleosomes in the promoter regions.

An important feature of the nucleosome is that it is not a static structure. Histones are constantly evicted and reassembled onto the DNA and can sometimes be

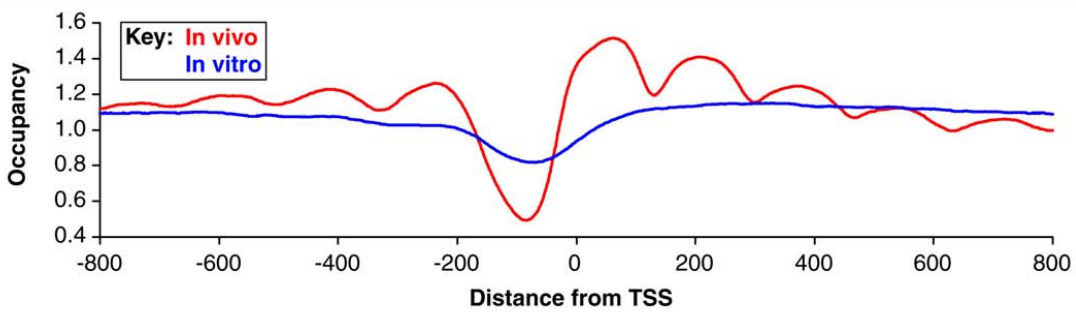

Figure 1.2: Stereotypical nucleosome architecture across TSSs in *S. cerevisiae*. Shown are the *in vivo* and *in vitro* nucleosome occupancy scores (defined subsection 1.2.2) averaged across *S. cerevisiae* genes where the nucleosome depleted region (NDR) and the regular array of nucleosomes are clearly recognizable on each sides of the TSSs. Figure borrowed from Bai and Morozov (2010).

replaced by histone variants and undergo post-translational modifications. The way nucleosomes are organised in the genome is believed to be determined both by *cis* and *trans* factors, i.e. by the affinity of histones for the DNA sequence and the effect of transcription factors and chromatin remodelers (Jiang and Pugh, 2009). The relative influence of these diverse factors remains incompletely understood and will be further discussed sections 2 and 4 as well as in chapter 3.

1.1.2.2 Histone Modifications

The nucleosome consists of two copies each of the core histones H2A, H2B, H3 and H4 that group into two H2.A/H2.B heterodimers and one H3/H4 heterotetramer. These relatively small proteins (atomic mass between 13 and 15 kDa), are highly conserved in eukaryotes and contain the "helix turn helix" motif capable of binding DNA (Alva et al., 2007). The H3 and H4 histones have long tails protruding from the nucleosome in which the epigenetic information is carried under the form of post-translational modifications. These modifications are com-

monly abbreviated as histone name, amino acid name and type of modification. For instance, the tri-methylation of lysine 4 in histone 3 is abbreviated H3K4me3.

Histone modifications have been associated with many cellular processes such as DNA repair, DNA replication, alternative splicing and transcriptional regulation. In this work we will only focus on this last process. While nucleosomes influence transcription by modulating the access of transcription factors to promoter regions, histone modifications are thought to influence transcription either by modifying the chromatin structure and thereby the affinity of transcription factors for their binding site (*cis*-effect), or by recruiting certain effector proteins (e.g. chromatin remodelers) able to change the chromatin structure (*trans*-effect) (Kouzarides, 2007). An example of *cis*-effect is the acetylation of lysine residues which neutralizes their charge and thereby reduces the interaction between the histone tail and DNA. This looser histone-DNA interaction leads the chromatin to a more open chromatin structure and facilitates the binding of transcription factors.

An important characteristic of histone marks is that they rarely occur alone at a single site of a single histone but rather in combination, at multiple sites in the nucleosome. A phenomenon of cross-talk among the different marks within and across histones has been highlighted, where some modifications are requisite to the establishment of other marks. H2BK120ub1 is for instance required for the tri-methylation H3K4me3 (Fuchs et al., 2009). This suggests that a single histone modification is unlikely to determine a cellular function alone but rather in combination with others. This idea is known as the "histone code" which hypothesizes

that instead of having a one modification / one function relationship, a series of "writing" and "erasing" events carried out by histone-modifying enzymes lead to a combinatorial pattern of histone modifications. These modifications are then interpreted by "reader" or "effector" proteins that bind to a specific subset of histone modifications and translate the histone code into biological outcomes such as activation or silencing of the transcription or other cellular responses (Jenuwein and Allis, 2001). In this way, the acetylation and trimethylation of H3K4, H3K36 and H3K79 are typically found in euchromatin, whereas heterochromatin is characterised by H3K9, H3K27 and H4K20 methylation (Portela and Esteller, 2010).

Recent studies have sought to characterise these histone signatures in various functional regions by running non-supervised learning methods on large collection of histone marks. For example, based on a collection of 21 histone methylations and 17 histone acetylations Ucar et al. identified 51 distinct patterns of histone modifications that coincide with different functional regions. This last point will be discussed more widely in chapter 4 (see in 4.1.3.2).

1.1.3 Epigenetics and Human Diseases

Epigenetic mechanisms have received much interest over the past years for the insight they provide into cellular processes under normal and pathological conditions. Cancers, neurological disorders, autoimmune diseases and ageing have been associated with changes in DNA methylation, histone modifications and expression of genes coding for chromatin-modifying enzyme (Portela and Esteller, 2010).

So far, the epigenetic research has mainly focused on cancer where loss of DNA methylation as well as hypermethylation of CpG islands at promoters have been characterised.

- Hypomethylation occurs principally in repetitive sequences, inducing chromosomal instability, loss of imprinting, activation of oncogenes and pseudogenes. For example, the loss of imprinting at the insulin-like growth factor 2 gene has been reported in various types of cancer including lung, liver, breast and colon cancer (Ito et al., 2008).

- Hypermethylation has been observed at specific CpG islands in promoter regions and is associated with gene inactivation. Genes targeted by hypermethylation are mainly implicated in important cellular pathways such as DNA repair, cell cycle control, apoptosis and others (Esteller, 2007).

Besides changes in DNA methylation, global losses of active H4K16ac, H3K4me3, and repressive H4K20me3 histone marks, as well as gains in repressive marks H3K9me3 and H3K27me3 have also been observed in various cancer cells. These alterations in the distribution of histone marks have been linked to abnormal expression patterns of both histone methyltransferases and histone demethylases (Chi et al., 2010). For this reason, histone-modifying enzymes are good candidates for anti-cancer therapy.

More details on epigenetic alterations occurring in cancer and other conditions can be found in Portela and Esteller (2010).

1.1.4 Questions

DNA methylation, histone post-translational modifications and nucleosomes positioning are regarded as the main components of the epigenetic code. They have been shown to be instrumental in the regulation of many cellular processes, however (i) the mechanisms driving the nucleosome organisation and the histone marks deposition, (ii) the pathways epigenetic factors follow to regulate cellular processes and (iii) the combination of histone marks involved in these processes remain limited.

Recently, large research programs such as the ENCODE / modENCODE consortium, the NIH Roadmap Epigenomics Program and the Human Epigenome project have mapped a large number of epigenetic features in a variety of tissues and species (Bernstein et al., 2010; E et al., 2007; Eckhardt et al., 2004). This wealth of data have prompted the development of computational and statistical approaches to gain insight into the determinants of chromatin structure as well as the functions associated with these various chromatin states (Ernst and Kellis, 2010; Karlic et al., 2010). Hopes are that integration of histone modifications, nucleosome positions, transcription factor binding, gene expression level and genome annotations will provide a general picture of the chromatin structure and its multiple functions.

In this work, we will focus on two of these epigenetic mechanisms :

- first, we will aim to understand the relative importance of *cis* and *trans* factors in nucleosome positioning in human (chapter 3) by comparing the

position of nucleosomes *in vitro* and *in vivo*. Because the former capture the sole influence of *cis*-factors and the latter reflect both the influence of *cis* and *trans*-factors, the comparison between *in vitro* and *in vivo* nucleosome positions allows to distinguish regions mainly driven by *cis*-factors (those where nucleosomes have similar positions *in vitro* and *in vivo*) from regions mainly driven by *trans*-factors (those where nucleosomes have different positions *in vitro* and *in vivo*, i.e. where the position variations are greater than the measurement uncertainty). A major challenge in addressing this question though is that the current experimental methods (i) introduce some bias and (ii) sample nucleosome positions from a cell population, making the precise identification of individual nucleosomes not straightforward (see the following section). For this reason, we will introduce in chapter 2 a statistical approach capable of identifying nucleosomes that show consistent positions in a population of cells.

- Based on the nucleosome positions obtained in chapter 2 and a collection of 39 histone modifications, we will then investigate how these epigenetic features are predictive of gene expression in chapter 4.

The remainder of this chapter is organised as follows, section 2 describes how nucleosomes are assayed as well as the measures commonly used to summarize the nucleosome architecture; section 3 gives an overview of the factors known to influence the nucleosome organisation; finally, section 4 summarizes studies conducted on various model organisms and presents the similarities and differences found in these species.

1.2 Experimental Protocol And Measure of Nucleosome Organisation

1.2.1 Experimental Protocol and Bias

The mapping of nucleosome positions consists of two mains steps, the isolation of DNA fragments bound to histones (nucleosomal DNA), followed by DNA measurement. Prior to these steps, the DNA is usually cross-linked to histones in order to maintain the nucleosome structure during the subsequent stages (Jackson, 1999). The isolation of nucleosomal DNA is obtained by treating the chromatin with micrococcal nuclease (MNase), an enzyme that preferentially digests sequences that are not wrapped into nucleosomes. The MNase concentration is chosen so that only regions depleted of nucleosome, i.e. linker DNA, are cleaved. For example, Schones et al. set this concentration to generate roughly 80% mononucleosomes and 20% dinucleosomes. Nucleosomal DNA fragments approximately 150 bp long are then isolated by agarose gel electrophoresis and identified either by microarray experiment or deep sequencing, hence the name of their associated assays, MNase-Chip and MNase-Seq. Beside the technique used to identify the DNA fragments, an important difference between these two assays is that in MNase-Chip, the nucleosomal DNA is hybridized against a control sample (naked genomic DNA), allowing thus to control for possible experimental biases.

The main bias in this experiment arises from the use of MNase known to preferentially cleave AT/TA dinucleotide motifs (Horz and Altenburger, 1981). The resulting nucleosome map thus reflects regions both depleted of nucleosomes and

rich in AT/TA dinucleotides. However, when nucleosomal DNA fragments are separated by agarose gel electrophoresis after MNase digestion, a clear band of approximately 150 bp associated with mononucleosomes is found (Kaplan et al., 2010), suggesting that nucleosome protection is the dominant factor and that the MNase bias only results in imprecise cleavage. This bias can nevertheless be corrected by estimating the specificity of MNase for AT/TA dinucleotides on naked DNA (control sample), as done in MNase-Chip. Unfortunately, in addition of being MNase target, AT-rich sequences are also the signature of antinucleosomal sequences (see subsection 1.3.1). In this way, true linker sites can also be cleaved in control sample, leading to similarities between nucleosome and control maps, as demonstrated in Flick et al. (1986). This problem can therefore decrease the nucleosome signal if the nucleosomal DNA is normalized against a control sample. Despite this known problem and the existence of alternative normalizing approaches (Kaplan et al., 2010), the use of naked DNA remains the method of choice in nucleosome studies.

1.2.2 Measure of Nucleosome Organisation

After the nucleosomal DNA fragments are identified by sequencing or microarray experiments and mapped against a reference genome (only in the case of MNase-Seq), then comes the task of deciphering the nucleosome organisation. Because nucleosomes are mapped from a cell population, the reads or intensities associated with a nucleosome rarely perfectly overlap. Depending on whether a nucleosome is consistently positioned in the cell population assayed, its associated reads (intensities) are normally distributed within $\sim$ 15 bp of the nucleosome centre when

highly phased, or randomly distributed, as shown figure 1.3. This variation in read position reflects both the positional heterogeneity across cells and the over-trimming or undertrimming of the DNA at nucleosome borders during the sample preparation. Since reads associated with phased nucleosomes spreads $\sim$ 30 bp, the uncertainty due to the DNA over/undertrimming by MNase is estimated at +/- 15 bp (Jiang and Pugh, 2009).

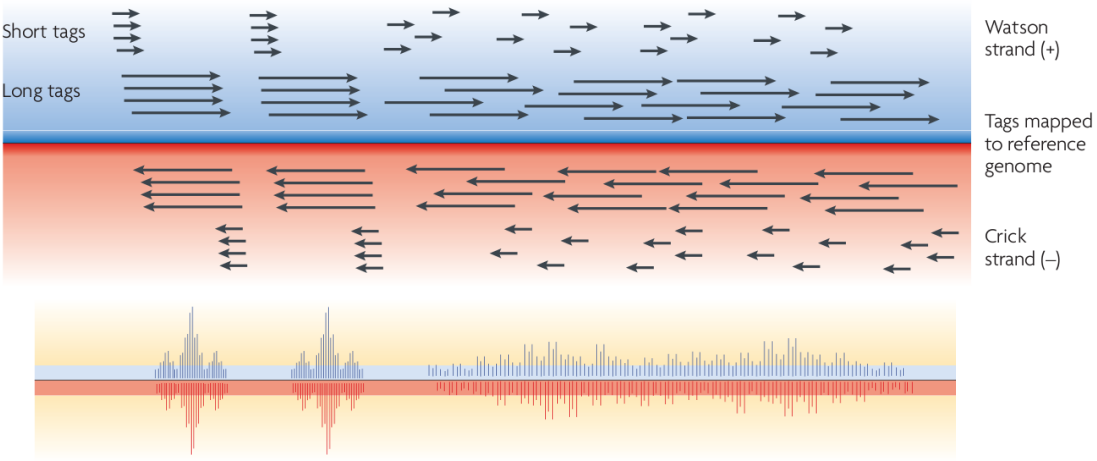

Figure 1.3: Distribution of reads associated with randomly distributed nucleosomes (right) and two nucleosomes consistently positioned (left) in a cell population. This figure, borrowed from Jiang and Pugh (2009), shows reads produced by short and long read sequencing platform mapped against a reference genome on both DNA strand (top). The 5' end position of the reads, corresponding to nucleosome edges, are shifted 75 bp down/up-stream (depending on which strand of the reference genome they map to) to localize nucleosome centres. These nucleosome centres are represented as bar graph where the two clusters of reads on the left side reflect the presence of two phased nucleosomes, whereas the randomly distributed reads (right) correspond to nucleosomes without positional preference.

These complex patterns of reads make the precise identification of nucleosomes non-trivial and raise the need for summary statistics or nucleosome calling methods. While studies based on MNase-Chip data have only use nucleosome calling

methods to identify the precise position of nucleosomes (Lee et al., 2007; Ozsolak et al., 2007; Yuan et al., 2005), MNase-Seq studies have mainly relied on either of two relatively simple statistics, nucleosome occupancy and nucleosome positioning (Kaplan et al., 2009; Valouev et al., 2011, 2008; Zhang et al., 2009).

1.2.2.1 Nucleosome Occupancy and Positioning

In MNase-Seq assays, the first x bp (depending on the platform used) of each nucleosomal fragment are sequenced and mapped against a reference genome. The fragments are then stored in a read vector r whose entries are the number of nucleosome centres (fragment 5' end position $\pm$ 75 bp) at each base pair.

Nucleosome occupancy measures the amount of nucleosomes bound at a given base pair in the cell population. It is defined as the sum of reads covering a given base pair, that is, the sum of nucleosome centres located within 75 bp :

$$\mathrm{NO}(x) = \sum_{i=x-75}^{x+75} r(i) \tag{1.1}$$

Where $\mathrm{NO}(x)$ is the estimated nucleosome occupancy at base pair x.

Nucleosome positioning (also called stringency, phasing) quantifies the extent to which nucleosome positions vary across the cell population assayed. It is defined as the probability that a nucleosome centre is found at base pair x. This probability can be estimated by dividing the number of nucleosome fragments found at base pair x by the number of nucleosome fragments located in the surrounding 150 bp window (Kaplan et al., 2010). However, because nucleosome

positions are measured at a resolution of ~ 30 bp, the resulting nucleosome map has important noise at the single-base-pair level. To overcome this issue, one solution is to smooth the nucleosome centres over the neighbouring positions :

$$\mathrm{NP}(x, w) = \frac{\mathrm{D}(x, w)}{\sum_{i=-75}^{75} \mathrm{D}(x+i, w)} \tag{1.2}$$

$$\mathrm{D}(x, w) = \sum_{i=-w}^{w} \mathrm{K}(i, w) \cdot r(x+i) \tag{1.3}$$

where $\mathrm{D}(x, w)$ represents a smoothed centres count, $\mathrm{K}(u, v)$ a smoothing function (e.g. Gaussian kernel) and w its bandwidth.

Weaknesses of these measures. Despite the use of a smoothing function to account for position uncertainty, the measure of nucleosome positioning remains sensitive to experimental noise because it measures the concentration of reads regardless of coverage, making it vulnerable to random isolated reads that can be interpreted as highly positioned. For example, consider two cases where (i) one read and (ii) $1,000$ reads are found at a position x surrounded by no other reads. Although in both situations the nucleosome positioning is equal to one, the addition of another read in the vicinity of x will change this value to 0.5 and 0.999 in case one and two, respectively (Kaplan et al., 2010).

Nucleosome occupancy, on the other hand, is highly sensitive to read number variations coming from sequencing platform. A recent study compared three major second-generation DNA sequencers with long-range polymerase chain reaction

(PCR) and found that regions extending more than 100 bp give anomalously high or low read numbers, on average approximately every 1000 bp (Harismendy et al., 2009).

Solutions suggested to overcome these weaknesses. To minimize the impact of noise in nucleosome measurement, Kaplan et al. (2010) propose to discard regions with low reads coverage for the estimation of nucleosome positioning and Stein et al. (2010) suggest to use DNA control for the estimation of nucleosome occupancy to correct for the background read variations.

Taken individually, these measurements capture partial aspects of the nucleosome architecture. It would therefore be desirable to combine the two concepts into one single statistic where nucleosomes could be identified without looking for regions with large nucleosome occupancy and positioning scores. To address this need, we introduce NucleoFinder in the following chapter.

1.2.2.2 Nucleosome Calling Methods

Beside nucleosome occupancy and positioning, methods specifically tailored to the detection of nucleosomes have been developed recently. They address the problem of nucleosome identification from a signal processing perspective where the purpose is to filter background noise and score peaks associated with well-positioned nucleosomes. These methods include Laplacian of Gaussian (Zhang et al., 2008b), Fourier Transform (Flores and Orozco, 2011), correlation with nucleosome footprint (two peaks on both strands separated from $\sim$ 150 bp, Weiner et al. 2010), mixed methods (Di Ges et al., 2009) for MNase-Seq data; Laplacian

of Gaussian (Ozsolak et al., 2007) and Hidden Markov Models (HMM) (Kuan et al., 2009; Lee et al., 2007; Yassour et al., 2008; Yuan et al., 2005) for MNase-Chip data.

As mentioned in 1.2.1, the design of MNase-Chip experiment accounts directly for the MNase bias by competitively hybridizing the nucleosomal DNA against a control sample. This implies that the methods tailored for MNase-Chip data do not need to take into account this bias. Conversely, MNase-Seq experiment assays nucleosomal and control DNA separately, which means that the MNase bias can only be corrected at the nucleosome calling stage. However, because the majority of the MNase-Seq studies do not include control samples (only Valouev et al. 2011, 2008; Zhang et al. 2009 do), none of the current MNase-Seq calling methods have been developed with this feature. The method we introduce in the next chapter addresses this bias.

As examples, and because these methods will be compared to our approach in the next chapter, we give a brief description of two of these methods designed for MNase-Seq data.

- TemplateFilter aims to characterise the pattern of forward and reverse strand reads at the two ends of a nucleosome by looking at their correlation with representative templates. A first set of nucleosomes is identified by running the method with Gaussian-shape templates. This set is then used to refine new templates, representative of nucleosome in the data. Due to possible over/under trimming of the DNA by MNase, the correlation between forward and reverse strand reads is calculated over a range of

distances (within a boundary define by the user), from which nucleosomes positions are inferred by choosing the largest correlation values (Weiner et al., 2010).

- NPS filters the background noise by removing high frequency components from a wavelet decomposition. Then, peaks are detected by Gaussian smoothing followed by Laplacian edge detection to identify inflection points in the data. For each peaks whose width is comprised between 80 and 250 bp, a p-value is calculated as the probability that the same or larger number of reads can be observed by chance within the peak region, given the coverage (Zhang et al., 2008b).

1.3 Determinants of Nucleosome Positioning

After reviewing the techniques currently used to assay and measure the nucleosome organisation, we now turn to the factors that determine this nucleosome architecture : (i) the affinity of nucleosome for the underlying DNA sequence (*cis*-factors), (ii) proteins competing with histones to bind DNA (*trans*-factors) and (iii) the steric hindrance between nucleosomes that constrain their lateral movement.

1.3.1 Sequence Based Factors

Eukaryotic genomes have evolved to facilitate nucleosome formation. An elegant study from Zhang et al. (2009) tested the hypothesis of whether the yeast genomes has evolved to favour nucleosome formation by comparing *in*

vitro chromatin assembly in *S. cerevisiae* and *E. coli*. In this study, random 5-10 kb fragments were generated from both genomes and then assembled with histones such that nucleosome positioning was only driven by sequence affinity. The resulting chromatin was digested with MNase to extract nucleosomes, from which the DNA was purified and sequenced. The mapped fragments showed that in the presence of both *S. cerevisiae* and *E. coli* DNA, histones are nine time more likely to form nucleosomes on *S. cerevisiae* than *E. coli* DNA. Another *in vitro* study conducted on mouse estimated that the range of affinity between DNA and histones is 1000 fold or greater (Thastrom et al., 1999). Together, these two studies suggest that eukaryotic genomes have evolved to facilitate the binding of histones and raise the question of whether eukaryotic genomes use a "nucleosome code" to control the position of nucleosome, in particular in regulatory regions.

DNA favouring and repelling motifs. Over 120 direct protein-DNA interactions and several hundred water mediated ones are present in a nucleosome (Davey et al., 2002) and although it can binds any DNA sequence, it tends to favour some sequences over others. Unlike transcription factors, nucleosome does not have a sequence specific binding motifs, however the constraint of wrapping DNA around nucleosome means that highly bendable DNA segments are more likely to bind the histone octamer. In this way, Segal and Widom (2009) showed that poly-dA/dT sequences (homopolymeric stretches of A or T, 10-20 bp long or even greater) have poor distortion properties resulting in nucleosome exclusion. In particular, the analysis of the presence or absence of unfavourable 5-mers in nucleosomes sequenced from many organisms indicated a strong role for A-rich motifs as antinucleosomal sequences (Field et al., 2008). Conversely, sequences

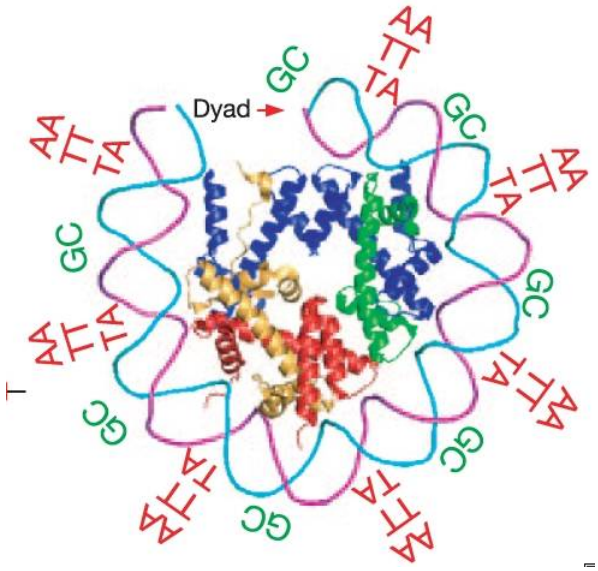

Figure 1.4: Three-dimensional structure of the histones along with the AA/TT/TA and G/C dinucleotides over-represented every 5 bp. This dinucleotide periodicity has been shown to facilitate the bending of DNA around the histone octamer. This crystal structure, first characterized by Luger et al. (1997), represents the nucleosome viewed from above in the superhelical axis (figure taken from Segal et al. 2006).

with 10 bp periodicity of AA/TT/TA and G/C dinucleotides offset by 5 bp have been shown to facilitate the sharp bending of DNA around the histone octamer (Anselmi et al., 1999), see figure 1.4. Despite a statistical enrichment of this periodic signal on the genome scale, these patterns are weak and occurs barely above a random distribution *in vivo* (Mavrich et al., 2008b).

DNA-histone interactions models are statistical models that predict nucleosome position from DNA sequence. They have largely contributed to (i) characterise nucleosome favouring and repelling motifs, and (ii) quantify the effect of *cis* factors *in vivo*. The general procedure followed by these DNA-histone interactions models consists in (i) gathering a training set of sequences associated with nucleosomes highly positioned along with their linkers, (ii) identifying the motifs that discriminate nucleosomes from linkers, and (iii) evaluating the power of discrimination of these motifs by cross-validation on the training set. The majority of these methods are based on position weight matrices (PWM) that model the distributions of k-mers along the nucleosome positions. In this way, Ioshikhes et al. (2006) modelled the variations of AA and TT frequencies along the nucleosome (150 $\times$ 2 PWM), Segal et al. (2006) used a PWM with all the 16 dinucleotides, and Field et al. (2008) combined the same 16 dinucleotides PWM with the distribution of 5-mers in linkers to capture the position-independent signal presents in linkers. This model will be referred to as Segal's model in the following. Beside these PWM based approaches, Peckham et al. (2007) and Gupta et al. (2008) derived a support vector machine (SVM) classifier based on the frequencies of k-mers (k = 1..6) found in nucleosome and linker sequences. Lastly, Yuan and Liu (2008) characterised periodic motifs in nucleosomal DNA

by wavelet decomposition of the dinucleotide frequencies.

On yeast data, these studies identified the 10 bp AA/TT/TA periodicity above mentioned, as well as an over representation of AT-rich and GC-rich motifs in linker and nucleosome sequences respectively. A comparison of these methods on a set of $\sim$ 200 well-positioned nucleosomes showed that Segal, Peckham and Yuan models discriminate nucleosome from linker sequences with similar ROC-scores[1], ranging from 0.8 to 0.9 (Yuan and Liu, 2008). Overall, these methods showed a better ability to predict nucleosome-depleted regions than those occupied by nucleosomes, suggesting that the intrinsic sequence signal is most commonly used to exclude nucleosomes from specific regions (such as the NDR upstream of TSSs) rather than to favour nucleosomes in nucleosome-rich regions (Radman-Livaja and Rando, 2010).

In section 1.4, we will present the results obtained with these models in different model organisms. In chapter 3, we will make use of similar PWMs to estimate the influence of *cis*-factors in human.

1.3.2 *Trans*-Regulatory Factors

Three different mechanisms have been proposed to explain the influence of *trans*-regulatory factors in the *in vivo* nucleosome organisation. First, because of their relatively high abundance, transcription factors (TFs) can take advantage of hi-

[1]The ROC-score is equal to the probability that the classifier under study ranks a randomly chosen positive instance higher than a randomly chosen negative one (Fawcett, 2006). It is thus equal to 1 for a perfect classification and 0.5 for random guessing. The ROC-score is obtained by measuring the area under the classifier ROC curve.

stone turnover or spontaneous "wrapping-unwrapping" of nucleosomal DNA to bind sites transiently uncovered (Choi and Kim, 2008; Newman et al., 2006). This passive binding can further be eased by the affinity of TFs for DNA, where in some extreme cases such as NF-κB p50, TFs can bind nucleosomal DNA with the same affinity as free DNA (Angelov et al., 2004). Also, despite their lack of enzymatic activity, these proteins can recruit chromatin remodelers capable of displacing neighbouring nucleosomes (Radman-Livaja and Rando, 2010).

Second, enzymes called chromatin remodelers use their ATPase domain to remove or slide nucleosomes along the DNA (Boeger et al., 2004; Fazzio and Tsukiyama, 2003), and thus have the ability to override the nucleosome sequence preferences. A well-documented example in yeast is the repressor Isw2, demonstrated to move nucleosomes from their "preferred" sequence to functionally important regions, and consequently to inhibit their regulatory activity. In this way, the comparison of genome-wide nucleosome positioning maps between wild-type and Isw2δ mutant cells revealed that a third of Isw2δ-bound sites resulted in a shift (15 bp on average) of the first nucleosome downstream of the TSS away from the NDR (Whitehouse et al., 2007). This suggests that Isw2 uses the energy produced by ATP hydrolysis to move nucleosomes over the TSS and the NDR, where most transcription factor binding sites lie.

Finally, in addition to *trans*-factors operating prior to the polymerase machinery assembly, the process of transcription by RNA polymerase II (Pol II) itself leads to important changes in chromatin structure (Mavrich et al., 2008b; Parnell et al., 2008; Schones et al., 2008). Components of the transcription initiation complex

are bulky and have the ability to bend DNA when binding to the promoter. Although the exact role of Pol II in reshaping chromatin remains unknown, the fact that nucleosome loss occurs despite the absence of Pol II binding on the mutated TATA box at PHO5 promoter in (Barbaric et al., 2007), or that Pol II is able to bind in the absence of normally required activating signals at an artificially created nucleosome depleted region in (Zhang and Reese, 2007) indicate that nucleosome removal at promoter is a pre-requisite for Pol II binding.

1.3.3 Statistical Positioning

The last factor contributing to the nucleosome organisation arises from the fact that the chromatin consists of highly compacted nucleosome arrays, imposing thus a restriction on nucleosomes lateral movements. Observing that beyond ~ 1 kb from promoter regions, nucleosome phasing dissipates and that this tendency increases with the distance to TSS; noting also that lower nucleosome density at intergenic regions leads to random nucleosome positioning, Kornberg and Stryer (1988) introduced the barrier model for statistical positioning of nucleosomes. According to this model, barriers located along a dense array of nucleosomes restrict the lateral movement of nucleosomes and results in well-positioned nucleosomes, regardless of sequence preferences. Given a nucleosome length and a linker width, this model calculates the probability of observing a nucleosome at a distance d from a barrier by taking into account all possible non-overlapping ways of placing nucleosomes in between. Barriers can be nucleosome excluding sequences like poly-dA/dT found in NDRs (*cis*-factors) or *trans*-factors such as those described in the previous sections. An example is given figure 1.5 where the

nucleosome under the arrow corresponds to the barrier and suffices to position the 9 nucleosomes on its right.

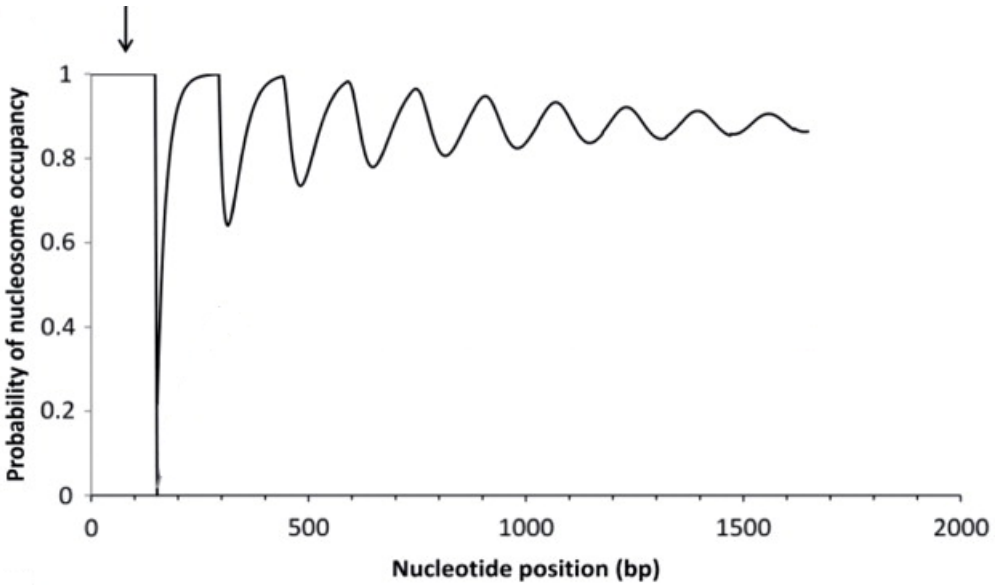

Figure 1.5: Effect of the steric hindrance on nucleosome positioning. Shown is the probability that a position x bp away from the barrier (under the arrow) is contained in a nucleosome. This probability is calculated by the barrier model with the assumption that linkers and nucleosomal DNA are 20 bp and 147 bp long respectively (values estimated in *S. cerevisiae*). With these settings, this figure shows that a single barrier can position $\sim$ 10 nucleosomes on both sides. Figure taken from Stein et al. (2010)

In addition to explaining the "fuzzy" nucleosomes positioning at locations distant from promoters, the Kornberg and Stryer's model also predicts that in genomes with short linkers and short distances between barriers, a fraction of positioned nucleosomes are sufficient to obtain high degree of positioning. Consistent with this idea, $\sim$ 80% of the yeast genome consists of positioned nucleosomes (Lee et al., 2007), where the average linker length is estimated at $\sim$ 20 bp. Conversely, the combination of large linkers in human ($\sim$ 50 bp) and worm ($\sim$ 30

bp) with lower gene density (12-15 genes/Mb in human) result in many fewer positioned nucleosomes in these two species : 16% and 20% of the nucleosomes in worm and human have been shown have a level of positioning equal or greater than $\sim$ 20% (Valouev et al., 2011, 2008).

1.4 Different Mechanisms Across Species

Numerous studies have been conducted in eukaryotic model organisms to measure the effects of *cis* and *trans*-factors on the nucleosome architecture. Different opinions as to which mechanisms primarily determine the nucleosome organisation have arisen. Furthermore, since eukaryotic genomes have different linker size and gene density (as mentioned above), it has been suggested that the relative importance of these mechanisms vary across species.

1.4.1 Yeast

Having the smallest genome among eukaryotic model organisms, *Saccharomyces cerevisiae* has been widely studied in the framework of DNA-histone interactions model. As mentioned in 1.3.1, various approaches have been proposed using PWM, SVM, wavelet decomposition and led to fairly different conclusions. On one hand, Segal and colleagues found that their model predictions (i) separate *in vivo* nucleosome from linker sequences with a ROC-score of 0.86 (Field et al., 2008) and (ii) correlate with *in vivo* nucleosome occupancy with a value of 0.75 (Kaplan et al., 2009), leading them to conclude that much of the nucleosomal organisation is due to *cis*-factors in yeast. On the other hand, Peckham et al. (2007) and Zhang et al. (2009) estimated that only $\sim$ 20% of nucleosomes are

positioned by DNA signals, pointing to a much modest role of the DNA sequence. In this subsection, we review three major studies that illustrate these different views.

Kaplan et al. argue in favour of a nucleosome organisation mainly driven by *cis*-factors. Segal and colleagues sought to determine whether the DNA sequence is the main determinant of the genome-wide nucleosome organisation in yeast (Kaplan et al., 2009). To capture the sole influence of DNA, they assembled chicken histones on the purified yeast genome (*in vitro* map) from which they trained their model (accounting both for the distribution of dinucleotides along the nucleosome and the distribution of 5-mers in the genome). Nucleosome occupancy predictions were then generated from the model and correlated against both *in vitro* and *in vivo* nucleosome maps. Remarkably high values ($R = 0.89$ and $R = 0.75$) were found, leading the authors to the conclusion that (i) their model captures most of the *in vitro* DNA-histone interaction features and (ii) the DNA sequence has a dominant role in the *in vivo* nucleosome organisation in yeast.

They further showed that low nucleosome occupancy was predicted at TSSs, TESs and regulatory sequences, suggesting that these functional sites are encoded in the yeast genome. However, the level of depletion was more pronounced *in vivo* than *in vitro* and in the model predictions, probably because of the action of transcription factors and chromatin remodelers. From their estimation that DNA accounts for $\sim 50\%$ ($R^2 = 0.56$) of the *in vivo* nucleosome occupancy, Kaplan et al. concluded that *in vivo*, nucleosome positions are determined primarily by DNA sequence and advocated for the existence of a nucleosome positioning

code.

Mavrich et al. support the statistical positioning hypothesis and suggest that *cis*-factors have a large contribution in the formation of the barrier. Addressing the same question from the perspective of nucleosome positioning, Mavrich et al. (2008a) came to a different conclusion. After generating an *in vivo* nucleosome map, Mavrich et al. observed that (i) nucleosomes are more densely packed over genes than intergenic regions, (ii) a high level of phasing occurs at TSSs and decays more swiftly in the upstream direction (toward intergenic regions) than in the downstream direction, and (iii) no sign of phasing is found outside regions surrounding TSSs. These observations led Mavrich et al. to conclude that rather than being encoded in the DNA sequence, the nucleosome organisation in yeast is likely to be driven by statistical positioning, where the NDR, acting as barrier, positions the surrounding nucleosomes. Also, they noted that poly-dA/dT motifs are enriched at NDRs, leading them to hypothesize that the underlying sequence has an important contribution in setting up the barrier necessary for statistical positioning at promoters.

Zhang et al. support the statistical positioning hypothesis and question the importance of DNA sequence in the formation of the barrier. Also supportive of the barrier model, Zhang et al. called into question Mavrich et al. hypothesis. Based on their *in vitro* and *in vivo* nucleosomes maps, they showed that *in vitro*, while nucleosome positioning is non-random, it was far below the highly structured organisation observed *in vivo* around TSS and TES. As displayed figure 1.6, the NDR and +1 nucleosome position are much more

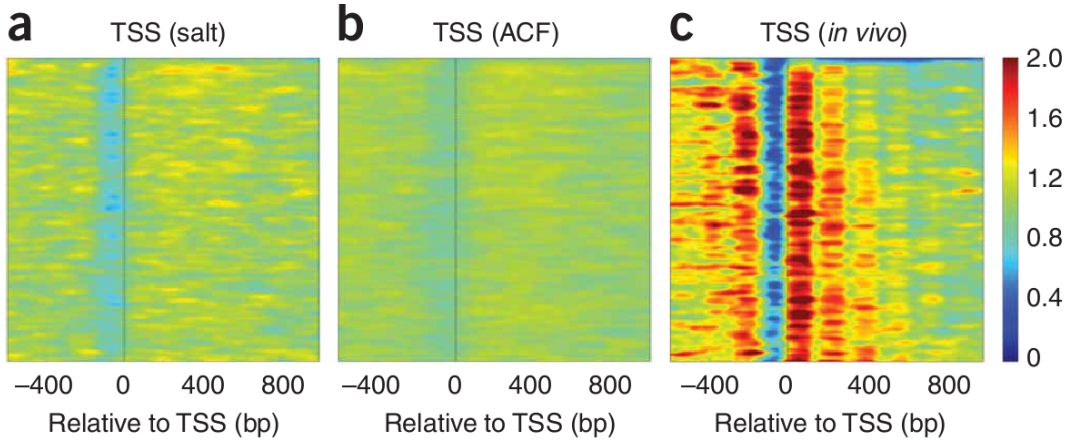

Figure 1.6: Nucleosome density profiles around TSSs for chromatin assembled by salt dialysis, ACF (*in vitro*) or *in vivo*. Figure taken from Zhang et al. (2009).

pronounced *in vivo* than *in vitro*.

