PhD degree in Molecular Medicine

European School of Medicine (SEMM)
University of Milano and University of Naples "Federico II"
Settore Disciplinare: BIO/10

Overlapping sequence features of mammalian enhancers coordinately control engagement of transcription factor consensus sites and nucleosomal occupancy

Iros Giacomo Barozzi

Supervisors:
Gioacchino Natoli, MD, PhD
Saverio Minucci, MD

Anno accademico 2012-2013

PhD degree in Molecular Medicine

Abstract

European School of Medicine (SEMM)

Settore Disciplinare: BIO/10

by Iros Giacomo Barozzi

In mammalian cells transcription factors (TFs) bind only to a small fraction of the available consensus sites in the genome. In particular, they prefer sites embedded in regions of computationally predicted high nucleosomal occupancy. This is compatible with non-exclusive mechanisms of nucleosome-driven TF-binding and nucleosome-mediated masking of TF binding sites, suggesting that TFs, and in particular pioneers, must overcome a strong barrier in order to engage binding. Exploiting the available information for the hematopoietic master regulator Pu.1, we applied machine learning approaches and uncovered the sequence-encoded information that discriminates engaged from non-engaged genomic consensus sites. We identified a minimal set of features which predicts Pu.1 binding with 78% accuracy, among which sequence determinants able to drive nucleosome occupancy were found. Consistent with this, while Pu.1 maintained nucleosome depletion at many thousand cell type-specific enhancers in macrophages, these site are otherwise occupied by nucleosomes in other cell types and in *in vitro* reconstituted chromatin. As predicted, engaged consensus sites showed higher sequence-encoded nucleosome occupancy compared to the myriad of non-occupied (and likely non-functional) consensus sites that randomly occur in mammalian genomes. The same sequence features selected in machine learning also explains up to 45% of the variability observed in the nucleosome occupancy in cells where Pu.1 is not expressed (a performance equal or better than what achieved by *ad hoc* models), suggesting that the same information contributes to nucleosome occupancy and positioning. These data reveal a basic organizational principle of mammalian enhancers whereby TF-engagement at its consensus sites and nucleosome occupancy are coordinately controlled by overlapping sequence features. This model also suggests that co-evolution of these features may be crucial to ensure cell-type specific

enhancer activation. The nucleosomal patterns at Pu.1-bound sites in macrophages were further characterized, uncovering distinct subtypes with different DNA sequence composition, which mirror distinctive nucleosomal configurations either in the presence or in the absence of Pu.1.

Contents

Abstract i

List of Figures v

List of Tables vii

Abbreviations viii

1 Introduction 1
- 1.1 Regulation of DNA transcription . . . 1
- 1.2 Taxonomy of *cis*-regulatory elements . . . 3
- 1.3 Transcriptional regulation: a combinatorial problem . . . 5
- 1.4 Chromatin organization . . . 7
  - 1.4.1 The nucleosome . . . 7
  - 1.4.2 Nucleosomal organization at *cis*-regulatory elements . . . 8
  - 1.4.3 Chromatin modifications at *cis*-regulatory elements . . . 9
    - 1.4.3.1 Post-translational modifications of the histones . . . 9
    - 1.4.3.2 DNA methylation . . . 11
  - 1.4.4 A wider picture: chromatin states . . . 12
  - 1.4.5 Sequence-specific TFs and their interplay with chromatin determine cell fate . . . 13
- 1.5 Determinants of transcription factor binding . . . 14
- 1.6 Predictions of transcription factor binding from the sequence . . . 15
- 1.7 Determinants of nucleosome occupancy and positioning . . . 16
- 1.8 Predictions of nucleosomal patterns from the DNA sequence . . . 20
- 1.9 Regulation of transcription in murine macrophages . . . 21
  - 1.9.1 TFs in the hematopoietic system . . . 21
  - 1.9.2 Pu.1: one of the master regulators of macrophage differentiation . . . 22
  - 1.9.3 How does Pu.1 select its binding sites *in vivo*? . . . 24
- 1.10 Experimental approaches to probe *cis*-regulatory elements and chromatin . . . 26
  - 1.10.1 DNA accessibility assays . . . 27
  - 1.10.2 Chromatin immuno-precipitation . . . 27
  - 1.10.3 Determination of nucleosome positions . . . 28

2 Methods 30
- 2.1 Pu.1 ChIP-seq in murine macrophages . . . 30
- 2.2 *in vitro* Pu.1 ChIP-seq data analysis . . . 31
- 2.3 A collective *cis*-regulatory repertoire bound by Pu.1 . . . 31
- 2.4 Genome-wide maps of regions putatively bound by Pu.1 . . . 32

2.5 Measuring features in DNA strings . . . 33
2.5.1 Position Weight Matrices . . . 35
2.6 Supervised learning using Support Vector Machines . . . 36
2.7 Mnase-seq data analysis . . . 39
2.8 Support Vector Regressors . . . 40
2.9 Data analysis upon Pu.1 depletion . . . 41
2.10 Chromatin-bound RNA-seq analysis . . . 42
2.11 Statistics and plots . . . 42
2.12 Experimental procedures . . . 42
2.12.1 Cell culture, retroviral infection and ChIP . . . 42
2.12.2 *In vitro* nucleosome assembly . . . 43
2.12.3 MNase digestion . . . 43
2.12.4 *In vitro* ChIP . . . 44

3 Results **45**
3.1 The *theoretical* cistrome of Pu.1 . . . 45
3.2 The collective *cis*-regulatory repertoire bound by Pu.1 *in vivo* . . . 46
3.3 Discrimination of engaged and non-engaged Pu.1-binding sites *in vivo* . . . 47
3.4 Nucleosomal organization at Pu.1 bound and unbound sites . . . 55
3.5 Pu.1-bound sites show spatial sequence constraints . . . 57
3.6 Nucleosomal organization at Pu.1 sites *in vitro* . . . 60
3.7 Predicting nucleosome occupancy from features instructive for Pu.1 binding . 63
3.8 A detailed evaluation of chromatin organization at Pu.1 binding sites in BMDMs 66
3.8.1 TSS-distal sites . . . 66
3.8.1.1 Sequence determinants . . . 71
3.8.1.2 Nucleosomal patterns in unrelated cell-types and *in vitro* . . . 73
3.8.1.3 *in vitro* ChIP against Pu.1 . . . 75
3.8.2 TSS-proximal sites . . . 77
3.9 Is Pu.1 required to maintain the nucleosomal organization at *cis*-regulatory elements in BMDMs? . . . 82

4 Discussion **86**

Bibliography **93**

List of Figures

1.1 Enhancers drive tissue-specific gene expression patterns 4
1.2 Nucleosome occupancy and positioning 17
1.3 Determinants of nucleosomal patterns 18
1.4 Master regulator TFs in hematopoiesis 23
1.5 Pu.1 mRNA levels 24

2.1 Inference of DNA shape features 35
2.2 Position weight matrices basics 37

3.1 Pu.1 binding preference 45
3.2 How many ChIP-seq determined binding events show a *canonical* binding site? 46
3.3 The collective *cis*-regulatory repertoire bound by Pu.1 47
3.4 Gene-centered annotation of the collective *cis*-regulatory repertoire bound by Pu.1 47
3.5 Bound and unbound *canonical* binding sites show significantly different affinities for the ETS site 48
3.6 SVM schema 49
3.7 SVM performances 50
3.8 Bound sites show a significantly higher number of binding sites for distinct families of TFs 51
3.9 Bound sites show a significantly higher C+G content than unbound sites . . . 51
3.10 Bound sites show a significantly higher theoretical nucleosome occupancy than unbound sites 51
3.11 Runx1 and AP-1 positional distribution around Pu.1 bound and unbound sites 53
3.12 Nucleosomal organization at TSS-distal Pu.1 binding sites 55
3.13 Nucleosomal organization at TSS-proximal Pu.1 binding sites 55
3.14 Nucleosomal organization at TSS-distal Pu.1 binding sites in BMDMs, ESCs, NPCs and MEFs 56
3.15 Nucleosomal organization at TSS-proximal Pu.1 binding sites in BMDMs, ESCs, NPCs and MEFs 56
3.16 Nucleosomal organization at TSS-distal Pu.1 binding sites in cell types other than BMDMs 57
3.17 Nucleosomal organization at TSS-proximal Pu.1 binding sites in cell types other than BMDMs 58
3.18 Nucleosome *container site* 59
3.19 Frequency of AA and CG-rich dinucleotides around TSS-distal Pu.1 binding sites 59
3.20 Frequency AA and CG-rich dinucleotides around TSS-proximal Pu.1 binding sites 60
3.21 AAAA frequency at TSS-distal Pu.1 binding sites 60

3.22 AAAA frequency at TSS-proximal Pu.1 binding sites 60
3.23 *in vitro* reconstituted nucleosomes at Pu.1-contacted sites in macrophages . . 61
3.24 Frequency of dinucleotides at *in vivo* and *in vitro* reconstituted strongly positioned nucleosomes 62
3.25 Support Vector Regressor schema 63
3.26 Details on the performances of the SVR in ESC and *in vitro* 64
3.27 Performances of the Support Vector Regressor 65
3.28 Nucleosome, hPTMs, TFs and polII heatmaps representation at TSS-distal Pu.1-bound sites in macrophages 67
3.29 Pu.1 ChIP-seq score (according to MACS) of the peaks in different deciles . . 68
3.30 Bulk signals of the nucleosome midpoints in each decile 68
3.31 Bulk H3K27ac and H3K4me1 ChIP-seq signals in each decile 69
3.32 Bulk H3K27ac and H3K4me1 ChIP-seq signals (normalized by nucleosome occupancy) in each decile 70
3.33 Bulk polII ChIP-seq signals in each decile 70
3.34 FPKM of nearest expressed genes in different deciles 71
3.35 C+G content of the regions in different deciles 71
3.36 Positional content in dinucleotides of the regions belonging to different deciles 72
3.37 Cumulative nucleosome profile in cells other than macrophages and *in vitro* for the regions in different deciles 73
3.38 Overall nucleosome occupancy in cells other than macrophages and *in vitro* for the regions in different deciles 74
3.39 Venn diagram showing the overlap between *in vitro* and *in vivo* Pu.1-bound sites 75
3.40 Nucleosome occupancy at sites bound by Pu.1 either *in vivo*, *in vitro* or in both conditions 76
3.41 ChIP-seq scores of the *in vitro* reconstituted Pu.1 peaks 76
3.42 Bulk signals of the nucleosome midpoints in each decile (TSS-proximal sites) 78
3.43 Distributions of distances from the nearest TSSs for the regions in different deciles 78
3.44 Bulk H3K27ac and H3K4me3 ChIP-seq signals in each decile (TSS-proximal sites) 79
3.45 Bulk H3K27ac and H3K4me3 ChIP-seq signals (normalized by nucleosome occupancy) in each decile (TSS-proximal sites) 80
3.46 Bulk polII ChIP-seq signals in each decile (TSS-proximal sites) 80
3.47 Decile-specific distributions of the FPKMs of the genes in the different deciles 81
3.48 Cumulative nucleosome profile in ESCs and *in vitro* for the regions in different deciles 81
3.49 Pu.1 protein quantification in Pu.1-depleted macrophages 82
3.50 Pu.1 ChIP-seq peaks in Pu.1-depleted macrophages 83
3.51 Nucleosomal patterns around Pu.1-bound sites in Pu.1-depleted macrophages 84
3.52 Nucleosome occupancy over Pu.1-bound sites in Pu.1-depleted macrophages . 85