To quantify the contribution of the DNA sequence in nucleosome positioning *in vivo*, Zhang et al. counted the number of *in vitro* nucleosomes falling within 10 bp (to account for measurement uncertainty) of the *in vivo* nucleosomes located downstream of TSSs. From this analysis, the authors estimated that intrinsic histone-DNA interactions account for $\sim 20\%$ of the *in vivo* nucleosome positions, and although this result was above expected from randomly positioned nucleosomes, it was far below this obtained using another *in vivo* sample. This value is in line with Peckham et al. (2007) who estimated that 22-25% of the *in vivo* nucleosomes are intrinsically encoded.

According to Zhang et al., this result along with the absence of statistical positioning *in vitro* (figure 1.6) suggest that *in vivo*, the barrier is not due to intrinsic DNA-histone interactions. Instead, they proposed that the NDR is largely influ-

enced by the underlying sequence which facilitates the transcription machinery to bind, but that the statistical positioning is mainly driven by the transcription initiation complex, arguing that unlike *in vitro*, *in vivo*, the +1 nucleosome is well-positioned, next to the TSS.

Shortly after, the same authors directly criticized Kaplan et al. for using nucleosome occupancy to measure the role of *cis*-factors. In this correspondence, Zhang et al. (2010) argued that the key biological problem in the study of nucleosomes is to identify the forces that drive well-positioned nucleosomes, because those nucleosomes provide information as to which regions are consistently occluded or accessible to *trans*-factors. According to them, nucleosome occupancy informs about the average histone level bound to a given region in a population of cells, but does not address the precise location of nucleosomes and cannot therefore be used to quantify the influence of *cis*-factors.

Overall, these studies (Mavrich et al. 2008a; Peckham et al. 2007; Zhang et al. 2009) point to an overestimation of the influence of *cis*-factors in the *in vivo* nucleosome organisation by Kaplan et al., and conclude that, although not the major determinant, DNA has a non-negligible role in the nucleosome architecture in yeast.

1.4.2 Chicken, Worm and Fly

In the same way as in yeast, a modest $\sim$ 10 bp periodicity of AA/TT/TA has been observed along nucleosomes in chicken, worm and fly *in vivo* (Mavrich et al.,

2008b; Satchwell et al., 1986; Valouev et al., 2008). In these last two species, genome-wide studies showed features similar to those found in yeast, i.e. a nucleosome depleted region flanked by well-positioned nucleosomes upstream of TSSs. This stereotypical promoter architecture has not been confirmed in chicken because of the limited size the only study conducted in this species.

Nucleosome positioning. Observations made in worm and fly point toward a lack of positioning signal for the majority of the genome. Valouev et al. (2008) estimated that 16% of the worm genome has a nucleosome positioning equal to or greater than 20%, with a peak at promoters. A comparison of nucleosome positioning across species would be of great interest to understand whether the influence of nucleosome on gene regulation is similar across species or not. Unfortunately, nucleosome positioning was calculated with different parameters in yeast and worm, and no such analysis was performed in fly and chicken.

Nucleosome occupancy. To test whether Segal's model reflects a common mechanism across eukaryote, nucleosome occupancy predictions were generated on the worm genome in Kaplan et al. (2009). A correlation of 0.6 was found between predicted and *in vivo* nucleosome occupancy, from which Kaplan et al. concluded that sequence determinants driving the chromatin structure are common in *S. Cerevisiae* and *C. Elegans*.

Even though these organisms have not been as extensively studied as in yeast, a common trend emerges from these studies : although similar sequence characteristics of nucleosomes are found in yeast, worm, chicken and fly, the important

genome size of these multicellular organisms results in a more fuzzy nucleosome structure, consistently with the barrier model introduced in 1.3.3.

1.4.3 Human

Four recent studies have shed light on the mechanisms driving the nucleosome architecture *in vivo* and *in vitro* in human (Gupta et al., 2008; Schones et al., 2008; Valouev et al., 2011; **?**). In contrast with yeast, the NDR and the phasing occurring downstream of TSSs are not found in the majority of promoters, but rather coincide with Pol II binding *in vivo*. More precisely, both the extent of nucleosome depletion and the level of phasing emanating from the NDR are correlated with the presence of Pol II (Valouev et al., 2011). Beside active promoter regions, a majority of nucleosomes does not show any sign of positioning : Valouev et al. estimated that 20% of the genome is covered by nucleosomes with stringency of at least 23% (lowest stringency level that can be statistically detected) *in vivo*.

Nucleosome specific motifs are enriched *in vitro* and *in vivo*. The study of *in vitro* data revealed that the same G/C-rich and A/T-rich motifs found at the centres and borders of well-positioned nucleosomes in yeast were also present in human. The regions having this particular profile were shown to be enriched *in vivo* by highly positioned nucleosomes, suggesting that DNA sequence does influence nucleosomes *in vivo* (Valouev et al., 2011). In agreement with these results, Gupta et al. found that their SVM trained on the 1,000 highest and lowest occupied regions *in vivo* could classify these regions with a ROC-score of 0.9. This large value points to a real influence of DNA onto nucleosome positioning *in vivo*,

without saying, however, what the size of this effect is genome wide.

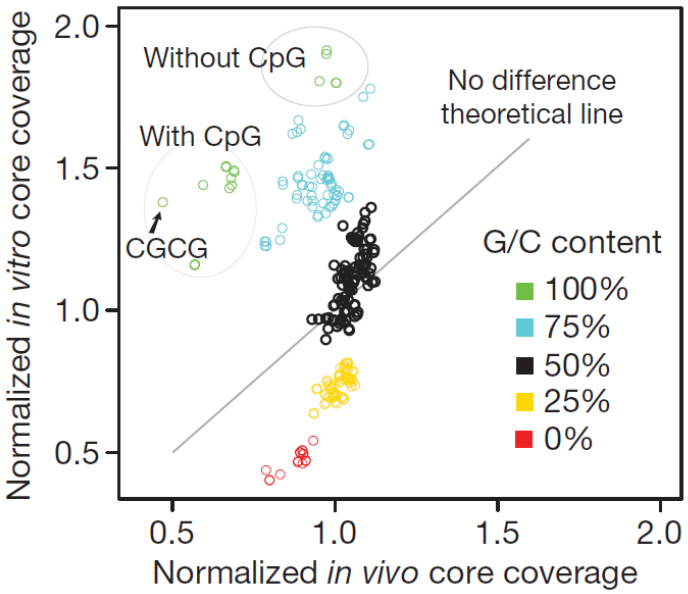

Figure 1.7: Normalized *in vitro* versus *in vivo* nucleosome coverage, where each dot represents the one of the 256 possible 4-mers (coloured in function of their G/C content). The vertical trend in the data shows that the relationship between G/C content and nucleosome occupancy *in vitro* almost disappears *in vivo*. Figure borrowed from Valouev et al. (2011).

***Cis* factors have a small influence *in vivo*.** This question was partly addressed in Valouev et al. (2011) who demonstrated that this influence is very limited and generally overridden by *trans*-factors : the observation of all the tetranucleotide combinations in the genome showed that the strong association between nucleosome occupancy and AT/GC content *in vitro* almost disappears *in vivo* (see figure 1.7). Valouev et al. also tested whether the nucleosome phasing occurring at promoters was intrinsically encoded by looking for regularly spaced nucleosomes *in vitro*. The absence of any sign of phasing *in vitro* indicated that, in accordance with the barrier model, nucleosome phasing is likely to be due to

steric hindrance.

To assess the overall influence of DNA on the genome scale, Tillo et al. compared the level of nucleosome occupancy *in vivo* with predictions from Segal's model. A correlation of 0.28 was found, much less than that (0.75) calculated in yeast, confirming a minor role for DNA sequence in human. In the same study, transcription factor binding sites were shown to have higher than average *in vivo* and predicted nucleosome occupancy, in contrast with yeast (Kaplan et al., 2009). According to the authors, the explanation lies in that most human regulatory sites act in a cell-type specific manner, and therefore, it may be advantageous to keep them occluded by nucleosome to avoid abnormal transcription. This idea was supported by Valouev et al. who showed that the sequences associated with *in vivo* CTCF and NRSF binding sites favour nucleosome formation.

1.5 Discussion

The packaging of DNA into nucleosomes allows to fit large eukaryotic genomes into the nucleus, thereby modulating the access of *trans*-factors to their cognate site. The precise knowledge of nucleosomes locations is thus crucial to (i) understand their role in downstream cellular processes such as transcription regulation and (ii) pinpoint the factors that determine the nucleosome organisation. To address these questions, a non-trivial step of nucleosome identification is required to distinguish well-positioned nucleosomes, hallmark of regulatory events in large eukaryotic genomes, from randomly positioned nucleosomes. Different approaches have been proposed using either summary statistics or nucleosome calling methods, and led to different conclusion regarding the importance of *cis* factors in yeast : while Kaplan et al. estimated that $\sim 50\%$ of the *in vivo* nucleosome organisation is intrinsically encoded, Zhang et al. found a modest value of 20%.

According to Zhang et al. (2010), this discrepancy arises from the use of nucleosome occupancy by Kaplan et al. to compare *in vivo* and *in vitro* nucleosomes. Noting that this measure reflects the average histone level bound to a given DNA region in a cell population, Zhang et al. argued that nucleosome occupancy fails to capture the key biological feature of the nucleosome organisation, i.e well-positioned nucleosomes that indicate the regions consistently occluded or accessible to *trans* factors.

In the previous section we have seen that thanks to the mapping of nucleosomes *in vitro* in yeast, much work has been done to quantify *cis*-effects in this species.

By contrast, fewer studies have been conducted in human, resulting in an incomplete understanding of these mechanisms in our species. The recent mapping of nucleosomes *in vitro* by Valouev et al. should however overcome this limitation, where the authors have already demonstrated that although enriched *in vivo*, nucleosome favouring motifs are generally overridden by *trans*-factors. This study provides a valuable insight into nucleosome organisation in human, with however two weaknesses : it does not precisely quantify the relative effect of *cis*-factors and it is based on nucleosomes positioning, which, as indicated in 1.2.2, is sensitive to technical noises.

For these reasons, we introduce in the next chapter NucleoFinder, a statistical approach that identify well-positioned nucleosomes while accounting for control data. We will show that our model has a better ability to identify well-positioned nucleosomes than other nucleosome calling methods. Based on our model, we will then quantify the influence of *cis*-factors in chapter 3.

Chapter 2

NucleoFinder : a Statistical Approach for The Identification of Positioned Nucleosomes

2.1 Introduction

The identification of nucleosome specific motifs as well as the changes in nucleosome organisation that accompany gene induction requires to know where nucleosomes are positioned with precision. As pointed out in 1.4.3, beside active regulatory regions, most of the human genome shows a weak level of nucleosome positioning, making their localization not straightforward. In the previous chapter, we saw that the summary statistics currently available capture important aspects of the nucleosome architecture but are sensitive to experimental noises and, like nucleosome calling methods tailored to MNase-Seq data, do not account for the known MNase bias toward AT/TA motifs. To overcome these

limitations, we introduce NucleoFinder, a pseudo-Bayesian approach capable of identifying nucleosomes that show both a high level of positioning and an enrichment relative to a control experiment. Although the use of a control sample may reduce the nucleosome signal (section 1.2.1), we still include it to account not only for the MNase bias, but also for other biases that may arise during the sample preparation, amplification and sequencing stages. After a brief description of the pre-processing steps, we introduce our model and assess its fit to the data. Finally, the last two sections address the questions of (i) the sensitivity of the predictions to the window parameter and (ii) the performance of our model compared with existing approaches based on well-characterised features of the nucleosome organisation.

2.2 Data Pre-Processing

The nucleosomes used in this study were isolated in (i) three haematopoietic cell lines (CD4+, CD8+ T cells and granulocytes), (ii) *in vitro* and (iii) in a control sample made of naked DNA (Valouev et al., 2011). They were then treated with micrococcal nuclease to release nucleosome cores, and deep sequenced using the SOLiD platform (Pandey et al., 2008). The resulting 25 bp reads derived from the 5' end of nucleosomal DNA were subsequently mapped against the human hg18 assembly and yielded a genomic coverage ranging between 16-28 $\times$. Before running NucleoFinder, the DNA libraries underwent four pre-processing step to account for experimental bias.

- The sequence tags (reads) mapping to more than one genomic region were discarded from the analysis.

- Due to possible errors during the amplification and sequencing stages (Harismendy et al., 2009), positions covered by an excessive number of reads, above what is warranted by the sequencing depth (binomial distribution p-value $< 10^{-5}$), were discarded, as suggested in Zhang et al. (2008a). For example, in the case of the *in vitro* library (608.7 million reads), given that the mappable human genome is $\sim 2.510^9$ bp long (Rozowsky et al., 2009), we do not expect to see more than 5 reads per base pair.

- Because nucleosome and control dataset are compared in NucleoFinder, read counts in the nucleosome set were linearly scaled so that control and nucleosome samples have equal number of reads.

- Since the reads come from the 5' end of nucleosomal DNA fragments, their positions were shifted to +/- 75 bp (according to their strand) to localize nucleosome centres. This value corresponds to half the estimated average nucleosome length *in vivo* (153 bp) and *in vitro* (147 bp) in Valouev et al. (2011).

2.3 NucleoFinder

2.3.1 Initial Idea

NucleoFinder is based on a similar idea as nucleosome positioning (presented in subsection 1.2.2) where, given a sliding 150 bp window, it looks for an enrichment of reads in a 30 bp window (to account for the MNase cleavage uncertainty) relative to its two 60 bp flanking windows.

The original purpose of NucleoFinder was to develop an automated approach that detects regions both significantly enriched relative to a control sample and presenting a high level of positioning, without having to set arbitrary thresholds on the level of enrichment and positioning. Our first attempt was inspired by peak-calling methods (Ji et al., 2008; Rozowsky et al., 2009) that measure the enrichment of reads in a ChIP sample relatively to a control sample : in a window w, the number of reads in the ChIP sample (x_n) given the sum of reads in the ChIP and control samples (x_c) is assumed to follow a binomial distribution when there is no enrichment :

$$x_c|(x_n + x_c) \sim \text{Bin}(x_n + x_c, \theta = 0.5) \tag{2.1}$$

This simple approach can be extended to nucleosome detection, where instead of looking at a single window w, we are interested in the pattern "*not enriched, enriched, not enriched*" in three consecutive segments 60, 30, 60 bp long. Furthermore, to distinguish relatively high background noise from local peaks in the two side windows, they can be divided into two 30 bp long windows (smaller windows are not considered due to the $\sim$ 30 bp measurement uncertainty).

2.3.1.1 Likelihood

For a given 150 bp region, let $\mathbf{x}^n = \{x_1^n, x_2^n, x_3^n, x_4^n, x_5^n\}$ and $\mathbf{x}^c = \{x_1^c, x_2^c, x_3^c, x_4^c, x_5^c\}$ denote the number of reads falling in each of the five 30 bp bins in the nucleosome and control sets. The bins are grouped into three segments $\mathbf{S} = \{S_1, S_2, S_3\}$ that include bins $\big\{\{1,2\},\{3\},\{4,5\}\big\}$ respectively. Each random variable x_i^n ($i \in 1,..,5$) is distributed according to a binomial with a probability of success θ_i.

A beta prior is put on θ_i with hyperparameters depending on the considered segment. The number of reads is assumed independent from one bin to another which allows the likelihood to be factorized. The likelihood of $\mathbf{x}^n$ is thus expressed as :

$$\pi(\mathbf{x}^n|\mathbf{x}^c, \boldsymbol{\theta}) = \prod_{i=1}^{5} \text{Bin}(x_i^n|x_i^c, \theta_i) \tag{2.2}$$

2.3.1.2 Priors

Whether segment j is background or enriched, the prior on θ_i is beta distributed with parameters :

- $\{\hat{\alpha}, \hat{\beta}\}$ estimated from the data by the method of moments (equation 2.7). The hyperparameters are in fact not directly estimated with $\text{Be}(\hat{\alpha}, \hat{\beta})$ but from the marginal distribution that we introduce below. Since most of the nucleosomes in the human genome show little or no sign of positioning, the data consist mainly of reads randomly distributed. Hence, the estimation of the hyperparameters from the full genome should allow to capture the background nucleosome distribution.

- $\{1, 1\}$, and thus is equivalent to a uniform distribution (equation 2.4). Unlike $\text{Be}(\hat{\alpha}, \hat{\beta})$ which has most of the mass concentrated between 0 and 0.5, this prior gives same weight to θ_{en} in $[0, 1]$.

$$\theta_{bg} \sim \text{Be}(\hat{\alpha}, \hat{\beta}) \tag{2.3}$$

$$\theta_{en} \sim \text{Be}(1,1) = \text{Unif}(0,1) \tag{2.4}$$

2.3.1.3 Empirical Bayes Priors

This way to determine the prior distributions directly from the data belongs to empirical Bayes methods. This type of approach contrasts with standard Bayesian methods in which the priors are fixed before observing the data. In fact, the two approaches are not so different since empirical Bayes can be viewed as an approximation of the full hierarchical Bayesian analysis where the hyper-parameters are set to their most likely values, rather than being integrated out. The use of the data to (i) set the priors and (ii) calculate the likelihood can be seen as an inappropriate double use of the data, but like other pseudo-Bayes approach, empirical Bayes turns out to be extremely effective in actual data analysis (Berger, 2006).

2.3.1.4 Marginal Likelihood

Our main interest is to find regions that match the profile of well-positioned nucleosomes, i.e. when S_2 is enriched and $\{S_1, S_3\}$ are background. To capture this pattern, we introduce eight models that cover all possible combinations of θ_{bg} and θ_{en} in S_1, S_2, S_3, as indicated table 2.1. In each 150 bp sliding region, the marginal likelihoods associated with these eight models are calculated. When $\mathcal{M}_1$

Model	Parameters		
	θ_1	θ_2	θ_3
$\mathcal{M}_0$	θ_{bg}	θ_{bg}	θ_{bg}
$\mathcal{M}_1$	θ_{bg}	θ_{en}	θ_{bg}
$\mathcal{M}_2$	θ_{en}	θ_{en}	θ_{en}
$\mathcal{M}_3$	θ_{en}	θ_{bg}	θ_{en}
$\mathcal{M}_4$	θ_{en}	θ_{en}	θ_{bg}
$\mathcal{M}_5$	θ_{bg}	θ_{en}	θ_{en}
$\mathcal{M}_6$	θ_{en}	θ_{bg}	θ_{bg}
$\mathcal{M}_7$	θ_{bg}	θ_{bg}	θ_{en}

Table 2.1: Models Priors : to precisely capture the profile of well-positioned nucleosomes, we introduce eight models that cover all possible combinations of θ_{bg} and θ_{en} in S_1, S_2, S_3.

has the largest marginal likelihood among the models, this region is considered as a well-positioned nucleosome. The reason for using eight models is that it allows a better discrimination between well-positioned nucleosomes and background or randomly enriched regions. Since the calculation of the marginal likelihood is similar across the models, we only illustrate it with $\mathcal{M}_1$, where the priors are as follows :

$$\pi(\boldsymbol{\theta}|\mathcal{M}_1) = \begin{cases} \text{Be}(\hat{\alpha}, \hat{\beta}) & ,\forall i \setminus \{3\} \\ \text{Be}(1,1) & ,i = 3 \end{cases}$$

$$\begin{aligned} \pi(\mathbf{x}|\mathcal{M}_1) &= \int \pi(\mathbf{x}|\boldsymbol{\theta}, \mathcal{M}_1)\pi(\boldsymbol{\theta}|\mathcal{M}_1)d\boldsymbol{\theta} \\ &= \int \text{Bin}(x_3^n|x_3^c, \theta_3)\text{Be}(1,1)d\theta_3 \cdot \prod_{i\in\{1,2,4,5\}} \int \text{Bin}(x_i^n|x_i^c, \theta_i)\text{Be}(\hat{\alpha}, \hat{\beta})d\theta_i \end{aligned} \tag{2.5}$$

$$\pi(\mathbf{x}|\mathcal{M}_1) = \text{Be-Bin}(x_3^n|x_3^c, 1, 1) \prod_{i\in\{1,2,4,5\}} \cdot\text{Be-Bin}(x_i^n|x_i^c, \hat{\alpha}, \hat{\beta}) \tag{2.6}$$

The number of reads $\mathbf{x}$ and the probability of success θ_i are independent across bins, which allows to factorize the marginal likelihood (equation 2.5). Since the binomial and beta distributions are conjugate, the marginal likelihoods in 2.6 can be computed analytically using the beta-binomial distribution (Gelman et al., 2004). As mentioned in the priors subsection, the hyperparameters are estimated with the marginal distribution from the data by the method of moments. The marginal being beta-binomial, $\hat{\alpha}$ and $\hat{\beta}$ are given by :

$$\begin{aligned} \hat{\alpha} &= \frac{nm_1 - m_2}{n(\frac{m_2}{m_1} - m_1 - 1) + m_1} \\ \hat{\beta} &= \frac{(n - m_1)(n - \frac{m_2}{m_1})}{n(\frac{m_2}{m_1} - m_1 - 1) + m_1} \end{aligned} \tag{2.7}$$

where m_1 and m_2 are the first two sample moments, $n = \sum_i \frac{x_i^n + x_i^c}{N}$ and N the number of observations.

2.3.2 Modifications

When fitting this first version of NucleoFinder, we realized that many good candidates were not detected because few reads (generally one), corresponding to background noise, were found in the side segments $\{S_1, S_3\}$ of the nucleosome sample but not in the control sample. These regions were interpreted as enriched and resulted in an over representation of models 2, 4 and 5. To circumvent this issue, instead of using a binomial distribution to measure the enrichment in the nucleosome sample, we decided to model the difference of reads between nucleosome and control samples. In this way, instead of comparing the number of

nucleosome and control reads in 30 bp windows, we now work with the difference of reads in the two samples. The difference is truncated to zero since negative values do not provide any additional information beside the fact that the considered region is not enriched. This truncated difference is then modelled using a Poisson distribution where prior distributions on the rate parameter are either gamma or uniform, similarly to what we did above.

2.3.3 NucleoFinder

2.3.3.1 Likelihood

Similarly to the first model formulation, given a 150 bp region, let $\mathbf{x} = \{x_1, x_2, x_3, x_4, x_5\}$ denotes the corrected number of nucleosome centres falling in each of the five 30 bp bins. Each random variable x_i is assumed Poisson distributed with parameter λ_i that depends on the considered segment. The likelihood of $\mathbf{x}$ is expressed as :

$$\pi(\mathbf{x}|\boldsymbol{\lambda}) = \prod_{i=1}^{5} \mathrm{Po}(x_i|\lambda_i) \tag{2.8}$$

2.3.3.2 Priors

Whether segment j is background or enriched, the prior on λ_i is either Gamma distributed with parameters also estimated by the method of moments from the data or uniformly distributed in the interval $[0, x_{max}]$, where x_{max} denote the maximum read counts in $\mathbf{x}$:

$$\pi(\lambda_{bg}) = \text{Ga}(\hat{\alpha}, \hat{\beta}), \tag{2.9}$$

$$\pi(\lambda_{en}) = \text{Unif}(0, x_{max}), \tag{2.10}$$

In this way, the background prior distribution matches the empirical distribution mainly made up of background regions. The prior associated with enriched regions gives the same weight to any count number, which, unlike the background prior, does not penalize large counts.

2.3.3.3 Marginal Likelihood

In the same way as the first version of NucleoFinder, our main interest is to find regions i such that S_2 is enriched and S_1 and S_3 are background. Using the same combination of priors, we introduce eight models that cover all possible combinations of λ_{bg} and λ_{en} in S_1, S_2, S_3. Here again, we illustrate the calculation of the marginal likelihood with $\mathcal{M}_1$, where the priors are as follows :

$$\pi(\boldsymbol{\lambda}|\mathcal{M}_1) = \begin{cases} \text{Ga}(\hat{\alpha}, \hat{\beta}) & ,\forall i \setminus \{3\} \\ \text{Unif}(0, x_{max}) & ,i = 3 \end{cases}$$

$$\begin{aligned} \pi(\mathbf{x}|\mathcal{M}_1) &= \int \pi(\mathbf{x}|\boldsymbol{\lambda}, \mathcal{M}_1)\pi(\boldsymbol{\lambda}|\mathcal{M}_1)d\boldsymbol{\lambda} \\ &= \int \text{Poi}(x_3|\lambda_3)\text{Unif}(0, x_{max})d\lambda_3 \cdot \prod_{i\in\{1,2,4,5\}} \int \text{Poi}(x_i|\lambda_i)\text{Ga}(\hat{\alpha}, \hat{\beta})d\lambda_i \\ \pi(\mathbf{x}|\mathcal{M}_1) &= \frac{1}{x_{max}} \int \frac{\lambda_3^{x_3} e^{-\lambda_3}}{x_3!} d\lambda_3 \cdot \prod_{i\in\{1,2,4,5\}} \text{Neg-Bin}(x_i|\hat{\alpha}, \hat{\beta}) \end{aligned} \tag{2.11}$$

The Poisson and gamma distributions being conjugate, the marginal likelihood can be calculated in closed-form using the negative-binomial distribution (Gelman et al., 2004). This is however not the case with the uniform prior, where the marginal likelihood can nevertheless be calculated by successive integrations by part. To avoid the calculation of this integral every 150 bp window, we compute it for $x_3 \in [0, x_{max}]$ and store the results during the initiation stage of NucleoFinder to speed-up the execution. Taken together, this precomputation and the use of the negative-binomial distribution make NucleoFinder a fast method, well-suited to be run on the genome-wide scale. On chromosome 20 for example, the execution time on a 3 GHz CPU is approximatively 10 minutes. When extrapolated to the size of the human genome ($\sim$ 3.1 Gb), the execution time is expected to be around 8 hours.

The estimation of the hyperparameters with the negative-binomial distribution leads to :

$$\hat{\alpha} = m_1 \cdot \hat{\beta} \tag{2.12}$$

$$\hat{\beta} = \frac{m_1}{-m_1^2 + m_2 - m_1} \tag{2.13}$$

As already mentioned, a nucleosome is detected when the marginal likelihood associated with $\mathcal{M}_1$ is the largest of all models. In addition to the information of the presence/absence of a nucleosome, it would be desirable to have a sense of how enriched and well-positioned are the detected regions. This will be particularly useful in chapter 3 where we will train position weight matrices on the 10,000 most positioned nucleosomes to identify nucleosome specific motifs. A

good measurement is provided by the Bayes factor between $\mathcal{M}_1$ and $\mathcal{M}_0$, that measure how much the data support $\mathcal{M}_1$ compared with the null model made of background priors only :

$$\text{BF} = \frac{\pi(\mathbf{x}|\mathcal{M}_1)}{\pi(\mathbf{x}|\mathcal{M}_0)} \tag{2.14}$$

2.3.4 Model Checking

Once the model constructed, we need to assess the fit of the model to the data to avoid possible misleading inferences. If the model fits, then data sampled from the model should look similar to the observed data. Conversely, any systematic differences between the simulations and the data would indicate that the model is inconsistent with the observed data. Model checking is commonly carried out by comparing the observed data $\mathbf{x}$ with the replicated data $\mathbf{x}^{rep}$, generated from the posterior predictive distribution of replicated data (Gelman et al., 1996) :

$$p(x^{rep}|\mathbf{x}) = \int p(x^{rep}|\lambda)p(\lambda|\mathbf{x})d\lambda \tag{2.15}$$

where $\mathbf{x}$ and λ represent the vector of observed data and parameters. To generate x^{rep} under the model, we first draw λ from the posterior distribution $p(\lambda|\mathbf{x})$ and then draws x^{rep} from the likelihood $\mathbf{x}^{rep}|\lambda \sim \text{Poi}(\lambda)$. In our case, we do not have one single model that explains the full dataset, but eight models that represent different aspects of the data. As pointed out in 2.3.1.2, most of the human genome is made up of nucleosomes randomly positioned and should consequently

look similar to data generated under the null model, $\mathcal{M}_0$.

Therefore, we perform model checking using $\mathcal{M}_0$ only. Since the number of reads is independent from one bin to another and that a same prior $\text{Ga}(\hat{\alpha}, \hat{\beta})$ is put on each segment, the comparison between observed and simulated data is carried out at the bin level. The posterior $p(\lambda|\mathbf{x})$ is given by :

$$\lambda|\mathbf{x} \sim \text{Ga}(\hat{\alpha} + N\bar{x}, \hat{\beta} + N) \tag{2.16}$$

with N the number of observations. Figure 2.1 displays the histograms and Q-Q plot of observed and simulated data. The observed data correspond to the number of reads found in 30 bp bins in chromosome 22, the simulated dataset consists of 10^4 read counts generated from the posterior predictive distribution. The almost perfect overlap between the histograms suggests that there is an excellent agreement between the two distributions. If we look at the Q-Q plot, it can indeed be noted that the points comprised between 0 and 5 (representing $\sim$ 97% and $\sim$ 99% of the mass in the observed and simulated dataset) lie along the $y = x$ line. For reads larger than 5, the Q-Q plot is steeper than the $y = x$ line, which indicates that the empirical distribution has a heavier tail.

These two plots therefore show that $\mathcal{M}_0$ fits a majority of regions in chromosome 22. The regions that do not fit well $\mathcal{M}_0$ have higher level of enrichment than expected under $\mathcal{M}_0$ and constitute good candidates for well-positioned nucleosomes if surrounded by "background" bins.

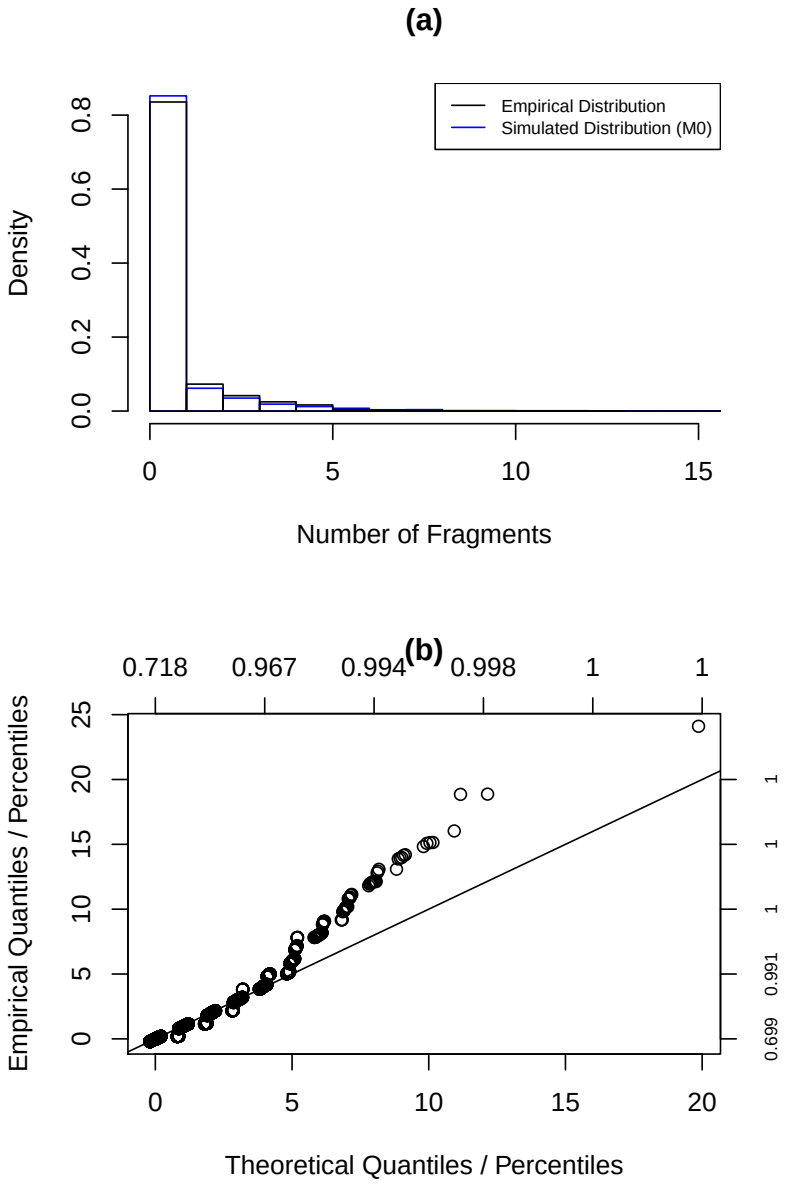

Figure 2.1: **(a)** Histograms and **(b)** Q-Q plot of observed and simulated read counts. Both quantiles and percentiles are showed in the Q-Q plot. The two distributions are almost identical between 0 and 5 where most of the mass is comprised. Above 5, the empirical distribution has heavier tail, indicating that highly enriched regions are not fitted by $\mathcal{M}_0$.

2.4 Sensitivity to the Window Size

NucleoFinder looks for well-positioned nucleosomes from the distribution of reads across five 30 bp windows which together span 150 bp. This last value has been well-documented in the nucleosome literature (Luger et al., 1997; Radman-Livaja and Rando, 2010) and is not called into question. By contrast, the choice of the central window width, defining the uncertainty of the nucleosome position, is not straightforward. It is motivated by the observation that reads at phased nucleosomes span $\sim$ 30 bp (Jiang and Pugh, 2009) but also because this same value was used both in the calculation of nucleosome positioning (Valouev et al., 2011, 2008) and in the detection of nucleosomes by Zhang et al. (2008b). As presented in subsection 1.2.1, this $\sim$ 30 bp uncertainty comes from the effect of DNA over/under-trimming at nucleosome borders during sample preparation. To assess the sensitivity of the central windows size on NucleoFinder predictions, NucleoFinder is run after binning the data into 10, 30 and 50 bp windows, making the division of the side and central windows as follows (7×10, 10, 7×10), (2×30, 30, 2×30), (50, 50, 50) base pairs. These values are chosen because they are all divisor of 150 and small enough to avoid fuzzy nucleosomes to be detected.

An important feature of the nucleosome organisation in most of the eukaryotic genomes is that nucleosome positions are variable across conditions, tissues, individuals (section 1.4). This implies that large-scale regions consistently occupied by nucleosomes on which nucleosome calling methods could be tested do not exist. Even in yeast, considered mainly populated by well-positioned nucleosomes (Radman-Livaja and Rando, 2010), substantial divergences have been

found across studies. An example is provided in Yassour et al. (2008) where the authors compared their nucleosome calling approach with a small set of experimentally-verified nucleosomes by Segal et al. (2006). The authors found that neither their predictions nor those of Yuan et al. (2005) and Lee et al. (2007) are consistent with this reference dataset. This means that there are no true positives available on which nucleosome calling methods can be tested.

Despite the absence of large-scale validated dataset, the window size can nevertheless be chosen in a way that NucleoFinder identifies well-characterised features of nucleosomes and shows a good robustness. We therefore assess the influence of the window size based on (i) the specificity of NucleoFinder after randomly permuting the data, (ii) the ability to identify the uniform nucleosomes spacing occurring near active promoters *in vivo* and (iii) the ability to detect the stereotypical A/T and G/C dinucleotide patterns present at well-positioned nucleosomes *in vitro* (Valouev et al., 2011). In order to keep the execution time reasonable, the first and third tests are carried out only on chromosome 22, where the number of nucleosome is sufficiently large to draw reliable conclusions. In addition to the sensitivity of the window size, we assess the effect of control data by performing the same tests without correcting for the MNase bias.

2.4.1 Specificity

For the three window sizes, NucleoFinder is run on chromosome 22 (CD4+T cells) before and after randomly permuting reads in 1 kb windows. Before permutation, the CD4+T dataset is both corrected and not corrected using the control

set. Even though the locations of well-positioned nucleosomes are unknown, the optimal window width w will be this for which the number of predictions is large before permutation and small after (suggesting both a good sensitivity and specificity). The specificity is calculated by considering the predictions after permutation as false positives (it is not excluded that nucleosome-like patterns occur by chance after permutation though) and the remaining regions as true negatives :

$$\text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}} \tag{2.17}$$

Table 2.4 summarizes these results where both before and after permutation, the number of prediction decreases as the window widens. This decrease is particularly pronounced when the window increases from 10 to 30 bp, whereas the number of predictions associated with $w = 30$ bp and $w = 50$ bp are roughly the same (72% and 79% of their predictions are common before and after permutation). NucleoFinder therefore tends to be more specific as the window width gets larger. One could now wonder which window sizes shows the best balance between sensitivity and specificity. This analysis cannot be done however since the number of true positives is unknown. Lastly, it can be noted that the use of control data have a marginal effect on the number of predictions.

Window width (bp)	Control Before permutation	No Control Before permutation	No Control After permutation	Specificity (%)
10	149,408	148,331	56,371	98.39
30	96,821	93,762	27,218	99.22
50	87,384	90,258	24,863	99.29

Table 2.2: Number of predictions found before (with and without controlling for MNase bias) and after permutation in chromosome 22 (CD4+T cells) along with specificity.

2.4.2 Inter-Nucleosome Distance Comparison

According to the barrier model (subsection 1.3.3), the steric hindrance imposed by RNA Pol II induces a compact and regularly spaced nucleosome array $\sim$ 1kb downstream of the TSS and progressively dissipates when moving away from it (Mavrich et al., 2008a). The spacing just downstream of the TSS of expressed genes is therefore expected to be similar to that estimated at active promoters (178-187 bp, Valouev et al. 2011).

For each window sizes, we estimate the inter-nucleosome distance in the 1 kb region downstream of TSSs associated with highly expressed genes (RPKM $>$ 10) in the whole genome. To account for missing predictions, nucleosomes are enumerated from the TSS to the 3' end of the 1 kb window, without constraining them to be consecutive, and a linear model is fitted between their positions and their associated number. All possible combinations are performed and the numbering associated with the model having the largest R^2 is kept for the estimation of nucleosome spacing. This is illustrated figure 2.2 where the four nucleosomes predicted with $w = 30$ bp are well-aligned when considering a missing nucleosome between the third and fourth predictions.

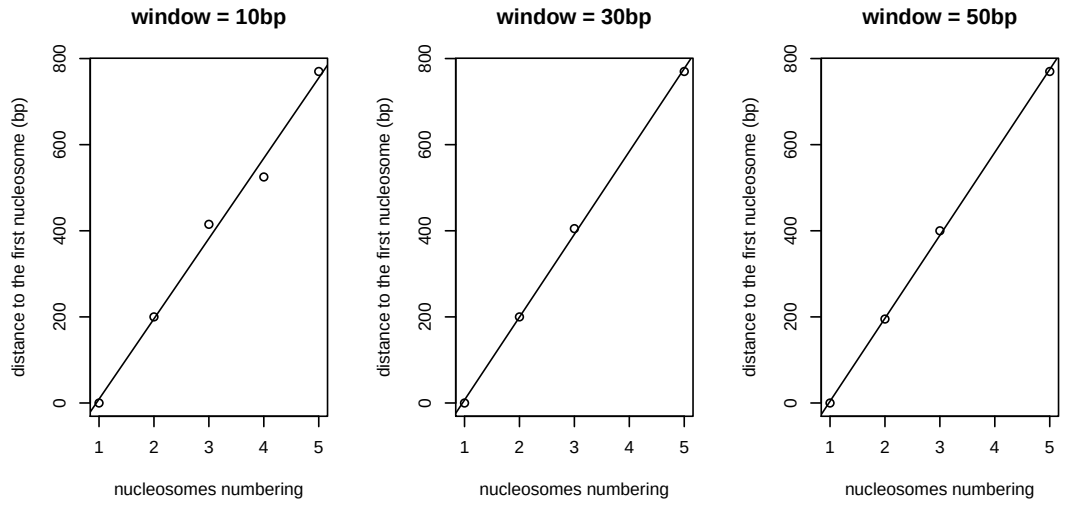

Figure 2.2: Estimation of the inter-nucleosome distance downstream the RCSD1 gene TSS : linear models are fitted between nucleosome positions and numberings. To account for missing nucleosomes, nucleosome numbers are not constrained to be consecutive.

From the inter-nucleosome distances estimated at 2,754 highly expressed genes, we compute the mean, standard deviation and average number of missing (expected minus predicted) nucleosomes, where the expected number of nucleosomes in a 1 kb window is calculated as $1000/182 \approx 5.5$ (i.e. region length / expected spacing). In agreement with the idea that the stringency of NucleoFinder increases with the window size, the nucleosome spacing and the number of missing nucleosomes are on average larger when $w = 30, 50$ bp than when $w = 10$ bp (table 2.3). Three arguments point toward a better estimation of the nucleosome spacing when $w = 30, 50$ bp : (i) unlike when $w = 10$ bp, the nucleosome spacings estimated with $w = 30, 50$ bp are consistent with Valouev et al.; (ii) the standard deviation of these estimates are smaller in the former cases than in the latter; (iii) in comparison with $w = 10$ bp, the distributions of R^2 associated with $w = 30, 50$

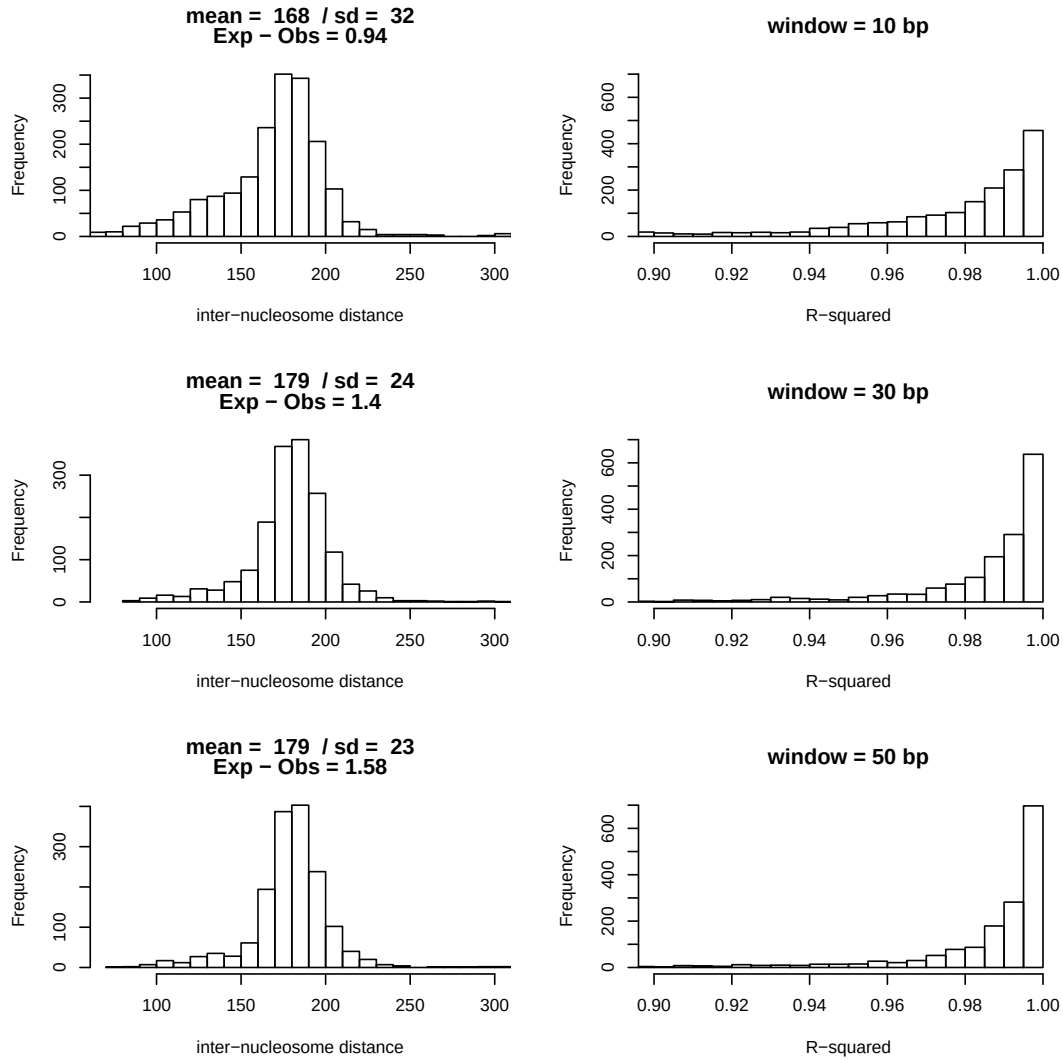

Figure 2.3: Estimation of the inter-nucleosome distance from the 1 kb window downstream of the TSS of expressed genes. Rows correspond to the window widths, columns to the distributions of inter-nucleosome distances (left) and associated R^2 (right). From the distribution of inter-nucleosome distances, the mean, standard deviation and average number of expected minus observed nucleosomes are displayed.

bp are shifted toward 1, suggesting that nucleosomes detected in these settings have a more regular spacing. These results therefore point toward a better accuracy and precision (smaller dispersion) in the estimation of nucleosome spacing when $w = 30, 50$ bp. Lastly, when running the same analyses without control, similar results are obtained (table 2.3).

Window width (bp)	inter-nucleosome distance		Average Number of Missing Nucleosomes
	Mean	Standard Deviation	
10	169	31	0.89
30	180	24	1.53
50	181	24	1.59

Table 2.3: Estimation of the inter-nucleosome distance (mean, standard deviation) and the expected minus predicted nucleosomes without control.

2.4.3 Ability to Retrieve Nucleosome Specific Dinucleotides

As introduced in chapter 1, *in vitro*, the nucleosome organisation is only determined by histone-DNA interactions. Valouev et al. showed that *in vitro*, well-positioned nucleosomes are characterised by an enrichment of G/C motifs and a depletion of A/T motifs at the nucleosome centre. In the framework of the window size sensitivity, it means that the optimal window size will be the one for which predictions have this clear GC/AT-rich pattern at the nucleosome centre. To quantify this pattern for each window size, we measure the difference between the GC and AA/TT frequencies at the centre of nucleosomes predicted in chromosome 22. To account for the two-fold symmetry axis of the nucleosome structure (Richmond and Davey, 2003), the GC and AA/TT frequencies

are calculated from both the sense and antisense sequences associated with the predictions. For each position of the nucleosome, the frequencies $\hat{f}_{GC}$ and $\hat{f}_{AA/TT}$ are calculated along with 95% confidence intervals (Goodman, 1965), as shown figure 2.4 :

$$f_i \in \hat{f}_i \pm \sqrt{\chi_1^2(1-\alpha/m)\frac{\hat{f}_i(1-\hat{f}_i)}{n}} \tag{2.18}$$

where $\chi_1^2(1-\alpha/m)$ is defined as the $100(1-\alpha/m)$th percentile of the chi-square distribution with 1 degree of freedom, n the number of sequences used for the estimation of $\hat{f}_i$ and m the number of dinucleotides. From Goodman (1965), we then compute a confidence interval for the difference $\Delta = f_{GC} - f_{AA/TT}$:

$$\Delta \in \hat{\Delta} \pm \sqrt{\chi_1^2(1-\alpha/m)\frac{\hat{f}_{GC} + \hat{f}_{AA/TT} - \hat{\Delta}^2}{n}} \tag{2.19}$$

where $\hat{\Delta} = \hat{f}_{GC} - \hat{f}_{AA/TT}$. Consistently with their role in promoting or repelling nucleosome, AA/TT and GC frequencies display U-like and bell-like shapes (figure 2.4). These frequencies do not show much variation across window sizes. However, when measuring the difference $\hat{f}_{GC} - \hat{f}_{AA/TT}$ at the nucleosome centre, larger values (0.0037 [0.0028 - 0.0046]) are obtained for $w = 30, 50$ bp than for $w = 10$ bp (0.0033 [0.0024 - 0.0042]). Although the difference between 0.0037 and 0.0033 is not significant, this result suggests a better ability of NucleoFinder to detect well-positioned nucleosomes when run with $w = 30, 50$ bp. Again, the same analysis was carried out without correcting for the MNase bias and

produced almost identical results : $\max(f_{GC} - f_{AA/TT}) = 0.0034, 0.0036, 0.0037$ when $w = 10, 30, 50$ bp respectively.

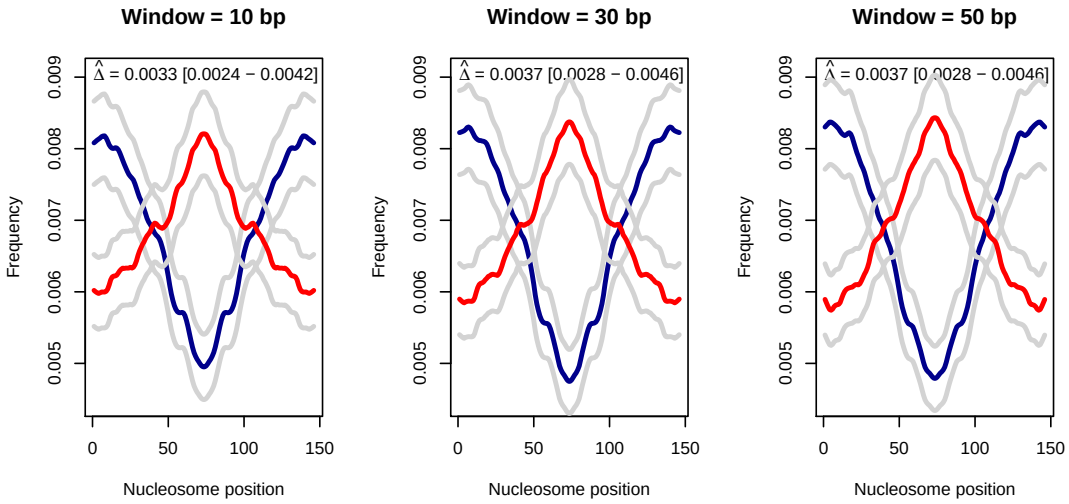

Figure 2.4: AA/TT and GC dinucleotide frequencies and 95% confidence intervals along nucleosome predictions *in vitro*. The differences $\hat{\Delta} = \hat{f}_{GC} - \hat{f}_{AA/TT}$ estimated at the nucleosome centre are displayed on the top of each panel. $\hat{\Delta}$ is larger when $w = 30, 50$ bp.

2.4.4 Summary

To sum up, the number of false positives generated after reads permutation demonstrated that NucleoFinder is more specific when the size of the central window increases. This increase in specificity is more pronounced between $w = 10$ and $w = 30$ than $w = 30$ bp and $w = 50$ bp. This higher specificity of NucleoFinder when $w = 30, 50$ bp was confirmed in the estimation of the inter-nucleosomes distance where larger estimate and larger number of missing nucleosomes were found. More importantly, unlike when $w = 10$ bp, $w = 30, 50$ bp led to unbiased estimations of nucleosome spacing. Finally the analysis of the G/C and A/T content from *in vitro* predictions supported the higher specificity of these latter window sizes for well-positioned nucleosomes. The typical U-like and bell-like shapes were detected more accurately with $w = 30$ bp and $w = 50$. Taken together, these results point to a better ability of NucleoFinder to capture important characteristics of the nucleosome organisation when w is set to 30 or 50 bp. In the rest of this work we therefore set w to 30 bp, in line with previous studies (Jiang and Pugh, 2009; Valouev et al., 2011, 2008; Zhang et al., 2008b). Lastly, the correction of the MNase bias by the control sample appeared to have a minor effect on the nucleosome detection.

2.5 Method Comparisons

Based on the same three criteria, we now compare the performance of NucleoFinder with two other (MNase-Seq) nucleosomes calling methods, NPS and TemplateFilter, introduced in 1.2.2.2. To rule out the possibility that the signature detected by these methods could be identified by simple clustering, we

also run K-means on the vector of reads ($\mathbf{x} = \{x_1, x_2, x_3, x_4, x_5\}$) by iteratively increasing the number N of clusters until one of them matches the pattern of well-positioned nucleosomes (i.e. when the number of read counts is maximum in the central window x_3). In the following analysis, $N = 4$.

Lastly, given that NPS and TemplateFilter (i) are designed to work with reads coming from the 5' end of nucleosomal DNA and (ii) do not account for control experiments, only the first two data preprocessing steps (i.e. discarding reads that map to more than one genomic region and those arising from amplification and sequencing error) are performed for these two approaches.

2.5.1 Endothelin-1 Promoter Region

Before comparing these four methods on a chromosome scale, we first propose a visual inspection of their predictions in the endothelin-1 promoter studied in Ozsolak et al. (2007). To understand what kind of read patterns are captured by these methods, the binned reads and the nucleosome occupancy score (showing the nucleosome envelop) are represented along their predictions figure 2.5. Beside one prediction by TemplateFilter, the four methods show excellent consistency with the ve nucleosomes observed by Ozsolak et al.. This only false positive might arise from the gap occurring in the centre of the leftmost nucleosome and may be interpreted as two sub peaks.

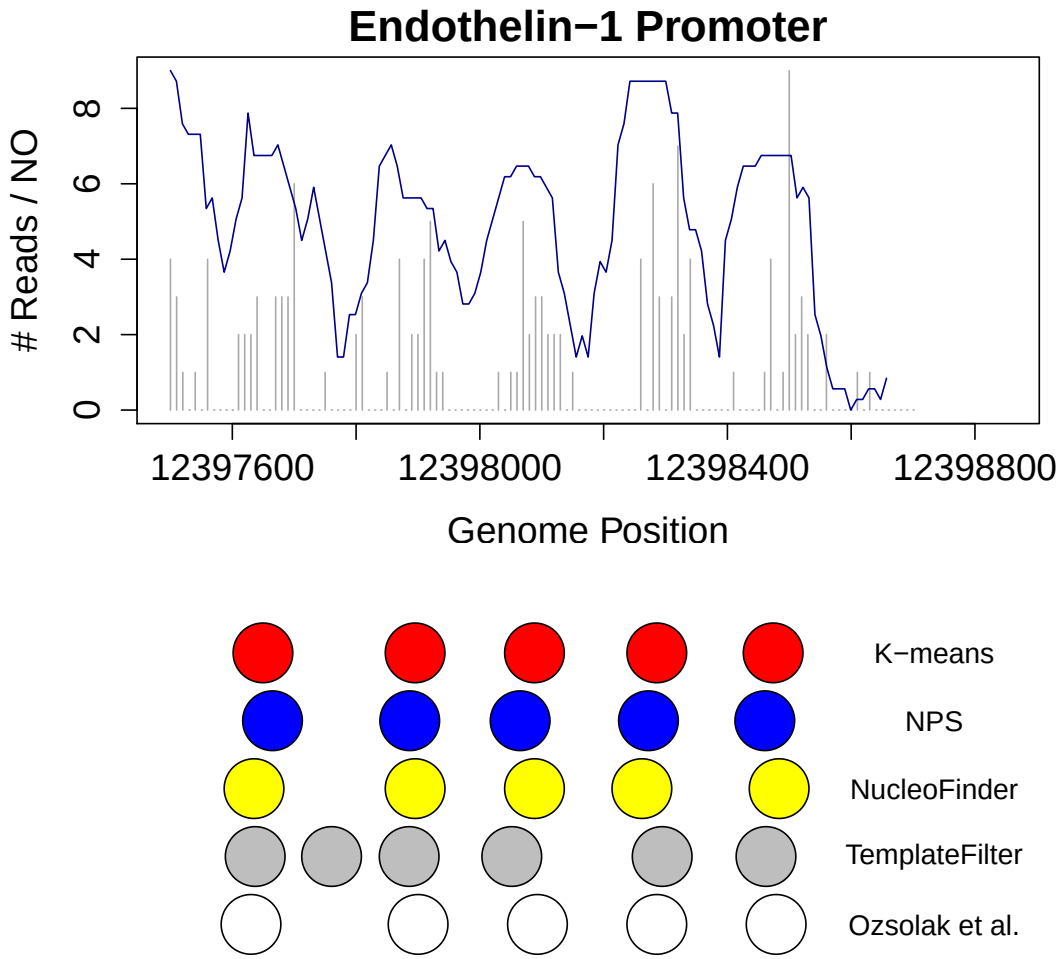

Figure 2.5: Shown are the four methods predictions on the Endothelin-1 Promoter, where beside one prediction by TemplateFilter, the four methods show a good agreement with the ve nucleosomes observed in Ozsolak et al. (2007). The read counts and the nucleosome occupancy score are displayed on the top panel.

2.5.2 Specificity

To have a general idea of how NucleoFinder relates to K-means, NPS and TemplateFilter, the four methods are then compared on chromosome 22 where common predictions across methods, defined as the predictions whose central 30 bp overlap, are counted. Out of 247,874 predicted regions, 23.5%, 18.8% and 12.7% are common to two, three or four methods, that is, more than half of them are predicted by at least two methods. These results indicate that, although agreeing on a large number of regions, these four approaches capture slightly different aspect of the nucleosome signature. When looking at each method individually, it can be noticed that the total number of predictions (table 2.4) as well as the proportion of method specific predictions is much larger in K-means and TemplateFilter than in NPS and NucleoFinder (figure 2.6a). These large numbers of predictions found with K-means and TemplateFilter suggests that they either have higher sensitivity or simply perform poorly.

Method	Before permutation	After permutation	Specificity (%)
K-means	163,584	203,976	94.20
NPS	70,787	68,099	98.06
NucleoFinder	96,821	27,218	99.22
TemplateFilter	161,116	205,768	94.14

Table 2.4: For each methods, the number of nucleosome predictions before and after permutation as well as the specificity computed in chromosome 22.

We then estimate the specificity of these four methods by randomly permuting reads in 1 kb windows along the chromosome 22. A total of 203,976, 68,099, 27,218 and 205,768 false positives are generated by K-means, NPS, NucleoFinder

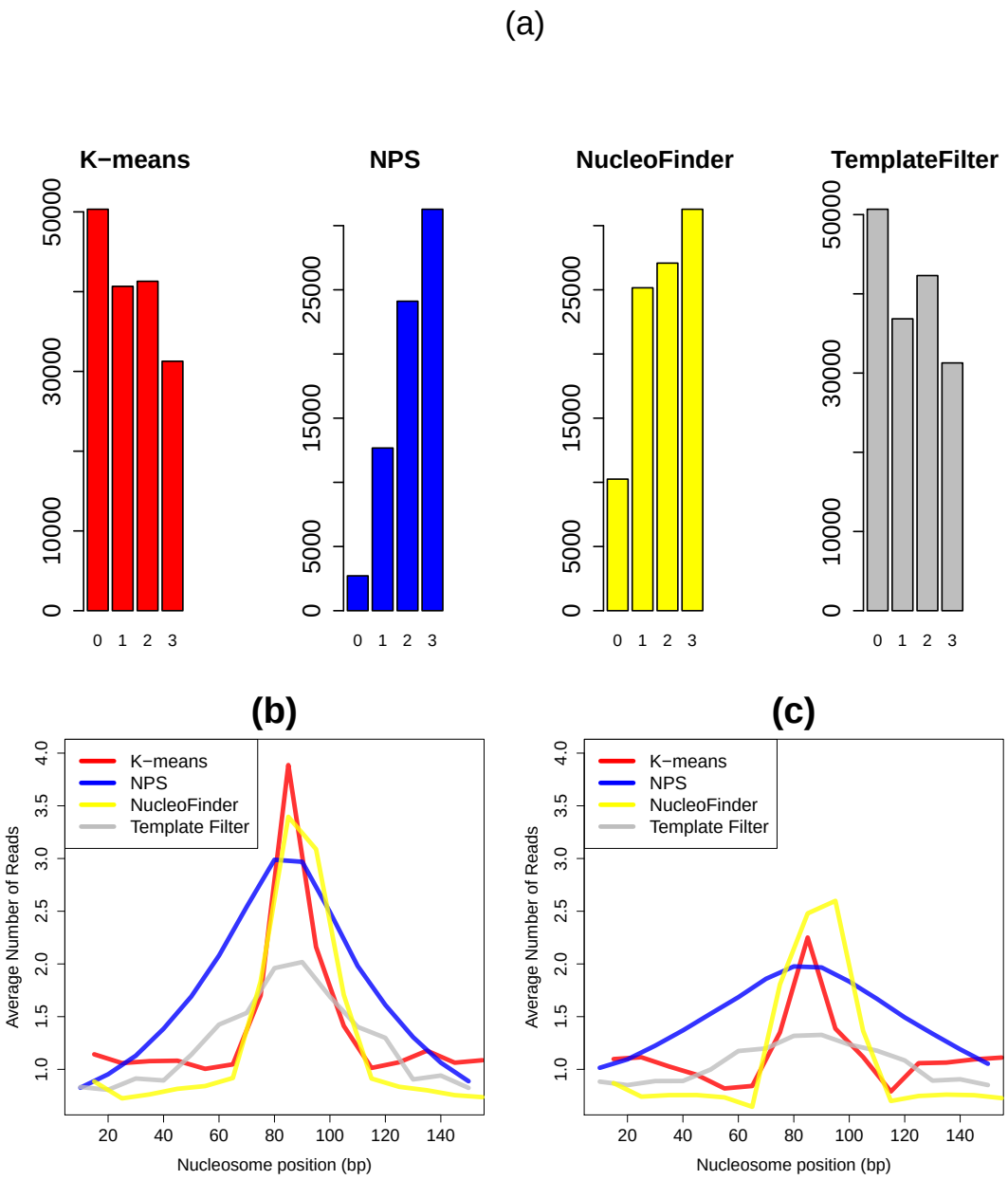

Figure 2.6: **(a)** For each method, the number of common predictions with 0,...,3 other methods is represented. The total number of predictions is much higher in Kmeans and TemplateFilter than NPS and NucleoFinder, suggesting a higher sensitivity of the two former methods. **(b,c)** Nucleosome profile built by averaging reads counts in 10 bp bins along the nucleosomes predictions before and after permutation in chromosome 22. After permutation, the level of enrichment decreases and the profiles flatten out.

and TemplateFilter, resulting in a higher specificity of NPS and NucleoFinder relatively to K-means and TemplateFilter (see table 2.4). If we compare the number of predictions before and after permutation, we observe a $\sim$ 25 % increase for K-means and TemplateFilter, a nearly identical result for NPS, and a decrease of $\sim$ 70 % for NucleoFinder. The fact that after permutation NucleoFinder has the smallest number of predictions and the largest decrease in prediction supports the idea that it is highly specific of well-positioned nucleosomes.

To understand the characteristic of these four methods, a stereotypical profile is built for each of them by averaging the reads count in 10 bp bins along their predictions, before and after permutation. Before permutation, the four methods result in fairly different bell-shaped profiles where unlike TemplateFilter, both K-means and NucleoFinder profiles reveal a sharp enrichment in the central 30 bp window and a low background in the side windows (figure 2.6b). After permutation, the average level of enrichment decreases substantially in all four methods and, except for K-means and NucleoFinder, their profiles flatten out. Although both having a pronounced enrichment in the central segment, these last two methods result in very different numbers of predictions (table 2.4). This difference is likely due to the higher stringency of NucleoFinder in the side segments, as the low average read counts shows figure 2.6b,c. This is consistent with the high specificity of NucleoFinder described earlier and supports the idea that NucleoFinder only detects the signature of well-positioned nucleosomes.

However, the results of K-means after permutation should be interpreted with caution as they rely on the assumption that the data consist of genuine nucleo-

some signal : after K-means has clustered the data into four groups, the label nucleosome is assigned to the cluster (its associated regions) whose profile matches that of nucleosome. This means that whether the input data contain good or poor quality data (permuted data), this approach always generates nucleosome predictions.

The large number of false positive found by TemplateFilter is, on the other hand, probably due to the variety of templates utilized by this method to detect nucleosomes. The representative templates, inferred from the authors data, are correlated with forward and reverse-strand reads for various distances between the two strands (the correlation threshold used to call a nucleosome can be adjusted by the user). The important variability across templates (one or two peaks with different widths, see supplementary figure S1, Weiner et al. 2010) suggests that they can capture various nucleosome profiles (high sensitivity), but at the same time, are more likely to pick up nucleosome like patterns (or background noise) present after permutation. This idea is coherent with the relatively flat nucleosome profiles found with TemplateFilter before and after read permutation (figure 1b,c, see explanation below), corresponding to a mixture of templates with different shapes. One possible explanation for which NucleoFinder suffers less from the permuted data is that it explicitly models the null and the alternative hypotheses (background and enriched regions).

2.5.3 Inter-Nucleosome Distance Comparison

We now turn to the estimation of the nucleosome spacing from the 1 kb regions downstream of highly expressed genes TSSs. In agreement with the probable higher sensitivity of K-means and TemplateFilter, the number of missing nucleosomes and the estimation of nucleosome spacing appear to be smaller in these two methods (0.62, 0.39) as compared with NPS (1.38) and NucleoFinder (1.57, table 2.5). Taking Valouev et al. value of nucleosome spacing as reference (182 bp), it can be noticed that K-means and TemplateFilter result in an under-estimation of ~ 20 bp and ~ 10 bp respectively, unlike NPS and NucleoFinder that provide unbiased estimations. Finally, in addition to being biased, K-means results in a standard deviation almost twice as large as those found in the other methods, indicating that this method perform poorly both in term of accuracy and precision.

Region / Method	Inter-nucleosome Distance		Average Number of Missing Nucleosomes
	Mean	Standard Deviation	
Active promoters	178 - 187	-	-
K-means	159	36	0.71
NPS	187	21	1.39
NucleoFinder	180	24	1.40
TemplateFilter	172	20	0.62

Table 2.5: Estimation of the inter-nucleosome distance (mean, standard deviation) and the expected minus predicted nucleosomes.

2.5.4 Ability to Recover Nucleosome Specific Dinucleotides

Lastly, we compare the four methods on their ability to detect the G/C enrichment and A/T depletion at the centre of well-positioned nucleosomes *in vitro*. In the same way as before, we extract the sequences associated with the methods predictions from which we calculate the AA/TT and GC frequencies along the nucleosome. Again, consistently with their role in promoting or repelling nucleosome, AA/TT and GC frequencies display U-like and bell-like shapes in the four methods (figure 2.7). This time, the variations observed across methods are more important than those across window sizes, where the difference $\hat{\Delta} = \hat{f}_{GC} - \hat{f}_{AA/TT}$ at the nucleosome centre varies from 0.0024[0.0018 - 0.0030] for K-means to 0.0037[0.0028 - 0.0046] for NucleoFinder. However, the four confidence intervals overlap which implies that none of these four methods significantly outperform the others. Nevertheless, the larger values found in NPS and NucleoFinder support the idea that these two approaches have a greater ability to detect well-positioned nucleosomes.

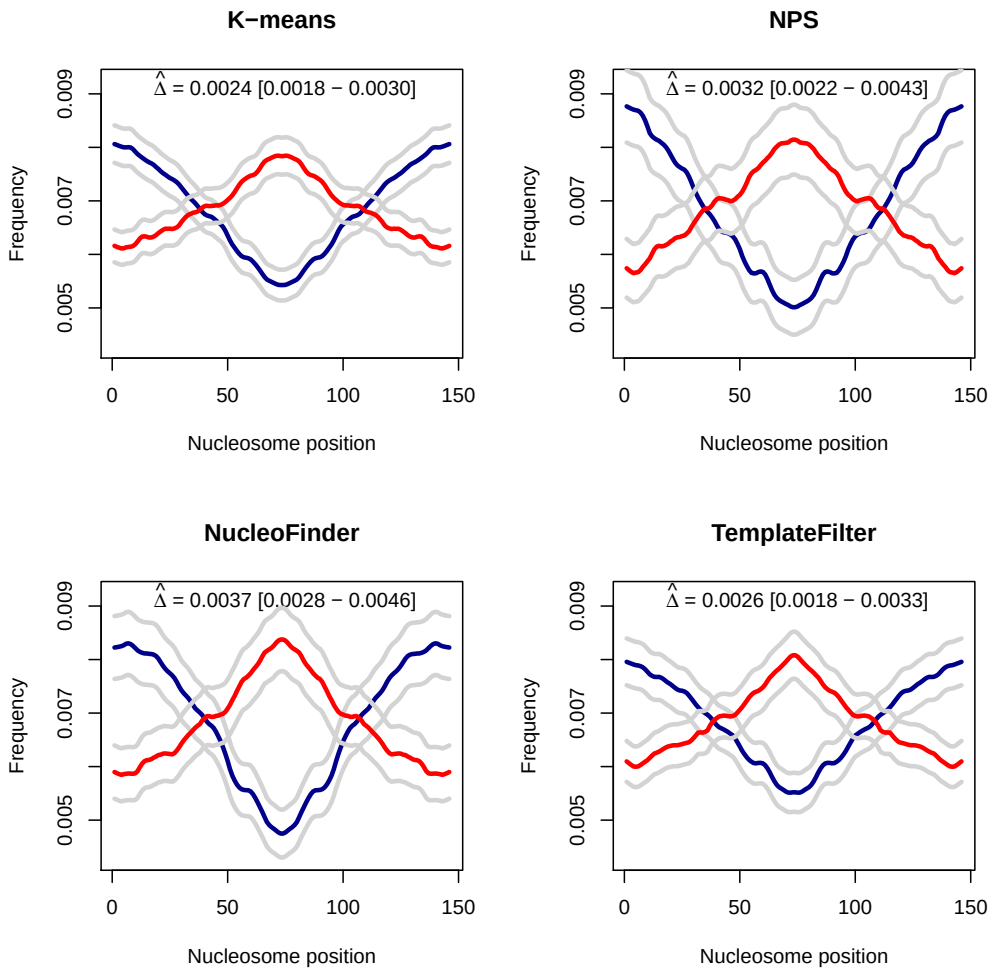

Figure 2.7: AA/TT and GC dinucleotide frequencies and 95% confidence intervals along nucleosome predictions in K-means, NPS, NucleoFinder and TemplateFilter. Larger values of $\hat{\Delta} = \hat{f}_{GC} - \hat{f}_{AA/TT}$ are found in NucleoFinder and NPS than K-means and TemplateFilter.

2.6 Discussion

2.6.1 General

In this chapter, we have introduced NucleoFinder, a statistical approach for the detection of well-positioned nucleosomes. After assessing the fit of the model to the data, the sensitivity of the results to the window size was evaluated. We found that the predictions made with $w = 30, 50$ bp have similar performance, where, in comparison with $w = 10$ bp, they led to a smaller number of false positives, unbiased inter-nucleosome distance estimate and a good ability to detect the stereotypical A/T and G/C dinucleotide patterns. Because $w = 30$ bp was used previously in the calculation of nucleosome positioning and in another nucleosomes calling methods Zhang et al. (2008b), we set the window parameter at this value.

Based on the same three criteria, the performances of NucleoFinder was then compared with K-means, NPS and TemplateFilter. While K-means and TemplateFilter appeared to have a high sensitivity leading to a large number of false positives, an underestimation of inter-nucleosome distance and a lower ability to detect the A/T and G/C dinucleotide patterns, NucleoFinder showed better performances than the three other methods. In particular, the fact that it generated almost ten times fewer false positives (after permutation) for ~ 40 % less positives (before permutation) than K-means and TemplateFilter suggests a good trade-off between sensitivity and specificity.

Surprisingly, predictions generated by NucleoFinder with and without correct-

ing for the MNase bias showed hardly any differences and therefore confirm that the nucleosome protection is the dominant factor in MNase digestion (subsection 1.2.1). Lastly, the poor performances of K-means relative to the other approaches indicates that methods specifically tailored for nucleosome detection do not identify obvious patterns that could simply be detected by unsupervised learning.

2.6.2 Sequence Coverage

NucleoFinder takes into account the sequence coverage through the empirical Bayes approach that estimates the hyperparameter from the data. This means that the classification of a region as background or enriched does not require to define an arbitrary threshold function of the coverage. : the fitting of the Negative Binomial distribution to the data directly influences this threshold by skewing the Negative Binomial toward zero or higher values, whether the sequence coverage is high or low.

However, in a low coverage setting, the predictions made by NucleoFinder are likely to be impacted by experimental noises : the negative binomial being skewed toward zero, the method will be more vulnerable to random noises.

2.6.3 Model Improvement

Lastly, despite the fact that the results obtained with NucleoFinder are consistent with the signature of well positioned nucleosomes, the model is based on the assumption that the truncated difference of two Poisson random variables is also Poisson distributed. To avoid to make this strong assumption, the model

could be modified by keeping nucleosome and control samples reads separated in each window and assuming they are Poisson distributed with rates λ^s and λ^c respectively.

Chapter 3

Are *in vivo* Nucleosomes Primarily Determined by the DNA Sequence in Human ?

3.1 Introduction

In the introduction chapter we have seen that the arrangement of nucleosomes along chromosomes influence the accessibility of DNA and thus provides regulatory mechanism for DNA-dependent processes such as transcription. The factors that determines the nucleosome architecture are known but their relative importance are yet to be determined with precision : are nucleosome positions primarily "encoded" by nucleosome specific motifs (*cis*-factors) or are they a consequence of the activity of chromatin remodelers and transcription factors (*trans*-factors) ? The study of well-positioned nucleosomes *in vitro* and *in vivo* should provides clues as to the mechanisms that determine the *in vivo* organisation : because *in*

vitro nucleosomes are influenced by the DNA sequence only, the comparison of *in vivo* and *in vitro* nucleosome positions will allow us to estimate the relative importance of *cis* and *trans*-factors.

In the introduction chapter we further saw that studies in yeast showed that (i) the sequence information can discriminate nucleosome from linker *in vivo* with a ROC-score of 0.71, and that (ii) $\sim 20 - 25\%$ of the nucleosome positioning *in vivo* is due to intrinsic sequence signal (Peckham et al., 2007; Zhang et al., 2009). In Human, Gupta et al. used the same SVM classifier as Peckham et al. but only on a small fraction of the Human genome. Valouev et al. (2011) observed that G/C-rich loci flanked by A/T-rich motifs were enriched *in vivo*, but did not estimate the magnitude of these *cis*-effects. Only Tillo et al. (2010) addressed this question genome-wide from the angle of nucleosome occupancy, where a correlation of 0.28 between predictions from Segal's model and *in vivo* nucleosome occupancy was calculated. However, as mentioned in the introduction chapter, nucleosome occupancy reflects the average histone levels bound to a given DNA region in a cell population and therefore fails to capture the key information contained in nucleosome positions.

For this reason, in this chapter we propose to estimate the contribution of *cis*-factors in the *in vivo* nucleosome organisation based on NucleoFinder predictions. The remainder of this chapter is organised as follow, section 2 (i) introduces the statistical model used to capture nucleosomes favouring motifs and (ii) assesses its accuracy, section 3 measures the effects of *cis*-factors *in vivo*.

3.2 Nucleosome-Specific Motifs : Identification and Assessment

This section focuses on *cis*-factors where, from *in vitro* NucleoFinder predictions, we first aim to capture nucleosome-specific motifs in position weight matrices (PWMs) and assess their ability to distinguish nucleosome from linker sequences. From these PWMs we will then (i) address the question of whether the nucleosome positioning signal is mainly due to specific motifs or to their combination at specific positions, (ii) investigate whether similar nucleosome favouring motifs are used in human and yeast, and (iii) see whether a relationship exists between the discriminative power of our fitted PWMs and the level of occupancy/spacing *in vitro*.

3.2.1 Data Pre-Processing

To characterise DNA motifs responsible for consistent positioning of nucleosomes, NucleoFinder is run genome-wide on the *in vitro* library. The resulting predictions are then sorted according to their Bayes Factor and the top 10,000 regions are retained. These regions are then successively :

1. centred on the base pair likely to be the true centre, i.e. this having the highest nucleosome positioning score within 30 bp (to account for MNase cleavage uncertainty),

2. extended on both side so that each region spans 250 bp : 150 bp for the nucleosome, 2×50 bp for the linkers (in agreement with Valouev et al. estimation).

3. For each region, the DNA sequence and its reverse complement are extracted to account for the twofold symmetry axis of the nucleosome structure (Richmond and Davey, 2003).

4. Nucleosome and linker sequences are split in two separated sets. We will refer to these two sets as the 10,000 nucleosomes set hereafter.

3.2.2 Identification of Nucleosome Specific k-mers

3.2.2.1 Learning Position Weight Matrix from Sequences

In a similar way to Segal et al. (2006) and Field et al. (2008) who modelled the nucleosome sequence preference using dinucleotides distribution along the nucleosome, we build seven position frequency matrices (PFM) based on the observed k-mers (k in 2,..,8) in the 10,000 nucleosome sequences. This type of approach is motivated by the fact that the structural properties of the DNA helix, such as flexibility or stability, depend on the interactions between neighbouring base pairs (Baldi and Baisnee, 2000; Goodsell and Dickerson, 1994). To identify motifs that best discriminate nucleosomes from linkers, these position frequency matrices are converted into PWM, $\mathbf{\Theta}^k$, by taking the log ratio of the position frequency matrices relative to the linker k-mers frequencies :

$$\mathbf{\Theta}^k_{i,j} = \log_2 \frac{f^k_{i,j}}{p^k_i} \tag{3.1}$$

where $f^k_{i,j}$ is the frequency of k-mer i at position j in the nucleosome set and p^k_i the frequencies of k-mer i (regardless of its position) in the linker set. A pseudo-

count of 1 is added to each cell of the position frequency matrix so that null values are eliminated before log-conversion. The underlying assumption when using a PWM is that any binding site $B = (b_1, ...b_J)$ made of J nucleotides, is a random observation from a product multinomial distribution $\mathbf{\Phi}$, where $\mathbf{\Phi} = (\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_J)$ and $\boldsymbol{\phi}_j = (\phi_{a,j}, \phi_{c,j}, \phi_{g,j}, \phi_{t,j})$ represents the probability of observing nucleotides A, C, G, T at the jth position ($\sum \phi_{.,j} = 1$ for j, $1 \leq j \leq J$, Chen et al. 2007). The score assigned by a PWM to a binding site B is then given by the likelihood of a product multinomial :

$$\pi(B|\mathbf{\Phi}) \propto \prod_{j=1}^{J} \phi_{b_j,j} \tag{3.2}$$

where $J = 150 - k + 1$. Since the coefficients of our PWMs $\mathbf{\Theta}^k$ are on the log scale, the product in 3.2 becomes a sum (hereafter referred to as log-odds score).