List of Tables

2.1 List of the Pu.1 ChIP-seq datasets collected from the literature 31
2.2 PWMs from the literature 33
2.3 Features considered as input for the selection 38
2.4 Sequencing statistics of the MNase-seq samples 40

3.1 SVM performaces along ten training-test randomizations 50
3.2 Frequently selected features during multiple initialization of training-test datasets. 52
3.3 Overlaps of the Pu.1-bound TSS-distal regions in each decile with the ChIP-seq datasets, the *canonical* Pu.1-binding sites and the CpGi 69
3.4 Overlaps of *in vitro* ChIP-seq against Pu.1 with the *canonical* Pu.1-binding sites and the TSSs of Ensembl genes 77
3.5 Overlaps of the Pu.1-bound TSS-proximal regions in each decile with the ChIP-seq datasets, the *canonical* Pu.1-binding sites and the CpGi 80

Abbreviations

3C	**C**hromatin **C**onformation **C**apture
5C	**C**hromatin **C**onformation **C**apture **C**arbon **C**opy
ATP	**A**denosine-5'-**t**ri**p**hosphate
AUC	**A**rea **U**nder the **C**urve
BAC	**B**acterial **A**rtificial **C**hromosome
bHLH	**b**asic **H**elix **L**oop **H**elix
BMDM	**B**one **M**arrow **D**erived **M**acrophages
bp	**b**ase **p**airs
ChIP	**Ch**romatin **I**mmuno-**P**recipitation
ChIP-seq	**ChIP** followed by HT-**seq**uencing
ChIP on chip	**ChIP** followed by **chip** hybridization
C+G	**C**ytosine + **G**uanine content of a DNA stretch
CpG	**CG** dinucleotide
CRM	**C**is-**R**egulatory **M**odule
CTCF	**C**CC**TC**-binding **F**actor
DBD	**D**NA **B**inding **D**omain
DNA	**D**eoxyribonucleic **A**cid
DNMT	**DN**A **M**ethyl **T**ransferase
EHP	**E**ndoderm/**h**epatic **P**rogenitor
EMSA	**E**lettroforetic **M**obility **S**hift **A**ssay
eRNA	**E**nhancer **R**NA
ESC	**E**mbryonic **S**tem **C**ell
FISH	**F**luorescence **i**n **s**itu **h**ybridization
FPKM	**F**ragments **P**er **K**ilobase of transcript per **M**illion mapped reads
FRAP	**F**luorescence **R**ecovery **A**fter **P**hotobleaching
GTF	**G**eneral **T**ranscription **F**actor
HSC	**H**ematopoietic **S**tem **C**ell
HT	**H**igh **T**hroughput
IP	**I**mmuno-**P**recipitation
IRF	**I**nterferon **R**egulatory **F**actor
kbp	**k**ilo **b**ase **p**airs

LCR	**L**ocus **C**ontrol **R**egion
LINE	**L**ong **I**nterspersed **N**uclear **E**lement
LTR	**L**ong **T**erminal **R**epeat
mRNA	**m**essanger **RNA**
MEF	**M**ouse **E**mbryonic **F**ibroblast
MPP	**M**ulti**p**otent **P**rogenitor
NDR	**N**ucleosome **D**epleted **R**egion
NFR	**N**ucleosome **F**ree **R**egion
NPC	**N**eural **P**rogenitor **C**ell
NPS	**N**ucleosome **P**ositioning **S**equence
nt	**n**ucleo**t**ide
PIC	**P**re **I**nitiation **C**omplex
polII	RNA **Pol**ymerase **II**
PCR	**P**olymerase **C**hain **R**eaction
PPI	**P**rotein **P**rotein **I**nteraction
hPTM	histone **P**ost **T**ranslational **M**odification
PWM	**P**osition **W**eight **M**atrix
rbf	**r**adial **b**asis **f**unction
RNA	**R**ibonucleic **A**cid
RRBS	**R**educed **R**epresentation **B**isulfite **S**equencing
TF	**T**ranscription **F**actor
TFBS	**T**ranscription **F**actor **B**inding **S**ite
TSS	**T**ranscriptional **S**tart **S**ite

Chapter 1

Introduction

1.1 Regulation of DNA transcription

Seminal studies in bacteria during the 1960s (Jacob and Monod, 1961, Englesberg et al., 1965, Gilbert and Müller-Hill, 1967) showed that regulation of transcription (namely how the information is transferred from a DNA to a RNA molecule) depends on the recognition of specific genomic DNA sequences (*cis*-regulatory elements, or simply regulatory elements) by particular proteins termed *trans*-factors. These DNA sequences mediate the maintainance and the re-organization of the transcriptional program of a cell, either in response to environmental (DeRisi et al., 1997) or developmental (Arbeitman et al., 2002) cues. Shaping the complex body plans of multi-cellular organisms require the coordinated transcription of thousands of genes in space and time, which is largely dependent on *cis*-regulatory elements (Zinzen et al., 2009).

Different types of regulatory elements can be distinguished in the genome. A core promoter is the minimal set of regulatory sequences sourrounding the transcriptional start site (TSS) of a gene. In order to be defined as such, a core promoter must be able to drive transcription *in vitro* (Smale and Kadonaga, 2003). Core promoters can be bound by general transcription factors (GTFs) resulting in the assembly of the pre-initiation complex (PIC), which helps positioning the RNA polymerase II (polII) at the TSS (Lenhard et al., 2012). Core promoters represent only a fraction of *cis*-regulatory elements in Metazoa. The remaining elements

spread from hundreds to millions of kilobases from them and act as platforms for the recruitment of multiple transcription factors (TFs), co-factors (activators and/or repressors) and chromatin remodeling complexes. Only the concerted binding of specific combinations of TFs at both core promoter and TSS-distal *cis*-regulatory elements is able to drive tissue-specific gene expression (see figure 1.1). In this context, multiple regulatory signals converge on the TSS of a gene, through a mechanism called DNA looping (Bulger and Groudine, 1999). At present, this mechanism is widely supported by experiments of chromatin conformation capture (3C) based techniques (de Wit and de Laat, 2012) and DNA fluorescence in situ hybridization (FISH) coupled with Super-resolution microscopy (van de Corput et al., 2012). Long-range regulatory interactions have been shown to be an extensive characteristic of the genomes of Metazoa, while playing a significantly less widespread role in other eukaryotes (Levine, 2010).

It should be note that these interactions are highly dynamic and subject to stochasticity (Coulon et al., 2013). At present, there are evidences showing that the residence time of a TF on a *cis*-regulatory element is the major determinant of the activity of the region (Lickwar et al., 2012). According to this report, shorter but frequent interactions could result in little or no effect compared to longer but less frequent ones. Similarly to single protein-DNA interactions, distinct tissue-specific combinations of TFs could be able to increase (or decrease) the probability of a looping interaction to be stabilized for an amount of time sufficient for functional consequences (see figure 1.1).

Spatial proximity is achieved and maintained through specific protein-protein and protein-DNA interactions. Although the underlying mechanisms are still poorly understood, the Mediator complex and Cohesin have been recently shown to form a ring able to constraint two DNA segments in space, thereby creating a loop (Kagey et al., 2010, Dorsett, 2011). Cohesin has also been shown to be an invariant component of the majority of clusters of TFs in human colorectal cancer (Yan et al., 2013). Interestingly, the same study showed that Cohesin remains bound to these clusters during mitosis while other factors tested (namely Klf5, Hnf4a and Myc) are evicted. These results suggest a role for Cohesin in bookmarking active regulatory elements after DNA replication and chromatin condensation. Ldb1 was also found to be indispensable to the Gata1-mediated looping of the locus control region (LCR) of the β-globin with its core promoter (Deng et al., 2012).

1.2 Taxonomy of *cis*-regulatory elements

During recent years it has been proposed that the promoters of higher eukaryotes can be categorized in a few different groups (Lenhard et al., 2012). These show correlated DNA sequence content and ability to drive initiation of transcription either from a single or from multiple nucleotide positions (usually referred as focused and dispersed initiation):

- Tissue-Specific: focused promoters, often display TATA-box and almost no CpG islands;
- Housekeeping: dispersed promoters, almost TATA-less and with short CpG islands;
- Developmentally-Regulated: dispersed promoters, large CpG islands, Polycomb-regulated.

TATA-box is a core promoter element located around 25 bp upstream of a TSS. It is present in 25% of the TSSs in *H. sapiens* (Yang et al., 2007). For a detailed description of CpG islands refer to paragraph 1.4.3.2.
On the other hand, *cis*-regulatory elements other than core promoters can be divided according to their effects on transcription: (Maston et al., 2006):

- Enhancers: interact with the TSS of one or more genes, increasing their transcriptional rate;
- Silencers: same as enhancers, but they act by decreasing the transcriptional rate of their target genes;
- Insulators: inhibit the activity of enhancers and silencers, by competition or blocking;
- Locus Control Regions (LCRs): clusters of regulatory elements, often affecting the transcriptional rate at locus containing more than one gene.

From an operational point of view, these elements are often divided into two groups according to their distance from the nearest TSS. As a rule of thumb, it is widely accepted the use of an arbitrary threshold of 2.5 kbp around annotated TSSs in order to distinguish TSS-proximal *cis*-regulatory elements from TSS-distal ones.
As anticipated in the previous paragraph, multiple regulatory signals converge on the TSS of a single gene to fine tune its transcriptional rate in a tissue-specific manner (see figure

1.1). It has been shown that up to tens of enhancers can be simultaneously engaged in the regulation of a single gene (Arnold et al., 2013). This was observed also for housekeeping genes, which by definition have similar levels of expression across cell types (Arnold et al., 2013). The high number of enhancers per gene could reflect combinatorial regulation as well as conferring robustness through redundancy, as testified by the report of *shadow* enhancers controlling developmental genes (Hong et al., 2008).