3.2.2.2 Identification of Important Motifs

To identify the k-mers carrying important information, the information content $\mathrm{I}_{\mathrm{i}}^{\mathrm{k}}$ (D'haeseleer, 2006), measuring how much the nucleosome distribution differs from the linkers distribution at position i, is calculated as follow :

$$\mathrm{I}_{\mathrm{j}}^{\mathrm{k}} = -\sum_{i=1}^{n_k} f_{i,j}^k \times \log_2(\frac{f_{i,j}^k}{p_i^k}) \tag{3.3}$$

where n_k is the total number of k-mers. Therefore, the information content is maximal when a k-mer is considerably over-represented in the nucleosome se-

quences relative to linker sequences. The information content associated with $\mathbf{\Theta}^k$ are shown figure 3.1(b), where it can be noted that the information is not uniformly distributed along the nucleosome, but concentrated in the middle. This is likely to be due to the over-representation of G/C-rich and under-representation of A/T-rich motifs at the nucleosome centre. Furthermore, up to $k = 6$, the larger k is, the higher the information content is, indicating that larger motifs have a better ability to discriminate nucleosomes from linkers. Above $k = 6$, the dataset is too small to accurately estimate the k-mers frequencies, i.e. large motifs are too infrequent in the 10,000 nucleosomes set to precisely estimate their frequency. Figure 3.1(c) shows the number of observations used to estimate each $f^k_{i,j}$ (for $k = 2, 6, 8$), where it can be seen that 2-mers (k = 2) frequencies are often estimated on several thousand observations, whereas 8-mers (k = 8) frequencies are generally estimated on less than ten observations. Consequently, smaller values of information content are obtained for $k = 7, 8$.

To visualize important motifs, particularly those located in the centre, a sequence logo is built by scaling the 6-mer normalized frequencies ($\frac{f^6_{i,j}}{p^6_i}$) by the information content at each position (Schneider and Stephens, 1990). To ease the comprehension of figure 3.1(a), G/C-rich motifs are coloured in red and A/T-rich motifs in blue (the classification is done using principal component analysis, see following). As observed with dinucleotides in subsection 2.4.3, G/C-rich motifs are highly enriched at the centre of the nucleosome, whereas A/T-rich motifs are barely distinguishable. Consistently, the most over-represented motif in the centre of nucleosome is GCGCGC. To the best of our knowledge, no particular biological function has been associated with this motif. However, a CGCG-box (the four

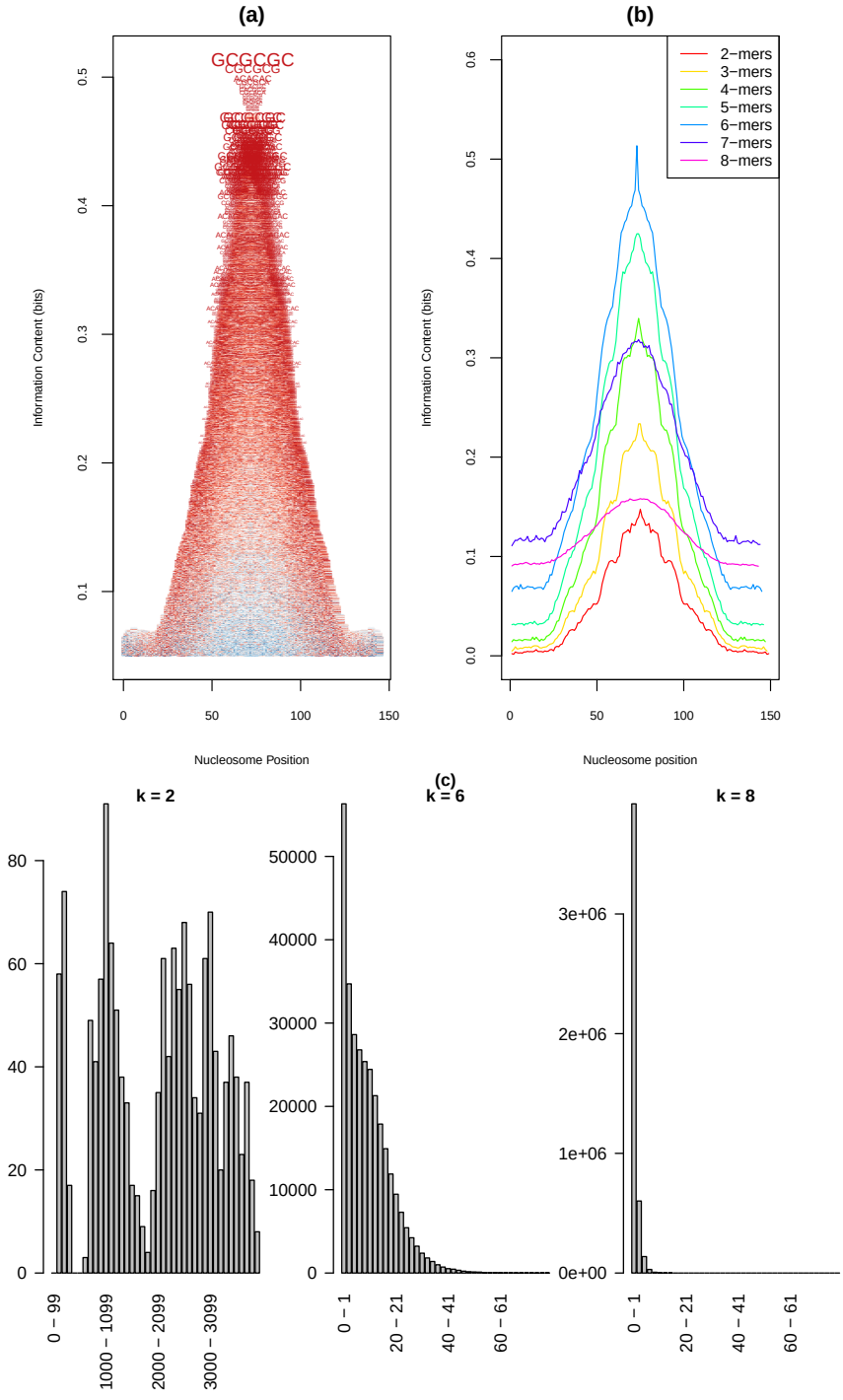

Figure 3.1: **(a)** 6-mers Sequence Logo : GC motifs are enriched at the centre of nucleosomes. **(b)** Information content associated with $\mathbf{\Theta}^k$, k in 2,...,8. The divergence from the nucleosome to the linker k-mers distributions is more pronounced in the central $\sim$ 50 bp region that in the flanks. **(c)** Histogram of the number of observations used for each $f^k_{i,j}$ estimation (for $k = 2, 6, 8$) : 2-mers frequencies are mainly estimated on several thousand observations, whereas 8-mers frequencies are estimated on less than ten observations in a large majority of cases.

central bases of GCGCGC) has been identified in promoters of genes involved in ethylene signaling, light signal perceptionas and other cell functions in Arabidopsis (Yang and Poovaiah, 2002). The transcription factors (AtSR, Arabidopsis thaliana signal-responsive) specific to this CGCG-box have been found to have $\sim 50\%$ identity with two predicted proteins in human, leading the authors to hypothesize that the CGCG-box may represent a new type of DNA-binding domain in Arabidopsis and human. If this DNA-binding domain does exist in human and are preferentially wrapped into nucleosomes, the binding of the AtSRs homologous proteins probably requires the eviction of nucleosomes to bind to their cognate site.

3.2.2.3 Principal Component Analysis

To recover this pattern, a principal component analysis is performed on the standardized $\mathbf{\Theta}^6$ to capture the relationships amongst motifs. A spectral decomposition is thus performed on the correlation matrix of $\mathbf{\Theta}^6$, where the variables are the 2080 6-mers and the units of observations are the 145 positions along the nucleosome (PCA can also be performed by singular value decomposition, as described section 4.3). Each resulting eigenvalues $(\lambda_1, ..., \lambda_{n_6})$ is proportional to the variation explained by their associated eigenvector (principal axis). More precisely, the sum of the eigenvalues is equal to the sum of squared distances between the observations and their multidimensional mean. Hence the proportion of the total variation explained by the mth principal component is :

$$\frac{\lambda_m}{\lambda_1 + ... + \lambda_{n_6}} \tag{3.4}$$

In figure 3.2(a), these normalized eigenvalues suggests that although none of the principal component explains the majority of the variation, the variance of the first principal component is much larger than the others and may thus represents important features in the data. In the case of a $n \times p$ matrix $\mathbf{X}$ whose entries are independent Gaussian variables, Johnstone showed that when suitably normalized, the largest eigenvalues of the sample covariance matrix $\mathbf{S} = (n-1)^{-1}\mathbf{X}'\mathbf{X}$ follows a Tracy-Widom distribution of order 1. Johnstone's theorem allows thus the use of the Tracy-Widom statistic for testing the significance of the k-th largest eigenvalues of a covariance matrix. To be valid however, this test requires that $n \leq p$. This condition being not met by $\mathbf{\Theta}^6$ ($n = 145$ and $p = 2080$), this test can unfortunately not be used on λ_1.

When projecting the variables onto the first two principal axes, it appears that the first one (horizontal axis figure 3.2b) coincides relatively closely to the A/T versus G/C content. To characterise this A/T versus G/C pattern in the sequence logo (figure 3.1), the 6-mers are coloured according to their projection onto the first principal axis, red, for positive values (G/C-rich motifs), blue, for negative values (A/T-rich motifs). This allows to clearly see that the nucleosome core is populated with G/C-rich 6-mers.

To observe more clearly this distribution of A/T versus G/C 6-mers along the

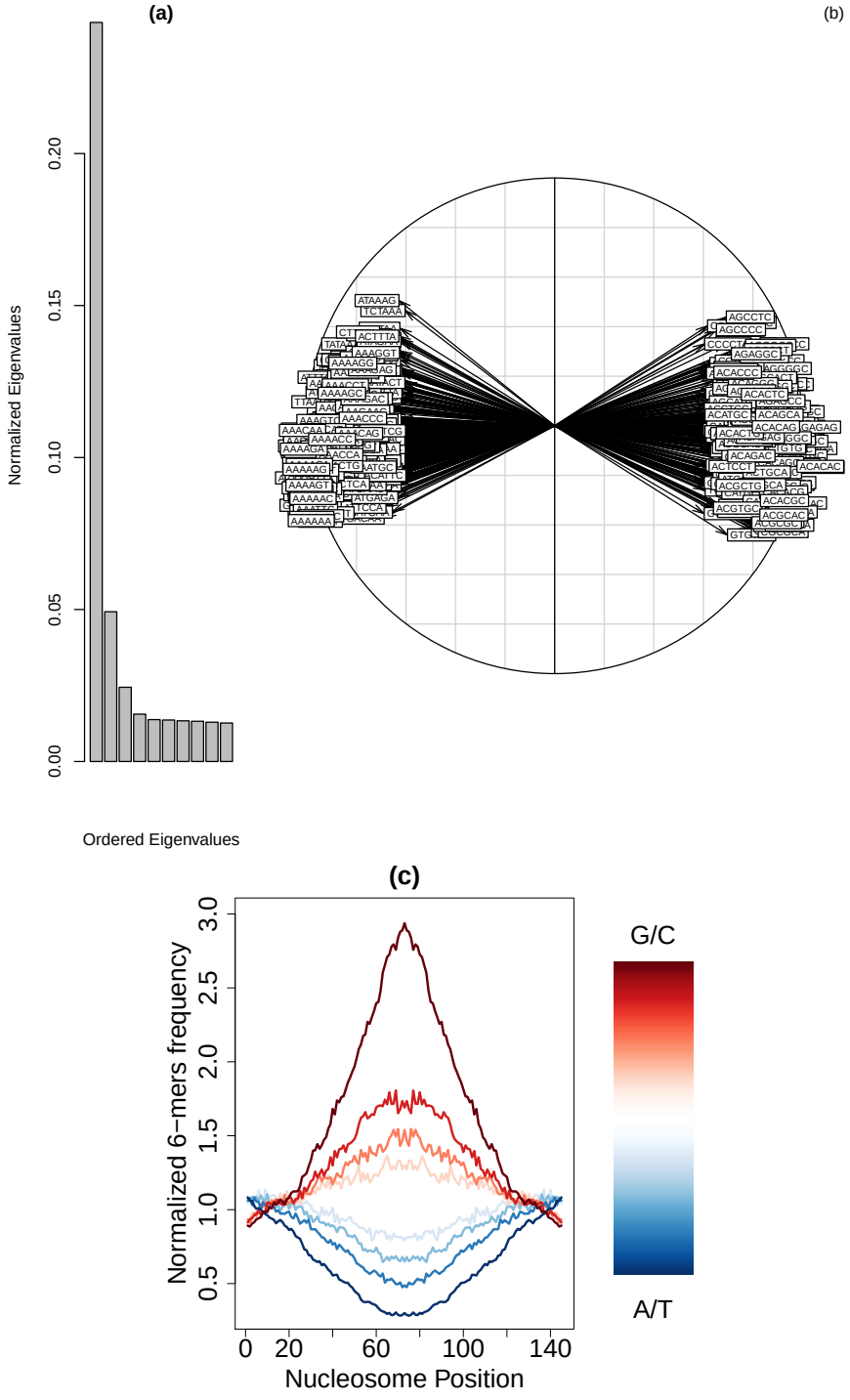

Figure 3.2: **(a)** Normalized Eigenvalues obtained by PCA on $\mathbf{\Theta}^6$. **(b)** Projection of 6-mers onto the first (horizontal) and second (vertical) principal axes, only motifs with large positive and negative projections are shown. The principal axis appears to coincide with the A/T versus G/C content. **(c)** The 6-mers are binned according to their projection onto the first principal axis. Their normalized frequencies are then averaged within each bins. The higher the A/T and G/C contents are, the more pronounced the U-like and bell-like shapes are, respectively.

nucleosome, (i) the 6-mers are binned according to their projection onto the first principal axis (208 6-mers per bin), (ii) for each position j in the nucleosome, their normalized frequencies ($\frac{f^6_{i,j}}{p^6_i}$) are averaged within each bin and plotted figure 3.2(c). The resulting curves, corresponding to the 8 bins, are then coloured according to the average projection values of the 6-mers that constitute them. The same U-like and bell-like shapes found in 2.4.3 are recovered, where, interestingly, the normalized frequencies of 6-mers become larger with the G/C content and smaller with the A/T content. The difference of frequency at the nucleosome centre between the larger G/C and A/T content bins is ~ 2.5, that is, 5 times larger than what we found in the previous chapter with GC and AA/TT dinucleotides. This larger difference shows that this G/C versus A/T pattern is much more pronounced when using (i) the best 10,000 genome-wide NucleoFinder predictions, (ii) 6-mers instead of dinucleotides and (iii) when controlling for linker frequencies rather than genome expectation. All together, these results show that $\mathbf{\Theta}^6$ captures the well-characterised nucleosome favouring and repelling motifs and demonstrate that the level of nucleosome ocupancy is directly linked to the G/C content.

3.2.3 PWM Performances

3.2.3.1 Cross-Validation on the 10,000 Nucleosomes Set

Once the k-mers distribution characterised, the performances of $\mathbf{\Theta}^k$ to discriminate nucleosomes from linkers is evaluated in a 10-fold cross-validation scheme on the 10,000 nucleosomes dataset. Since linker sequences are 50 bp long, i.e. three time shorter than $\mathbf{\Theta}^k$ width, one linker is randomly sample with replacement in

each linker pair to form a 150 bp long sequence. The 10,000 nucleosomes and linker pairs are then randomly partitioned into ten equally sized groups. 10 iterations of training and validation are performed such that within each iteration a different group is held-out for validation while the remaining 9 groups are used for training $\mathbf{\Theta}^k$. For each $\mathbf{\Theta}^k$, a ROC curve is then obtained by varying the log-odds score threshold separating nucleosomes from linkers. The performance of each $\mathbf{\Theta}^k$ are finally computed by measuring the area under the ROC curves (ROC-scores).

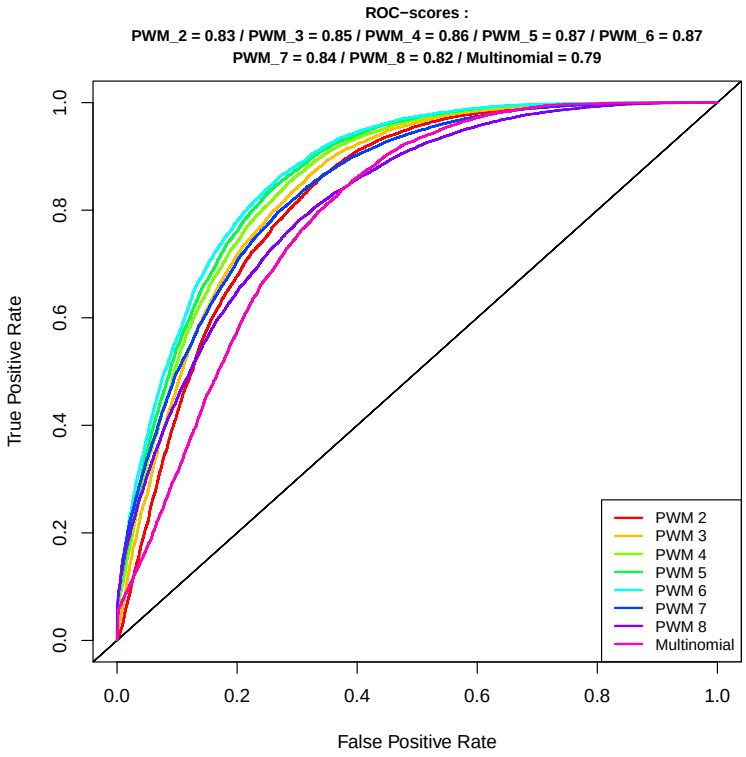

Figure 3.3: The performances of $\mathbf{\Theta}^k$ and $\boldsymbol{\theta}^6$ are evaluated by cross-validation. The ROC-score increases with k.

The obtained ROC-scores range between 0.82 and 0.87 (figure 3.3), indicative of a good discriminative power of $\mathbf{\Theta}^k$ when trained and tested on the 10,000 nucleosomes set. Consistently with the higher information content found with larger motifs, the ROC-score increases with k up to $k = 6$. For $k = 7, 8$, smaller values of ROC-scores are obtained probably because of the same lack of accuracy in the estimation of k-mers frequencies highlighted in the calculation of the information content.

As presented in the introduction chapter (section 1.3), other DNA-histone interactions models based on PWM, support vector machine, wavelet decomposition have also been developed from *in vivo* data in yeast (Peckham et al., 2007; Segal et al., 2006; Yuan and Liu, 2008) and human (Gupta et al., 2008). A comparison of the yeast models led to ROC-scores comprised between 0.8 and 0.9 (Yuan and Liu, 2008). Interestingly, while the features utilized by some of these approaches are fairly different from ours (support vector machine, wavelet decomposition versus PWM), their performances are remarkably similar and suggests that although having a high discriminative power *in vitro*, nucleosome positions cannot be predicted with certainty from the sequence.

3.2.3.2 Spatial Distribution of Motifs

The PWMs used so far account for both the over/under representation of k-mers at each position as well as the variation of this k-mers distribution along the nucleosome. Thus, we do not know whether the performances of $\mathbf{\Theta}^k$ come from (i) specific motifs, (ii) their spatial distribution along the nucleosome (i.e. combination of motifs at specific locations), (iii) or both. To assess the importance

of this spatial distribution we fit a simple multinomial model, $\boldsymbol{\theta}^6$, and compare its ability to discriminate nucleosome from linker sequences with $\boldsymbol{\Theta}^k$ (product multinomial model). $\boldsymbol{\theta}^6$ assumes that every 6-mer are uniformly distributed along the nucleosome, loosing therefore the spatial information. $\boldsymbol{\theta}^6$ is estimated in a similar way as $\boldsymbol{\Theta}^k$ from the 10,000 nucleosome set :

$$\boldsymbol{\theta}_i^6 = log_2 \frac{f_i^6}{p_i^6} \tag{3.5}$$

where f_i^6 and p_i^6 are the frequencies of 6-mer i in the nucleosome and linker sets. As before, $\boldsymbol{\theta}^6$ log-odds scores are calculated in 10-fold cross-validation on the 10,000 nucleosomes set and lead to a ROC-score of 0.79. As shown figure 3.3, this score is lower than any of those obtained using $\boldsymbol{\Theta}^k$. To test whether $\boldsymbol{\theta}^6$ and $\boldsymbol{\Theta}^k$ (k in 1,..,6) ROC-scores are significantly different, 95% confidence intervals are calculated via bootstrap by randomly sampling with replacement from the original dataset. To reduce the computational burden, only 100 bootstrap samples are generated for each model, and the the 2nd and 98th percentiles of the bootstrap distribution are used as limits of the confidence interval. As displayed table 3.1, the ROC-score associated with $\boldsymbol{\theta}^6$ is significantly lower than any of those obtained using $\boldsymbol{\Theta}^k$.

This result indicates that the spatial information contained in $\boldsymbol{\Theta}^k$ is important for distinguishing nucleosome from linker sequences and that the loss of this information leads to a decrease in ROC-score of ~ 0.08. On the other hand, the

Model	ROC-scores		
	Estimate	Percentiles	
		2nd	98th
$\boldsymbol{\theta}^6$	0.796	0.792	0.799
$\boldsymbol{\Theta}^2$	0.829	0.826	0.832
$\boldsymbol{\Theta}^3$	0.845	0.842	0.849
$\boldsymbol{\Theta}^4$	0.858	0.855	0.860
$\boldsymbol{\Theta}^5$	0.867	0.864	0.870
$\boldsymbol{\Theta}^6$	0.874	0.871	0.876
$\boldsymbol{\Theta}^7$	0.843	0.841	0.846
$\boldsymbol{\Theta}^8$	0.817	0.815	0.820

Table 3.1: For each $\boldsymbol{\Theta}^k$, shown are the ROC-score estimations along with their confidence intervals derived from 100 bootstrap samples.

discriminative power of $\boldsymbol{\theta}^6$ being not too far from $\boldsymbol{\Theta}^k$, this result also implies that 6-mers alone have an important role in nucleosome positioning.

3.2.3.3 Common Patterns With Yeast

Histones are amongst the most highly conserved proteins in eukaryotes and consequently result in similar mechanisms of nucleosome packaging across eukaryotes. To determine whether these mechanisms are common in yeast and human, i.e. whether similar motifs are responsible for nucleosome positioning, Gupta et al. (2008) trained two SVMs on *in vivo* human and yeast data. Predictions were subsequently generated from both SVMs on the *in vivo* human dataset and a correlation of 0.86 was found between the two sets. The authors concluded that this high correlation supports the idea that human and yeast (and more generally eukaryotes) do share common nucleosome positioning mechanisms.

If *in vivo*, common nucleosome favouring motifs can be detected in human and

yeast, we expect that these similarities should be even more pronounced *in vitro* where the effects of *trans* factors are removed. Similarly to what we did in human, we (i) run NucleoFinder on the yeast *in vitro* data produced by Kaplan et al. (2009), (ii) select the 10,000 regions with the largest Bayes Factor and (iii) fit a 6-mers PWM ($\mathbf{\Theta}^6_{yeast}$) from nucleosomes and linkers sequences.

From each model, $\mathbf{\Theta}^6_{yeast}$ and $\mathbf{\Theta}^6_{human}$ (the PWM previously derived from human sequences), log-odds scores are calculated on the human 10,000 nucleosomes set, from which a correlation of 0.44 is found between the two sets. From the $\mathbf{\Theta}^6_{yeast}$ log-odds scores, a ROC-score of 0.73 is obtained, that is ~ 0.15 lower than that calculated with $\mathbf{\Theta}^6_{human}$. These surprisingly small values of correlation and ROC-score are not in line with Gupta et al. who support the idea that human and yeast share similar nucleosome positioning mechanisms. Questioning our result, we went back to Kaplan et al. (2009) and Valouev et al. (2011) articles to compare the protocols used in the preparation of *in vitro* nucleosomes and realized that they were reconstituted at 40% and 100% of the physiological histone to DNA ratio in Kaplan et al. (2009) and Valouev et al. (2011) respectively. Because of this important difference in protocol, we interpret our results with caution.

3.2.3.4 The Level of Positioning is Linked to the Presence of Nucleosome Favouring Motifs

The 10,000 sequences used so far, selected on the basis of their high Bayes Factor and thus wrapped into strongly positioned nucleosomes, have shown a good ability to discriminate nucleosome from linker *in vitro*. One can now wonder whether $\mathbf{\Theta}^6$ has a similar power of discrimination on all the NucleoFinder predictions, in-

cluding those having lower Bayes Factor. To address this question, NucleoFinder is run on a randomly chosen 10 Mb region in chromosome 22, from which 25,309 predictions are obtained. The choice of a relatively small region is motivated by the time it takes to extract the sequences and compute log-odds scores. In the same way as previously done in 3.2.3.1, linkers are merged to form 150 bp long sequences and log-odds scores are computed from $\mathbf{\Theta}^6$ on the nucleosome and linker sequences. Figure 3.4 displays the log-odds scores distributions (a) in this 10 Mb region and (b) in the 10,000 nucleosomes set, where the separation between the nucleosome and linker distributions is much clearer in the latter case than in the former, as reflected by their associated ROC-score of 0.67 and 0.87.

This poor accuracy may be due to the fact that the motifs used by $\mathbf{\Theta}^6$ to predict nucleosomes are fairly infrequent in regions detected with lower Bayes Factor. To test this possible relationship between the presence of nucleosome-specific motifs (reflected by the log-odds scores) and the Bayes Factor with which nucleosomes are detected, the 25,309 predictions are binned into five groups according to their Bayes Factor, from which ROC-scores are computed and plotted against the averaged Bayes Factor. The number of groups is chosen so that the subsamples used in the 10-fold cross-validation are large enough. Figure 3.4(c) confirms this relationship between Bayes Factor and ROC-score.

This interesting result suggests that (i) the presence of well-positioned nucleosomes is directly related to the presence of nucleosome-specific motifs and (ii) the human genome is populated by motifs having a wide range of affinity for nucleosomes and consequently excludes the idea that it is divided in regions either

nucleosome-specific or linker-specific. Therefore, because the nucleosome-specific motifs characterised from the 10,000 nucleosomes set are relatively infrequent in the human genome, $\mathbf{\Theta}^6$ has a limited power of discrimination in the 10 Mb region used here.

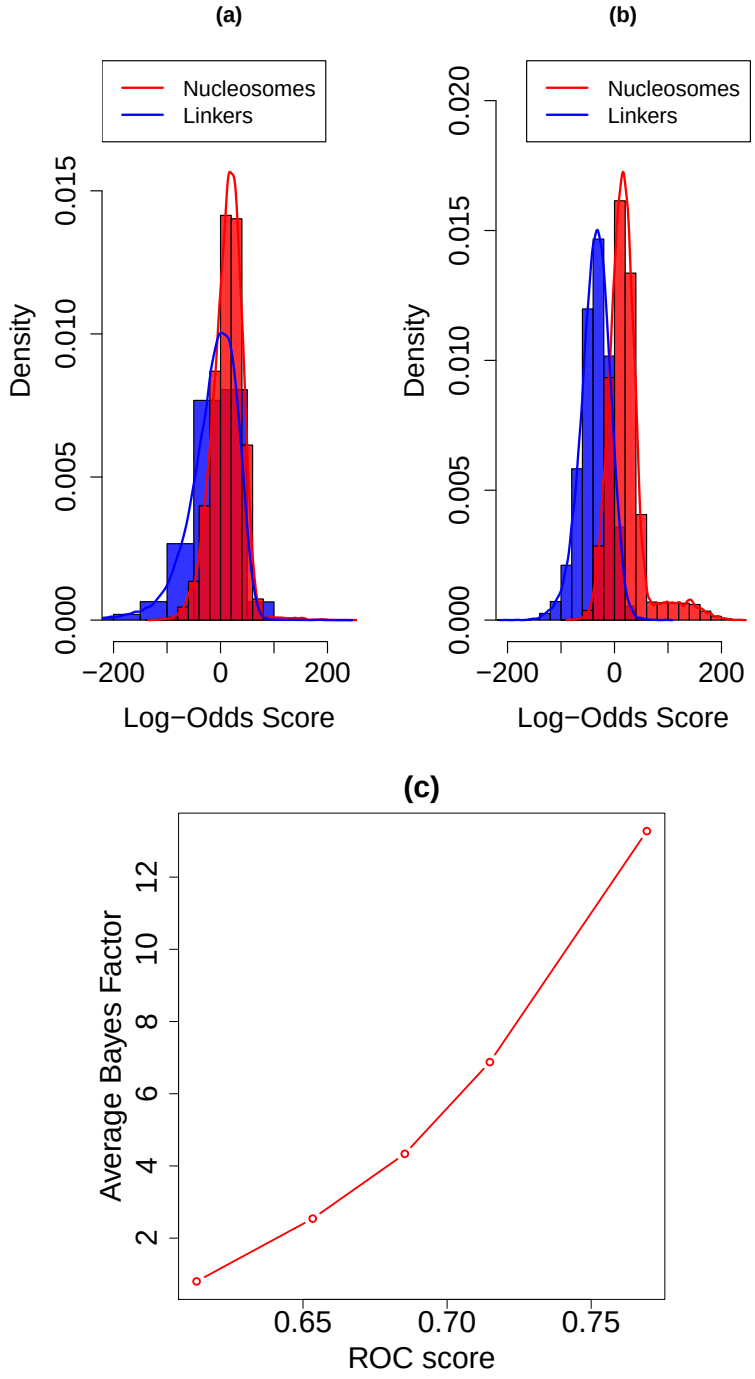

Figure 3.4: Distributions of log-odds scores associated with nucleosomes and linkers calculated in **(a)** a 10 Mb region randomly chosen chromosome 22 (ROC-score = 0.67), **(b)** the 10,000 nucleosomes set (ROC-score = 0.87). **(c)** Relationship between the level of nucleosome positioning/occupancy (Bayes Factor) and the presence of nucleosome-specific motifs, reflected by the ability of $\mathbf{\Theta}^6$ to discriminate nucleosomes from linkers (ROC-score) in a 10 Mb region randomly chosen in chromosome 22.

3.3 Influence of *Cis*-Factors *In vivo*

In the previous section, we showed that nucleosome-specific motifs have a large influence on the nucleosome architecture *in vitro* but are generally infrequent. We also saw that, *in vitro*, the Bayes Factor with which nucleosomes are detected is directly linked to the ability of these motifs to discriminate nucleosomes from linkers sequences. In this section, we seek to measure the influence of *cis*-factors *in vivo*. As we cannot exclude the possibility that more complex mechanisms, not captured in $\mathbf{\Theta}^6$ (for instance involving long range interactions), are acting on nucleosome positioning, we directly compare the positions of *in vitro* with *in vivo* nucleosomes.

As mentioned in the introduction chapter (section 1.4), this strategy has already been followed in :

- Kaplan et al. (2009), where *in vitro* and *in vivo* nucleosome occupancy scores were correlated;
- Zhang et al. (2009), where the number of *in vitro* reads falling within 10 bp of *in vivo* nucleosome centers were counted.

In this section, instead of working on the reads level or on the nucleosome occupancy level, we decide to take advantage of NucleoFinder and work on the nucleosome level, where the positions of *in vivo* and *in vitro* nucleosomes are compared. More specifically, we want to test whether the number of overlaps between *in vivo* and *in vitro* nucleosomes is significantly larger than expected given the nucleosome distribution of nucleosome *in vivo*. This null distribution can be

estimated by sampling positions in the neighborhood of the loci of interest (to account for the non-homogeneous distribution of nucleosomes *in vivo*, cf. sub-section 1.1.2) and counting the number of overlap occurring by chance between these positions and *in vivo* nucleosomes.

Before carrying out such comparison on a chromosome scale to have a general idea of the influence of *cis*-factors *in vivo* (sub-section 3.3.3), we first compare *in vivo* and *in vitro* nucleosomes on the 10,000 nucleosomes set (carrying nucleosome-specific motifs) to ensure that we can repeat the results already reported in the literature (nucleosome-specific motifs are enriched in well-positioned nucleosomes *in vivo*, see sub-section 1.4.3) with NucleoFinder predictions. We will also investigate in this section whether a similar relationship between the power of discrimination of these motifs and the enrichment of nucleosomes *in vitro* also exists *in vivo*.

3.3.1 Nucleosome-Specific Motifs are Enriched *In vivo*

Following on Valouev et al. observations that regions with strong G/C cores and A/T flanks are enriched both *in vitro* and *in vivo*, we expect the 10,000 nucleosomes set, carrying these motifs, to be enriched *in vivo*.

To test this hypothesis, we look at whether nucleosomes in the 10,000 nucleosomes set (carrying nucleosome-specific motifs) are significantly more enriched *in vivo* than 10,000 regions randomly sampled within 1 kb of each *in vitro* nucleosome (to account for the non-uniform distribution of nucleosomes along the genome). We remind that this random set allows to estimate how often *in vitro*

nucleosomes can overlap *in vivo* nucleosomes by chance. The enrichment is then calculated as the number of overlap between the *in vitro* and the random sets with NucleoFinder predictions in 1, 2 or 3 in vivo dataset (CD4+T, CD8+T and granulocyte). To account for imprecision in nucleosome position, windows ranging from 5 to 50 bp across the nucleosome centres are considered.

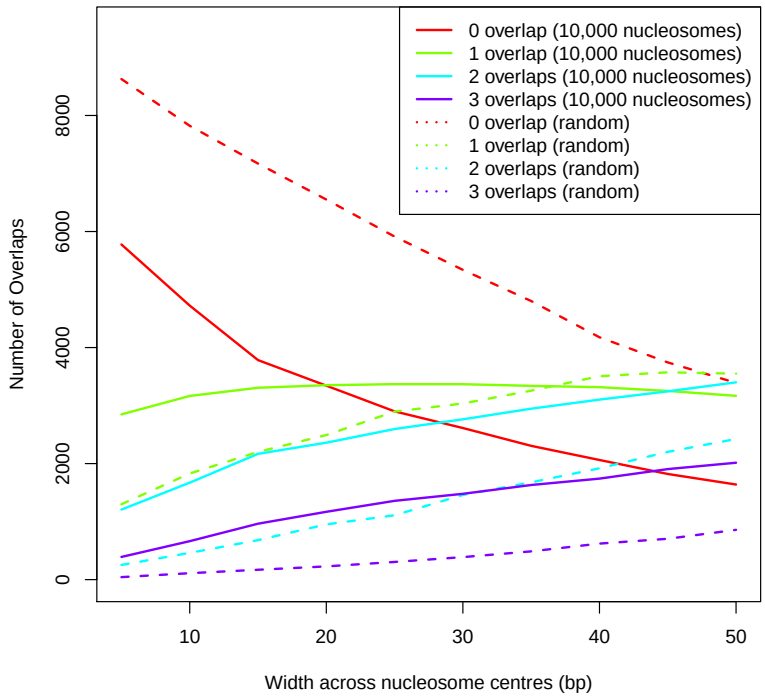

Figure 3.5: Number of overlaps between the two *in vitro* sets (the 10,000 nucleosomes set and random set) and NucleoFinder predictions in the three *in vivo* dataset (CD4+T, CD8+T, granulocyte). To account for the uncertainty in nucleosome positioning, windows varying between 5 and 50 bp are considered across the prediction centres.

Unsurprisingly, the total number of overlaps in both sets increases as w becomes

larger due to overlaps with more distant nucleosomes (figure 3.5). For each window width, we then calculate the proportion of 0, 1, 2 and 3 overlaps in the 10,000 nucleosomes and random set. When looking for instance at $w = 30$ bp, i.e. the uncertainty chosen in NucleoFinder, $\sim$ 32.9%, 27.0% and 14.5% of the 10,000 nucleosomes set are found to overlaps predictions with 1, 2 and 3 *in vivo* dataset, that is, $\sim$ 3.2%, 12.7% and 10.7% above what is found in the random set (figure 3.5). To test whether these proportions differ between the 10,000 and random nucleosomes set, a Pearsons chi-squared test is performed :

$$X^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} \tag{3.6}$$

where X^2 is distributed as a χ^2 with 3 degrees of freedom, O_i, the number of $i-1$ overlaps in the 10,000 nucleosome set, and E_i, the number of $i-1$ overlaps in the random set. This test is typically used to measure the goodness of fit between observed and expected values in the context of categorical data. For a type I error rate of 5%, we reject the hypothesis that the proportion of overlaps is equal in the 10,000 and the random nucleosomes sets if X^2 is larger than 7.81, the 95th percentile of $\chi^2_{(3)}$. As displayed table 3.2, the values obtained for X^2 are much larger than this value, indicating that the proportions differ significantly between the two sets. This means that nucleosome-specific sequences are indeed significantly more enriched *in vivo* than randomly sampled regions, and thus confirms the influence of nucleosome favouring motifs in the nucleosome organisation *in vivo*.

w (bp)	X^2	$-\log_{10}$(p-value)
5	9202.9	4597.1
10	8126.1	4058.8
15	9110.9	4551.1
20	7833.2	3912.3
25	7241.8	3616.7
30	5658.8	2825.3
35	4947.1	2469.5
40	3831.6	1911.9
45	3559.4	1775.8
50	2900.3	1446.4

Table 3.2: Pearsons chi-squared statistics X^2 and $-\log_{10}$(p-value) calculated for various window width w across nucleosome predictions.

The next question that comes naturally is how often one expects these motifs to be occupied by well-positioned nucleosomes *in vivo*, that is, what is the size of this effect. To do so, we could simply look at the proportion p of the 10,000 *in vitro* nucleosomes overlapping at least one nucleosome *in vivo*. Since a non-negligible number of overlaps occurs by chance, this would probably lead to an over-estimation of the size effect though. To address this problem, the number of overlaps in the 10,000 nucleosomes set is corrected by the number of overlaps in the random set :

$$p = \frac{\sum_{i=2}^{4}(O_i - E_i)}{\sum_{i=1}^{4} O_i} \tag{3.7}$$

At $w = 30$ bp, $p = 0.27$, which means that, while almost always occupied *in vitro* (ROC-score = 0.87), these motifs (found at highly positioned nucleosomes *in vitro*) are occupied in less than a third of the cases *in vivo*. Although modest, this

proportion suggests that strong nucleosome-specific motifs have a non-negligible influence *in vivo*.

3.3.2 *In vivo* Enrichment is Related to the Presence of Nucleosome-Specific Motifs

In 3.2.3.4, we showed that the Bayes Factor with which nucleosomes are detected *in vitro* is linked to the presence of nucleosome favouring motifs. Since these motifs also influence the *in vivo* architecture, we now investigate whether a similar relationship exists between the enrichment across multiple cell lines (*in vivo*) and these same motifs.