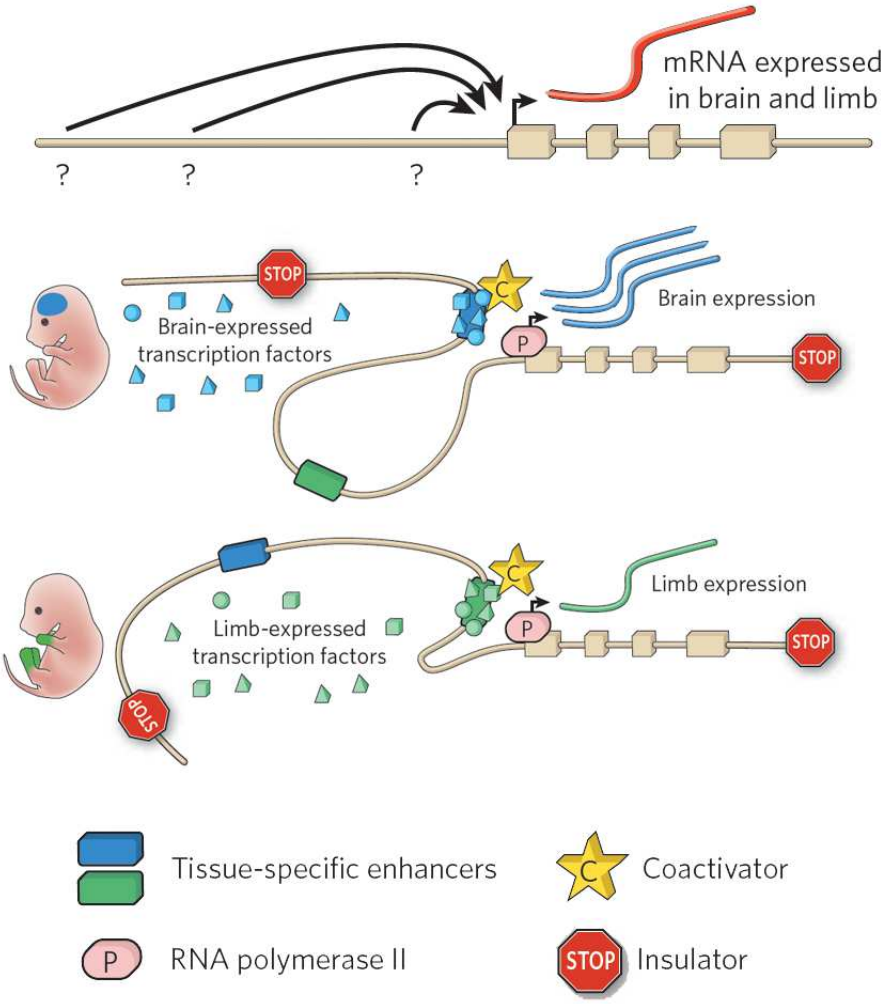

FIGURE 1.1: The information encoded in the core promoter itself can be insufficient to drive an expression pattern that is tissue-specific (upper panel). The coordinated action of TFs is required to promote co-factors recruitment and enhancer-promoter looping, which in turn regulates tissue specific levels of gene expression (middle and lower panel). Insulator elements function to restrict the activity of enhancers to defined chromatin domains by preventing aspecific enhancer-promoter interactions. Adapted from Visel et al., 2009.

1.3 Transcriptional regulation: a combinatorial problem

Transcriptional regulation is achieved through the binding of combinations of TFs, which are encoded in the genomic sequence as dense clusters of transcription factor binding sites (TFBSs). These are generally referred to as *cis*-regulatory modules (CRMs). TFBSs are short (usually 6-12 nt) DNA recognition sequences (motifs); they are degenerated and usually not sufficient to predict the binding of a TF on their own (Wasserman and Sandelin, 2004).
A recent study (Ravasi et al., 2010) split human and murine TFs into two broad categories based on their mRNA expression pattern across tissues. Two groups were defined, one of widely expressed and another of lineage-specific TFs. The widely expressed TFs were termed *facilitators*, given that their expression across different tissues could facilitate transcriptional programs, while the lineage-specific were named *specifiers*. For example, TFs like Jun, Fos, Myc and Tp53 were among *facilitators* while Myod1 and Gata1 were among *specifiers*. Among the specifiers many master regulators were found. TFs defined as such play a pivotal role into the differentiation towards a specific lineage. In line with this other studies demonstrated that widely expressed TFs can bind to completely different sets of promoters and enhancers in presence of different master regulators. Smad3 has been shown to be co-opted by Oct4 in embryonic stem cells (ESCs) by Myod1 in myotubes and by Pu.1 in pro-B cells at an almost completely different set of *cis*-regulatory elements (Mullen et al., 2011). In this scenario, the different concentrations of TFs coupled with the genomic information (CRMs) are able to generate the complex combinatorial patterns responsible for cell identity (Neph et al., 2012a). This combinatorial regulation can be achieved through PPI among TFs or indirect mechanisms of cooperativity (Spitz and Furlong, 2012):

- Transcriptional sinergy: two non-interacting TFs bound in proximity along DNA can be stabilized by PPI to form a complex (through Mediator or co-factors like p300/Cbp);
- Passive enhancer priming: one TF can act as a place-holder and displace a nucleosome. This allows another TF to access the *cis*-regulatory element, that otherwise would not be able to bind;
- Local DNA bending: one TFs is specifically able to induce a change in the local shape of the chromatin fiber, so that another TF acquire the ability to bind;

- Chromatin remodeling dependent on a pioneer TF: a TF that is able to invade chromatin is called a pioneer (Magnani et al., 2011). Recruitment of chromatin remodeling complexes by pioneers might increase the accessibility to the locus to other TFs that do not possess pioneering activity.

At present, in order to achieve cooperativity among TFs different models have been proposed. These are not mutually exclusive in the sense that they account for different way of regulation that co-exist in the same genome. The three main models have been recently reviewed in (Spitz and Furlong, 2012):

- Enhanceosome: there is a strict grammar (namely precise spacing and orientation) over the motif composition of the regulatory element as all the TFs must be expressed at the proper concentration for its activity; in practice, this precise organization has been observed in very few cases (Merika and Thanos, 2001), among which the most studied is an enhancer of the Ifn-β gene (Maniatis et al., 1998);

- Billboard: the majority of the regulatory elements in Metazoan genomes do not follow an enhanseosome model; in the billboard model the accent is on the players involved. Namely there is no strict grammar in the sequence but a core of motifs is required for activity. In this context, additive and independent binding takes place, while in the enhanseosome model a strict cooperativity is required in order to reach the only conformation that is functional;

- TF Collective: a core set of TFs is bound to the regulatory element with looser sequence requirements compared to the billboard model; TFs can bind either through DNA-protein or protein-protein interactions, with no strict requirement on the presence of the motifs.

As anticipated, there exist *cis*-regulatory elements where strict grammar is essential to ensure proper activity (Maniatis et al., 1998, Senger et al., 2004), but they represent a minority. In fact, evidences in *D. melanogaster* suggest that while the identity of the motifs into the module is conserved, the architecture is flexible. This is supported by functional experiments in which enhancers that diverged over a 100 millions years ago were demonstrated to drive the same expression pattern during development, while the underlying motifs being almost completely

reshuffled (Liberman and Stathopoulos, 2009, Hare et al., 2008). Similar conclusions were drawn in independent computer simulations (Lusk and Eisen, 2010).

This picture gets a further level of complexity when considering that regulatory processes take place in the context of chromatin. Chromatin status is influenced by binding of TFs and vice versa. TFs are able to recruit protein complexes that can read the chromatin status of *cis*-regulatory elements as well as enzymes able to modify it, leaving a trace of previous activities (Kouzarides, 2007). The next section gives a general overview of chromatin organization and frame TFs into this context.

1.4 Chromatin organization

Genomic DNA is organized in the nucleus of eukaryotic cells in a protein-DNA complex called chromatin. Chromatin is not only a way of compacting genomic information in a restricted space. Transcription, replication, recombination and repair are all intrinsecally related to this reversible property (the condensation/de-condensation).

1.4.1 The nucleosome

The nucleosome is the basic building block of chromatin (Kornberg, 1974). The nucleosome core particle consists of 146-147 bp of DNA wrapped in 1.67 left-handed superhelical turns around a histone octamer, which consists of two H2A-H2B dimers and a H3-H4 tetramer (Luger et al., 1997). Core particles are connected by stretches of "linker DNA", which can vary in length based on the species or even on the tissue considered (Valouev et al., 2011). The linker histone H1 is not part of the core particle, and it has been implicated mainly in chromatin compaction. The amino acid sequence of histones is almost completely under negative selection and conserved from *S. cerevisiae* to *H. sapiens* (Kornberg and Lorch, 1999). Studies in *S. cerevisiae* and *in vitro* showed that histones can act as general repressors (Kornberg and Lorch, 1999). In recent years, it has emerged a more complex picture in which nucleosomes are plastic and can acquire different states (Zentner and Henikoff, 2013), controlled through covalent post-translational modification of the N-terminal tails of histones

(hPTMs or "marks") and ATP-dependent remodeling (Kornberg and Lorch, 1999). The availability of variants of the histone genes adds a further level of complexity.

1.4.2 Nucleosomal organization at *cis*-regulatory elements

Nucleosomal organization at regulatory elements has been the subject of intense studies, especially at TSSs and enhancers. Active and poised (namely those in which polII is engaged but is not elongating into the gene body) TSSs show a particular configuration in which a nucleosome free region (NFR) is flanked by two well-positioned nucleosomes (named +1 and -1 according to the direction of transcription), followed by a nucleosomal array overlapping the initial portion of the gene body (Jiang and Pugh, 2009). The function of as well as the mechanism that originate this structure are still unclear. It has been shown that the NFR is occupied by two divergent PICs in *S. cerevisiae* (Rhee and Pugh, 2012), which fits with the hypothesis that the two positioned nucleosomes could play a role in polII pausing at TSSs. Nevertheless, this role for the +1 nucleosome has been recently challenged in *D. melanogaster* (Kwak et al., 2013). PolII has been long hypothesize to be responsible for the maintenance of this NFR. A recent study reported that α-amanitin treatment (which causes the release and degradation of the polII complex from chromatin) on human T-cells has only a minor effect on the NFR and the surrounding nucleosomal configuration (Fenouil et al., 2012). Although the polII complex does not seem to be directly implicated in the mechanism of NFR maintenance (unless it leaves some kind of memory), deletion of ATP-dependent chromatin remodeling factors have been reported to affect positioning of the +1 nucleosome in *S. cerevisiae* (van Bakel et al., 2013). Besides, once thought to work indiscriminately across the entire genome, recent observations point to a position-specific role (relative to the NFR) of many of the chromatin remodelers (Yen et al., 2012).

Considering the three main classes of core promoters in Metazoan (Lenhard et al., 2012), focused promoters often show a more disordered nucleosomal pattern compared to dispersed promoters. Considering instead the bulk of *cis*-regulatory elements (core promoters, TSS-proximal and TSS-distal elements) in murine ESCs, NPCs and MEFs, an intricate pattern was observed. Different TFs were associated to distinct behaviors (Teif et al., 2012), namely:

- TFs residing in regions with high nucleosome occupancy in cells types where they are not expressed but undergo a decrease in nucleosome occupancy in the cell types where the TF is expressed (and bound);
- TFs residing in regions of general high nucleosome occupancy;
- TFs residing in regions of general low nucleosome occupancy.

Among the first class, tissue-specific CTCF binding events correlated with tissue-specific nucleosomal patterns. Sites in common between ESCs and MEFs showed the same organization, namely a lower nucleosome occupancy at the CTCF binding site, flanked by two positioned nucleosomes. ESCs-specific sites shown instead an increase in nucleosome occupancy over the CTCF site in MEFs (Teif et al., 2012). A recent study extended these observations by showing an extensive asymmetry regarding the positions of nucleosomes around binding sites of the majority of human TFs (Kundaje et al., 2012).

Beside showing differences in their levels and positions at regulatory elements, histones can be regulated by a layer of PTMs, giving rise to the so-called histone code (Kouzarides, 2007). This expression was conceived upon the observation that distinct combinations of hPTMs correlate with a different DNA function. This code represents a trace of past events in chromatin (for example transcriptional elongation) and thus serve as an epigenetic memory because it can be propagated through cell divisions (Kouzarides, 2007). This is particularly well established for X chromosome inactivation and heterochromatin formation. Another well-established epigenetic mark is cytosine methylation (Smith and Meissner, 2013), which together with histone marks mediate the cross talk among TFs and the DNA. On the other hand, some TFs can invade chromatin, modifying the accessibility of the underlying DNA molecule to other effectors among which factors able to read and write these modifications (Gardner et al., 2011).

1.4.3 Chromatin modifications at *cis*-regulatory elements

1.4.3.1 Post-translational modifications of the histones

Although PTMs have been detected at more than 60 different residues on the histones tails, this still represents an underestimate of the real number. The residues involved are mainly

lysines and arginines (and to a lower extent serines, threonines and tyrosines). The most studied modifications are acetylation and methylation (which can be mono-, di- or tri- for lysines and mono- or di- for arginines), although many others have been found to play roles in chromatin condensation, transcription, repair and replication, with phosphorylation, ubiquitylation, sumoylation, ADP-ribosylation, and proline isomerization among the most notable. Some chromatin marks are mutually exclusive (e.g. the lysine 27 of the Histone H3, referred to as H3K27, can be methylated or acetylated) but in general the number of combinations of marks a single histone can acquire is virtually a much greater number than that observed so far (Kouzarides, 2007).

The availability of high-throughput (HT) approaches aimed at measure hPTMs genome-wide paved the way to the unfolding of the histone code. The very first map of chromatin marks in human T-cells highlighted a link between the chromatin status of a TSS of a gene and its transcriptional activity (Barski et al., 2007). Given population-averaged measurements (ChIP-seq), higher the transcriptional level of a gene higher the levels of H2A.Z (a variant of H2A) and H3K4me3 (and to some extent also of H3K4me2 and H3K4me1) around its TSS and also higher the H3K36me3 and H3K20me1 levels on its body. On the contrary, lower transcriptional activity is mirrored by higher H3K27me3 and H3K9me (di- and tri-) levels over the corresponding TSSs.

Heintzman and collegues (Heintzman et al., 2009) showed that distal enhancers and TSSs can be distinguished by the ratio between mono- and tri- methylation of H3K4. While high levels of H3K4me1 and low levels of H3K4me3 ($H3K4me1^{high}/me3^{low}$) are characteristics of distal regulatory elements, the opposite ratio ($H3K4me1^{low}/me3^{high}$) marks TSSs. Nevertheless, a study reported a positive correlation between the level of H3K4me3 and enhancer activity (Pekowska et al., 2011). Besides, histone acetylation (in particular H3K27ac) has been shown to distinguish active from poised enhancers (Creyghton et al., 2010). Poised elements show a methylation signature of enhancers but no sign of activity, either because of lack of activating signals or because of active repression. They can also be a footprint of previous activities (Ostuni et al., 2013). The concept of poised regulatory element was first denfined for the so-called bivalent TSSs. Co-occurrence of activating H3K4me3 and repressive H3K27me3 marks at TSSs was observed at lowly transcribed developmental genes in ESCs (Bernstein et al., 2006). Being in a transiently poised transcriptional state in ESCs, these genes can

either lose the activating or the repressive mark during differentiation, turning on or off their transcription (Mikkelsen et al., 2007). All these transitions requires the active egagement of enzymes able to modify histone tails, namely methyl-transferases and de-methylases as well as acetylases and de-acetylases (Gardner et al., 2011).

Pioneer studies showed that polII can often be recruited at *cis*-regulatory elements other than TSSs (De Santa et al., 2010, Kim et al., 2010). These studies showed that localization of polII at these sites was not only the result of polII being engaged in enhancer-promoter loops, but that it was productively transcribing enhancer RNAs (eRNAs). Since these products show almost no measurable evolutionary conservation, it remains unclear to what extent they are functional or if they are byproducts of polII engagement onto chromatin. It cannot be also excluded that the engagement of polII rather than its product could itself be of functional relevance. Further studies (Ørom et al., 2010, Lam et al., 2013, Li et al., 2013) used targeted degradation to demonstrate that at least some eRNAs are themselves directly implicated in gene activation.

Taken together, these evidences suggest that the transcriptional profile of regulatory elements is more complicated than initially thought. Since very recently core promoters and enhancers were defined over almost exclusive features. The current emerging picture points instead to a continuum among these two classes of regulatory elements.

Similar to TFs, the specific function of hPTMs is likely to be quite context-specific and also not as clear-cut as initially thought. For example the role H3K4me3 as an activating mark has been recently challenged by a study in *S. cerevisiae* which shows that Set1-dependent H3K4 methylation acts as a gene repressor upon stress (Weiner et al., 2012).

1.4.3.2 DNA methylation

Cytosine residues in DNA can be methylated *in vivo* resulting in 5-Methylcytosine (5mC). This reaction is mediated by enzymes possessing DNA methyltransferase (DNMT) activity (Smith and Meissner, 2013). DNA methylation is considered a *bona fide* epigenetic mark, which can be faithfully inherited through cell divisions. Although most of the cytosine methylation occurs at CpG dinucleotides, cytosine can also be methylated in non-CpG context, as observed in ESCs (Lister et al., 2009).

Mammalian genomes are largely depleted of CpGs, and among them 60-80% are methylated.

Clusters of relatively high CpG density (termed CpG islands) are instead largely unmethylated throughout organism development. Nearly half of the islands overlaps known TSSs, while half are orphans (at least given the current level of annotation of the mammalian genomes). Only 21.8% of CpG islands, especially the TSS-distal ones, undergo dynamic changes in methylation during development (Ziller et al., 2013). On the contrary, dynamic changes have been found to occur more frequently in diseases, like cancer (Aran and Hellman, 2013). CpG methylation at TSSs has been associated with down-modulation of transcription and long-term gene silencing. This is in line with the pivotal role of CpG methylation in suppressing the transcription of transposable elements, thereby inhibiting their ability to spread (Smith and Meissner, 2013). Compared to promoters, enhancers show narrow tissue-specific activity which is reflected by characteristic patterns of hPTMs (Heintzman et al., 2009). Similarly, DNA methylation at enhancers exhibit tissue-specific patterns and it can directly exert its effects on TF binding, as shown for the Glucocorticoid Receptor (Wiench et al., 2011). In line with its importance, the methylation level of enhancers has been recently shown to be better correlated to aberrant expression of target genes in cancer, as compared to TSS methylation (Aran and Hellman, 2013).

1.4.4 A wider picture: chromatin states

The increasing availability of genome-wide maps of hPTMs and DNA methylation allowed the unsupervised segmentation of the cell-types specific epigenomes in dozens of functionally distinct compartments called chromatin states (Ernst and Kellis, 2010). As for single hPTMs, distinct chromatin states showed different extent of cell-type specificity (Ernst et al., 2011). Similar results were found profiling chromatin regulators (Ram et al., 2011). As expected, different combinations of histone marks co-occur with proteins that read, deposit and erase them. Interestingly, this study confirmed a previous observation in which counteracting enzymes (e.g. acetylases and de-acetylases) are found on overlapping sets of regions (e.g. acetylated active regulatory elements) (Wang et al., 2009), suggesting a deeper level of fine-tuning of hPTMs than previously thought.

1.4.5 Sequence-specific TFs and their interplay with chromatin determine cell fate

The majority of inferences made so far concerning chromatin states, transcription factors occupancy and transcriptional outputs are correlative. Strong efforts have to be made in order to understand to what extent the observed chromatin states represent not only a consequence of but also a prerequisite for recruitment (or exclusion) of the transcriptional machinery (Gardner et al., 2011).

In a way that is largely independent of the state of chromatin, at least a specific class of TFs is able to invade it: the pioneers (Magnani et al., 2011). As mentioned, they are able to recognize their sites even in chromatinized context and to recruit chromatin regulators such as ATP-chromatin remodeler factors. In principle, compared to TFs not showing this activity, pioneers should exhibit longer residence time on chromatin. From an experimental point of view, this can be assessed by FRAP, in which the fluorescence-tagged TF is irreversibly bleached by a laser pulse from an area of interest and the time needed for recovery is measured. Although this has been done for a limited amount of TFs, it has been demonstrated that FoxA1, which is central in endoderm specification (Sekiya et al., 2009) shows a recovery time higher than the other TFs tested but lower than the histone linker. According with this, Foxa2 has also been shown to be responsible for chromatin remodeling at nucleosome-occupied regulatory elements marked by H2A.Z during differentiation of ESC to endoderm/hepatic progenitors (EHP) (Li et al., 2012). Interestingly, around 15% of FoxA1 binding sites in interphase are not evicted during mitosis Caravaca et al., 2013. Unless to a lesser extent compared to Cohesin (Yan et al., 2013), this suggests a role also for pioneer factors (and not for other factors, e.g. Klf5, Hnf4a and Myc, see Caravaca et al., 2013) in mitotic bookmarking of regulatory elements. This is also consistent with the strong enrichment of motifs for pioneer factors (e.g. Fox- and Ets- related factors) observed in the sequence of the Cohesin clusters (Yan et al., 2013).

The pioneering activity of at least a fraction of them candidates TFs as main drivers in the determination of the transcriptional landscape of a cell, which in turns governs its fate. In fact, the over-expression of a single cell-type specific TF or a combination of them has been demonstrated to be sufficient in order to induce re-programming of cells from a lineage to another (Pereira et al., 2012). Nevertheless, it has been shown that they are not sufficient

to completely erase the epigenetic memory of the cell of origin (Lister et al., 2011), which in turn can affect TF binding (e.g. DNA methylation) (Wiench et al., 2011).

If TFs rule cell fate, this implies that determinants of TF binding are at its foundation. Even though they are sequence-specific by definition, given the size of mammalian genomes, the sites that are bound *in vivo* are only a minor fraction compared to the hundreds of thousands of sequences matching high-affinity TF recognition sites (Pan et al., 2010). This is apparently in contrast to their non-ambiguous binding profiles. Using information theory (Wunderlich and Mirny, 2009) it was shown that TF-binding motifs are not instructive enough to avoid spurious hits to the mammalian genomic background, calling for the presence of additional genetic or epigenetic features in order to achieve their specificity.