To do so, the same analysis as before is run on the 10 Mb region chromosome 22 used in 3.2.3.4, where after randomly sampling positions within 1kb of the *in vitro* predictions, the number of overlaps between the predictions *in vitro* and *in vivo* (CD4+T, CD8+T and granulocyte) is computed with $w = 30$ bp. The 25,309 *in vitro* predictions are then binned into the same five groups as before according to their Bayes Factor. For each group, the X^2 statistics and the ROC-scores are calculated. On figure 3.6, we can see that similarly to the *in vitro* Bayes Factor, X^2 increases with the ROC-score. Thus, not only nucleosome-specific motifs are enriched *in vivo*, but this *in vivo* enrichment is directly related to the motifs' ability to discriminate nucleosome from linker. It can also be noticed that, although above the $\chi^2_{(3)}$ 95th percentile, the values of X^2 obtained in this 10 Mb region are more modest than those in the 10,000 nucleosomes set.

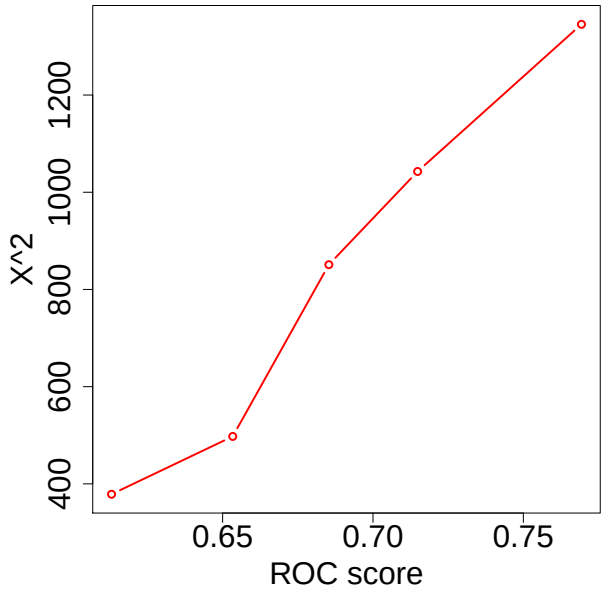

Figure 3.6: *In vivo*, the nucleosome enrichment (measured by X^2) is directly related to the presence of motifs that discriminate nucleosome from linker sequences (reflected by the ROC-score).

3.3.3 The Relative Influence of *Cis* and *Trans*-Factors Differ Across Functional Regions

As described subsection 1.3.2, the cellular environment (*trans*-factors) can move nucleosomes to sequences that are not intrinsically favourable to being occupied. Depending on their function, regions across the genome are led to interact more or less with these *trans*-factors. The extent to which nucleosomes are moved *in vivo* from their "intrinsically encoded position" *in vitro* thus reflect the relative importance of *cis* and *trans*-factors on the nucleosome organisation. Assuming that the differences between *in vivo* and *in vitro* nucleosome positions result only from the action of *trans*-factors, we seek to estimate the relative contribution of *cis* and *trans*-factors by computing X^2 and p in functionally different regions. In this way chromosome 22 is divided in five different regions : (i) highly expressed promoters, (ii) non-expressed promoters (defined as the 2kb regions upstream of TSSs of genes having RPKM > 10, < 0.02 respectively), (iii) exons, (iv) introns and (v) intergenic regions (defined as the regions left after the four other categories selected). The locations of TSSs, exons and introns are provided by the RefSeq database (Pruitt, 2004).

Region	X^2	$-\log_{10}$(p-value)	p
Unexpressed promoters	98.2	47.0	0.127
Expressed promoters	23.9	10.6	0.047
Exons	177.1	86.2	0.080
Introns	6616.0	3303.8	0.139
Intergenic regions	8274.8	4133.1	0.146

Table 3.3: Pearsons chi-squared statistics X^2, $-\log_{10}$(p-value) and corrected proportion p calculated in functionally different regions, chromosome 22.

Table 3.3 summarizes these results where it can first be noticed that the values of p are much lower than those computed in the 10,000 nucleosomes set ($p = 0.27$), which again support the relationship between nucleosome favouring motifs and *in vivo* enrichment. Substantial differences of p can otherwise be noticed across regions : on one hand intergenic regions and unexpressed promoters display large values suggestive of a less pronounced influence of *trans*-factors in these regions. This is consistent with the idea that, generally lacking functions, these regions are less prone to interact with transcription factors or being modified by chromatin remodelers. On the other hand, expressed promoters have the smallest proportion value, $p = 4.7$ %, in agreement with the idea that many interactions with the transcription machinery occur in these regions.

It can otherwise be noted that the influence of *trans*-factors in transcribed regions varies substantially depending on whether it is an exon or an intron. Previous studies have demonstrated that *in vivo*, exons are more frequently occupied by stably positioned nucleosomes than their surrounding introns (Tilgner et al., 2009). This is indeed what we observe in the CD4+T set where an average of 1.5 nucleosomes / kb in exons and 0.15 nucleosomes / kb in introns (i.e. 10 time less) is detected by NucleoFinder in chromosome 22. The authors put forward the hypothesis that the presence of a stably positioned nucleosome in exons reduces the transcription rate, allowing the recognition of the splice signals by splicing factors. Our results refine this observation by showing that these well-positioned nucleosomes in exons are mainly driven by *trans*-factors.

Overall, the fact that the (corrected) proportion p of overlaps between *in vitro* and *in vivo* nucleosomes ranges between 0.04 and 0.15 points toward a major role for *trans*-acting factors in the nucleosome architecture *in vivo*.

3.4 Discussion

The study of well-positioned nucleosomes *in vitro* in human have allowed us to recover the known pattern of A/T-rich and G/C-rich motifs at the flank and the centre of nucleosomes. In line with previous studies, we showed that these motifs are able to distinguish well-positioned nucleosomes from their linkers with high accuracy. We also demonstrated that PWMs have a significantly better discriminative power than multinomial model. However, the difference between the two models is relatively small which suggests that the motifs rather than their spatial distribution are important in nucleosome positioning.

Because nucleosome specific motifs are fairly infrequent in the human genome, the majority of *in vitro* nucleosomes cannot be predicted accurately from them (ROC-score = 0.67). We further showed that the presence of these particular motifs is linked (i) to the Bayes factor with which nucleosomes are detected *in vitro* and (ii) to the nucleosome enrichment in different cell lines *in vivo*.

From the comparison of *in vitro* and *in vivo* nucleosomes, we then estimated the influence of *cis*-factors in the *in vivo* nucleosome organisation at $\sim$ 10-15%, i.e. $\sim$ 10% less than in yeast. Interestingly, while an important discrepancy was found between nucleosome occupancy and positioning in the estimation of *cis*-factors in yeast, our result is not too far from Tillo et al. (2010) estimate (R^2 =

0.08), based on nucleosome occupancy. All together, these results support the idea that DNA has a modest role in nucleosome positioning in human, as pointed out in Valouev et al. (2011).

Chapter 4

Transcription Regulation by Nucleosomes and Histone Modifications

4.1 Introduction

4.1.1 General

As introduced in chapter 1, epigenetic factors have been shown to be instrumental in the regulation of many cellular processes, including gene transcription. The epigenetic information is embedded into the chromatin structure and have commonly been classified into three categories according to the chromatin layer it belongs to :

- the DNA level where the methylation of cytosine nucleotides leads to transcriptional silencing,

- the nucleosome level, where the nucleosome position, the use of histone variants and post-translational modifications of histones can induce or silence transcription,
- higher structural level where the chromatin itself interacts with other components present in the nucleus such as the nuclear lamina.

The mechanisms by which epigenetic factors are modified, the pathways followed to regulate downstream cellular processes and the combination of marks involved in these cellular processes remain incompletely understood. In the previous chapter, the focus was on the first point where we measured the relative influence of *cis* and *trans* factors on the nucleosome architecture and demonstrated that the former have a minor role. This chapter, on the other hand, is concerned with the last point, that is, understanding the effects of histone marks and nucleosome positioning on gene regulation.

Histone modifications have been shown to alter the chromatin structure by two mechanisms that can function either individually or in combination. The first is the reduction of the DNA/histones interaction caused by the variation of charge induced by the histone modification. The second is the recruitment of non-histone protein complexes able to remodel the chromatin (Kouzarides, 2007).
Prior studies have allowed the association of specific modifications with gene activation or repression as well as with various genomic region including promoters, enhancers, insulators and transcribed regions (Barski et al., 2007; Wang et al., 2008). In this way, acetylation and phosphorylation are generally associated with transcription, sumoylation and deimination with gene silencing, methylation

and ubiquitination are found both in activation and repression of transcription (Kouzarides, 2007).

The large number of histone modifications as well as the high level of correlation among some of them have given rise to the "histone code" which hypothesizes that the combination of different histone marks induces a particular chromatin state and subsequently allows a given cellular processes to be performed (Jenuwein and Allis, 2001).
In order to (i) understand the downstream effects associated with various sets of histone modifications, and (ii) gain insight into the "histone code", computational and statistical approaches integrating different data types have been developed. In this way, two directions have been followed in the literature :

- the first aims to identify the epigenetic factors (mainly histone modifications) involve in gene regulation and measure how much of the variation in gene expression they explain;
- the second seeks to identify combinatorial patterns of epigenetic factors associated with different functional regions.

Before presenting these two directions, we review the patterns of epigenetic modifications characterised at different functional locations in the genome.

4.1.2 Histone Modifications across Functional Regions

Promoters. The epigenetic state at promoter region have received much attention for the insight it provides into the regulation of transcription. Different

modes of regulation have been linked to high and low GC content promoters. A study in human stem cells showed that, while the characteristic modifications of active promoter H3K4me3 is found in a majority of high CpG-content promoters, this mark is tightly associated with the level of expression in low CpG-content promoters (Kim et al., 2005). However, these observations in stem cells do not all apply in differentiated cells. In particular, a large proportion of high CpG-content promoters, non essential in differentiated cells, are made inactive during differentiation by the deposition of the H3K27me3, H3K9me3 marks (Mikkelsen et al., 2007). The H3K4me3 rich sites are accompanied by features of accessible chromatin, including histone acetylations, occupancy by the H3.3, H2A.Z histone variants, hypersensitivity to DNase I digestion and DNA hypomethylation (Wang et al., 2008). Many promoters carrying this particular open structure have been shown to be enriched in Pol II reflecting either transcriptional initiation or active transcription (Ramirez-Carrozzi et al., 2009). Interestingly, in accordance with the phenomon of multivalency, Wang et al. (2008) showed that a set of 17 "backbone" modifications (H2A.Z, H2BK5ac, H2BK12ac, H2BK20ac, H2BK120ac, H3K4ac, H3K4me1/2/3, H3K9ac, H3K9me1, H3K18ac, H3K27ac, H3K36ac, H4K5ac, H4K91ac) are present in a majority of active promoter in CD4+T cells.

Gene bodies. Mammalian genes are characterised by numerous short exons distributed between large introns. Four methylation (H2BK5me1, H3K27me1, H3K36me3, H3K79me1/2) and six acetylation (H2BK20ac, H2BK120ac, H3K4ac, H4K5ac, H4K8ac and H4K16ac) marks have been linked to active transcription where a majority of them have been shown to correlate with the level of gene

expression (Barski et al., 2007; Wang et al., 2008). The enrichment in these marks appeared to be more pronounced in exons than introns, which could possibly be due to a higher abundance of nucleosomes in these regions (Tilgner et al., 2009).

Enhancers. Enhancers are located from few hundred base pairs to several hundred kilobase pairs upstream or downstream their associated transcript. They can also be found in intronic region as well as 5 and 3 untranslated region. By observing the histone marks at known enhancers, (Heintzman et al., 2009) noticed a relative enrichment in H3K4me1 along with a depletion in H3K4me3. Other marks such as H2K27ac, H3K36me1, H3K4me2, H3K27me1, H3K9me1 and H2BK5me1 were also found in another study (Hon et al., 2009), suggesting again a certain level of redundancy across the histone marks. Using these histone marks to find new enhancers, Heintzman et al. found that the chromatin patterns at enhancers are more variable and cell type specific than those observed at promoters.

Large-scale repression. As the cell differentiates from a totipotent to a specialized state, many genes are not expressed and large proportions of the genome become stably repressed. These regions are marked with the silencing histone modifications H3K9me2/3, H3K27me2/3, H3K79me3 and H4K20me3 (Wen et al., 2009) and have been shown to interact with the nuclear lamina (filaments located in inner face of the nuclear) in approximatively half of human genome in fibroblast (Wen et al., 2009). This repressive chromatin environment results in reduced expression level for the genes lying in these regions (Reddy et al., 2008).

Summary. Based on Barski et al. (2007); Wang et al. (2008) observations, the histone modifications are classified by genome location and function. Enhancer

modifications are not included as they are not used in the remainder of this chapter.

Active Marks			Silencing
TSS	Genic region	Common	Marks
H3K4me3	H2BK5me1	H2BK12ac	H3K9me2/3
H2A.Z	H3K27me1	H2BK20ac	H3K27me2/3
H2AK9ac	H3K36me3	H2BK120ac	H4K20me3
H2BK5ac	H4K16ac	H3K4ac	
H3K9ac		H3K4me1/2	
H3K18ac		H3K9me1	
H3K27ac		H3K79me1/2/3	
H3K36ac		H4K5ac	
H4K91ac		H4K8ac	
		H4K12ac	
		H4K20me1	

Table 4.1: Classification of histone marks by region and function according to Barski et al. (2007); Wang et al. (2008).

4.1.3 Previous Work

Now that the patterns of epigenetic modifications in different functional regions have been presented, we now turn to the two directions explored in the literature to characterise histone modifications.

4.1.3.1 Link Between Gene Expression Level and Histone Modifications

The first direction, gene centered, focuses on the relationship between gene expression level and epigenetic pattern at promoter and/or gene body. The goal

is essentially to relate the level of gene expression to the enrichment in histone marks at promoters (or transcribed regions) and identify the marks which are the most predictive of gene expression. Various machine learning approaches have been used to address these questions, including Bayesian network (Cheng et al., 2011; Yu et al., 2008), support vector machine (Cheng et al., 2011), linear and non linear regressions (Cheng et al., 2011; Karlic et al., 2010; Xu et al., 2010). These studies share a similar setup where genes are used as units of observations, enrichment in histone modifications as predictors and gene expression level as response variable.

A general picture is now emerging where (i) histone marks measured at promoter and gene body are highly predictive of gene expression (Pearson correlation = 0.7-0.8, Hoang et al. 2011). (ii) When fitting models between gene expression and histone marks in small regions (e.g. 1 kb) upstream, across and downstream of genes, large R^2 are found across the whole transcribed regions and extend few kilo bases upstream of the TSS and downstream of the TES (Cheng and Gerstein, 2012). This result indicates that histone modifications present at promoter, transcribed and TES regions are highly predictive of gene expression level. It does not specify however whether all the marks have the same predictive power throughout these regions or whether the correlation between gene expression and histone marks is driven by different sets of marks in different regions. (iii) The addition of the transcription factor binding information to the histone marks are redundant for the prediction of gene expression (Cheng and Gerstein, 2012). (iv) In promoter regions, an almost identical prediction accuracy can be obtained with a handful of these marks. For instance, Karlic et al. found that four histone

marks (H4K20me1, H3K27ac, H3K79me1, H2BK5ac) have an almost identical predictive power as a set of 39 marks, suggesting a high degree of redundancy among histone modifications at promoter. (v) Slight differences in histone signatures are found between high and low CpG content promoters (Karlic et al., 2010). (vi) The relationship between histone modifications and gene expression is general and does neither depend on the tissue / cell-type, nor on the type of gene (coding/non-protein-coding genes, Zhang and Zhang 2011).

4.1.3.2 Functional Elements are Marked with Specific Histone Modifications

The second direction is somehow similar in the sense that it aims to characterise chromatin signatures at functional elements. However, it differs in that regions under study are not restricted to the genes regions and that the objective is to identify the chromatin patterns itself, not to predict gene expression. As rightly said in Ernst and Kellis (2010), the goal is not to uncover causal effects between chromatin signatures and regulatory processes but rather to identify the signatures existing in different functional regions that could potentially be used in future experiments to pinpoint new functional elements.

Supervised classification methods have first been introduced to identify chromatin patterns at known functional sites (Heintzman et al., 2007; Liu et al., 2005). However, because (i) much of our current knowledge of the human genome is concentrated on genes and promoters, these studies have ignored a majority of the genome, and because (ii) different regulatory elements are used in different conditions and that their location are not necessary known *a priori*, unsupervised

learning methods appeared to be better suited for this type of problem.

So far, four approaches have been introduced to characterise chromatin signatures at different functional regions. They proceed in a similar manner where common combinatorial patterns are identified with clustering techniques and then characterised by mapping the occurrences of each clusters to known functional regions.

Hon et al. introduced *ChromaSig*, a progressive alignment approach that iteratively (re-)defines chromatin motifs and localizes the loci that match these motifs. From a collection of 21 histone modifications assayed in Barski et al. (2007), Hon et al. identified 14 distinct chromatin states at promoters, 11 at enhancers and 7 at loci distal to known regulatory regions. Beside recovering known patterns at promoters and enhancers, they found that H3K36me3 is found at exon 5' ends and is associated with exon expression and alternative splicing. They also characterised different chromatin signatures in repeat sequences that associate with distinct modes of gene repression (e.g. lamina-associated domains).

Jaschek and Tanay approached this problem using an HMM, considering that this family of models are well suited for large datasets where the spatial dependency among functional regions can be used to improve the clustering (e.g. transcribed regions are expected to occur after promoters). Using multivariate distributions to model chromatin signatures over the same set of 21 histone modifications as Hon et al., Jaschek and Tanay inferred 16 distinct hidden states. The distributions of histone marks associated with each state allowed them to retrieve

known patterns of histone marks at TSS, transcribed and large-scale repressive regions.

Ernst and Kellis also opted for an HMM approach where, unlike Jaschek and Tanay (2009), the histone mark were binarized in 200 bp windows to reduce the number of parameters and avoid making any assumption on their distributions. Based on the same dataset as Hon et al. augmented with 17 additional histone modifications (mainly acetylations), Ernst and Kellis characterised 51 distinct chromatin states that could be grouped into four categories according to their genomic positions : promoter, transcribed, active intergenic, repressive and repetitive regions.

Ucar et al. introduced a biclustering approach that identified 843 distinct clusters, among which 721 were enriched for at least one of the eight functional regions under study (CpG island, conserved sequence, insulator, DNAse hypersensitivity site, enhancer, long intersperse non-coding RNA, promoter and gene body). To characterise these functional regions from the obtained clusters, the authors identified the histone modifications that frequently co-occur in each set of region-specific clusters. These "core chromatin modifications" along with the emission probabilities inferred by Ernst and Kellis will be utilized in section 4.4 to characterise clusters of histone marks found in promoter and transcribed regions.

Overall, even if these methods have resulted in very different number of clusters, their findings are consistent with prior knowledge on the location and function of histone modifications presented in the previous section (e.g. H3K4me3 is found

at active promoters).

4.1.4 Question

In this introduction, we have seen that two main directions have been explored to characterise histone modifications : the first seeks to measure the effect of histone marks on transcriptional regulation and identify the marks that are the most predictive. The second aims to identify the combinatorial patterns of histone marks that characterise distinct functional regions. In this chapter we aim at bridging the two approaches by (i) clustering the histone marks in promoter and transcribed regions using gene expression level, and (ii) measuring their individual effect on gene expression along genes. The first step is similar to the second direction, with the difference that in addition to clustering histone marks on their co-occurrence in the same functional regions, we make use of gene expression to select the histone marks that account for most of the gene expression variation. The second step is then similar the first direction where the influence of each cluster on transcription level is measured along the genes. This will allow us to :

- identify clusters of histone marks associated with gene activation or repression in promoter and transcribed regions and see whether they are region-specific or shared across the two regions;
- link clusters to gene function (cell-specific versus housekeeping genes);
- understand how broad promoter and transcript specific signals are;
- see whether the combination of promoter and transcript specific signals can improve the prediction of gene expression;

- lastly, as mentioned in the introduction chapter, we will see whether the size of the nucleosome depleted region have a significant influence on gene expression level.

The remainder of this chapter is organised as follow, section 2 introduces the data, the way they are processed and how the promoter/transcribed region boundary is chosen; section 3 describes how histone modifications are clustered; section 4 aims to characterise the function associated with the clusters; section 5 combines the promoter and transcript clusters signals in order to improve the prediction of gene expression, finally section 6 examines whether the nucleosome organisation across the TSS has an influence on transcription.

4.2 Data Description and Pre-Processing

4.2.1 Data Description

Histone modifications. The ChIP-Seq tags used in this analysis were taken from two genome-wide studies of (i) 20 lysine and arginine histone methylations, 1 histone variant (H2A.Z) (Barski et al., 2007) and (ii) 19 histone acetylations (Wang et al., 2008). Both studies were measured in human CD4+T cells and produced by a same group (Zhao et al.). From the former, the ChIP-Seq tags of the transcription polymerase II (Pol II) were also retrieved in order to test the influence of nucleosome position on both the transcription and the binding of the transcription polymerase II (Pol II). Similarly to what was done in 2.2, tags mapping more than one genomic region were discarded, the remaining ones were then corrected for possible amplification errors and shifted to +/- 75 bp

(according to their strand) to localize nucleosome centres.

Nucleosomes. Since the histone modifications used in Barski et al. (2007) and Wang et al. (2008) were assayed in CD4+T cells, we could have re-used the nucleosome data from Valouev et al. (2011) also measured in CD4+T cells and already processed in the previous chapters. However, the group who generated the histone modification datasets (Zhao et al.) also studied nucleosome in human CD4+T cells using a similar protocol and a same sequencing platform (Solexa, i.e. different from this utilized in Valouev et al. 2011). Therefore, to minimize the inter-study and inter-platform variations, we made use of Zhao's nucleosome dataset (Schones et al., 2008) in this analysis. The same normalizing and nucleosome calling procedures (NucleoFinder) used in chapter 1 were then performed.

Gene expression. The gene expression dataset came from Zhao's nucleosome study (Schones et al., 2008) who made use of an Affymetrix Human Genome U133 Plus 2.0 GeneChip. Raw expression values were normalized with the RMA algorithm (Bolstad et al., 2003) and averaged over the replicates. During this normalization step, the microarray probes were mapped to 18,982 Ensembl genes. Because our aim is to relate gene expression level with histone modifications at specific genomic regions (promoter, gene body), mapping the probes with Ensembl transcripts would have been preferable to distinguish alternative transcripts. However, given that the probes of U133 Plus 2.0 GeneChip target the 3' end of transcripts, this distinction cannot be done.

Promoter selection. Given that in human, most of the genes are alternatively spliced, we need to ensure that the gene body and promoter at which histone

modification are measured correspond to the right gene expression measurements. To do so, we mapped the Ensembl genes to their Ensembl transcript sets, and for each set, the transcript with the highest averaged enrichment across the 39 histone modifications at promoter is selected (as proposed in Karlic et al. 2010). Each histone mark is standardized beforehand so they all have the same influence in the choice of promoters.

4.2.2 Summarization of Histone Modification Enrichment

In 4.1.2, we saw that histone modification enrichment is strongly dependent on the genomic region considered, some histone marks are promoter specific, others gene specific. In this way, to understand the differences between these two adjacent regions we need to find the boundary that separates the signals specific to each region and choose a measurement that summarizes the signal accurately in each region. This last point has been addressed in two ways, while the first two studies on this topic simply counted the number of ChIP-Seq tags in a window few kilo bases large centered on the TSS (Karlic et al., 2010; Yu et al., 2008), Xu et al. (2010) and Hoang et al. (2011) considered that part of the histone modification information is encoded in its spatial distribution. To take into account this spatial distribution, these authors suggested to (i) characterise this distribution ("template") for each mark by averaging the histone modification signal along all the (scaled) genes, and (ii) estimate a "gene-dependent amplitude" by regressing the histone modification signal along each (scaled) gene on its associated "template". When fitting MARS models using these two methods, Hoang et al. found that the spatial approach resulted in a 0.02 increase in R^2 over the counting

approach, corresponding to an improvement of 4% in the model fit. This result suggests that the spatial distribution does not necessarily need to be accounted for when predicting gene expression. For this reason, in the following, we will follow the counting approach.

4.2.3 Variables Transformation

Before analyzing and fitting models to the data, transformations of the response variable (gene expression) and predictors might be required to improve the fit and correct possible violations of model assumptions such as normality and constant variance of the error.

4.2.3.1 Transformation of the predictors

Given that the size of genes greatly differ from one another, the count of Chip-Seq tag in a given region (promoter, gene body, see 4.2.4 for the determination of the boundary between these two regions) is normalized by the number of base pairs in this region so they can be compared. The distributions of normalized tag count in a 4kb window centred on TSSs are shows figure 4.1(a), where it can be seen that most of the mass is concentrated around zero with few larger values as indicated by the long tails. When fitting a linear model, the points in the tails will have large leverage and potentially distort the regression estimate. To reduce the effect of these points, a log or square root transformation is commonly applied to reduce the range of the data. Given that the normalized tag counts are close to zero, the $log(1 + x)$ transformation has little effect as shown 4.1(b). The square root, on the other hand, successfully reduce the long tails and spread out the mass around zero.

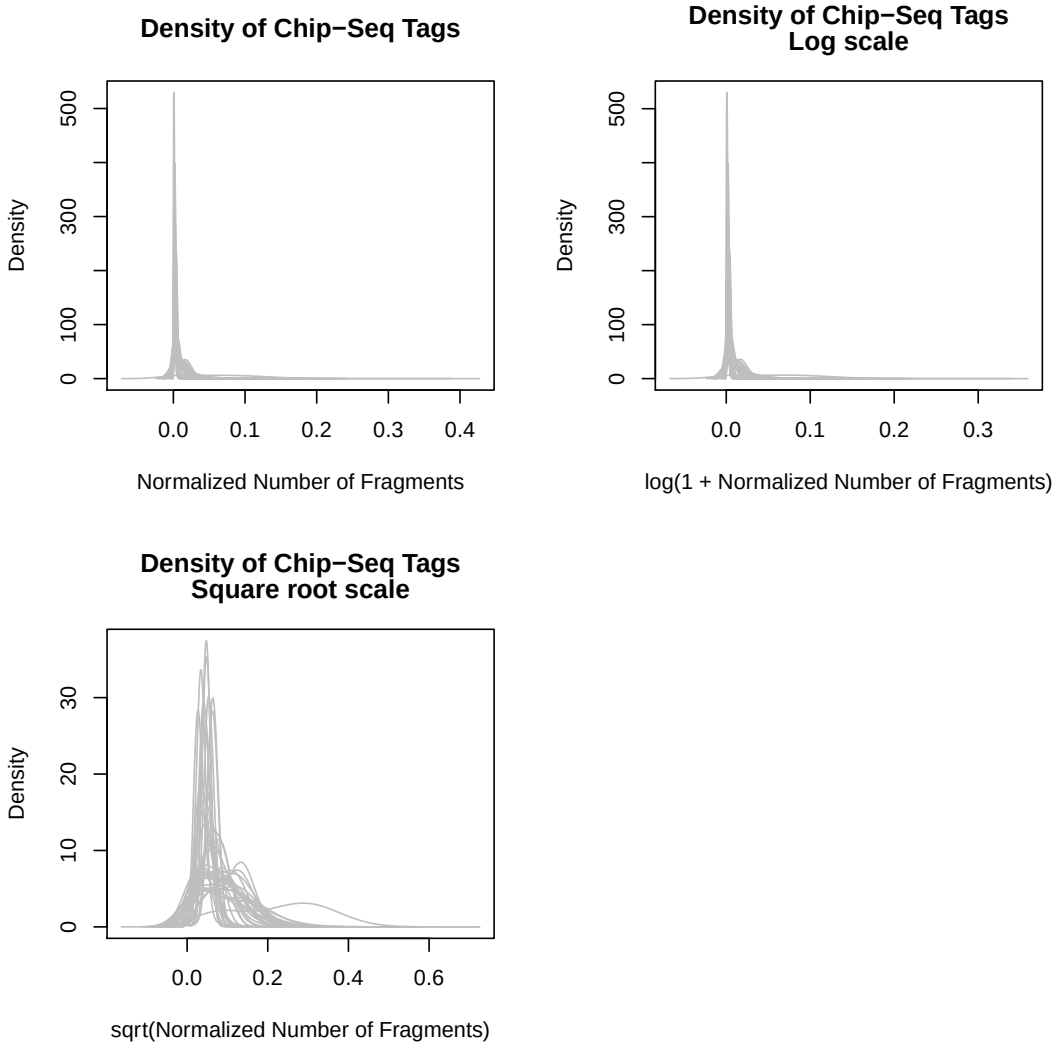

Figure 4.1: Distributions of normalized tag count in a 4kb window centered on TSSs, associated with the 39 histone modifications on the natural scale **(a)**, log scale **(b)** and square root scale **(c)**. The square root transformation allows to reduce the long tails and spread out the mass around zero.

4.2.3.2 Transformation of the response variable

In this chapter we will make use of linear models to regress either histone marks (to determine the promoter/gene boundary) or their principal components (to determine how groups of histone marks are predictive of gene expression) $\mathbf{X} = (x_{ij})$, onto the gene expression level of n gene $\mathbf{y} = (y_1, ..., y_n)^T$:

$$\mathbf{y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\epsilon} \tag{4.1}$$

where $\boldsymbol{\beta}$ is the vector of coefficients associated with the histone marks (or their principal components) and $\boldsymbol{\epsilon}$ is the vector of random error. $\boldsymbol{\beta}$ is commonly estimated by least-squares and lead to the solution :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{4.2}$$

The underlying assumptions when using a linear model are :

- the mean of the responses, $E(\mathbf{y})$, is a linear function of the predictors, and
- $\boldsymbol{\epsilon} = \mathcal{N}(0, \sigma^2\mathbf{I})$, that is, the errors are independent, normaly distributed and have equal variance.

The estimation of the coefficients, their confidence interval and associated test are valid if the model fit the data correctly. In this way a diagnostic procedure is commonly performed prior to the analysis and consists in checking the assumptions of constant variance and normality of the error, as well as removing outliers.

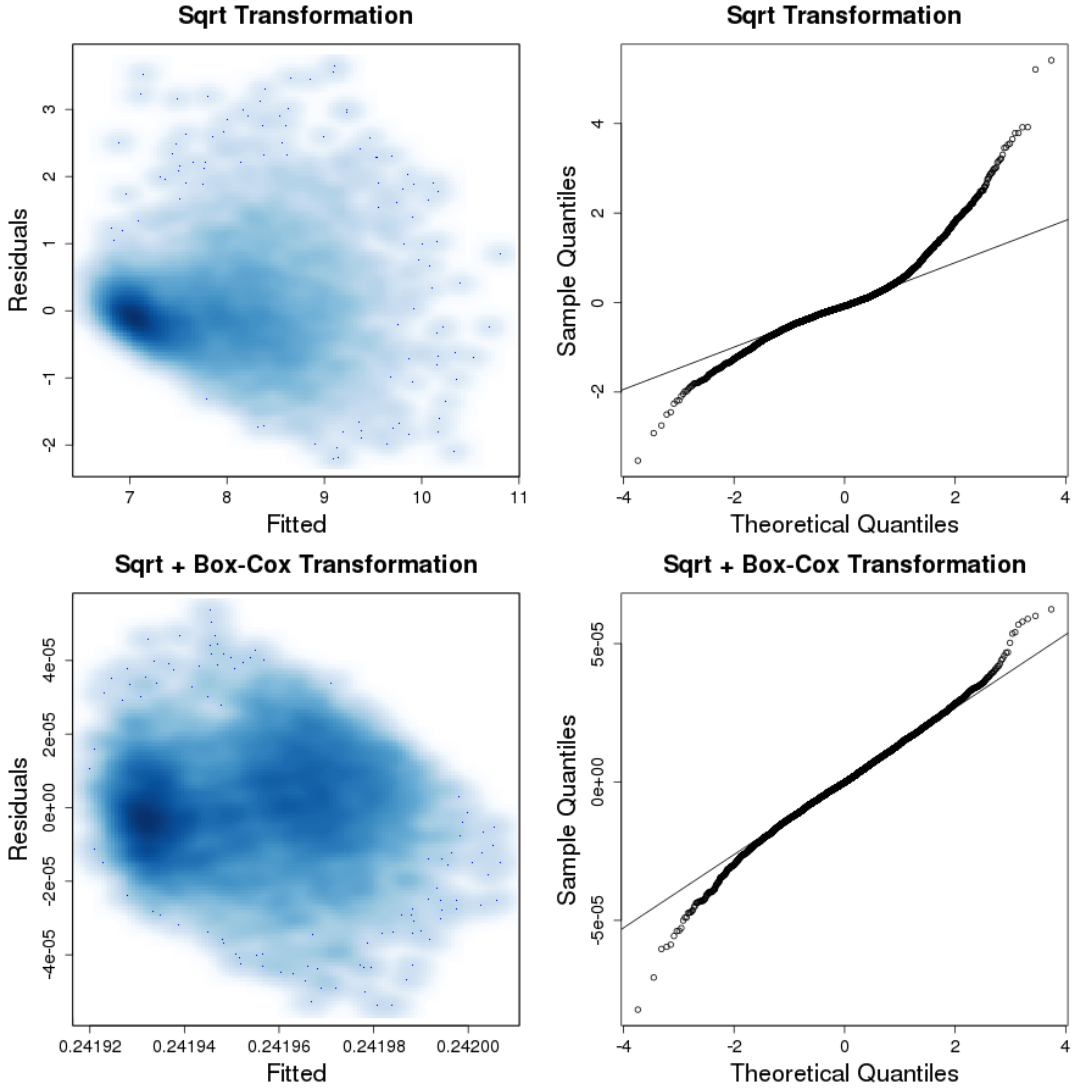

Figure 4.2: Diagnostic plots : residuals versus fitted plots (left) and normal Q-Q plots of the residual (right) computed on in the promoter region (4 kb across the TSS) on the natural scale (top), after Box-Cox transformation (middle) and after both Box-Cox and square root transformation (bottom).

To do so, we (i) plot the residuals ($\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\boldsymbol{\beta}}\mathbf{X}$) against the predicted values ($\hat{\mathbf{Y}} = \hat{\boldsymbol{\beta}}\mathbf{X}$) and check whether the error is homogeneous across the predictions, and (ii) assess the normality of the residuals using a Q-Q plot. For the sake of simplicity, this procedure is illustrated in the promoter region (4kb window centered on TSSs) where the 39 histone marks are used as predictors. A similar procedure is also performed in transcribed regions and by using the principal components as predictors. After square root transformation of the predictors, the residuals versus fitted plot suggests that the variance is not constant across the predictions and increases with the fitted values (top left, figure 4.2). Also, the Q-Q plot shows a large deviation of the data from the normal distribution on both tails (top right, figure 4.2). To deal with non-constant variance and non-normality, the Box-Cox transformation is commonly performed. The Box-Cox method transforms the response $\mathbf{y}$ into $t_\lambda(\mathbf{y})$ where the family of transformations indexed by λ is :

$$t_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ log(y) & \lambda = 0 \end{cases} \tag{4.3}$$

The parameter λ is estimated by maximum likelihood. The same two diagnostic plots are generated after Box-Cox transformation (bottom row, figure 4.2), where the residual shows higher homogeneity across the fitted values than previously. Also, even if the points do not lie perfectly along the $y = x$ line in the Q-Q plots, a substantial improvement is visible compared to what obtained before. Taken together, these diagnostic plots therefore suggest that after square root (predictors) and Box-Cox (response) transformations, the models fit the data

relatively well and can be used for further inference.

4.2.4 Determination of the Promoter/Gene Separation

Next, to find the boundary that separates promoter from gene body we look at the width of the histone enrichment across TSSs. Given that genes length vary over four orders of magnitude, the enrichment found at promoters may also vary with the gene size, making its estimation potentially difficult. However, when dividing the Ensembl transcripts into two subsets according to their length and aligning the histone signals at the TSS, no striking differences can be noticed (figures 4.3a,b). This suggests that the histone enrichment at promoter is independent of the gene length and that a same window width can be apply across TSSs to identify the enrichment at promoters. Looking at figures 4.3(a,b), the promoter specific peaks seem to be contained in a window 4-6 kb across the TSS.

To determine this window width, we fit linear models between gene expression level and the 39 histone modifications (square root of the normalized Chip-Seq count) using four different window width (2, 4, 6 and 8 kb) across the TSSs. Similarly, linear models are fitted in gene body regions where whole gene and whole gene minus 1, 2, 3 and 4 kb at the 5' end are considered. The R^2 associated with these nine models are shown figure 4.3 where, unlike gene models which are not affected by the exclusion of 1-4 kb in the 5' end regions, promoter models have better fits when increasing the window size across TSSs. This increase is particularly visible between 2 and 4 kb than 4 and 6, or 6 and 8, which raises the question of whether these different R^2 are significantly different. To answer this

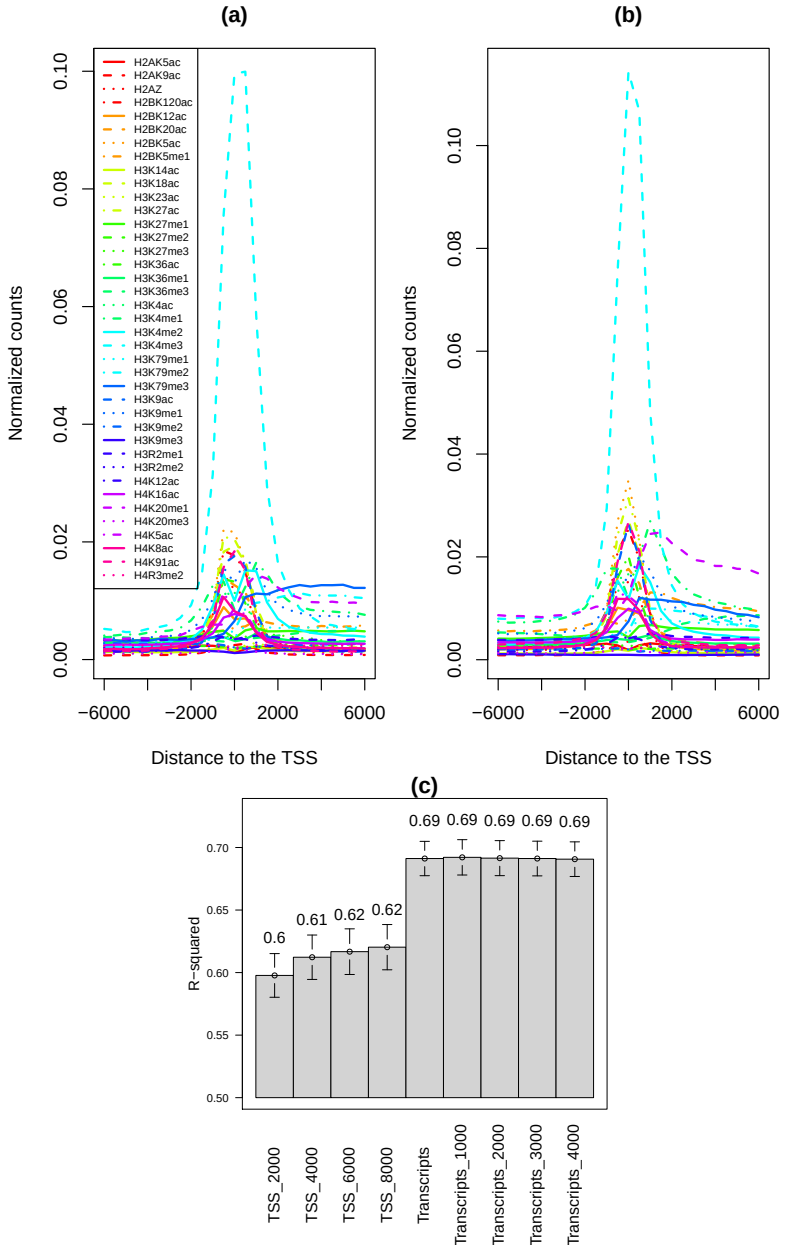

Figure 4.3: Normalized count of histone modifications across TSSs in the 50% larger **(a)** and smaller **(b)** Ensembl transcripts. Comparison of R^2 associated with linear models fitted at promoter regions (2, 4, 6 and 8 kb across TSSs) and in gene bodies where whole gene and whole gene minus 1, 2, 3 and 4 kb at the 5' end are considered **(c)**.

question, 95% confidence intervals are produced for each R^2 values by bootstrapping 1,000 times the two datasets. The resulting confidence intervals associated with promoter models overlap each others, suggesting that the differences among R^2 are not significant. However, because the 4,6,8 kb models seem very similar and larger than the 2kb one, we define the "TSS dataset" as the enrichment in histone modifications in the 4 kb region surrounding TSSs and the "transcript dataset" the histone marks enrichment within transcribed regions minus 2 kb at the 5' end. This boundary at 2kb appears as a good compromise between capturing promoter specific signal without overlapping too much with transcribed regions.

4.2.5 Collinearity Among Predictors

Before examining the differences between the histone marks at TSS and transcript regions, a look at the following correlation heatmaps (figure 4.4) built from the TSS and transcript datasets allows us to get a sense of the relationships amongst histone modifications and gene expression level (GE). The first observation that can be made is that a large number of these histone marks are highly positively correlated, indicating an important redundancy among them. Different block of marks can be distinguished in each heatmap :

- In the TSS dataset, the variables can roughly be divided into three groups. From the top to the bottom of the heatmap we have (i) a set of marks highly correlated between each other (particularly the histone acetylations) and also with the level of gene expression (GE), (ii) a set of five methylations (H3K36me3, H4K20me3, H3K9me2, H4R3me2, H3K27me2) which show lit-

tle if any correlation with the first block (including gene expression) but are positively correlated with the last block, (iii) two methylations (H3K9me3, H3K27me3) which are negatively correlated with the first group (including gene expression) and positively with the second.

- In the transcript dataset four groups can be distinguished : (i) the same block of highly correlated acetylations found in TSS (containing less marks though) that weakly correlates with gene expression, (ii) nine marks that, on average, show low positive correlation with any other histone modifications both within and outside the block, (iii) seventeen marks, highly correlated with gene expression and (iv) a set of six modifications that are negatively correlated with the first and third block (including gene expression) and positively with the second.

One way to address this problem of collinearity is to reduce the number of predictors by removing those either redundant or with low power of prediction. As the model becomes less complex, it is less able to adapt to complicated structures and result in an increased bias. At the same time, because more general, this model will suffer less from overfitting and have lower variance. This is the well-known bias-variance trade-off dilemma where the objective is to choose an estimator with the smallest mean squared error :

$$\begin{aligned} MSE(\hat{\beta}) &= E[(\hat{\beta} - \beta)^2] \\ &= Var(\hat{\beta}) + Bias(\hat{\beta}, \beta)^2 \end{aligned} \tag{4.4}$$

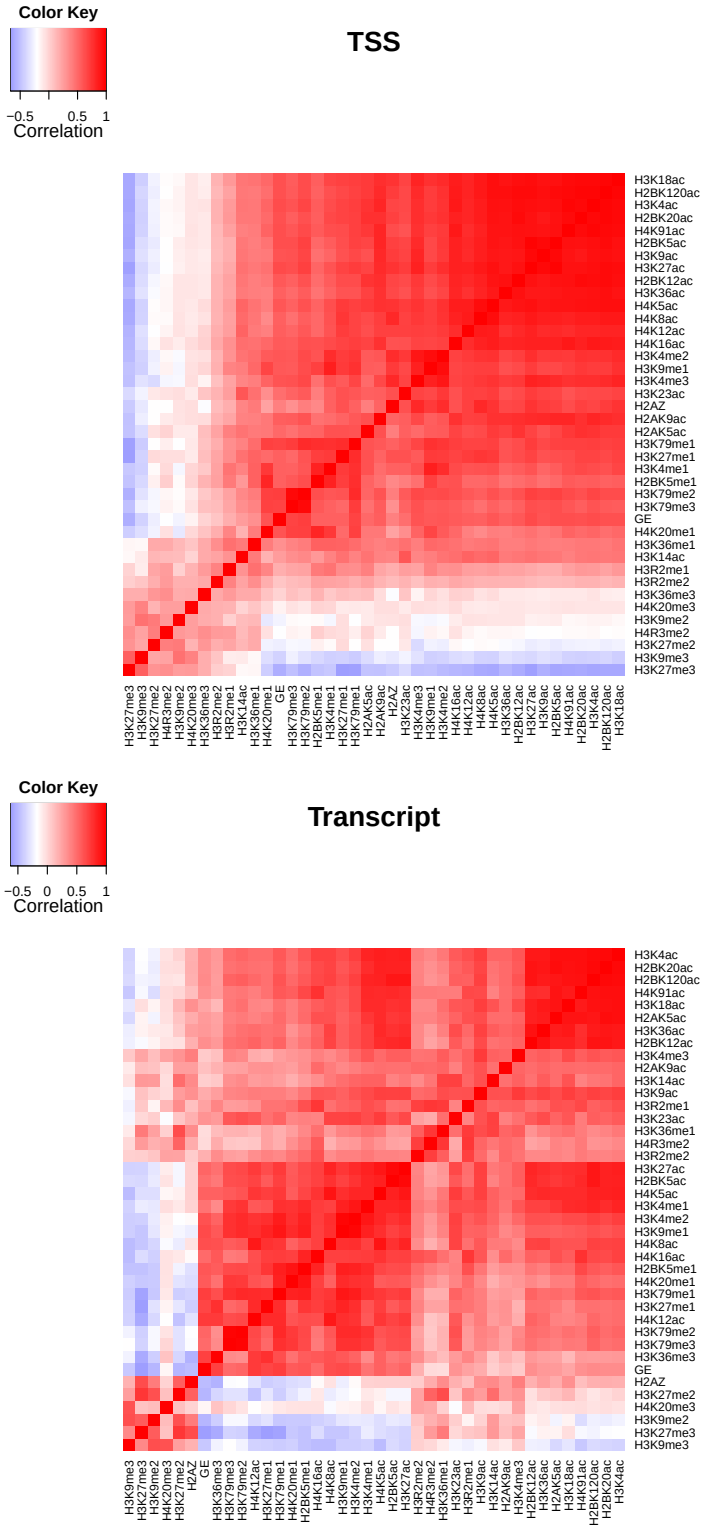

Figure 4.4: Heatmap of correlation among histone modifications and gene expression level (GE) in the TSS (top) and transcript (bottom) datasets.

Among the large number of approaches to variable selection and coefficient shrinkage that have been proposed (Hastie et al., 2003), we choose principal component regression in the following for its ability to both identify sets of correlated predictors and measure the contribution of these sets to the variation of the response variable. To ease the interpretation of the principal components, we make use of an alternative approach called sparse PCA. Sparse PCA aims to express the principal axes with as few original variables as possible while accounting for most of the variation in the explanatory variables. A shrinkage parameter allows to trade-off between prediction accuracy and number of explanatory variables used in each principal axes. This approach allows therefore to cluster correlated histone modifications into few principal axes and then evaluates their influence on gene expression level.

4.3 Principal Component Regression

4.3.1 Description

Principal component regression (PCR) proceeds in the same way as linear regression except that $\mathbf{y}$ is regressed on the principal components of $\mathbf{X}$ instead of $\mathbf{X}$ directly. This method is well suited when many predictors are highly correlated and can be represented by a small number principal components. The principal components can be obtained by decomposing $\mathbf{X}$ into orthogonal scores (principal components) $\mathbf{U}$ and loadings (principal axes) $\mathbf{V}$ by singular value decomposition :

$$\mathbf{X} = \mathbf{UDV}^T \tag{4.5}$$

where $\mathbf{U}$ and $\mathbf{V}$ are $n \times p$ and $p \times p$ orthogonal matrices, with the columns of $\mathbf{U}$ and $\mathbf{V}$ spanning the column spaces and the row spaces of $\mathbf{X}$ respectively, $\mathbf{D}$ is a $p \times p$ diagonal matrix whose values $d_1 \geq d_2 \geq ... \geq d_p \geq 0$ are the singular values of $\mathbf{X}$. The principal components $\mathbf{U}$ therefore give the positions of the original data in the new system of principal axes.

PCA can been viewed as a rotation under the constrains that each new axis $\mathbf{u_i}$, is orthogonal to the previous one(s) and lies in the direction of the greatest variation left after removing the effect of the previous components. In this way the first principal component (PC) $\mathbf{u_1}$ yields the largest sample variance among all possible linear combinations of the columns of $\mathbf{X}$ and can be expressed as :

$$Var(\mathbf{u_1}) = \frac{d_1^2}{n} \tag{4.6}$$

Since PCA looks for linear combination of the original data with high variance, this technique depends on the scaling of the data. To prevent variables with large variance to dominate the principal components the data are typically standardized. Given that all the histone modifications used in this analysis come from Zhao's group and have been produced with the same protocol, we only centered them. After running a principal component decomposition on $\mathbf{X}$, $\mathbf{y}$ is regressed onto the principal components instead of the predictors where the coefficients are

given by :

$$\hat{\boldsymbol{\beta}} = \mathbf{V}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{y} = \mathbf{V}\mathbf{D}^{-1}\mathbf{U}^T\mathbf{y} \tag{4.7}$$

4.3.2 Sparse PCA

The goal of PCA is to summarize correlated variables into a small number of principal components, see how much of the variations they explain and read the contribution of the original variables to each principal components from the loadings. Since most of the variability is generally captured by the first few principal components, one typically discard those associated with small eigenvalues d_i to make the interpretation simpler. To ease the interpretation of the loadings, and because many chromatin states involve only a few histone modifications (Schones and Zhao, 2008), a further simplification can also be done by forcing some of them to zero.

In factor analysis, this problem is addressed by finding a rotation that reduces the number of non-null loadings. In PCA, Zou et al. (2006) propose an approach inspired from shrinkage regression that induce sparsity in the loading matrix. The idea is that since PCA can be expressed as a regression problem, a shrinkage method can be used instead of the ordinary least squares, setting some loadings to zero. Zou et al. opted for the elastic net penalty (Zou and Hastie, 2005) which combines the norm 1 and 2 penalties of the coefficients such that the first encourages sparse loadings while the second encourages highly correlated predictors to

be averaged. In this way, we replace the classical PCA by Zou et al.'s method in the principal component regression to induce sparsity in the loadings and make them easier to interpret.

Suppose we consider the first k principal components, let $\mathbf{X}_i$ denote the ith row vector of the matrix $\mathbf{X}$ and $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ be $p \times k$ matrices. For any norm 1 and 2 penalties $\lambda_1 > 0$, $\lambda_2 > 0$, Zou et al. showed that the solution of the optimization problem :

$$\begin{aligned}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) = &\underset{\alpha,\beta}{\operatorname{argmin}} \sum_{i=1}^{n} |\mathbf{X}_i - \boldsymbol{\alpha}\boldsymbol{\beta}^T \mathbf{X}_i|^2 + \lambda_2 \sum_{j=1}^{k} |\boldsymbol{\beta}_j|^2 + \sum_{j=1}^{k} \lambda_1 |\boldsymbol{\beta}_j|_1 \\ &\text{subject to } \boldsymbol{\alpha}^T \boldsymbol{\alpha} = \mathbf{I}_k. \end{aligned} \tag{4.8}$$

leads to $\hat{\boldsymbol{\beta}} \propto \mathbf{V}$. In this way, the larger λ_1, the sparser the $\mathbf{V}$s are. To determine this λ_1 parameter and the optimal number of principal components to keep in PCR, we evaluate the R^2 and the mean squared error of prediction (MSEP) in 10-fold cross validation for various values of λ_1 and k :

$$MSEP = \frac{1}{n} \sum_{k=1}^{10} \sum_{i \in \kappa(k)} (f^{-k}(\mathbf{x}_i) - y_i)^2 \tag{4.9}$$

where κ is an indexing function $\kappa : \{1, ..., n\} \mapsto \{1, ..., 10\}$ that indicates which partition (fold) observation i belongs to, and f^{-k} the PCR trained on all partition but k.

On figure 4.5, the values of MSEP and R^2 are displayed for various number

of principal components as a function of the number of non-null loadings per principal component (which is directly related to λ_1). In both set, the R^2 curves increase markedly for small number of non-null loadings (1 - 10) and then reach a plateau for larger values. This can be explained by the redundancy among histone marks where few well-chosen marks representative of a given function are sufficient to explain most of the variability. Important improvements in model fit can be observed when the number of PCs is larger than 4 and 5 in the transcript and TSS sets respectively, which motivate us to select the models with :

- 5 PCs and 5 non-null loadings per PC in the transcript set,
- 6 PCs and 7 non-null loadings per PC in the TSS set,

as indicated by the black rectangles figure 4.5. The R^2 associated with these two models corresponds to 91% and 97% of the full models fits and thus indicate that most of the information is kept after dimension reduction in both sets.

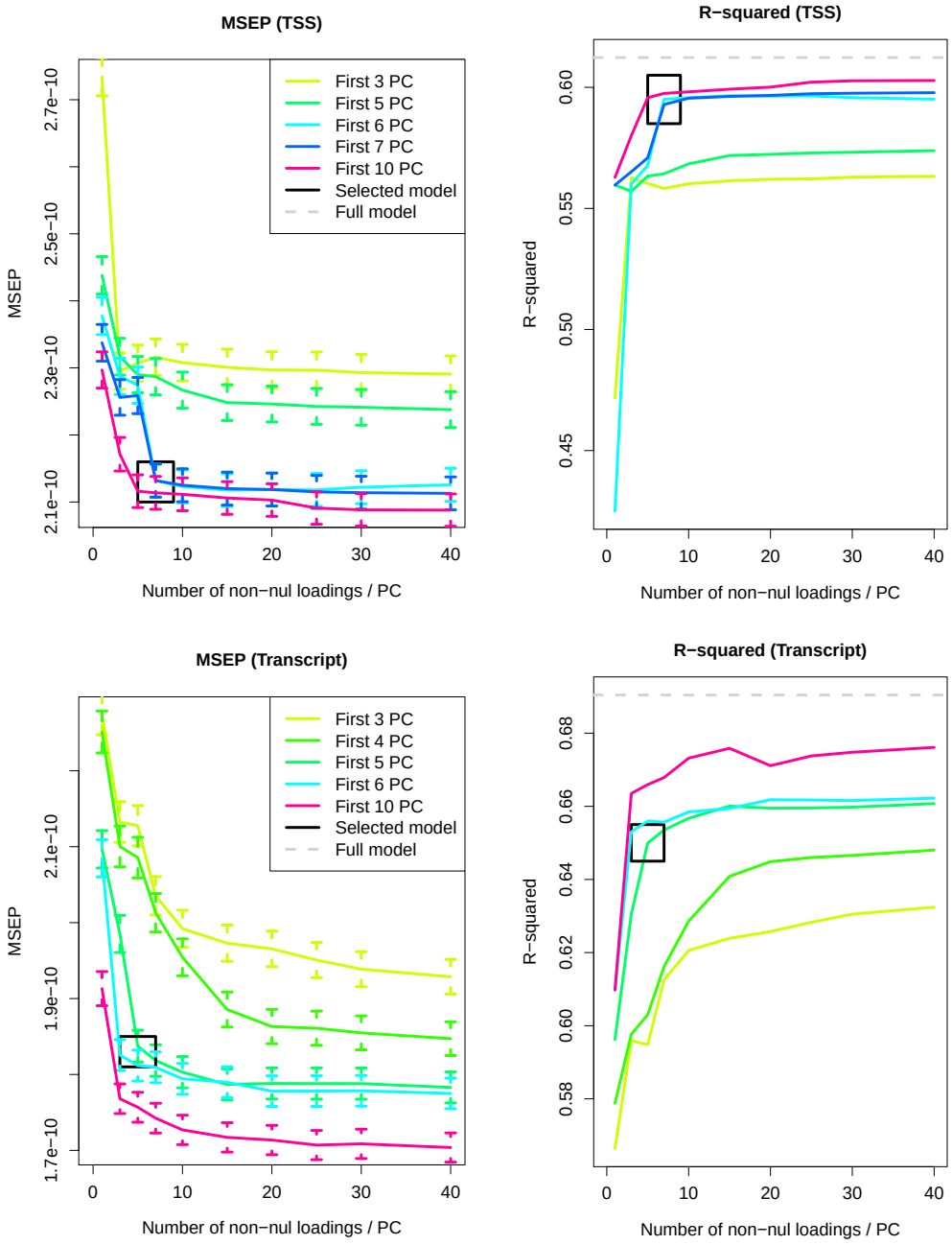

Figure 4.5: Tenfold cross-validation estimate of mean-squared error prediction (along with ± two standard error bars) and R^2 against the number non-null loadings per principal component in TSS (top) and transcript (bottom). In the TSS set, a good trade-off between accuracy and interpretability is reached when PCR is fitted with the first 6 principal components and 7 non-null loadings per PC (indicated by the black rectangle). Similarly, we consider that a good trade-off is reached in the transcript set when the first 5 principal components and 5 non-null loadings per PC are considered. The full models R^2 are shown by the gray dashed lines.

4.4 Characterization of the Principal Components

4.4.1 Principal Axes

The loadings found in the TSS and transcript sets are displayed figure 4.6 and represent the contribution of the histone marks to the principal axes in both datasets. The action of sparse PCA is clearly visible, as shown by the small number of loadings with non-null values in each principal axis. These non-null loadings represent clusters of histone modifications that may be associated with different functions.

The remainder of this section is dedicated to the characterisation of the PCs' function and their spatial relationship with gene expression level. To do so, we (i) relate the histone marks that constitute the principal axes to chromatin signatures identified in previous studies, (ii) check whether common chromatin patterns are detected in promoter and transcribed regions, (iii) look at the sign and the strength of the association between the PCs and gene expression level, the variation of this association (iv) between cell-type specific and housekeeping genes and (v) spatially, across the TSS.

Because the function of the principal components will be refined as we advance through the chapter, the following notation will be used to keep track of their function : $\mathrm{PC}_{\mathrm{Function}}^{\mathrm{dataset-PC\_Number}}$, where function is one of *Tr* (gene activation in transcribed regions), *Pr* (gene activation in promoter regions), *Co* (gene activa-

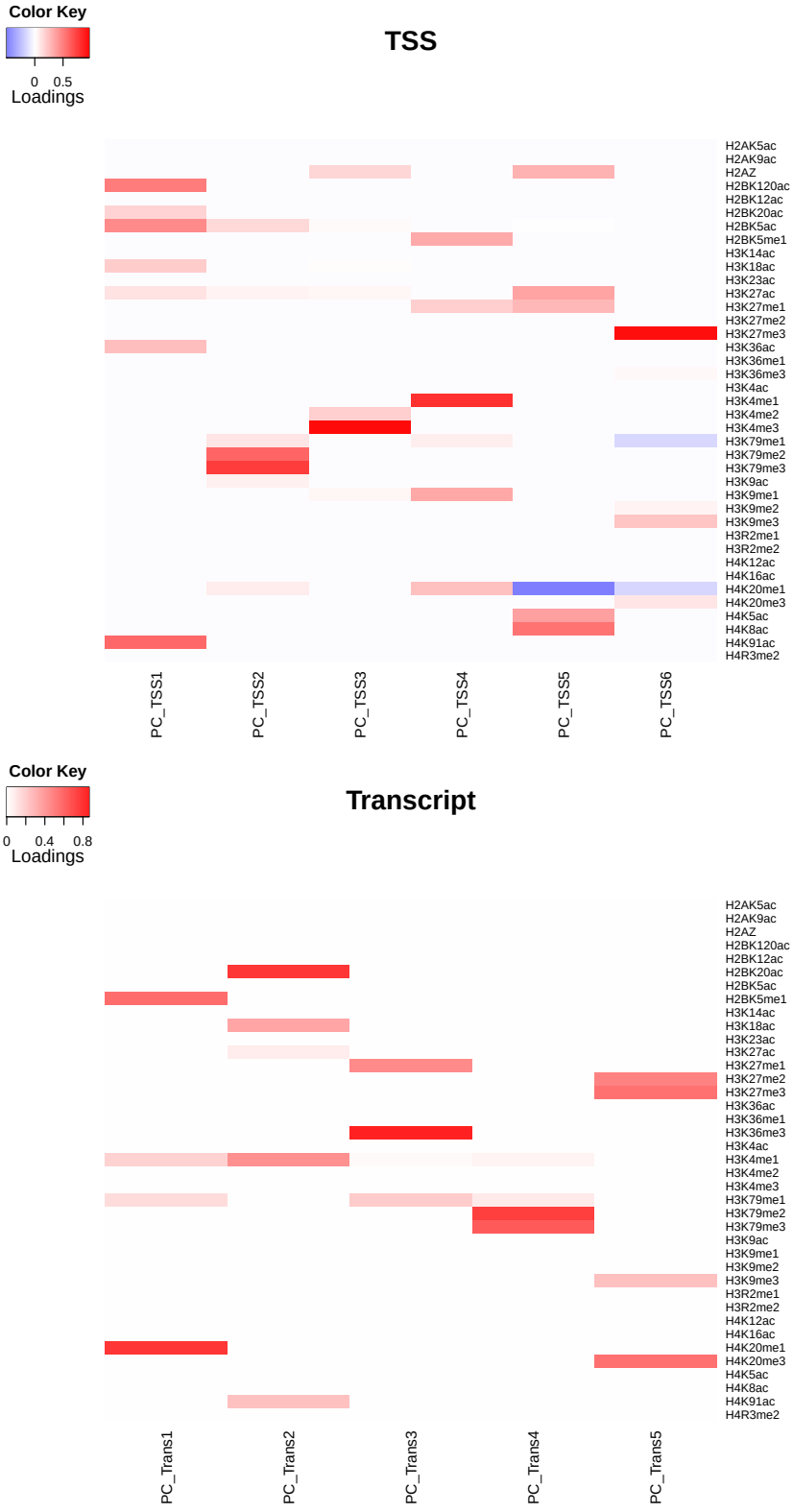

Figure 4.6: Loadings computed by the sparse PCA method in TSS and transcript datasets.

tion in promoter and transcribed regions), *Si* (gene silencing), *Ti* (tissue specific) and *Un* (undetermined). Each time a PC function is refined, the new function will be reflected in its name.

4.4.2 Characterization Based on Previous Studies

To characterise the PCs, we first aim to relate the histone marks associated with their principal axes to the chromatin signatures found in the studies presented in the introduction section. These signatures are either in the form of groups of histones modifications (as shown table 4.1) or emission probabilities associated with the chromatin states inferred by Ernst and Kellis HMM. For the first category, we measure the similarity between principal axes and chromatin signatures by Manhattan (l_1) distance after binarizing the loadings :

$$d_1(\mathbf{a}^j, \mathbf{pa}^k) = ||\mathbf{a}^j - \mathbf{pa}^k||_1 = \sum_{i=1}^{p} |a_i^j - pc_i^k| \tag{4.10}$$

where a_i^j and pa_i^k are dummy variables indicating the presence or absence (1/0) of the ith histone mark in the jth annotation set and the kth principal axis. For the second category, the similarity is given by the angle between each couple of principal axis and vector of emission probabilities.

The results corresponding to the TSS and transcript sets are summarized in table 4.2 and 4.3 respectively. For each PC, the Manhattan distances with the chromatin signatures characterised in Barski et al. (2007); Ucar et al. (2011); Wang

Source	Category	PC^{TSS1}	PC^{TSS2}	PC^{TSS3}	PC^{TSS4}	PC^{TSS5}	PC^{TSS6}
Barski & Wang	Promoter	**6**	14	**10**	19	9	18
	Gene Body	11	**11**	11	**6**	**8**	7
	Common	17	13	17	12	16	19
	Silencing	12	12	12	11	11	**2**
Ucar	Promoter	**6**	14	10	19	9	18
	Gene Body	7	**7**	7	**12**	**6**	**11**
	Enhancer	7	11	7	**12**	8	**11**
	Insulator	7	**7**	**5**	**12**	8	**11**
Ernst		Active Inter.	Trans.	Prom.	Trans.	Trans.	Silencing.
	Majority	Prom.	Trans.	Prom.	Trans.	Trans.	Silencing

Table 4.2: Characterization of PCs' function in the TSS set based on previous annotations. Shown are the Manhattan (l_1) distances calculated between each PC and chromatin signature annotated in Barski et al. (2007); Ucar et al. (2011); Wang et al. (2008). For each PC, the smallest l_1 distance is in bold. The angles between the principal axes and the emission probabilities inferred by Ernst and Kellis are measured. The hidden state associated with the smallest angle (highest collinearity) is displayed. Each PC is then classified based on the majority vote of the 3 sources of annotations.

Source	Category	PC^{Trans1}	PC^{Trans2}	PC^{Trans3}	PC^{Trans4}	PC^{Trans5}
Barski & Wang	Promoter	18	10	16	18	17
	Transcription	**5**	**9**	**5**	**9**	8
	Common	13	15	15	**9**	18
	Silencing	10	10	10	10	**1**
Ucar	Promoter	18	10	16	18	17
	Gene Body	**11**	**9**	**9**	**11**	**10**
	Enhancer	**11**	**9**	**9**	**11**	**10**
	Insulator	**11**	**9**	**9**	**11**	**10**
Ernst		Trans.	Active Inter.	Trans.	Trans.	Silencing.
	Majority	Trans	Undetermined	Trans.	Trans.	Silencing

Table 4.3: Characterization of PCs' function in the transcript set in the same fashion as table 4.2.

et al. (2008) as well as the hidden state whose emission probabilities have the highest collinearity with the associated principal axis are displayed. The last row of both tables proposes a classification for each PC based on the majority vote of the 3 sources. As the annotation from Barski et al.; Wang et al. and Ernst and Kellis show very good consistency, the proposed classification is mostly driven by them. Out of 6 PCs in the TSS set, 2 are classified as promoter, 3 as transcript and 1 as silencing. Similarly, the 5 PCs found in the transcript set can be split into 3 transcript, 1 gene silencing and 1 unknown PCs.

In this classification we can note that (i) transcript-specific PCs are overrepresented in the two sets, suggesting that the signal associated with gene body is detected in a close proximity of TSSs; (ii) more than one TSS and transcript specific signatures are found in the promoter and both sets respectively, which is consistent with previous studies that identified multiple clusters of histone marks in these two regions.

4.4.3 Collinearity Among Principal Axes

The annotations from previous studies provided a useful starting point for the characterisation of the PCs. We now wonder whether among the silencing and the 3 transcript-specific PCs present in each set, some of them are shared across the two regions. Similarly to the previous subsection, we measure the angle between each pair of principal axes (PA) and identify those with a high degree of collinearity. This measure has been shown to be directly linked to the correlation among the vectors' coordinates by the cosine function (Legendre and Legendre,

1998) :

$$r_{PA_i,PA_j} = cos(\theta_{PA_i,PA_j}) \tag{4.11}$$

for any principal axes PA_i and PA_j. Since these two measures are equivalent, correlations instead of angles are used in this subsection to facilitate the comprehension.

Figure 4.7 reveals that out of five principal axes in the transcript set, only one ($\mathrm{PA}_{\mathrm{Tr}}^{\mathrm{Trans3}}$) does not seem related to those in the TSS set. The first two principal axes in the transcript set, $\mathrm{PA}_{\mathrm{Tr}}^{\mathrm{Trans1}}$ and $\mathrm{PA}_{\mathrm{Un}}^{\mathrm{Trans2}}$, show modest correlations with $\mathrm{PA}_{\mathrm{Tr}}^{\mathrm{TSS4}}$ and $\mathrm{PA}_{\mathrm{Tr}}^{\mathrm{TSS5}}$, and $\mathrm{PA}_{\mathrm{Pr}}^{\mathrm{TSS1}}$ and $\mathrm{PA}_{\mathrm{Tr}}^{\mathrm{TSS4}}$ respectively. If we compare the categories of these correlated principal axes, we can note that $\mathrm{PA}_{\mathrm{Tr}}^{\mathrm{Trans1}}$, $\mathrm{PA}_{\mathrm{Tr}}^{\mathrm{TSS4}}$ and $\mathrm{PA}_{\mathrm{Tr}}^{\mathrm{TSS5}}$ are all annotated transcript-specific, whereas $\mathrm{PA}_{\mathrm{Un}}^{\mathrm{Trans2}}$ correlates both with a promoter-specific ($\mathrm{PA}_{\mathrm{Pr}}^{\mathrm{TSS1}}$) and a transcript-specific ($\mathrm{PA}_{\mathrm{Tr}}^{\mathrm{TSS4}}$) principal axes which suggests that it contains histone marks associated with both regions. This could explains its classification as undetermined.

The last two PAs in the transcript set ($\mathrm{PA}_{\mathrm{Tr}}^{\mathrm{Trans4}}$ and $\mathrm{PA}_{\mathrm{Si}}^{\mathrm{Trans5}}$) correlate with only one PA in the TSS set each, with larger correlation values than the two previous ones. The almost perfect correlation between $\mathrm{PA}_{\mathrm{Tr}}^{\mathrm{Trans4}}$ and $\mathrm{PA}_{\mathrm{Tr}}^{\mathrm{TSS2}}$, both classified as transcript-specific, indicates they are identical and capture the same signal (H3K79me1/2/3, as shown on the heatmap of loadings). Similarly, the substantial correlation between the two silencing $\mathrm{PA}_{\mathrm{Si}}^{\mathrm{TSS6}}$ and $\mathrm{PA}_{\mathrm{Si}}^{\mathrm{Trans5}}$, suggests

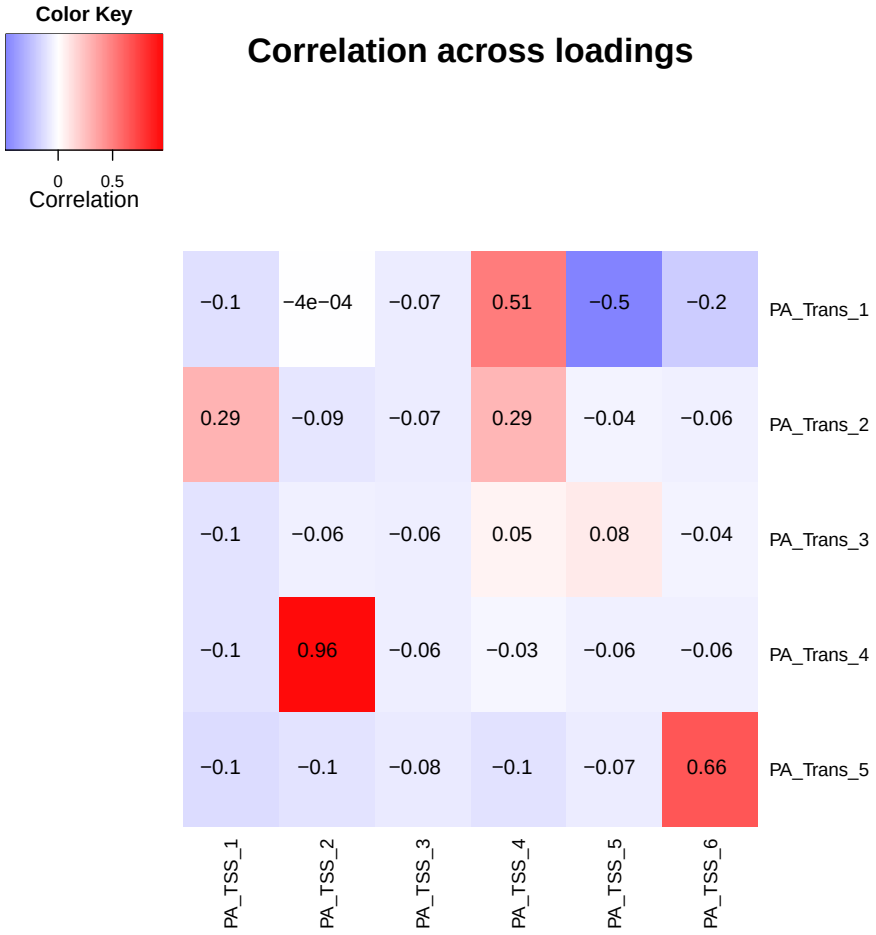

Figure 4.7: Correlation among the principal axes in the TSS (vertical) and transcript (horizontal) sets.

that the same histone marks are involved in gene silencing in both regions.

Overall these results supplement the classification made previously by revealing that :

- the principal axes $\mathrm{PA}_{\mathrm{Tr}}^{\mathrm{TSS3}}$ and $\mathrm{PA}_{\mathrm{Tr}}^{\mathrm{Trans3}}$, annotated as TSS and transcript specific, show little or no correlation with any PA found in the other set, supporting the fact they are region specific;
- conversely, each pair of PAs, $\mathrm{PA}_{\mathrm{Tr}}^{\mathrm{TSS2}}$ and $\mathrm{PA}_{\mathrm{Tr}}^{\mathrm{Trans4}}$, $\mathrm{PA}_{\mathrm{Si}}^{\mathrm{TSS6}}$ and $\mathrm{PA}_{\mathrm{Si}}^{\mathrm{Trans5}}$, classified as transcript-specific and gene silencing respectively, show high degree of collinearity, suggesting that each pair detect the same signal;
- the uncharacterised $\mathrm{PA}_{\mathrm{Un}}^{\mathrm{Trans2}}$ correlates with two PAs classified as promoter and a transcript specific, implying that it consists of histone marks associated with both regions ($\mathrm{PA}_{\mathrm{Co}}^{\mathrm{Trans2}}$).

4.4.4 Variable Importance

To gain further insight into the characterisation of the PCs, we now aim to understand the strength and the sign of their relationship with gene expression.

4.4.4.1 CAR Scores

We are interested in the quantification of each PC's contribution to gene expression variation, that is, we would like to decompose the explained variance of gene expression by PC. When the predictors are uncorrelated, the assessment of each predictor's contribution is simply given by the R^2 from its univariate regression,

all univariate R^2 values adding up to the full model R^2. In our case, because sparse PCA is used to simplify the interpretation of loadings, the PCs are no longer orthogonal, making the full model R^2 not straightforward to break down into PCs. To address this problem, different measures of relative importance have been proposed among which we choose the one introduced by (Zuber and Strimmer, 2010) based on the CAR score :

$$\boldsymbol{\omega} = \mathbf{P}^{-1/2}\mathbf{P}_{\mathbf{Xy}} \tag{4.12a}$$

$$= Corr(\mathbf{P}^{-1/2}\mathbf{X}_{std}, \mathbf{y}) \tag{4.12b}$$

The CAR score can be interpreted as the marginal correlations between response and predictors ($\mathbf{P}_{\mathbf{Xy}}$) adjusted by the correlation among predictors ($\mathbf{P}^{-1/2}$), or equivalently, as the correlation between the response and the decorrelated predictors ($\mathbf{P}^{-1/2}\mathbf{X}_{std}$, where $\mathbf{X}_{std}$ are the standardized predictors). The associated measure of variable importance is given by the square of $\boldsymbol{\omega}$ and presents the interesting properties that (i) it is non-negative, (ii) $\boldsymbol{\omega}^2$ decomposes the proportion of variance explained by the predictors ($\sum_{i=1}^{p} \omega_i^2 = R^2$), (iii) ω_i^2 reduces to the univariate R^2 when the explanatory variables are uncorrelated, and (iv) the decomposition respects orthogonal subgroups. Furthermore, this measure is closely related to partial correlation, where instead of calculating the correlation of $\mathbf{y}$ and $\mathbf{X_i}$ after removing the linear effect of the remaining $p-1$ predictors $\mathbf{X}_{-i}$, the response $\mathbf{y}$ is left unchanged in CAR score (4.12b).

In the rest of this chapter, the CAR score will be used to measure the contribution of each PC to gene expression level. To ensure this measure is consistent

with other methods, we compared the CAR score with two others metrics **lmg** and **pmvd** presented in Gromping (2006) and found almost identical results.

4.4.4.2 Application to the TSS and Transcript Sets

We now compute the CAR scores for each PC in the TSS (figure 4.8, top left) and transcript sets (bottom right). Instead of using the CAR score directly, we calculate $\text{sign}(\text{CAR}) \times \text{CAR}^2$ which has the advantage of giving both the direction of the relationship and the contribution of the PC to the full model's R^2. In this way, the last PCs in both sets have large negative scores, consistently with their silencing role. If we now examine the scores in the TSS set, we can notice that the largest score is obtained by $\text{PC}_{\text{Co}}^{\text{TSS2}}$, annotated as transcript-specific, while those classified as specific to this regions, $\text{PC}_{\text{Pr}}^{\text{TSS1}}$ and $\text{PC}_{\text{Pr}}^{\text{TSS3}}$, have more modest scores. This somehow surprising result either calls into question the classification of $\text{PC}_{\text{Tr}}^{\text{TSS2}}$ or suggests that the histone modifications that influence the most gene expression in the TSS region are common to those in of gene body.
In the transcript set it is $\text{PC}_{\text{Tr}}^{\text{Trans3}}$, annotated as specific to this region, which dominates the other transcript-specific $\text{PC}_{\text{Tr}}^{\text{Trans1}}$, $\text{PC}_{\text{Tr}}^{\text{Trans4}}$ and the undetermined $\text{PC}_{\text{Co}}^{\text{Trans2}}$ principal components.