Considering chromatin determinants, two studies (Nili et al., 2010, Tillo et al., 2010) spotted a positive correlation between *in vitro* local nucleosome occupancy (predicted as well as experimentally verified) and engagement of binding sites for multiple human TFs. Namely, contacted sites showed an intrinsic propensity for the region to be wrapped into a nucleosome, compared to similar sites never found to be contacted by the TFs *in vivo*. More recently, the correlation of *in vitro* binding preferences with the in vitro nucleosome occupancy of the same DNA stretch has been tested for 137 sequence-specific DNA-binding proteins in *S. cerevisiae* (Charoensawan et al., 2012). 98 out of 137 have been found to be positively correlated, with transcriptional activators among the most strongly correlated. Another piece of evidence came recently from Winter et al., 2013, in which the authors observed that open Dnase I hypersensitive sites are often occupied by rotationally stable nucleosomes in cell types where the same site is not accessible. Along with the observation that Progesterone Receptor contacts sites pre-marked by a nucleosome (Ballaré et al., 2012), these studies suggest that particular nucleosomal configuration (occupancy and positioning of short genomic regions) and TF-binding sites which can be productively engaged might be intrinsically imposed by the genomic sequence.

1.5 Determinants of transcription factor binding

Understanding to what extent the sequence information encoded in the genomic DNA itself specifies its regulatory properties is not only a challenging task. The answer to such question

would shed light on the relative contribution of genetics and epigenetics to regulation. As mentioned in the previous paragraph, given the size of mammalian genomes, the sites that are bound *in vivo* correspond only to a small fraction of the hundreds of thousands of sequences matching high-affinity TF recognition sites (Pan et al., 2010).

A general observation is that, in order to be bound, a motif should be available to the TF. What does available mean? Many studies have shown the high power of DNA hypersensitivity to nuclease digestion in predicting TF binding (Kaplan et al., 2011, Arvey et al., 2012). Despite the encouraging results, DNA hypersensitivity represents a readout of some previous remodeling event. Not being independent on the binding itself, it is not correct considering hypersensitivity as a determinant of binding, especially in case of pioneer factors and master regulators (the former being able to invade inaccessible chromatin, the latter being expressed very early during differentiation). Availability of a certain binding site to its cognate TF reflects the contribution of the sequence itself (i.e. presence of binding sites for partner TFs) as well as of the epigenetic signature of the region, which is a legacy coming from a previous stage of differentiation or exposure to environmental insults (Ostuni et al., 2013).

Multiple evidences (described in the next paragraph) suggest that engaged binding sites reside in a peculiar sequence context that can be responsible for directing TF-binding. If this is true, it should be possible to distinguish engaged from non-engaged sites using predictors trained on the genomic sequence alone.

1.6 Predictions of transcription factor binding from the sequence

The problem of predicting transcription factor binding starting from the local sequence of mammalian genomes has been successfully addressed in two recent papers (Yáñez-Cuna et al., 2012, Arvey et al., 2012). While both papers tackled cell type specificity of binding for different TFs and co-regulators (e.g. the acetyltransferase p300) only Arvey and collegues (Arvey et al., 2012) compared real binding events to a negative set (namely nearby regions) but applied the method only to the best 1,000 ChIP-seq determined contacted sites.

Yáñez-Cuna and collegues (Yáñez-Cuna et al., 2012) used motifs representing published binding preferences for known TFs. Arvey and collegues (Arvey et al., 2012) applied instead a

completely unbiased approach which considers degenerated k-mers. The first method reached AUCs between 0.62 and 0.95 in discriminating cell-type specific binding, depending on the dataset considered. The second one got AUCs between 0.5 and 0.95 in discriminating real binding to a negative set.

None of the approaches explicitly considered other features of the genomic sequence that have been found to be related to TF engagement at its consensus sequences. A recent study (Kwon et al., 2011) highlighted a higher C+G content in muscle-specific CRMs that could be validated in functional assays (compared to predicted candidates that could not be validated), which was also reflected at the level of C+G-rich di-nucleotides. This is in accordance with the observation that higher C+G content at engaged elements compared to non-engaged ones favours higher nucleosome occupancy.

Predictions of the three-dimensional DNA shape induced only by sequence has been recently shown to improve the correct classification of binding sites contacted *in vivo* by bHLH TFs in *S. cerevisiae* (Gordân et al., 2013). The authors found that DNA shape features are able to recapitulate the boost in predictive power that can be achieved using positional 2-mers and 3-mers. The great advantage of using DNA shape is that it can capture the same information in a relatively small number of features compared to the whole set of positional 2/3-mers. Generally speaking, DNA shape is strictly related to correct protein-DNA recognition (Rohs et al., 2009). The inclusion of these kind of measurements in the prediction captures the fact that degenerated binding motifs can form very similar three-dimensional shapes, an information that otherwise will be missed.

1.7 Determinants of nucleosome occupancy and positioning

Nucleosome positions are usually described through occupancy and positioning. Given a certain strecth of DNA in a population of cells, the occupancy defines the probability it is wrapped into a nucleosome. Given similar occupancy, the DNA can slide along the histone octamer, resulting in different conformations. The less conformations the nucleosome can assume, the better its positioning, and vice versa (see figure 1.2). Similarly to TFs selecting their binding sites, it was hypothesized that a set of rules governing the nucleosome conformations across a genome must be in place.

In the current view rules are governed by three main variables (Struhl and Segal, 2013):

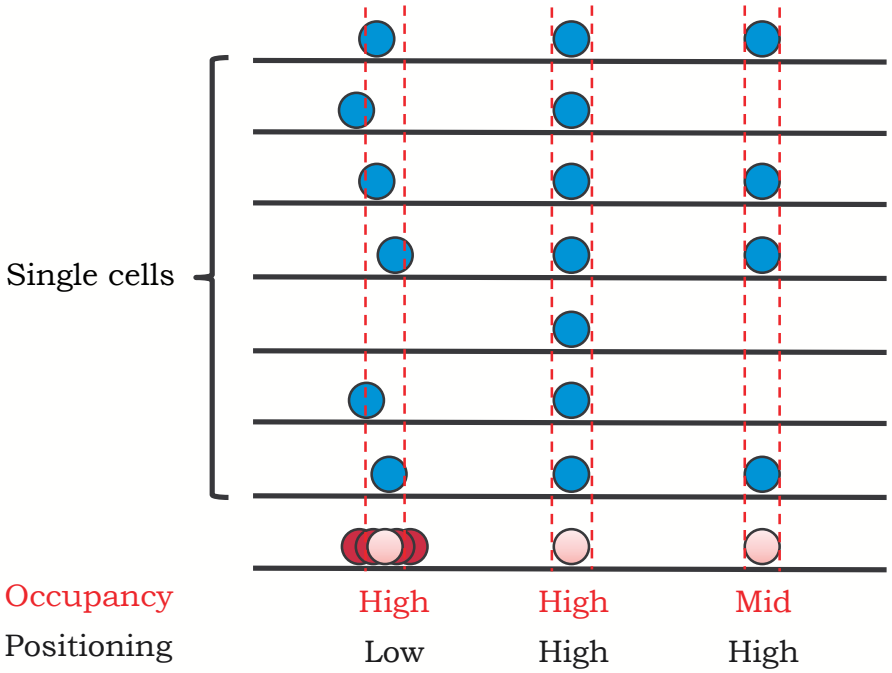

FIGURE 1.2: Nucleosome occupancy and positioning are two descriptors of nucleosomal DNA. Three situations are depicted in the figure. On the left, a region showing a high probability of been wrapped into a nucleosome, while the histone octamer can assume different rotational positions along the DNA fiber (high occupancy, low positioning). In the middle, a region displaying a similar occupancy but in which the histone octamer is not able to rotate along the DNA fiber is reported (high occupancy and positioning). On the right, a region with lower occupancy compared to the other two regions, but with high capability to position nucleosomes.

DNA sequence, *trans*-acting factors (including TFs and the transcriptional machinery) and ATP-dependent chromatin remodeling enzymes (see figure 1.3). The role of DNA sequence in nucleosome occupancy has been the object of a long controversy (reviewed in Struhl and Segal, 2013) concerning the relative role of nucleotide composition (Segal et al., 2006), DNA-bound barriers (Mavrich et al., 2008) and remodelers-driven nucleosome packing against barriers (Zhang et al., 2011) in determining nucleosome patterns *in vivo*. It is now demonstrated that each of these mechanisms contributes to the control of nucleosomal organization and that sequence-driven nucleosome assembly can be overcome by *trans*-acting factors at specific locations in the genome, e.g. at the +1 nucleosome relative to the TSS (Zhang et al., 2011).

Pioneer studies showed a much larger contribution of the sequence to occupancy (Segal et al., 2006, Segal and Widom, 2009) than positioning (Zhang et al., 2009). According to these findings, while the information for NFR formation at TSSs is encoded in the sequence, predictions of the exact positions are only modest (Yuan and Liu, 2008), pointing to the

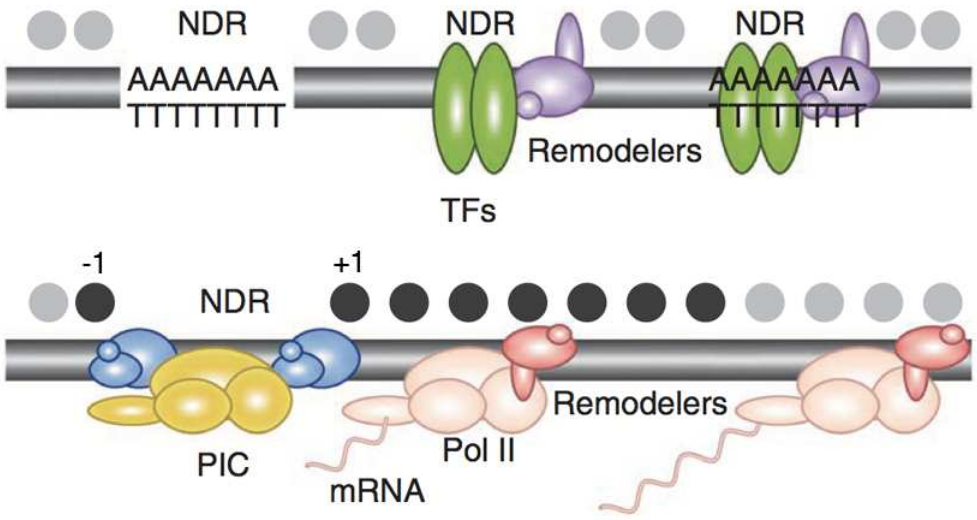

FIGURE 1.3: Gray circles indicate nucleosomes, dark grey ones represent better positioned nucleosomes. poly(dA:dT) tracts and/or transcription factors can generate NDRs (upper panel). By statistical positioning and action of chromatin remodeling complexes, arrays of nucleosomes are positioned at the sides of NDRs. At TSSs (lower panel) PIC-associated ATP-dependent chromatin remodeling factors are able to constraint positions of the +1 and -1 nucleosomes (Zhang et al., 2011). The positioned nucleosomes into the gene body are also dependent on chromatin remodelers complexed with the elongating polII (Yen et al., 2012). Adapted from (Struhl and Segal, 2013).

importance of trans-factors in fine-tuning nucleosome positions.

The pattern of dinucleotides is correlated with (and often responsible for) nucleosomal organization at different resolution. At a fine-grained level a 10 bp periodicity of AA/TT/TA that oscillate in phase with GC dinucleotides affects the DNA-histone octamer contacts (Satchwell et al., 1986) while at a coarse-grained level (hundreds of bp) an overall increase in GC content is in general a favorable condition for nucleosome formation (Tillo and Hughes, 2009), but not for positioning. Besides, a recent study split the human promoters into homogeneous groups by C+G content, and analyzed their *in vitro* capability to assemble into nucleosomes (Fenouil et al., 2012). Interestingly, going from low to intermediate CpG content the regions show increasing capacity of nucleosome assembly, which drops dramatically at highest C+G content. Another coarse-grained feature are Poly(dA:dT) tracts, which are virtually nucleosome-excluding sequence. They form stiff structures unable to bend around the histone octamer (Nelson et al., 1987, Segal and Widom, 2009). This in fact accounts for nucleosome depletion at poly(dA:dT) sequences commonly found in *S. cerevisiae* gene promoters. In human cells, nucleosome-repelling poly(dA:dT) tracts flanking moderately (dG:dC)-rich regions delimit *container sites*, defined as sequences able to accommodate positioned nucleosomes in *in vitro* assembly experiments (Valouev et al., 2011).

Beside sequence itself, fixed barriers on chromosomes generate adjacent ordered arrays of nucleosomes, first described under the term statistical positioning (Kornberg, 1981). A positioned nucleosome or another DNA-binding protein or complex as well as a repelling poly(dA:dT) tract can act as barriers. Nucleosome positiosing sequences and poly(dA:dT) tracts upstream of TSSs were first ascribed as the barriers responsible for the positioning of the +1 nucleosome, from which an array of regularly spaced nucleosomes emanate (Yuan and Liu, 2008). While the NFR upstream TSSs is largely encoded by the sequence (Yuan and Liu, 2008), *in vitro* reconstitution of chromatin only partially recapitulate the *in vivo* pattern. It has also been recently shown that the polII complex is not responsible for the maintainance of this pattern. In fact α-amanitin treatment (which causes the release and degradation of the polII complex from chromatin) showed only minor effects on human T-cells (Fenouil et al., 2012). Proper nucleosome positioning, spacing, and occupancy levels at 5' ends of most yeast genes was achieved by adding nuclear extract and ATP to the reaction of *in vitro* reconstitution of chromatin (Zhang et al., 2011). The same authors recently showed a position-specific role (relative to the NFR) for many ATP-dependent chromatin remodelers (Yen et al., 2012). Arrays of positioned nucleosomes have also been shown to emanate from sites bound by TFs (Kundaje et al., 2012). Although available data point to the coordinated action of TFs as barriers and DNA sequence constraints on nucleosome positions, the contribution of co-factors (which can alter the local chromatin environment through hPTMs) and ATP-dependent chromatin remodeling at TSS-distal *cis*-regulatory elements still remains to be investigated.

Although during recent years the determinants at the foundation of genomic nucleosome patterns have started to be elucidated, the debate is still in its very infancy about the fraction of nucleosomes in mammalian genomes showing reproducible positions. Valouev and collegues (Valouev et al., 2011) concluded that the majority of the human genome showed substantial flexibility of nucleosome positions. Using an unprecedented amount of data, a recent study (Gaffney et al., 2012) challenged this view and found that most nucleosomes have more consistent positioning than expected by chance and around 9% of them show moderate to strong positioning. Complex questions like this one will be better tackled as soon as the experimental procedures will be more standardized, the data throughput as well as the number of organisms and cell types studied will increase, and the computational approaches will be

more powerful.

1.8 Predictions of nucleosomal patterns from the DNA sequence

Computational models using sequence features to predict nucleosome occupancy have been described. While these have demonstrated that DNA sequence alone specify nucleosomal preferences, its quantitative contribution to *in vitro* and *in vivo* patterns is still under debate (Struhl and Segal, 2013).

Segal and collaborators (Segal et al., 2006) used a collection of 199 mononucleosome (142-152 bp in length) DNA sequences to construct a probabilistic model representing the DNA sequence preferences of the histone octamer in *S. cerevisiae*. This model is slightly more complicated than a PWM for TFs but it is able to recapitulate the most important features of chromatin structure of yeast. Being learned from *in vivo* data, this model could be influenced by the sequence preferences of other factors and by chromatin remodeling activities. The same authors devised a refined model (Kaplan et al., 2008) completely derived from *in vitro* data and applied it to the prediction of nucleosome occupancy *in vivo*. Performances in cross-validation achieved Pearson correlation coefficients of 0.89 and 0.75 for the *in vitro* and *in vivo* maps respectively. It is important to point out that a correlation of 0.75 correspond to a coefficient of determination (R^2) of around 0.56, namely the learned *in vitro* preferences are able to explain about 56% of the variability in the nucleosome patterns observed *in vivo*. Interestingly, the use of C+G content alone was later shown to give an R^2 of 0.5 (Tillo and Hughes, 2009), not far from the performances of the more complex model. When moving from *S. cerevisiae* to mammalian genomes, the model performances drop to a correlation coefficient of 0.28 for human CD4 T-cells (Tillo et al., 2010). Namely the *in vitro* model can explain less than 10% of the variability observed in the human nucleosome pattern. More recently, a statistical mechanics model was shown to outperform those methods on *in vivo* occupancy data from *S. cerevisiae* and to be able to recognize known NPSs (van der Heijden et al., 2012).

SVMs (Peckham et al., 2007) as well as models taking advantage of wavelet analysis to

extract spatially periodic signals (Yuan and Liu, 2008) were successfully applied in discriminating NPSs from NFRs in *S. cerevisiae*. Comparative genomics from six *Saccharomyces* genomes was also successfully employed to derive nucleosome positioning sequence patterns (Ioshikhes et al., 2006).

While these studies showed that DNA sequence is sufficient to partially predict nucleosome occupancy *in vitro* and *in vivo*, the contribution of the sequence to positioning is still ambiguous. Despite a recent report (Gaffney et al., 2012) and excluding some precise genomic locations (e.g. the +1 relative to TSSs) even the fraction of *in vivo* well-positioned nucleosomes remains elusive (Peckham et al., 2007, Zhang et al., 2009, Valouev et al., 2011).

1.9 Regulation of transcription in murine macrophages

Bone marrow derived macrophages (BMDMs) from *M. musculus* represent a very suitable system to study regulation of transcription. They can be differentiated from bone marrow giving rise to a very homogeneous population that can be polarized *in vitro* under pro- or anti- inflammatory stimulation, with little variability across single cells compared to other systems. This results in massive reorganization of chromatin and transcription on a very short time scale (Lawrence and Natoli, 2011).

1.9.1 TFs in the hematopoietic system

Hematopoiesis (see figure 1.4) is the process of proliferation, differentiation and maturation of all blood cells types. The primary organs involved in hematopoiesis during embryogenesis are the yolk sac and later the aorta-gonad mesonephros (AGM) region, the placenta and the fetal liver (Orkin and Zon, 2008). The fetal (or primary) hematopoiesis progresses toward the definitive hematopoiesis when the hematopoietic stem cells (HSC) migrate toward newly developed long bones. In *H. sapiens* the bone marrow is the primary organ of hematopoiesis. Although this process is conserved throughout vertebrate evolution the sites of primitive and definitive hematopoiesis differ among species (Orkin and Zon, 2008). HSCs are characterized by self-renewal, and the capacity to proliferate and differentiate into progenitors of each of the blood cell lineages. Recent studies have questioned the classical hierarchical organization

in which progenitors arise in an orderly manner from a HSC. HSCs are more plastic than previously thought and can be seen as groups of cells with varying developmental potentials (Orkin and Zon, 2008).

The balance among lineages is ensured by the expression of combinations of few TFs (see figure 1.4). These factors go under the name of master regulators for their ability of switching on or off cell-type specific transcriptional programs. Pax5 is for example required for proper B cell differentiation (Nutt and Kee, 2007). In its absence, pro-B cells are not committed to the B cell lineage but instead become capable of differentiating into a broad spectrum of hematopoietic cell types. Besides, cell fate determination has also been shown to be a function of TFs concentration. High-levels of Pu.1 (the corresponding gene is named SPI1 in *H. sapiens* and Sfpi1 in *M. musculus*) promote macrophage differentiation, whereas low-levels direct B cell formation (DeKoter and Singh, 2000). Besides, Pu.1 is expressed in other specialized populations of cells in many different tissues (as reported in figure 1.5). Pu.1 is not only capable of lineage conversion among blood cells. It has also been shown to be capable of direct reprogramming of fibroblasts to macrophages when combined with C/EBPα/β (Feng et al., 2008).

1.9.2 Pu.1: one of the master regulators of macrophage differentiation

We have recently shown that Pu.1 is not only a trigger for macrophage differentiation but it does that through the supervision of most of the cistrome (Ghisletti et al., 2010). This expanded the view about master regulators, which are not only TFs responsible for cell fate (switching on or off cell-type specific transcriptional programs) but are also supervisors of the majority of regulatory elements (enhancers as well as promoters) in the genome (Natoli, 2010). Similar findings were obtained in other models of differentiation, e.g. considering MyoD in skeletal muscle (Cao et al., 2010).

Pu.1 is expressed from very early stages of hematopoietic differentiation (Back et al., 2005). As mentioned, its effect on cell fate are context- and dose- dependent (DeKoter and Singh, 2000). Pu.1 expression increase along the myeloid lineage, reaching its maximum level in monocytes. While essential for macrophage identity, it also plays a central role in B- and early T- cells differentiation (Zhang et al., 2012). The reciprocal regulation between Pu.1 and Gata1 is instead responsible for the priming of multipotent progenitors to myelolymphoid or