To pinpoint region-specific PCs, we now project the two datasets onto the principal axes derived from the other set. The reason is as follows. The principal components $\mathbf{U}$ of a dataset $\mathbf{X}$ are obtained by projecting the row vectors $\mathbf{X}_i$ onto the principal axes $\mathbf{V}$ that correspond to the directions of maximum variation through $\mathbf{X}$. Among the principal axes in a given set, some of them are likely to represent region-specific variations in the data unlikely to be present in the other

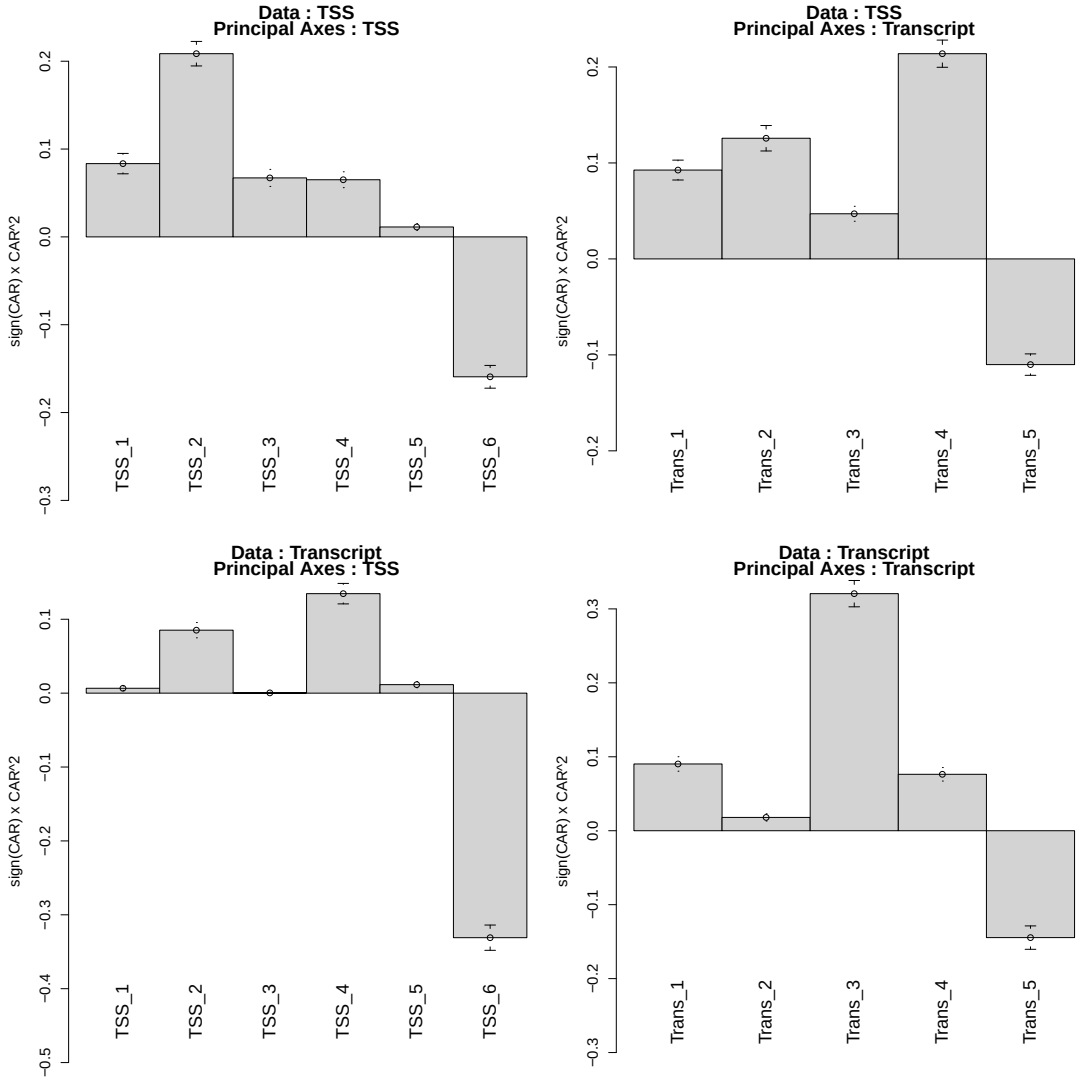

Figure 4.8: Car scores associated with the principal components in the TSS and transcript datasets. To pinpoint region-specific PCs, the TSS and transcript datasets are projected onto the principal axes derived from the transcript (top right) and the TSS sets respectively (bottom left). For each CAR score, a 95% confidence interval is produced by bootstrapping 1,000 times the two datasets.

set. Therefore, a simple way to identify these region-specific axes is to project one dataset onto the principal axes of the other dataset and look for those with small or null projections. Principal axes can therefore be viewed as filters that let the signals associated with their function unchanged and remove those which are not specific.

Instead of directly looking at the projections of the transcript dataset onto the TSS principal axes, we measure the association of these pseudo principal components with gene expression level using the CAR scores. If the projections onto the TSS-specific principal axes are small or null, as expected, their corresponding scores should also follow the same trend. Figure 4.8, bottom left, displays the CAR scores associated with the projections of the transcript set onto the TSS principal axes. The null scores obtained with $\mathrm{PA}_{\mathrm{Pr}}^{\mathrm{TSS1}}$ and $\mathrm{PA}_{\mathrm{Pr}}^{\mathrm{TSS3}}$ suggest that these two principal axes are TSS-specific. This result is in phase with their previous annotations and their weak collinearity with the principal axes from the transcript set. Conversely, $\mathrm{PA}_{\mathrm{Tr}}^{\mathrm{TSS2}}$, classified as transcript-specific, has a smaller score than obtained in the TSS set but non-null, suggesting that this principal axe contain signal common to both set.

The same procedure is performed with the TSS set that we project onto the principal axes derived from the transcript set (figure 4.8, top right). $\mathrm{PA}_{\mathrm{Tr}}^{\mathrm{Trans3}}$, annotated as transcript-specific, has a much smaller score than in the transcript set, confirming thus its transcript-specific classification. $\mathrm{PA}_{\mathrm{Co}}^{\mathrm{Trans2}}$ and $\mathrm{PA}_{\mathrm{Tr}}^{\mathrm{Trans4}}$ on the other hand, have larger score than in the TSS set. Since a near zero score was obtained in the transcript set for $\mathrm{PA}_{\mathrm{Co}}^{\mathrm{Trans2}}$, this result point to a higher speci-

ficity of this PA for TSS regions. This observation is consistent with the fact that PA_{Pr}^{Trans2} was related to the TSS-specific PA_{Pr}^{TSS1} in the previous subsection. In the previous subsection we also highlighted that PA_{Tr}^{Trans4} and PA_{Co}^{TSS2} were collinear, which implies that the lack of specificity of PA_{Co}^{TSS2} for either regions also applies for PA_{Co}^{Trans4}.

Lastly, little variations across the TSS and transcript sets have been observed for PA_{Tr}^{Trans1} and PA_{Tr}^{TSS4}, which suggest they are common to both regions.

4.4.5 Tissue/Cell-Type Specific versus Housekeeping Genes

So far, we have sought to classify the principal components based on their genomic location (promoter, transcribed regions) or their influence on gene expression (activating, silencing), but other categories such as genomic features or gene function could help us refining their understanding. For instance, Karlic et al. found that two distinct sets of histone marks decorate low and high CpG content promoters. Using the same criterion as Karlic et al., we separated the set of Ensembl genes into those with low and high CpG content at promoter and tested whether some PCs showed an association with this feature. The examination of the PCs' CAR scores in both sets did not reveals any clear relationship between any PCs and the CpG content.

Two other studies showed that genes involved in tissue/cell-type specific pathways display different patterns of histone marks at promoters as compared with housekeeping genes (Pekowska et al., 2010; Zhang and Zhang, 2011). In particular,

Pekowska et al. identified a H3K4me2 peak spanning few kilo bases downstream of the TSS in tissue-specific genes that is (i) independent of gene expression level and (ii) enriched in DNase I hypersensitivity sites, hallmark of *cis*-regulatory elements. Zhang and Zhang, on the other hand, identified three histone marks, H3K4me3, H3K27ac, H3K79me3, that are preferentially found at tissue-specific genes.

Similarly to what we did with CpG content, we test whether some PCs are associated with tissue-specific genes by computing their CAR scores in tissue-specific and housekeeping genes separately. To do so, we make use of the 454 CD4+T cell-specific and 630 housekeeping genes selected by Zhang and Zhang on the basis of their gene expression's entropy across different tissues. After projecting these two datasets onto the TSS and transcript principal axes, CAR scores are computed and 95% confidence intervals are generated from 1,000 bootstrap samples. On figure 4.9, we can notice that the confidence intervals are much larger than those obtained before on the full TSS and transcript datasets and make the comparison between CAR scores almost impossible. This problem is probably due to the smaller number of genes used in the cell-specific and housekeeping sets. An important difference can nevertheless be noted at $\mathrm{PC}_{\mathrm{Pr}}^{\mathrm{TSS3}}$ between the two sets of genes, where the CAR score is null in housekeeping genes but not in the tissue-specific genes.

Despite the fact that this difference is not significant (the two confidence intervals overlap), when looking at the heatmap of loadings figures 4.6, it can be noted that $\mathrm{PA}_{\mathrm{Ti}}^{\mathrm{TSS3}}$ is made up of three histone marks, H3K4me2/3 and H2AZ, all

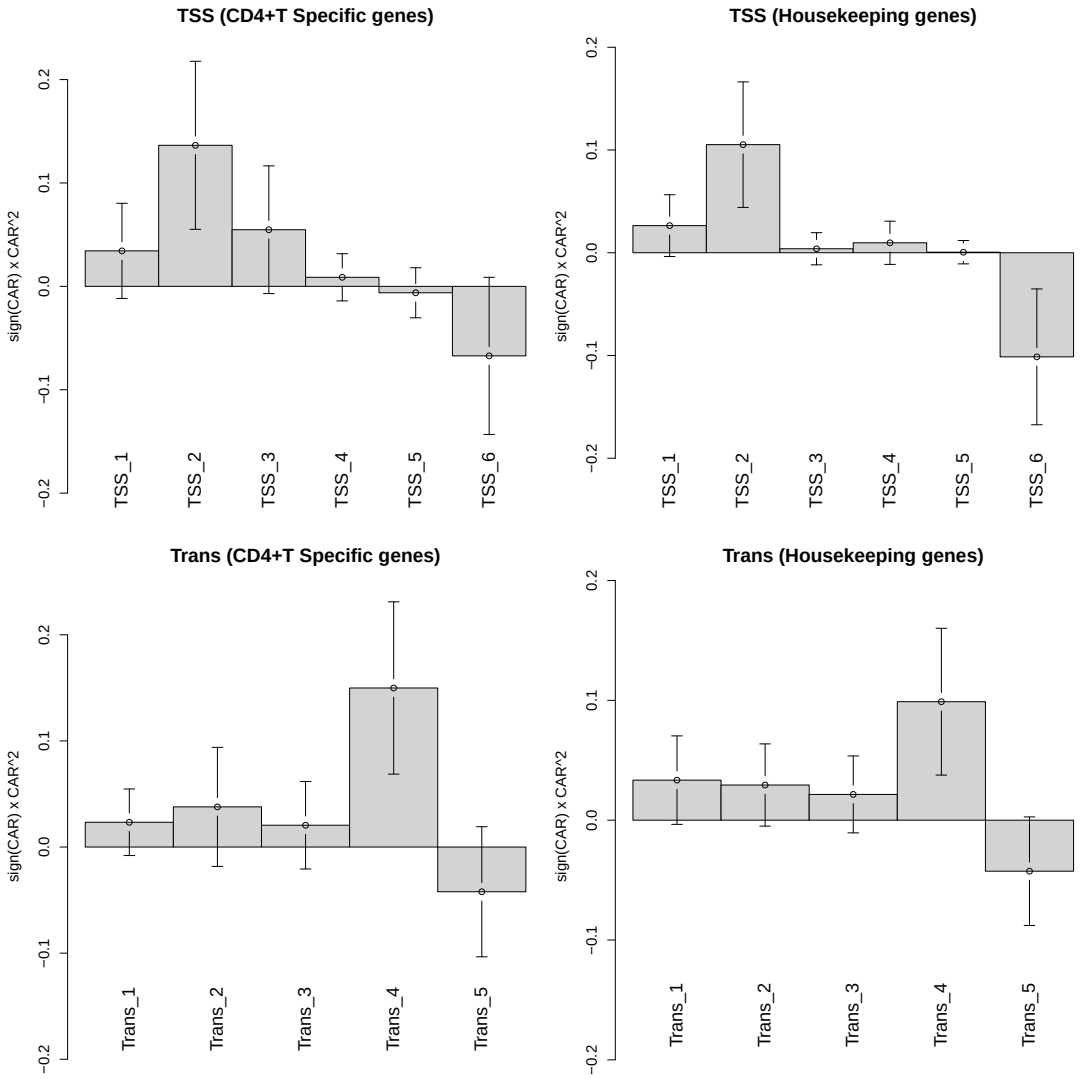

Figure 4.9: Car scores calculated in housekeeping and CD4+T specific gene datasets. PC_{Pr}^{TSS3} shows large CAR score in CD4+T specific genes only, suggesting that they are associated with tissue-specific genes. For each CAR score, a 95% confidence interval is produced by bootstrapping 1,000 times the cell-specific and housekeeping genes datasets.

characterised as tissue-specific in Pekowska et al. (2010) and Zhang and Zhang (2011). Furthermore, consistently with the location of the H3K4me2 peak found downstream of TSSs (Pekowska et al., 2010), this PC is classified as TSS-specific.

4.4.6 Spatial Distribution

The projection of the TSS and transcript datasets onto their principal axes have allowed us to distinguish region-specific PCs from those that are common to both regions. To confirm these results, we now explore the spatial effect of the PCs on gene expression and see whether they are consistent with the classification previously made. Previous studies have also investigated the spatial distribution of histone modifications and their predictive power along genes (Cheng and Gerstein, 2012; Cheng et al., 2011). They found that (i) the patterns of correlation between the histone marks and gene expression level mirror directly their ChIP-seq profiles, (ii) all together, the histone modifications are highly predictive of gene expression from the TSS to the TES, with a maximum in the ~ 2 kb region downstream of the TSS. Although very important, this second observation does not specify the relative importance of the histone marks and their variation along genes.

This question is addressed figures 4.10 where the CAR scores of the 6 PC^{TSS} (top left) and 5 PC^{Trans} (top right) are shown in a 60 kb window centered on TSSs (only genes longer than 30kb and whose 5' end is at least 30kb away from the previous gene 3' end are selected). Consistently with the previous results, the promoter-specific PC_{Ti}^{TSS3},PC_{Pr}^{Trans2}, PC_{Pr}^{TSS1} are correlated with gene expression in

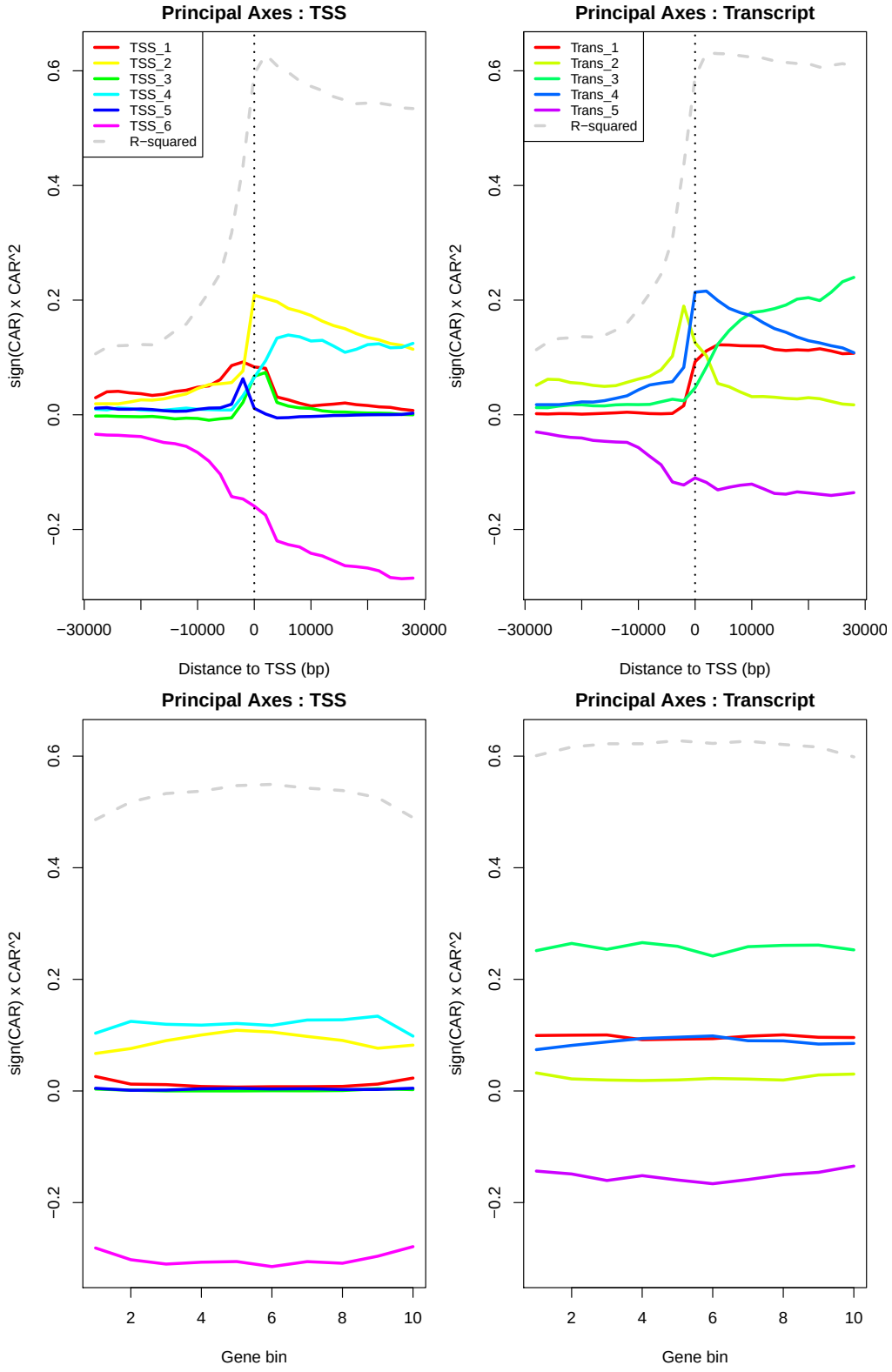

Figure 4.10: CAR scores associated with the PAs derived from the TSS (left) and transcript (right) datasets across TSSs (top) and along genes (bottom).

the region overlapping the TSS, whereas the transcript-specific PC_{Tr}^{Trans3} shows small and large CAR scores upstream and downstream of the TSS respectively. Only the transcript-specific PC_{Tr}^{TSS5} shows a correlation pattern inconsistent with its annotation : instead of having a non-null association with gene expression in transcribed region only, it peaks just upstream of the TSS, questioning thus its classification. However, if we go back to the annotations table 4.2, we can note that the Manhattan distances between (i) PC_{Tr}^{TSS5} and (ii) promoter and gene body annotation in Barski & Wang are very close (9 and 8 respectively). This point suggests that the original classification of this PC was weakly supported by Barski and Wang's annotations.

These two figures are otherwise useful for precisely delineating the regions where the PCs are associated with gene expression. In this way, all classified promoter-specific, PC_{Pr}^{TSS1}, PC_{Ti}^{TSS3} and PC_{Pr}^{Trans2} display distinct spatial distributions, where for instance, the tissue-specific PC_{Ti}^{TSS3} peaks downstream of the TSS, in agreement with Pekowska et al. observations. Also, it can be noted that the distinction previously made between transcribed and common (TSS and transcribed) regions is somehow artificial since both categories follow a similar pattern of correlation across the TSS : weak CAR scores upstream of the TSS, followed by higher values downstream of it. The only difference arises from the location at which the transition between these two modes occurs : upstream (PC_{Co}^{TSS2}) and across (PC_{Co}^{TSS4}) the TSS for PCs common to both regions, or downstream of the TSS for the transcript-specific PC_{Tr}^{Trans3}.

Lastly, the CAR scores of PC_{Co}^{TSS2}, PC_{Si}^{Trans5} and PC_{Tr}^{Trans3} show steady decrease

(PC_{Co}^{TSS2}, PC_{Si}^{Trans5}) and increase (PC_{Tr}^{Trans3}) in transcribed regions, raising the question of whether extremums are reached in a particular region of genes. To address this question, the Ensembl genes are split into 10 bins of equal length and the CAR score is calculated in each of them. Small variations of score can be observed throughout the transcribed regions (figures 4.10, bottom row), suggesting that PC_{Co}^{TSS2}, PC_{Si}^{Trans5} and PC_{Tr}^{Trans3} have in fact fairly constant values in transcribed regions.

4.5 Combination of TSS and Transcript Signals

As shown in the previous section, both the TSS model and the transcript model are highly predictive of gene expression. The question that can then be asked is whether the combination of the signals associated with both dataset can further improve the predictive power ? To answer this question, we compare the R^2 obtained with (i) the TSS model, (ii) the transcript model and (iii) a linear model using the eleven PCs derived from the TSS and transcript sets as predictors. In comparison with the models performances calculated in the original datasets ($R^2_{TS} = 0.595 \pm 0.019$, $R^2_{Tr} = 0.650 \pm 0.015$), an non significant improvement ($R^2_{both} = 0.674 \pm 0.014$) is obtained by combining them, suggesting that the signals contained in the TSS and transcript datasets are generally redundant for gene expression. This result does not mean that TSS and transcript specific signals are redundant, but since they are present in both datasets, the combination of these datasets does not lead to any substantial improvement in gene expression prediction.

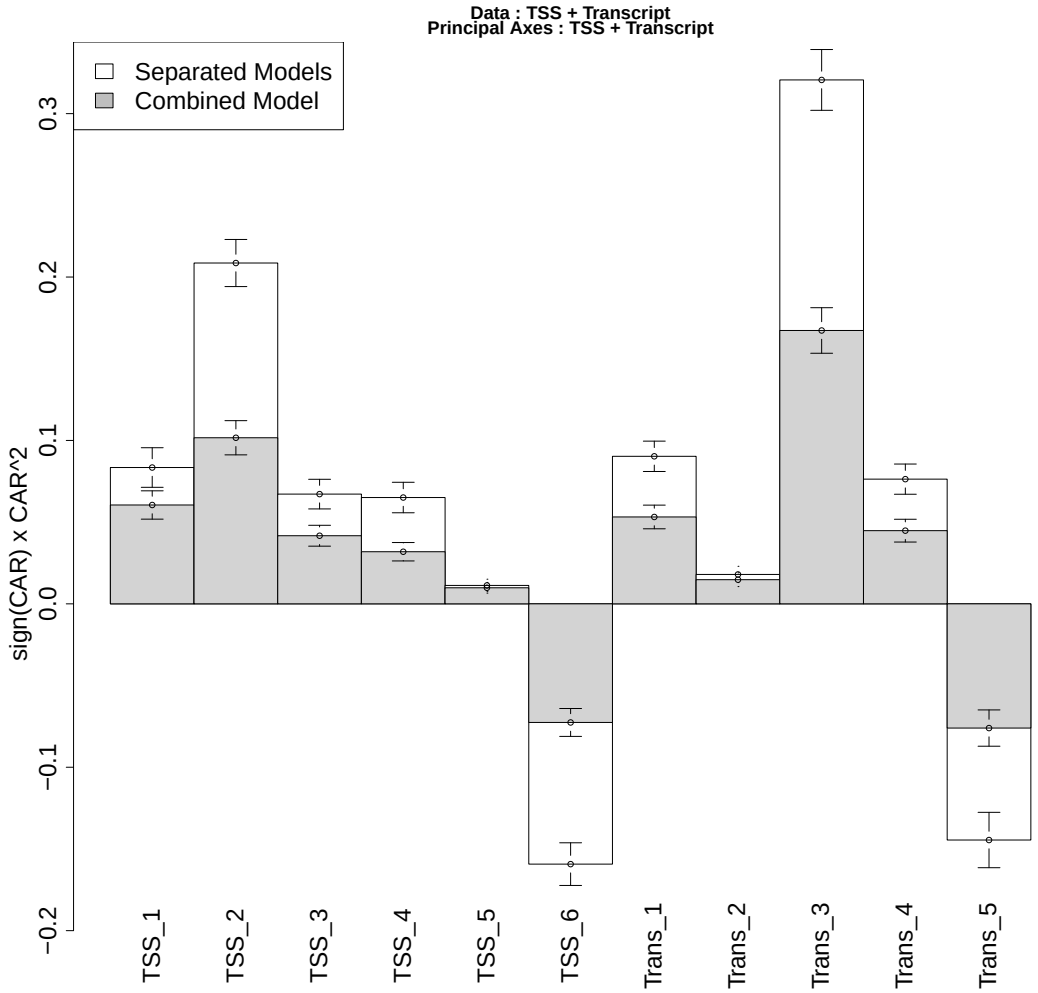

Figure 4.11: PCs CAR scores calculated either in the TSS and transcript datasets (white) or the combined (TSS + transcript) dataset (gray). Except PC_{Tr}^{TSS5} and PC_{Pr}^{Trans2}, all the PCs show important variations in CAR scores between their original dataset and the combined dataset. For each CAR score, 95% confidence intervals are produced from 1,000 bootstrap samples.

We then wonder whether this "overlap" of predictive information results in a uniform decrease of all PCs' CAR scores or whether some of them are impacted more than others. The first case would suggest that the signal captured by each PC is also detected by another PC in the other set and that, consequently, none of them are dataset specific. The second would on the other hand suggest that some PCs are highly redundant while others are dataset specific. In figure 4.11, we can see that apart from PC_{Tr}^{TSS5} and PC_{Pr}^{Trans2} whose values are close to zero, the combination of the TSS and transcript datasets results in a significant (the confidence intervals do not overlap) decrease in CAR score for all the PCs, supporting thus the first hypothesis.

4.6 Effect of the Nucleosome Position on Transcription

Nucleosomes located in the vicinity of TSSs have been shown to take different conformations, depending on whether they are associated with active or inactive promoters. In human, Valouev et al. (2011) found that the nucleosome depletion region (NDR) upstream of the TSS and the level of phasing emanating from this NDR are related to the binding of Pol II and to some extent to gene expression. However, this relationship between nucleosome positioning and gene expression is not systematic since a significant fraction of genes are associated with poised or stalled Pol II at their promoter (Schones et al., 2008). Typically, the promoters associated with these genes contain a NDR and exhibit very low levels of expression.

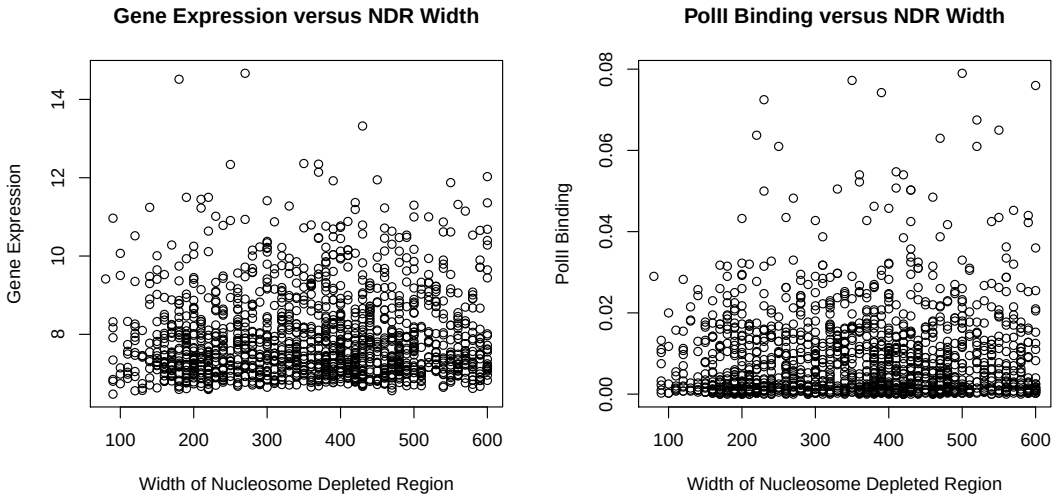

Figure 4.12: Scatter plots of gene expression and PolII binding versus NDR width : no relationship can be observed.

To investigate this relationship between the level of Pol II binding or gene expression and the width of the NDR, we make use of scatter plots (figure 4.12) and linear models (table 4.4). Consistent with Zaugg and Luscombe (2012) definition, the width of the NDR is defined as the distance between the nucleosomes immediately upstream and downstream of the TSS. The coefficients estimated in both models (using gene expression or PolII binding as response variable) are very small and not significant, as indicated by the large p-values table 4.4. In the same way, the examination of figure 4.12 suggests that there is no link between the amount of gene expression and PolII binding and the NDR width. These results suggest that either the effect is too weak to be detected or, in line with Zaugg and Luscombe (2012) observations in *Saccharomyces cerevisiae*, that there is no such association between NDR width and transcription level. Indeed, Zaugg

and Luscombe (2012) observed that nucleosome positions follow discrete configurations at promoters that determine the active and silent transcriptional states of a gene, but not its transcriptional level.

Response	Gene Expression			Pol II Binding		
	Estimate	Std. Error	Pr(> |t|)	Estimate	Std. Error	Pr(> |t|)
Intercept	7.868e+00	5.464e-02	¡2e-16	8.331e-03	5.263e-04	¡2e-16
NDR width	1.449e-04	8.552e-05	0.0902	1.503e-06	8.238e-07	0.0681

Table 4.4: The effects of the NDR width on gene expression and Pol II binding is estimated by fitting two linear models. The effect is not significant in both models, as shown by the large p-values associated with their coefficients.

4.7 Discussion

In this chapter, we have aimed to bridge two related areas that are concerned with (i) the characterisation of histone marks specific to distinct regions and (ii) the effects of these cluster of histone modifications on transcription. The approach followed was to first identify clusters of histone modifications in promoter and transcribed regions and then characterise their functions from (i) previous annotations, (ii) their association with gene expression, the variation of these associations (iii) between cell-specific and housekeeping genes and (iv) spatially across the TSS. Thanks to sparse PCA the clusters are made of a small number of histone marks (five to seven), which will make their identification in future studies easier. Table 4.5 summarizes the classification of the eleven clusters of histone marks.

Promoter		Transcript	Common	Silencing
Tissue Specific	-			
PC_{Ti}^{TSS3}	PC_{Pr}^{TSS1}	PC_{Tr}^{Trans3}	PC_{Co}^{TSS2}	PC_{Si}^{TSS6}
	PC_{Pr}^{Trans2}	PC_{Tr}^{TSS5}	PC_{Co}^{TSS4}	PC_{Si}^{Trans5}
			PC_{Co}^{Trans1}	
			PC_{Co}^{Trans4}	

Table 4.5: Summary of PCs categories.

From these analyses, we found that :

- Among eleven clusters identified, three are promoter specific, two are transcript specific, four are common to both regions and two are associated with gene silencing.

- Among the three promoter specific clusters, one (PC_{Ti}^{TSS3}) is associated with cell-specific processes. This result is supported both by (i) the association of PC_{Ti}^{TSS3} with gene expression level in cell-specific genes but not in housekeeping genes, and (ii) the fact that PC_{Ti}^{TSS3} is made up of histone marks characterised as tissue-specific in previous studies.

- TSS and transcript specific clusters are found in both datasets (TSS and transcribed regions). Thus, the combination of both datasets does not lead to an increase in gene expression prediction. Figure 4.10, showing the spatial distribution of CAR scores across TSSs, provides an insight into this somewhat surprising result : the CAR scores curves, reflecting the enrichments in histone marks, show that (i) promoter specific clusters extend as far as 5 kb on each side of the TSS and that (ii) transcript specific clusters start from the TSS. Consequently, there is a $\sim$ 5 kb region where both promoter and transcript specific clusters are present. Therefore, the way to capture promoter and transcript specific clusters independently would be to remove this $\sim$ 5 kb region from both datasets.

- Clusters specific of transcribed and common (TSS and transcribed) regions have similar pattern of correlation with transcription level, the only difference being the location at which the transition between weak and higher value of correlation occur.

- Lastly, unlike what was reported in Valouev et al. (2011), we have not found any relationship between the level of gene expression and the width of the nucleosome depleted region. As pointed out earlier, this lack of association is in line with a recent study from Zaugg and Luscombe (2012) who showed

that nucleosome positions determine the transcriptional states of a gene, but not its transcriptional level.

Chapter 5

Discussion

Epigenetics is the study of the mechanisms that control gene expression independently of changes in the DNA sequence. In multicellular eukaryotic, these mechanisms occur principally during the development stage where totipotent stem cells are turned into fully differentiated cells by modulating the access of DNA to DNA-binding proteins through chromatin structure modifications. These changes are preserved during cell division throughout one individual's life. The epigenetic information is present at three different levels of the chromatin, DNA, nucleosome and higher levels where the chromatin itself interacts with other components in the nucleus. In this work we only focused on the nucleosome level, where we (i) sought to determine the relative influence of *cis* and *trans*-factors in the nucleosome organisation in human, and (ii) investigated the relationship between histone post-translational modifications and transcription regulation.

In the introduction chapter, we presented how *cis* and *trans* factors act on nucleosome positioning and how the choice of summary statistics led to different conclusions regarding their relative influence *in vivo*. In Zhang et al. (2009, 2010), the authors discussed the use of nucleosome occupancy and nucleosome

positioning and showed that, unlike the latter, the former fails to capture the key biological feature of the nucleosome organisation, i.e. the precise knowledge of nucleosome positions. Since both of these statistics are sensitive to technical noise and that none of the nucleosome calling methods account for control sample, we introduced in chapter 2 a statistical approach able to detect nucleosomes that show (i) a high level of positioning and (ii) an enrichment relative to a control sample. The comparison of NucleoFinder with two other nucleosome calling methods showed that our approach is highly specific of well-positioned nucleosome and have equal or better ability to identify (i) the regular nucleosome spacing downstream of active gene TSSs and (ii) the stereotypical A/T, G/C dinucleotide patterns at well-positioned nucleosomes *in vitro*.

A large number of studies have been conducted in yeast to measure the relative influence of *cis* and *trans*-factors, from which a consensus has emerged on the fact that $\sim$ 20 % of the nucleosomes are intrinsically encoded in yeast. In human, on the other hand, fewer studies have been done and no such estimate have been obtained apart from this in Tillo et al. (2010), based on nucleosome occupancy. To fill this gap, we made use of the recent mapping of nucleosomes *in vivo* and *in vitro* in human by Valouev et al. Using a PWM based on 6-mers, we showed that nucleosome specific motifs are highly predictive of nucleosomes *in vitro*, but because fairly infrequent, the majority of the *in vitro* nucleosomes cannot be predicted accurately from the DNA sequence (ROC-score = 0.67). We further showed that the presence of these particular motifs is linked (i) to the Bayes factor with which nucleosomes are detected *in vitro* and (ii) to the nucleosome enrichment *in vivo*. Because the mechanisms responsible for nucleosome

positioning may not be fully captured with a PWM, we also directly compared *in vitro* with *in vivo* nucleosomes. We estimated that *cis*-factors account for $\sim$ 10-15% of the *in vivo* nucleosome organisation. This influence was larger at intergenic regions and unexpressed promoters than at expressed promoters, in agreement with the idea that transcription factors and chromatin remodelers interact more often with the latter than the formers.

In chapter 4, we aimed to (i) identify histone marks that cluster in different functional regions and (ii) measure their association with gene expression level. Thanks to previous annotations, the sign and the spatial distribution of the association between these clusters and gene expression, we found that among the eleven clusters, three are promoter specific, two are transcript specific, four are common to promoter and transcribed regions and two are associated with gene silencing. Prior to this analysis, we expected to find different groups of histone modifications across TSSs and in transcribed regions that, when combined together, would lead to an improvement in gene expression prediction. Instead, we found that both datasets contained TSS and transcript specific signals and that their combination did not lead to any significant improvement in gene expression prediction. Otherwise, based on CD4+T specific and housekeeping genes selected in Zhang and Zhang (2011), we identified one cluster of histone marks involve in cell-type processes. Finally, unlike expected, we did not see any associations between Pol II binding and (i) the width of nucleosome depletion upstream of TSSs and (ii) the level of phasing downstream of TSSs.

We see future development of the work in this thesis in two aspects. On sev-

eral occasions throughout this work, we argued that the identification of well-positioned nucleosomes is fundamental in the understanding of the nucleosome organisation because they indicate regions consistently occluded or accessible to *trans*-factors. In order to identify the determinant of the nucleosome architecture in chapter 3, we used NucleoFinder predictions to estimate the influence of *cis*-factors in the nucleosome organisation. Another interesting direction would be to investigate whether well-positioned nucleosomes away from genic regions, especially those driven by *cis*-factors, contain functional elements.

Otherwise, in chapter 4 we identified patterns of histone marks present at TSS and transcribed regions and measured their association with transcription level using data from a single cell-line of a single individual. To improve our understanding of cell-specific processes and get a sense of the inter-tissue / inter-individual variations, it would be of great interest to perform a similar association analysis across tissues and individuals on a gene basis.