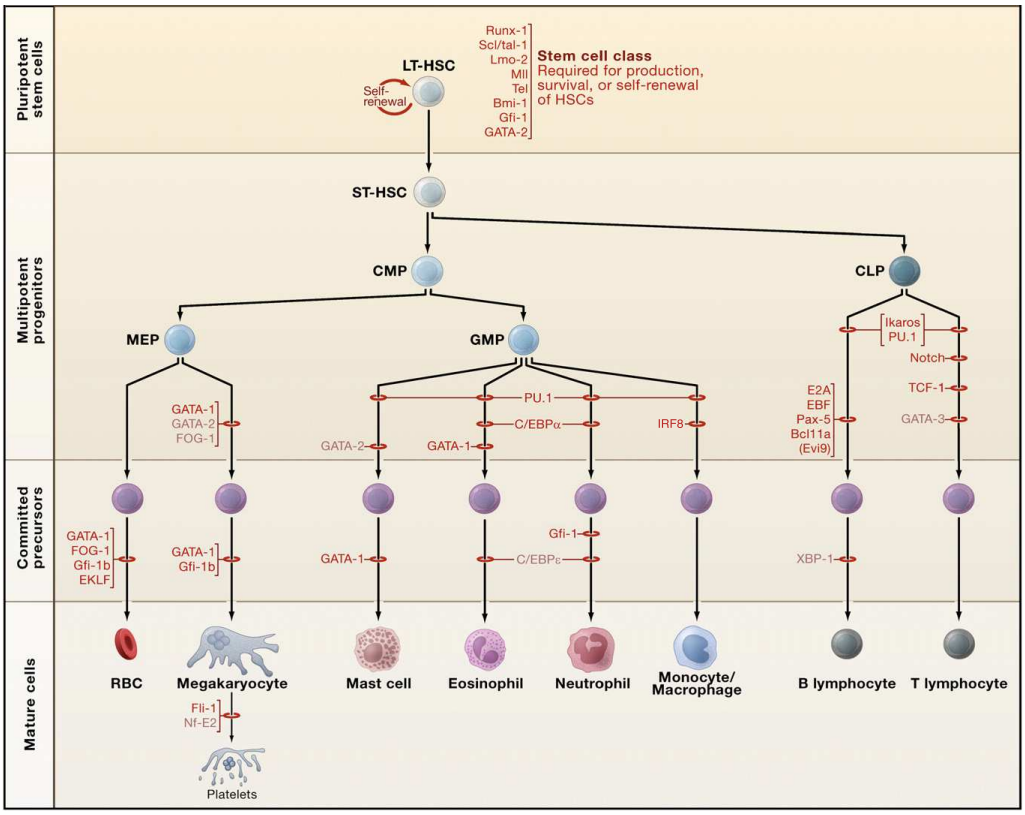

FIGURE 1.4: In the current view, hematopoiesis is hierarchically organized. Master regulators through different lineages are highlighted in red. Red bars indicate the stages at which hematopoietic development is blocked by the knockout of a given TF. Abbreviations: LT-HSC, long-term hematopoietic stem cell; ST-HSC, short-term hematopoietic stem cell; CMP, common myeloid progenitor; CLP, common lymphoid progenitor; MEP, megakaryocyte/erythroid progenitor; GMP, granulocyte/macrophage progenitor; RBCs, red blood cells. Adapted from Orkin and Zon, 2008.

myeloerythroid progenitor populations (Arinobu et al., 2007).

Given its very early expression a pioneering activity of Pu.1 can be envisioned. Although this has never been formally demonstrated as for other factors (Sekiya et al., 2009), indirect evidences suggest that Pu.1 can act as a pioneer factor. By ectopic expression in fibroblasts, Pu.1 is able to drive partial reprogramming to macrophages (Feng et al., 2008). In this context regions devoid of H3K4me1 but acquiring Pu.1 upon ectopic expression gain this mark (Ghisletti et al., 2010). These data are in agreement with an independent study, in which Pu.1 was fused to the estrogen receptor ligand-binding domain. After 24h of tamoxifen treatment, 43% of the induced Pu.1 sites gained H3K4me1, 32% consisted of induced sites that were marked by pre-existing H3K4me1 and 25% were H3K4me1 negative (Heinz et al., 2010). The same study also showed that induction of Pu.1 led to nucleosome remodeling, resulting in further expansion of the NDR centered on the Pu.1 binding site. Nevertheless, the observed remodeling occurs in regions showing an already partial nucleosomal organization,

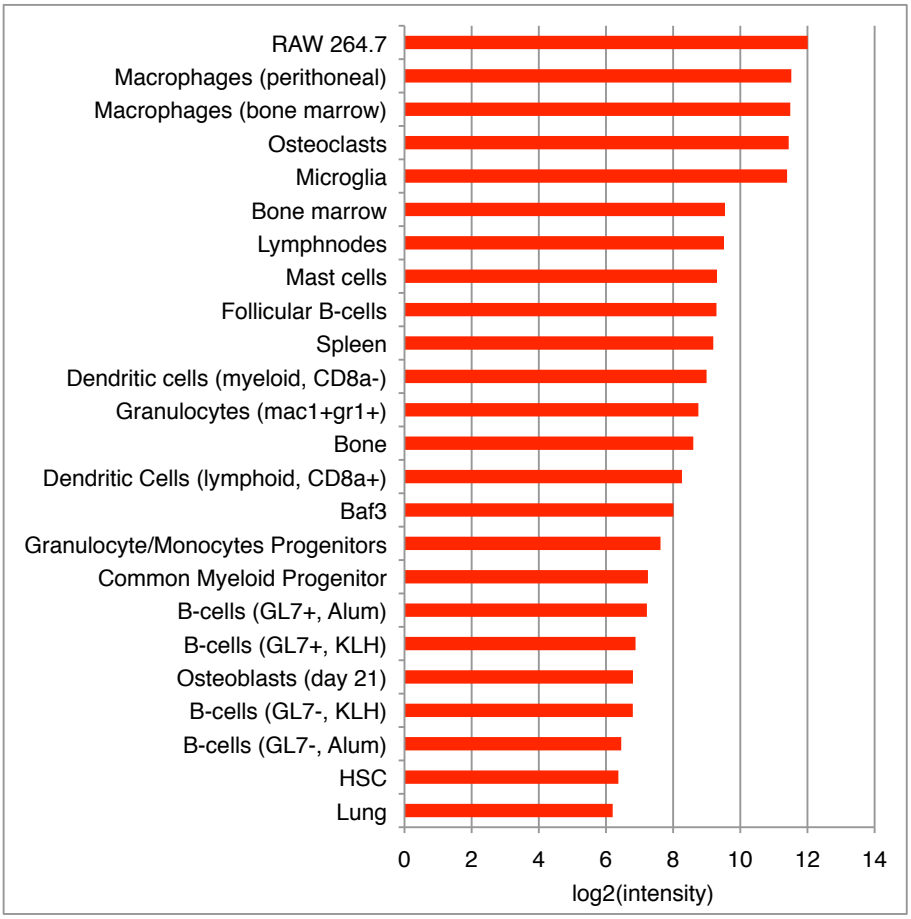

FIGURE 1.5: mRNA levels for Pu.1 in murine tissues/cell lines expressing it (those showing a microarray measured log2-intensity value of at least 6). Data from Wu et al., 2013.

possibly due to pre-existing low levels of PU.1 binding (Heinz et al., 2010). These findings need to be confirmed in a more physiological setting.

Of note, although the major expansion of the regulatory landscape driven by Pu.1 takes place during differentiation, we showed that in macrophages this landscape is plastic and can be expanded upon environmental insult (Ostuni et al., 2013).

1.9.3 How does Pu.1 select its binding sites *in vivo*?

While a fraction of TFs bind only to a pre-determined landscape (Lefterova et al., 2010, Mullen et al., 2011), other factors are able to determine this landscape. Master regulators, some of which have demonstrated pioneering activity (Magnani et al., 2011), are among those. Nevertheless, there are exceptions. Foxp3, the master regulator of regulatory T-cells, has been recently shown to bind to a pre-determined repertoire of enhancers (Samstein et al.,

2012). Given their ability to identify their target sites in a yet unmodified chromatin context, it is intriguing to hypothesize the binding of master regulators as mainly driven by sequence determinants. In line with this hypothesis, Pu.1 binding maps could be used to estimate to what extent these events are driven by sequence determinants.

Pu.1 binds a 10-nt long core (consensus sequence: AGAGGAAGTG) that has been experimentally determined using an *in vitro* microwell-based assay (Wei et al., 2010). Considering high-affinity binding sites there are 5 to 10 times more of them in the genome compared to sites bound *in vivo* (Pham et al., 2013). In this context, how can Pu.1 achieve specificity in engaging its consensus sites? How much of this information is encoded in the local sequence context?

In this thesis, we address these questions using a collections of contacted sites from multiple cell-types (namely we looked for the general sequence determinants of binding at the expense of the cell-type specific information). This means we gathered the *in vivo* available data (ChIP-seq) in *M. musculus* in order to get the wider set of Pu.1-contacted sites irrespectively of the cell type. We then calculated features in the sequence of the nearby bound and unbound Pu.1 binding sites and used this information for machine learning. The more information encoded in the sequence, the higher the accuracy of the approach. This means that our result represents only a lower bound as we expect more accurate predictions as soon as our ability in measuring new relevant sequence features improves.

A recent paper (Pham et al., 2013) shed some light into the mechanisms of Pu.1 binding sites selection in *H. sapiens*. Authors found that those canonical ETS sites showing the highest affinity have a higher probability of representing autonomous binding events. These sites also show a low DNase I accessibility. On the contrary, lower affinity sites are more accessible to DNase I digestion and are more likely to be found in clusters of motifs for partner TFs. Interestingly, unbound sites are enriched in gene deserts, suggesting a role for higher order chromatin structure.

1.10 Experimental approaches to probe *cis*-regulatory elements and chromatin

The main function of TFs is to affect transcriptional rate of their target genes, namely driving tissue-specific quantitative change. In case of core promoters, RNA is a direct readout of its functionality. In case of other regulatory elements, like enhancers or silencers, the problem is complicated by the precise identification of their targets. Chromosome Conformation Capture Carbon Copy (5C) data from 1% of the human genome (Sanyal et al., 2012) estimated that only 27% of the distal elements have an interaction with the nearest TSS, and 47% of elements have interactions with the TSS of the nearest expressed gene. From a quantitative point of view, this is complicated by the fact that any enhancer can be engaged in multiple regulatory loops.

In this context, surrogate assays to probe the effects of putative regulatory elements on transcription have been employed. For example, in the luciferase reporter assay, the region of interest is cloned in a plasmid, upstream of a luciferase gene under the control of a minimal promoter (which is constitutively expressed). After transfection in a cell line luciferase activity is assessed and used as a readout of the enhanced or reduced transcriptional rate compared to the constitutive promoter alone. STARR-seq (self-transcribing active regulatory region sequencing) represent an evolution of the luciferase assay, in the sense that it is HT and allows direct measurement of the RNA (Arnold et al., 2013). A genome-wide reporter library from randomly sheared genomic DNA of *D. melanogaster* was placed downstream of a minimal promoter, such that active enhancers transcribe themselves. In this way, the strength of each enhancer can be directly inferred by the number of sequences mapping to it (Arnold et al., 2013). A less quantitative but more physiological assay use stable transgene reporter. This has been successfully applied to the validation of CRMs predicted to drive precise patterns of expression during development in *D. melanogaster* (Zinzen et al., 2009).

In order to get a detailed description of a single regulatory element as well as genome-wide regulatory maps, different techniques have been flourished during the years. Since many of them recently moved from single-locus to genome-wide analyses, Computational Biology has become essential to handle and interpret the large amount of data generated.

1.10.1 DNA accessibility assays

A positive correlation between sensitivity to endonuclease and regulatory DNA has been known since almost 30 years ago (Gross and Garrard, 1988). Increased sensitivity is often referred as to increased accessibility. Euchromatic (including regulatory) regions are in general less condensed (and thus more accessible to endonucleases) than heterochromatic ones. More specifically, DNase I hypersensitive sites (DHSs) are regions of chromatin which are sensitive to cleavage by the DNase I enzyme (Gross and Garrard, 1988, Thurman et al., 2012). Sensitive regions generate smaller digested fragments (<<1kbp), which can be purified and subjected to HT-sequencing. In this way, regulatory elements can be mapped genome-wide with basically no bias. Increasing the throughput from tens to hundreds of millions sequenced fragments allows a further increase in resolution, namely the identification of the so-called footprints. These are short (tens of bp) stretches of DNA that were protected from digestion by the presence of TFs or more in general by DNA-binding proteins. They virtually encompass the entire *cis*-regulatory repertoire (either active or poised elements) of a given cell, allowing the characterization of its regulatory network (Neph et al., 2012b).