References

Vikram Alva, Moritz Ammelburg, Johannes Soding, and Andrei N Lupas. On the origin of the histone fold. *BMC Structural Biology*, 7:17, 2007. ISSN 1472-6807. doi: 10.1186/1472-6807-7-17. URL http://www.ncbi.nlm.nih.gov/pubmed/17391511. PMID: 17391511. 7

Dimitar Angelov, Francois Lenouvel, Fabienne Hans, Christoph W. Muller, Philippe Bouvet, Jan Bednar, Evangelos N. Moudrianakis, Jean Cadet, and Stefan Dimitrov. The histone octamer is "invisible" when NF-kB binds to the nucleosome. *The Journal of Biological Chemistry*, page M407235200, July 2004. doi: ⟨p⟩10.1074/jbc.M407235200⟨/p⟩. URL http://www.jbc.org/cgi/content/abstract/M407235200v1. 25

C Anselmi, G Bocchinfuso, P De Santis, M Savino, and A Scipioni. Dual role of DNA intrinsic curvature and flexibility in determining nucleosome stability. *Journal of Molecular Biology*, 286(5):1293–1301, March 1999. ISSN 0022-2836. doi: 06/jmbi.1998.2575. URL http://www.sciencedirect.com/science/article/pii/S002228369892575X. 23

Lu Bai and Alexandre V. Morozov. Gene regulation by nucleosome positioning. *Trends in Genetics*, 26(11):476–483, November 2010. ISSN 0168-9525. doi:

10.1016/j.tig.2010.08.003. URL http://www.sciencedirect.com/science/article/B6TCY-510101X-2/2/ea8c5b33d5dbbfb97911cb72610bb2ea. 6, 7

Pierre Baldi and Pierre-Franois Baisnee. Sequence analysis by additive scales: DNA structure for sequences and repeats of all lengths. *Bioinformatics*, 16(10): 865 –889, October 2000. doi: 10.1093/bioinformatics/16.10.865. URL http://bioinformatics.oxfordjournals.org/content/16/10/865.abstract. 78

Slobodan Barbaric, Tim Luckenbach, Andrea Schmid, Dorothea Blaschke, Wolfram Horz, and Philipp Korber. Redundancy of chromatin remodeling pathways for the induction of the yeast PHO5 promoter in vivo. *Journal of Biological Chemistry*, 282(38):27610 –27621, 2007. doi: 10.1074/jbc.M700623200. URL http://www.jbc.org/content/282/38/27610.abstract. 26

Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129 (4):823–837, May 2007. ISSN 0092-8674. doi: 10.1016/j.cell.2007.05.009. URL http://www.ncbi.nlm.nih.gov/pubmed/17512414. PMID: 17512414. 106, 109, 110, 113, 116, 117, 137, 138, 139

J. Berger. The case for objective bayesian analysis. *Bayesian Analysis*, 1(3): 385402, 2006. 44

Bradley E Bernstein, John A Stamatoyannopoulos, Joseph F Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, Marco A Marra, Arthur L Beaudet, Joseph R Ecker, Peggy J Farnham, Martin Hirst, Eric S Lander, Tarjei S Mikkelsen, and James A Thomson. The NIH roadmap

epigenomics mapping consortium. *Nature Biotechnology*, 28(10):1045–1048, October 2010. ISSN 1546-1696. doi: 10.1038/nbt1010-1045. URL `http://www.ncbi.nlm.nih.gov/pubmed/20944595`. PMID: 20944595. 11

Hinrich Boeger, Joachim Griesenbeck, J Seth Strattan, and Roger D Kornberg. Removal of promoter nucleosomes by disassembly rather than sliding in vivo. *Molecular Cell*, 14(5):667–673, June 2004. ISSN 1097-2765. doi: 10.1016/j.molcel.2004.05.013. URL `http://www.ncbi.nlm.nih.gov/pubmed/15175161`. PMID: 15175161. 25

B M Bolstad, R A Irizarry, M Astrand, and T P Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, January 2003. ISSN 1367-4803. URL `http://www.ncbi.nlm.nih.gov/pubmed/12538238`. PMID: 12538238. 117

Roberto Bonasio, Shengjiang Tu, and Danny Reinberg. Molecular signals of epigenetic states. *Science*, 330(6004):612–616, October 2010. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1191078. URL `http://www.sciencemag.org/content/330/6004/612`. 2

Deborah Bourchis and Olivier Voinnet. A small-RNA perspective on gametogenesis, fertilization, and early zygotic development. *Science*, 330(6004):617–622, October 2010. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1194776. URL `http://www.sciencemag.org/content/330/6004/617`. 2

Xin Chen, Lingqiong Guo, Zhaocheng Fan, and Tao Jiang. Learning position weight matrices from sequence and expression data. *Computational Systems Bioinformatics / Life Sciences Society. Computational Systems Bioinformatics*

Conference, 6:249–260, 2007. ISSN 1752-7791. URL http://www.ncbi.nlm.nih.gov/pubmed/17951829. PMID: 17951829. 79

Chao Cheng and Mark Gerstein. Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Research*, 40(2):553–568, January 2012. ISSN 1362-4962. doi: 10.1093/nar/gkr752. URL http://www.ncbi.nlm.nih.gov/pubmed/21926158. PMID: 21926158. 111, 150

Chao Cheng, Koon-Kiu Yan, Kevin Y Yip, Joel Rozowsky, Roger Alexander, Chong Shou, and Mark Gerstein. A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biology*, 12(2):R15, 2011. ISSN 1465-6906. doi: 10.1186/gb-2011-12-2-r15. URL http://genomebiology.com/2011/12/2/R15/additional. 111, 150

Ping Chi, C David Allis, and Gang Greg Wang. Covalent histone modifications–miswritten, misinterpreted and mis-erased in human cancers. *Nature Reviews Cancer*, 10(7):457–469, July 2010. ISSN 1474-1768. doi: 10.1038/nrc2876. URL http://www.ncbi.nlm.nih.gov/pubmed/20574448. PMID: 20574448. 10

Jung Kyoon Choi and Young-Joon Kim. Epigenetic regulation and the variability of gene expression. *Nature Genetic*, 40(2):141–147, February 2008. ISSN 1061-4036. doi: 10.1038/ng.2007.58. URL http://dx.doi.org/10.1038/ng.2007.58. 25

Curt A. Davey, David F. Sargent, Karolin Luger, Armin W. Maeder, and Timothy J. Richmond. Solvent mediated interactions in the structure of the nucleo-

some core particle at 1.9 resolution. *Journal of Molecular Biology*, 319(5):1097–1113, 2002. ISSN 0022-2836. doi: 10.1016/S0022-2836(02)00386-8. URL http://www.sciencedirect.com/science/article/pii/S0022283602003868. 21

Patrik D'haeseleer. What are DNA sequence motifs? *Nature Biotechnology*, 24 (4):423–425, April 2006. ISSN 1087-0156. doi: 10.1038/nbt0406-423. URL http://www.ncbi.nlm.nih.gov/pubmed/16601727. PMID: 16601727. 79

Vito Di Ges, Giosue Lo Bosco, Luca Pinello, Guo-Cheng Yuan, and Davide F V Corona. A multi-layer method to study genome-scale positions of nucleosomes. *Genomics*, 93(2):140–145, February 2009. ISSN 1089-8646. doi: 10.1016/j.ygeno.2008.09.012. URL http://www.ncbi.nlm.nih.gov/pubmed/18951969. PMID: 18951969. 18

Birney E, Stamatoyannopoulos JA, Dutta A, Guig R, Gingeras TR, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447, 447(7146, 7146):799, 799, June 2007. ISSN 0028-0836. doi: 10.1038/nature05874. URL http://ukpmc.ac.uk/articles/2212820. 11

Florian Eckhardt, Stephan Beck, Ivo G Gut, and Kurt Berlin. Future potential of the human epigenome project. *Expert Review of Molecular Diagnostics*, 4 (5):609–618, September 2004. ISSN 1473-7159. doi: 10.1586/14737159.4.5.609. URL http://www.ncbi.nlm.nih.gov/pubmed/15347255. PMID: 15347255. 11

Sarah CR Elgin. Heterochromatin and gene regulation in Drosophila. *Current Opinion in Genetics and Development*, 6(2):193–202, 1996. ISSN 0959-

437X. doi: 16/S0959-437X(96)80050-5. URL http://www.sciencedirect.com/science/article/pii/S0959437X96800505. 4

Jason Ernst and Manolis Kellis. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, 28(8):817–825, 2010. ISSN 1087-0156. doi: 10.1038/nbt.1662. URL http://dx.doi.org/10.1038/nbt.1662. 11, 112, 114, 137, 138, 139

Manel Esteller. Epigenetic gene silencing in cancer: the DNA hypermethylome. *Human Molecular Genetics*, 16 Spec No 1:R50–59, April 2007. ISSN 0964-6906. doi: 10.1093/hmg/ddm018. URL http://www.ncbi.nlm.nih.gov/pubmed/17613547. PMID: 17613547. 10

T. Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8): 861874, 2006. 24

Thomas G Fazzio and Toshio Tsukiyama. Chromatin remodeling in vivo: evidence for a nucleosome sliding mechanism. *Molecular Cell*, 12(5):1333–1340, November 2003. ISSN 1097-2765. URL http://www.ncbi.nlm.nih.gov/pubmed/14636590. PMID: 14636590. 25

Gary Felsenfeld and Mark Groudine. Controlling the double helix. *Nature*, 421 (6921):448–453, January 2003. ISSN 0028-0836. doi: 10.1038/nature01411. URL http://dx.doi.org/10.1038/nature01411. 4

Yair Field, Noam Kaplan, Yvonne Fondufe-Mittendorf, Irene K. Moore, Eilon Sharon, Yaniv Lubling, Jonathan Widom, and Eran Segal. Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS*

Computational Biology, 4(11):e1000216, November 2008. doi: 10.1371/journal.pcbi.1000216. URL http://dx.doi.org/10.1371/journal.pcbi.1000216. 21, 23, 28, 78

J T Flick, J C Eissenberg, and S C Elgin. Micrococcal nuclease as a DNA structural probe: its recognition sequences, their genomic distribution and correlation with DNA structure determinants. *Journal of Molecular Biology*, 190 (4):619–633, August 1986. ISSN 0022-2836. URL http://www.ncbi.nlm.nih.gov/pubmed/3097328. PMID: 3097328. 14

Oscar Flores and Modesto Orozco. nucleR : a package for non-parametric nucleosome positioning. *Bioinformatics*, 27(15):2149–2150, August 2011. ISSN 1367-4811. doi: 10.1093/bioinformatics/btr345. URL http://www.ncbi.nlm.nih.gov/pubmed/21653521. PMID: 21653521. 18

Stephen M Fuchs, R Nicholas Laribee, and Brian D Strahl. Protein modifications in transcription elongation. *Biochimica Et Biophysica Acta*, 1789(1):26–36, January 2009. ISSN 0006-3002. doi: 10.1016/j.bbagrm.2008.07.008. URL http://www.ncbi.nlm.nih.gov/pubmed/18718879. PMID: 18718879. 8

Andrew Gelman, Xiao-li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *STATISTICA SINICA*, pages 733—807, 1996. URL http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.142.9951. 50

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2004. ISBN 9781584883883. 46, 49

E R Gibney and C M Nolan. Epigenetics and gene expression. *Heredity*, 105 (1):4–13, May 2010. ISSN 0018-067X. doi: 10.1038/hdy.2010.54. URL `http://www.nature.com/hdy/journal/v105/n1/abs/hdy201054a.html`. 2

Mary Grace Goll and Timothy H Bestor. Eukaryotic cytosine methyltransferases. *Annual review of biochemistry*, 74:481–514, 2005. ISSN 0066-4154. doi: 10.1146/annurev.biochem.74.010904.153721. PMID: 15952895. 3

Leo A. Goodman. On simultaneous confidence intervals for multinomial proportions. *Technometrics*, 7(2):247–254, May 1965. ISSN 0040-1706. doi: 10.2307/1266673. URL `http://www.jstor.org/stable/1266673`. ArticleType: research-article / Full publication date: May, 1965 / Copyright 1965 American Statistical Association and American Society for Quality. 60

David S. Goodsell and Richard E. Dickerson. Bending and curvature calculations in b-DNA. *Nucleic Acids Research*, 22(24):5497 –5503, January 1994. doi: 10.1093/nar/22.24.5497. URL `http://nar.oxfordjournals.org/content/22/24/5497.short`. 78

U. Gromping. Relative importance for linear regression in r: the package relaimpo. *Journal of Statistical Software*, 17(1):127, 2006. 144

Lars Guelen, Ludo Pagie, Emilie Brasset, Wouter Meuleman, Marius B Faza, Wendy Talhout, Bert H Eussen, Annelies de Klein, Lodewyk Wessels, Wouter de Laat, and Bas van Steensel. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, 453(7197):948–951, June 2008. ISSN 1476-4687. doi: 10.1038/nature06947. URL `http://www.ncbi.nlm.nih.gov/pubmed/18463634`. PMID: 18463634. 4

Shobhit Gupta, Jonathan Dennis, Robert E. Thurman, Robert Kingston, John A. Stamatoyannopoulos, and William Stafford Noble. Predicting human nucleosome occupancy from primary sequence. *PLoS Computational Biology*, 4(8): e1000134, 2008. doi: 10.1371/journal.pcbi.1000134. URL `http://dx.doi.org/10.1371/journal.pcbi.1000134`. 23, 34, 76, 87, 89, 90

Olivier Harismendy, Pauline C Ng, Robert L Strausberg, Xiaoyun Wang, Timothy B Stockwell, Karen Y Beeson, Nicholas J Schork, Sarah S Murray, Eric J Topol, Samuel Levy, and Kelly A Frazer. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology*, 10(3):R32, 2009. ISSN 1465-6914. doi: 10.1186/gb-2009-10-3-r32. URL `http://www.ncbi.nlm.nih.gov/pubmed/19327155`. PMID: 19327155. 18, 41

T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, corrected edition, July 2003. ISBN 0387952845. 129

Nathaniel D Heintzman, Rhona K Stuart, Gary Hon, Yutao Fu, Christina W Ching, R David Hawkins, Leah O Barrera, Sara Van Calcar, Chunxu Qu, Keith A Ching, Wei Wang, Zhiping Weng, Roland D Green, Gregory E Crawford, and Bing Ren. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, 39(3):311–318, March 2007. ISSN 1061-4036. doi: 10.1038/ng1966. URL `http://dx.doi.org/10.1038/ng1966`. 112

Nathaniel D. Heintzman, Gary C. Hon, R. David Hawkins, Pouya Kheradpour, Alexander Stark, Lindsey F. Harp, Zhen Ye, Leonard K. Lee, Rhona K. Stuart, Christina W. Ching, Keith A. Ching, Jessica E. Antosiewicz-Bourget, Hui

Liu, Xinmin Zhang, Roland D. Green, Victor V. Lobanenkov, Ron Stewart, James A. Thomson, Gregory E. Crawford, Manolis Kellis, and Bing Ren. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–112, 2009. ISSN 0028-0836. doi: 10.1038/nature07829. URL `http://dx.doi.org/10.1038/nature07829`. 109

Stephen A Hoang, Xiaojiang Xu, and Stefan Bekiranov. Quantification of histone modification ChIP-seq enrichment for data mining and machine learning applications. *BMC Research Notes*, 4:288, 2011. ISSN 1756-0500. doi: 10.1186/1756-0500-4-288. URL `http://www.ncbi.nlm.nih.gov/pubmed/21834981`. PMID: 21834981. 111, 118

Gary Hon, Wei Wang, and Bing Ren. Discovery and annotation of functional chromatin signatures in the human genome. *PLoS Computational Biology*, 5(11):e1000566, November 2009. doi: 10.1371/journal.pcbi.1000566. URL `UR-http://dx.doi.org/10.1371/journal.pcbi.1000566,http://dx.doi.org/10.1371/journal.pcbi.1000566`. 109, 113, 114

W Horz and W Altenburger. Sequence specific cleavage of DNA by micrococcal nuclease. *Nucleic Acids Research*, 9(12):2643–2658, June 1981. ISSN 0305-1048. PMID: 7279658 PMCID: 326882. 13

Ilya P Ioshikhes, Istvan Albert, Sara J Zanton, and B Franklin Pugh. Nucleosome positions predicted through comparative genomics. *Nature Genetics*, 38(10):1210–1215, October 2006. ISSN 1061-4036. doi: 10.1038/ng1878. URL `http://dx.doi.org/10.1038/ng1878`. 23

Yoko Ito, Thibaud Koessler, Ashraf E K Ibrahim, Sushma Rai, Sarah L Vowler,

Sayeda Abu-Amero, Ana-Luisa Silva, Ana-Teresa Maia, Joanna E Huddleston, Santiago Uribe-Lewis, Kathryn Woodfine, Maja Jagodic, Raffaella Nativio, Alison Dunning, Gudrun Moore, Elena Klenova, Sheila Bingham, Paul D P Pharoah, James D Brenton, Stephan Beck, Manjinder S Sandhu, and Adele Murrell. Somatically acquired hypomethylation of IGF2 in breast and colorectal cancer. *Human Molecular Genetics*, 17(17):2633–2643, September 2008. ISSN 1460-2083. doi: 10.1093/hmg/ddn163. URL `http://www.ncbi.nlm.nih.gov/pubmed/18541649`. PMID: 18541649. 10

V Jackson. Formaldehyde cross-linking for studying nucleosomal dynamics. *Methods*, 17(2):125–139, February 1999. ISSN 1046-2023. doi: 10.1006/meth.1998.0724. URL `http://www.ncbi.nlm.nih.gov/pubmed/10075891`. PMID: 10075891. 13

Rami Jaschek and Amos Tanay. Spatial clustering of multivariate genomic and epigenomic information. In Serafim Batzoglou, editor, *Research in Computational Molecular Biology*, volume 5541 of *Lecture Notes in Computer Science*, pages 170–183. Springer Berlin / Heidelberg, 2009. ISBN 978-3-642-02007-0. URL `http://www.springerlink.com/content/m72w427572851271/abstract/`. 113, 114

T Jenuwein and C D Allis. Translating the histone code. *Science*, 293(5532):1074–1080, August 2001. ISSN 0036-8075. doi: 10.1126/science.1063127. URL `http://www.ncbi.nlm.nih.gov/pubmed/11498575`. PMID: 11498575. 9, 107

Hongkai Ji, Hui Jiang, Wenxiu Ma, David S Johnson, Richard M Myers, and Wing H Wong. An integrated software system for analyzing ChIP-chip and

ChIP-seq data. *Nature Biotechnology*, 26(11):1293–1300, November 2008. ISSN 1087-0156. doi: 10.1038/nbt.1505. URL http://dx.doi.org/10.1038/nbt.1505. 42

Cizhong Jiang and B. Franklin Pugh. Nucleosome positioning and gene regulation: advances through genomics. *Nature Review Genetics*, 10(3):161–172, March 2009. ISSN 1471-0056. doi: 10.1038/nrg2522. URL http://dx.doi.org/10.1038/nrg2522. 7, 15, 53, 62

Iain M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, April 2001. ISSN 0090-5364. doi: 10.1214/aos/1009210544. URL http://projecteuclid.org/euclid.aos/1009210544. Mathematical Reviews number (MathSciNet): MR1863961; Zentralblatt MATH identifier: 1016.62078. 83

Noam Kaplan, Irene K. Moore, Yvonne Fondufe-Mittendorf, Andrea J. Gossett, Desiree Tillo, Yair Field, Emily M. LeProust, Timothy R. Hughes, Jason D. Lieb, Jonathan Widom, and Eran Segal. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, 458(7236):362–366, March 2009. ISSN 0028-0836. doi: 10.1038/nature07667. URL http://dx.doi.org/10.1038/nature07667. 16, 28, 29, 32, 33, 36, 37, 90, 94

Noam Kaplan, Timothy R Hughes, Jason D Lieb, Jonathan Widom, and Eran Segal. Contribution of histone sequence preferences to nucleosome organization: proposed definitions and methodology. *Genome Biology*, 11(11):140, 2010. ISSN 1465-6906. doi: 10.1186/gb-2010-11-11-140. URL http://genomebiology.com/2010/11/11/140/abstract. 14, 16, 17, 18

Rosa Karlic, Ho-Ryun Chung, Julia Lasserre, Kristian Vlahoviek, and Martin Vingron. Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences*, 2010. doi: 10.1073/pnas.0909344107. URL `http://www.pnas.org/content/early/2010/01/28/0909344107.abstract`. 11, 111, 112, 118, 147

Tae Hoon Kim, Leah O. Barrera, Ming Zheng, Chunxu Qu, Michael A. Singer, Todd A. Richmond, Yingnian Wu, Roland D. Green, and Bing Ren. A high-resolution map of active promoters in the human genome. *Nature*, 436(7052): 876–880, 2005. ISSN 0028-0836. doi: 10.1038/nature03877. URL `http://dx.doi.org/10.1038/nature03877`. 108

Roger D. Kornberg and Lubert Stryer. Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. *Nucleic Acids Research*, 16 (14):6677 –6690, July 1988. doi: 10.1093/nar/16.14.6677. URL `http://nar.oxfordjournals.org/content/16/14/6677.abstract`. 26

Tony Kouzarides. Chromatin modifications and their function. *Cell*, 128(4):693–705, February 2007. ISSN 0092-8674. doi: 10.1016/j.cell.2007.02.005. URL `http://www.ncbi.nlm.nih.gov/pubmed/17320507`. PMID: 17320507. 8, 106, 107

Pei Fen Kuan, Dana Huebert, Audrey Gasch, and Sunduz Keles. A non-homogeneous hidden-state model on first order differences for automatic detection of nucleosome positions. *Statistical Applications in Genetics and Molecular Biology*, 8(1):Article29, 2009. ISSN 1544-6115. doi: 10.2202/1544-6115.1454.

URL http://www.ncbi.nlm.nih.gov/pubmed/19572828. PMID: 19572828. 19

William Lee, Desiree Tillo, Nicolas Bray, Randall H Morse, Ronald W Davis, Timothy R Hughes, and Corey Nislow. A high-resolution atlas of nucleosome occupancy in yeast. *Nature Genetics*, 39(10):1235–1244, October 2007. ISSN 1061-4036. doi: 10.1038/ng2117. URL http://dx.doi.org/10.1038/ng2117. 16, 19, 27, 54

Pierre Legendre and Louis Legendre. *Numerical Ecology*. Elsevier, November 1998. ISBN 9780444892508. 139

Chih Long Liu, Tommy Kaplan, Minkyu Kim, Stephen Buratowski, Stuart L Schreiber, Nir Friedman, and Oliver J Rando. Single-nucleosome mapping of histone modifications in S. cerevisiae. *PLoS Biology*, 3(10):e328, 2005. doi: 10.1371/journal.pbio.0030328. URL UR-http://dx.doi.org/10.1371/journal.pbio.0030328,http://dx.doi.org/10.1371/journal.pbio.0030328. 112

K Luger, A W Mader, R K Richmond, D F Sargent, and T J Richmond. Crystal structure of the nucleosome core particle at 2.8 a resolution. *Nature*, 389(6648):251–260, September 1997. ISSN 0028-0836. doi: 10.1038/38444. URL http://www.ncbi.nlm.nih.gov/pubmed/9305837. PMID: 9305837. 4, 22, 53

Travis N Mavrich, Ilya P Ioshikhes, Bryan J Venters, Cizhong Jiang, Lynn P Tomsho, Ji Qi, Stephan C Schuster, Istvan Albert, and B Franklin Pugh. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Research*, 18(7):1073–1083, July 2008a. ISSN

1088-9051. doi: 10.1101/gr.078261.108. URL http://www.ncbi.nlm.nih.gov/pubmed/18550805. PMID: 18550805. 30, 32, 56

Travis N Mavrich, Cizhong Jiang, Ilya P Ioshikhes, Xiaoyong Li, Bryan J Venters, Sara J Zanton, Lynn P Tomsho, Ji Qi, Robert L Glaser, Stephan C Schuster, David S Gilmour, Istvan Albert, and B Franklin Pugh. Nucleosome organization in the Drosophila genome. *Nature*, 453(7193):358–362, May 2008b. ISSN 1476-4687. doi: 10.1038/nature06929. URL http://www.ncbi.nlm.nih.gov/pubmed/18408708. PMID: 18408708. 23, 25, 32

Tarjei S. Mikkelsen, Manching Ku, David B. Jaffe, Biju Issac, Erez Lieberman, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–560, 2007. ISSN 0028-0836. doi: 10.1038/nature06008. URL http://dx.doi.org/10.1038/nature06008. 108

John R. S. Newman, Sina Ghaemmaghami, Jan Ihmels, David K. Breslow, Matthew Noble, Joseph L. DeRisi, and Jonathan S. Weissman. Single-cell proteomic analysis of s. cerevisiae reveals the architecture of biological noise. *Nature*, 441(7095):840–846, June 2006. ISSN 0028-0836. doi: 10.1038/nature04785. URL http://dx.doi.org/10.1038/nature04785. 25

Fatih Ozsolak, Jun S Song, X Shirley Liu, and David E Fisher. High-throughput mapping of the chromatin structure of human promoters. *Nature Biotechnology*, 25(2):244–248, 2007. ISSN 1087-0156. doi: 10.1038/nbt1279. URL http://dx.doi.org/10.1038/nbt1279. 16, 19, 63, 64

Vicki Pandey, Robert C. Nutter, and Ellen Prediger. Applied biosystems SOLiD system: Ligation-based sequencing. In Michal Janitz, editor, *Next Generation*

Genome Sequencing, page 2942. Wiley-VCH Verlag GmbH and Co. KGaA, 2008. ISBN 9783527625130. URL http://onlinelibrary.wiley.com/doi/10.1002/9783527625130.ch3/summary. 40

Timothy J Parnell, Jason T Huff, and Bradley R Cairns. RSC regulates nucleosome positioning at pol II genes and density at pol III genes. *EMBO Journal*, 27(1):100–110, January 2008. ISSN 0261-4189. doi: 10.1038/sj.emboj.7601946. URL http://dx.doi.org/10.1038/sj.emboj.7601946. 25

Heather E. Peckham, Robert E. Thurman, Yutao Fu, John A. Stamatoyannopoulos, William Stafford Noble, Kevin Struhl, and Zhiping Weng. Nucleosome positioning signals in genomic DNA. *Genome Research*, 17(8):000, July 2007. doi: 10.1101/gr.6101007. URL http://genome.cshlp.org/content/early/2007/07/01/gr.6101007.abstract. 23, 28, 31, 32, 76, 87

Aleksandra Pekowska, Touati Benoukraf, Pierre Ferrier, and Salvatore Spicuglia. A unique H3K4me2 profile marks tissue-specific gene regulation. *Genome Research*, 20(11):1493–1502, November 2010. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.109389.110. URL http://genome.cshlp.org/content/20/11/1493. 147, 148, 150, 152

Anna Portela and Manel Esteller. Epigenetic modifications and human disease. *Nature Biotechnology*, 28(10):1057–1068, October 2010. ISSN 1546-1696. doi: 10.1038/nbt.1685. URL http://www.ncbi.nlm.nih.gov/pubmed/20944598. PMID: 20944598. 9, 10

K. D. Pruitt. NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Re-*

search, 33(Database issue):D501–D504, December 2004. ISSN 1362-4962. doi: 10.1093/nar/gki025. URL http://nar.oxfordjournals.org/content/33/suppl_1/D501.full?ijkey=06NkMzN3kUcez&keytype=ref. 101

Marta Radman-Livaja and Oliver J Rando. Nucleosome positioning: how is it established, and why does it matter? *Developmental Biology*, 339(2):258–266, March 2010. ISSN 1095-564X. doi: 10.1016/j.ydbio.2009.06.012. URL http://www.ncbi.nlm.nih.gov/pubmed/19527704. PMID: 19527704. 24, 25, 53

Vladimir R. Ramirez-Carrozzi, Daniel Braas, Dev M. Bhatt, Christine S. Cheng, Christine Hong, Kevin R. Doty, Joshua C. Black, Alexander Hoffmann, Michael Carey, and Stephen T. Smale. A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. *Cell*, 138 (1):114–128, July 2009. ISSN 0092-8674. doi: 10.1016/j.cell.2009.04.020. URL http://www.cell.com/abstract/S0092-8674(09)00445-0. 108

K. L. Reddy, J. M. Zullo, E. Bertolino, and H. Singh. Transcriptional repression mediated by repositioning of genes to the nuclear lamina. *Nature*, 452(7184): 243–247, March 2008. ISSN 0028-0836. doi: 10.1038/nature06727. URL http://dx.doi.org/10.1038/nature06727. 109

Timothy J. Richmond and Curt A. Davey. The structure of DNA in the nucleosome core. *Nature*, 423(6936):145–150, May 2003. ISSN 0028-0836. doi: 10.1038/nature01595. URL http://dx.doi.org/10.1038/nature01595. 59, 78

Joel Rozowsky, Ghia Euskirchen, Raymond K Auerbach, Zhengdong D Zhang,

Theodore Gibson, Robert Bjornson, Nicholas Carriero, Michael Snyder, and Mark B Gerstein. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature Biotechnology*, 27(1):66–75, January 2009. ISSN 1087-0156. doi: 10.1038/nbt.1518. URL http://dx.doi.org/10.1038/nbt.1518. 41, 42

S C Satchwell, H R Drew, and A A Travers. Sequence periodicities in chicken nucleosome core DNA. *Journal of Molecular Biology*, 191(4):659–675, October 1986. ISSN 0022-2836. URL http://www.ncbi.nlm.nih.gov/pubmed/3806678. PMID: 3806678. 33

Alka Saxena and Piero Carninci. Long non-coding RNA modifies chromatin. *Bioessays*, 33(11):830–839, November 2011. ISSN 0265-9247. doi: 10.1002/bies.201100084. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3258546/. PMID: 21915889 PMCID: PMC3258546. 2

T D Schneider and R M Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18(20):6097–6100, October 1990. ISSN 0305-1048. PMID: 2172928 PMCID: 332411. 80

Dustin E. Schones and Keji Zhao. Genome-wide approaches to studying chromatin modifications. *Nature Reviews Genetics*, 9(3):179–191, March 2008. ISSN 1471-0056. doi: 10.1038/nrg2270. URL http://www.nature.com/nrg/journal/v9/n3/full/nrg2270.html. 131

Dustin E Schones, Kairong Cui, Suresh Cuddapah, Tae-Young Roh, Artem Barski, Zhibin Wang, Gang Wei, and Keji Zhao. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5):887–898, March 2008.

ISSN 1097-4172. doi: 10.1016/j.cell.2008.02.022. URL http://www.ncbi.nlm.nih.gov/pubmed/18329373. PMID: 18329373. 6, 13, 25, 34, 117, 155

Eran Segal and Jonathan Widom. Poly(dA:dT) tracts: major determinants of nucleosome organization. *Current Opinion in Structural Biology*, 19(1):65–71, February 2009. ISSN 1879-033X. doi: 10.1016/j.sbi.2009.01.004. URL http://www.ncbi.nlm.nih.gov/pubmed/19208466. PMID: 19208466. 21

Eran Segal, Yvonne Fondufe-Mittendorf, Lingyi Chen, AnnChristine Thstrom, Yair Field, Irene K Moore, Ji-Ping Z Wang, and Jonathan Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104):772–778, August 2006. ISSN 1476-4687. doi: 10.1038/nature04979. URL http://www.ncbi.nlm.nih.gov/pubmed/16862119. PMID: 16862119. 22, 23, 54, 78, 87

Sushma Shivaswamy, Akshay Bhinge, Yongjun Zhao, Steven Jones, Martin Hirst, and Vishwanath R Iyer. Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biology*, 6(3):e65, March 2008. doi: 10.1371/journal.pbio.0060065. URL http://dx.doi.org/10.1371/journal.pbio.0060065. 6

Arnold Stein, Taichi E. Takasuka, and Clayton K. Collings. Are nucleosome positions in vivo primarily determined by histone-DNA sequence preferences ? 38(3):709–719, January 2010. ISSN 0305-1048. doi: 10.1093/nar/gkp1043. PMID: 19934265 PMCID: 2817465. 18, 27

A Thastrom, P T Lowary, H R Widlund, H Cao, M Kubista, and J Widom. Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. *Journal of Molecular Biology*, 288(2):

213–229, April 1999. ISSN 0022-2836. doi: 10.1006/jmbi.1999.2686. URL http://www.ncbi.nlm.nih.gov/pubmed/10329138. PMID: 10329138. 21

Hagen Tilgner, Christoforos Nikolaou, Sonja Althammer, Michael Sammeth, Miguel Beato, Juan Valcarcel, and Roderic Guigo. Nucleosome positioning as a determinant of exon recognition. *Nature Structural and Molecular Biology*, 16(9):996–1001, 2009. ISSN 1545-9993. doi: 10.1038/nsmb.1658. URL http://dx.doi.org/10.1038/nsmb.1658. 102, 109

Desiree Tillo, Noam Kaplan, Irene K Moore, Yvonne Fondufe-Mittendorf, Andrea J Gossett, Yair Field, Jason D Lieb, Jonathan Widom, Eran Segal, and Timothy R Hughes. High nucleosome occupancy is encoded at human regulatory sequences. *PloS One*, 5(2):e9129, 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0009129. URL http://www.ncbi.nlm.nih.gov/pubmed/20161746. PMID: 20161746. 36, 76, 103, 162

Duygu Ucar, Qingyang Hu, and Kai Tan. Combinatorial chromatin modification patterns in the human genome revealed by subspace clustering. *Nucleic Acids Research*, 39(10):4063–4075, May 2011. ISSN 1362-4962. doi: 10.1093/nar/gkr016. URL http://www.ncbi.nlm.nih.gov/pubmed/21266477. PMID: 21266477. 9, 114, 137, 138

A. Valouev, S. M. Johnson, S. D. Boyd, C. L. Smith, A. Z. Fire, and A. Sidow. Determinants of nucleosome organization in primary human cells. *Nature*, 474 (7352):516–520, Jun 2011. 16, 19, 28, 34, 35, 36, 38, 40, 41, 53, 54, 56, 57, 59, 62, 76, 77, 90, 95, 104, 117, 155, 159, 162

Anton Valouev, Jeffrey Ichikawa, Thaisan Tonthat, Jeremy Stuart, Swati Ranade,

Heather Peckham, Kathy Zeng, Joel A Malek, Gina Costa, Kevin McKernan, Arend Sidow, Andrew Fire, and Steven M Johnson. A high-resolution, nucleosome position map of c. elegans reveals a lack of universal sequence-dictated positioning. *Genome research*, 18(7):1051–1063, July 2008. ISSN 1088-9051. doi: 10.1101/gr.076463.108. PMID: 18477713. 16, 19, 28, 33, 53, 62

Zhibin Wang, Chongzhi Zang, Jeffrey A Rosenfeld, Dustin E Schones, Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Weiqun Peng, Michael Q Zhang, and Keji Zhao. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genetics*, 40(7):897–903, July 2008. ISSN 1061-4036. doi: 10.1038/ng.154. PMID: 18552846 PMCID: 2769248. 106, 108, 109, 110, 116, 117, 137, 138, 139

Assaf Weiner, Amanda Hughes, Moran Yassour, Oliver J Rando, and Nir Friedman. High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome Research*, 20(1):90–100, January 2010. ISSN 1549-5469. doi: 10.1101/gr.098509.109. URL `http://www.ncbi.nlm.nih.gov/pubmed/19846608`. PMID: 19846608. 18, 20, 68

Bo Wen, Hao Wu, Yoichi Shinkai, Rafael A Irizarry, and Andrew P Feinberg. Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. *Nature Genetics*, 41(2):246–250, 2009. ISSN 1061-4036. doi: 10.1038/ng.297. URL `http://dx.doi.org/10.1038/ng.297`. 109

Iestyn Whitehouse, Oliver J Rando, Jeff Delrow, and Toshio Tsukiyama. Chromatin remodelling at promoters suppresses antisense transcription. *Nature*, 450

(7172):1031–1035, December 2007. ISSN 1476-4687. doi: 10.1038/nature06391. URL http://www.ncbi.nlm.nih.gov/pubmed/18075583. PMID: 18075583. 25

Michael Wigler, Daniel Levy, and Manuel Perucho. The somatic replication of DNA methylation. *Cell*, 24(1):33–40, April 1981. ISSN 0092-8674. doi: 10.1016/0092-8674(81)90498-0. URL http://www.sciencedirect.com/science/article/pii/0092867481904980. 2

Xiaojiang Xu, Stephen Hoang, Marty W Mayo, and Stefan Bekiranov. Application of machine learning methods to histone methylation ChIP-Seq data reveals H4R3me2 globally represses gene expression. *BMC Bioinformatics*, 11:396, 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-396. URL http://www.ncbi.nlm.nih.gov/pubmed/20653935. PMID: 20653935. 111, 118

Tianbao Yang and B W Poovaiah. A calmodulin-binding/CGCG box DNA-binding protein family involved in multiple signaling pathways in plants. *The Journal of biological chemistry*, 277(47):45049–45058, November 2002. ISSN 0021-9258. doi: 10.1074/jbc.M207941200. PMID: 12218065. 82

Moran Yassour, Tommy Kaplan, Ariel Jaimovich, and Nir Friedman. Nucleosome positioning from tiling microarray data. 24(13):i139–i146, July 2008. ISSN 1367-4803. doi: 10.1093/bioinformatics/btn151. PMID: 18586706 PMCID: 2718629. 19, 54

Hong Yu, Shanshan Zhu, Bing Zhou, Huiling Xue, and Jing-Dong J Han. Inferring causal relationships among different histone modifications and gene ex-

pression. *Genome Research*, 18(8):1314–1324, August 2008. ISSN 1088-9051. doi: 10.1101/gr.073080.107. URL http://www.ncbi.nlm.nih.gov/pubmed/18562678. PMID: 18562678. 111, 118

Guo-Cheng Yuan and Jun S Liu. Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Computational Biology*, 4(1):e13, January 2008. ISSN 1553-7358. doi: 10.1371/journal.pcbi.0040013. URL http://www.ncbi.nlm.nih.gov/pubmed/18225943. PMID: 18225943. 23, 24, 87

Guo-Cheng Yuan, Yuen-Jong Liu, Michael F. Dion, Michael D. Slack, Lani F. Wu, Steven J. Altschuler, and Oliver J. Rando. Genome-scale identification of nucleosome positions in S. cerevisiae. *Science*, 309(5734):626 –630, 2005. doi: 10.1126/science.1112178. URL http://www.sciencemag.org/content/309/5734/626.abstract. 16, 19, 54

Judith B. Zaugg and Nicholas M. Luscombe. A genomic model of condition-specific nucleosome behavior explains transcriptional activity in yeast. *Genome Research*, 22(1):84–94, January 2012. ISSN 1088-9051. doi: 10.1101/gr.124099.111. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3246209/. PMID: 21930892 PMCID: PMC3246209. 156, 159

Hesheng Zhang and Joseph C. Reese. Exposing the core promoter is sufficient to activate transcription and alter coactivator requirement at RNR3. *Proceedings of the National Academy of Sciences*, 104(21):8833 –8838, 2007. doi: 10.1073/pnas.0701666104. URL http://www.pnas.org/content/104/21/8833.abstract. 26

Yong Zhang, Tao Liu, Clifford A Meyer, Jerome Eeckhoute, David S Johnson,

Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu. Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137, 2008a. ISSN 1465-6914. doi: 10.1186/gb-2008-9-9-r137. URL `http://www.ncbi.nlm.nih.gov/pubmed/18798982`. PMID: 18798982. 41

Yong Zhang, Hyunjin Shin, Jun S Song, Ying Lei, and X Shirley Liu. Identifying positioned nucleosomes with epigenetic marks in human from ChIP-Seq. *BMC Genomics*, 9:537–537, 2008b. doi: 10.1186/1471-2164-9-537. PMID: 19014516 PMCID: 2596141. 18, 20, 53, 62, 72

Yong Zhang, Zarmik Moqtaderi, Barbara P Rattner, Ghia Euskirchen, Michael Snyder, James T Kadonaga, X Shirley Liu, and Kevin Struhl. Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. *Nature Structural and Molecular Biology*, 16(8):847–852, 2009. ISSN 1545-9993. doi: 10.1038/nsmb.1636. URL `http://dx.doi.org/10.1038/nsmb.1636`. 16, 19, 20, 28, 30, 31, 32, 37, 76, 94, 161

Yong Zhang, Zarmik Moqtaderi, Barbara P Rattner, Ghia Euskirchen, Michael Snyder, James T Kadonaga, X Shirley Liu, and Kevin Struhl. Evidence against a genomic code for nucleosome positioning reply to "Nucleosome sequence preferences influence in vivo nucleosome organization". *Nature Structural and Molecular Biology*, 17(8):920–923, 2010. ISSN 1545-9993. doi: 10.1038/nsmb0810-920. URL `http://www.nature.com/nsmb/journal/v17/n8/full/nsmb0810-920.html`. 32, 37, 161

Zhihua Zhang and Michael Q Zhang. Histone modification profiles are predictive

for tissue/cell-type specific expression of both protein-coding and microRNA genes. *BMC Bioinformatics*, 12:155, 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-155. URL http://www.ncbi.nlm.nih.gov/pubmed/21569556. PMID: 21569556. 112, 147, 148, 150, 163

Vicky W. Zhou, Alon Goren, and Bradley E. Bernstein. Charting histone modifications and the functional organization of mammalian genomes. *Nature Review Genetics*, 12(1):7–18, January 2011. ISSN 1471-0056. doi: 10.1038/nrg2905. URL http://dx.doi.org/10.1038/nrg2905. 3, 5

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, April 2005. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2005.00503.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2005.00503.x/full. 131

Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15 (2):265–286, June 2006. ISSN 1061-8600, 1537-2715. doi: 10.1198/106186006X113430. URL http://pubs.amstat.org/doi/abs/10.1198/106186006X113430?journalCode=jcgs. 131, 132

V. Zuber and K. Strimmer. High-dimensional regression and variable selection using CAR scores. *Arxiv preprint*, 2010. URL http://arxiv.org/abs/1007.5516. 143