Similarly, digestion with restriction enzymes has been combined with HT-sequencing (NA-Seq), providing an alternative method to monitor genome-wide the status of chromatin during differentiation and disease (Gargiulo et al., 2009).

Another approach termed Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE) (Giresi et al., 2007) was shown to be complementary to DNase I digestion. It is in fact also able to capture TSS-distal regulatory regions that DNase I enzyme cannot easily digest (Song et al., 2011). Nevertheless, footprints are precluded to FAIRE, which allows only a coarse-grained view of the accessibility landscape.

1.10.2 Chromatin immuno-precipitation

Chromatin immuno-precipitation (ChIP) is a technique aimed at probing DNA-protein interactions *in vivo*. This allows the study of the state of chromatin or the binding of transcription factors to it. Chemical cross-linking by formaldehyde is followed by precipitation with a specific antibody (Orlando et al., 1997). After de-cross-linking, enrichment at a single locus can be assessed by PCR, or the resulting material can be either hybridized to a custom array

(ChIP on chip) or sequenced (ChIP-seq). The advent of HT-sequencing technologies allowed extensive characterization of the regulatory DNA in mammals (Dunham et al., 2012). From a technical point of view, the main limiting factor of the technique is the availability of high quality antibodies.

Due to the very nature of the assay, ChIP is showing population- and time-averaged signals and thus is not suited to capture the dynamics of the events. Consider for example a CRM found to be bound by two different TFs. From a mechanistic point of view, the co-occupancy of TFs does not necessarily mean that a cooperative binding is occurring there. It could well be the result of a series of dynamic events (which could even be related to each other by some memory signal, e.g. a first factor reside on chromatin for a short time but leave a long lasting modification that gives directionality to the sequence of events). ChIP just freezes a picture of many different metastable states. Integration with co-immunoprecipitation, measurement of residence time on chromatin and imaging techniques like FISH is needed in order to gain a better understanding of the real scenario.

ChIP is in principle a quantitative technique. Given a population of cells, the more a certain DNA region is likely to be complexed with a given protein, the higher the signal. The relation among occupancy and functional engagement is far from being thoroughly investigated. A recent study in *S. cerevisiae* (Lickwar et al., 2012) showed that in case of Rap1 stronger enrichments not necessarily correlate with function. Rap1 residence time on chromatin has been instead linked to transcriptional activation. TF-occupancy measured by ChIP is only poorly correlated to residence time ($R^2 = 0.14$) and thus not predictive for a region to be functional.

1.10.3 Determination of nucleosome positions

Micrococcal nuclease (MNase) preferentially cuts linker DNA rather than DNA wrapped in nucleosomes. The digestion of native chromatin by MNase has become the standard approach to cut down the chromatin fiber into single mononucleosome. Nevertheless, a more accurate chemical method has been recently proposed (Brogaard et al., 2012), which also overcome another issue related to MNase, namely its slight cleavage preference for TA/AT dinucleotide which could bias the precise determination of nucleosome positions.

DNAse I is also been shown to digest nucleosomeal DNA with a 10 bp periodicity, according to

the exposure of the minor groove as it wraps around histones. DNAse I digestion patterns are indeed correlated to the positions of nucleosomes. Leveraging this principle over a pool of 49 samples (originating a set of 1.5 billions of short reads) coupled to a machine learning approach revealed that around 30% of the human genome is associated with regions of nucleosomal stability (Winter et al., 2013). As already mentioned, the fraction of the genome that is able to restrict nucleosome rotational positions is under dispute. Estimates from previous studies ranged from 20% (Valouev et al., 2011) to almost all of the genome being constrained (Gaffney et al., 2012).

Chapter 2

Methods

2.1 Pu.1 ChIP-seq in murine macrophages

Pu.1 ChIP-seq in bone marrow derived murine macrophages (BMDMs) was performed in the lab (refer to section 2.12.1 for details on the experimental protocol) using an anti-Pu.1 rabbit polyclonal antibody generated in-house against the N-terminus of the murine Pu.1 (aa. 1-100; NP_035485.1) and affinity purified (Ostuni et al., 2013).

After quality filtering, 51 nt long reads were aligned to the mm9 release of the murine genome using Bowtie v0.12.7 (Langmead et al., 2009). Only unique alignment were retained, allowing up to two mismatches compared to the reference genome (options -m 1 -v 2). Peak calling was performed using MACS v1.4 (Zhang et al., 2008) using a bandwidth (bw parameter) of 100 (bp). Cell type specific input was used as control. A golden set was defined by filtering peaks with a p-value lower than or equal to 1e-10. This dataset was annotated over Ensembl genes (Flicek et al., 2012) using GIN (Cesaroni et al., 2008) (*priority* set to "gene" and *promoter definition* to "-20,000"). The coordinates of the genes were downloaded from the UCSC genome browser (Fujita et al., 2011) on 2011, July 7^{th}. Peaks within +/- 2.5 kbp from TSSs were considered as TSS-proximal while all the others were defined as TSS-distal. In order to visualize the raw profiles on the Genome Browser (Flicek et al., 2012), wiggle files were generated with MACS v1.4 and converted to bigWig.

2.2 *in vitro* Pu.1 ChIP-seq data analysis

Analyses were performed as described in section 2.1 but considering a lower statistical threshold for the peak calling (p <= 1e-5). Nucleosomal occupancy over the sites was calculated as the number of paired-end fragments (determined by Mnase digestion followed by HT-sequencing) spanning the experimentally determined Pu.1 summits.

2.3 A collective *cis*-regulatory repertoire bound by Pu.1

Every murine ChIP-seq dataset of sufficient quality available in the literature was downloaded from the Gene Expression Omnibus (Barrett et al., 2013) (see table 2.1) and analyzed as described in section 2.1. Cell type specific inputs were used as control (see table 2.1). Genomic tracks were generated using MACS (Zhang et al., 2008) and normalized to the same sequencing depth for visualization. All the ChIPs considered were carried out with the same antibody (Santa Cruz SC-352), with the exception of the BMDMs-derived dataset generated in our lab (section 2.1), which was also included in the following analysis.

In order to define regions bound by Pu.1 in at least one of the seven cell types considered,

IP	Input/IgG	Decription	Cell type	Mouse strain
GSM538017	GSM537988	BMDM	BMDM	C57BL/6
GSM537983	GSM537988	ThioMac	Perithoneal Macrophages	C57BL/6
GSM774291	GSM774298	FLDN1	Thymocytes (FLDN1)	C57BL/6
GSM774292	GSM774299	FLDN2a	Thymocytes (FLDN2a)	C57BL/6
GSM774293	GSM774300	FLDN2b	Thymocytes (FLDN2b)	C57BL/6
GSM539537/8	GSM539550	ProB (2 rep)	Pro-B cells	38B9 (cell line)
GSM537989	GSM537993	Bcells	B-cells	C57BL/6

TABLE 2.1: List of the Pu.1 ChIP-seq datasets collected from the literature. The first column refers to GEO accession numbers (Barrett et al., 2013). FLDN stands for Fetal Liver Precursor Derived.

the binding events from different experiments were combined. First of all, the enriched regions were further split into their components (dense homotypic clusters, which often span up to few kilobases, are recognized by MACS as a single highly-enriched region). To this aim, PeakSplitter (Salmon-Divon et al., 2010) was run on the individual ChIP-seq profiles, considering only enriched regions with a p-value <= 1e-5 (as determined by MACS) and using the following parameters: -c 5 -f -v 0.7. Only subpeaks with 20 or more reads spanning

their summit in at least one ChIP-seq were considered for further analysis (those under this threshold were defined as *low-affinity* sites). Irrespective of their cell type of origin, coordinates of Pu.1-bound regions from different cell types were merged if their summits were found within 250 bp from each other. These regions were then annotated as TSS-proximal or TSS-distal as described in section 2.1.

2.4 Genome-wide maps of regions putatively bound by Pu.1

FIMO (Grant et al., 2011) (MEME version 4.6.1) was used to identify DNA stretches that could be potentially bound by Pu.1 (these sequences will be referred to as *canonical* binding sites or bound/w sites). FIMO was run at a p-value threshold of 1e-4, with default parameters except that no q-value was calculated. A published PWM for Spi1 DNA-binding domain (DBD) (Wei et al., 2010) was used as representative of Pu.1 binding preferences.

Some of the regions identified could have been missed by HT-sequencing because of mappability issues. Intuitively, the longer the read, the lower the probability to map to more than one place in the genome. Since the shortest reads in the datasets under investigation (see table 2.1) are 36 nt long, mappability scores computed for 36 nt reads were used. Scores were extracted from bigWig tracks (Derrien et al., 2012) downloaded from the UCSC Genome Browser (Fujita et al., 2011). For a given region (considered as 50 bp upstream and downstream of each *canonical* binding site identified), the highest mappability score was retrieved using custom scripts. Any region showing at least one bp with a mappability score of 1 was further analyzed.

Using this procedure, 613,210 putative Pu.1-binding sites were identified. Among those, 41,472 overlap a bound site of the cell-type a-specific Pu.1 cistrome (see section 2.3), meaning that 571,738 (93.2%) of the sites are never contacted by Pu.1 *in vivo*. On the other hand, among the bound sites, 41,472 show a canonical high-affinity Pu.1 binding site (Wei et al., 2010) within 50 bp from the peak summit (see figure 3.2), accounting for 42,9% of the total (the bound sequences without a canonical binding sites will be referred to as bound/wo).

2.5 Measuring features in DNA strings

Features were assessed in a 300 bp window (unless specified differently) centered on the summit of the ChIP-seq peaks in case of bound regions, and to the invariant GGAA core of the Pu.1 binding site in case of the unbound ones.

These features can be divided into five broad categories, namely PWMs, k-mers, repetitive elements, DNA shape and nucleosome theoretical occupancy. Each group is described in details.

- PWMs provide quantitative descriptions of the known binding sites for a TF (Wasserman and Sandelin, 2004). They can be used to assess putative binding in any DNA string (see section 2.5.1). PWMs were collected from the literature (see table 2.2). FIMO (Grant et al., 2011) (version included in Meme 4.6.1) scans an input region of

Reference	# PWMs
Portales-Casamar et al., 2010	146
Jolma et al., 2013	843
Jolma et al., 2010	26
Hallikas et al., 2006	4
Badis et al., 2009	104
Berger et al., 2008	177
Wei et al., 2010	27
Kulakovskiy et al., 2013	481

TABLE 2.2: List of publications and corresponding number of PWMs derived from them.

  DNA for occurrences of a PWM. It computes a log-likelihood ratio score (see paragraph 2.5.1 for details) with respect to each sequence position and converts these scores to p-values. FIMO was run on the regions of interest (using a 300 bp as well as a 100 bp window) and the corresponding p-values were transformed according to the formula -log10(p-value). Only p-values equal or lower than 1e-4 were retained, otherwise a p-value of 1 was assigned to the region. In case of multiple results for the same region, only the best p-value was considered. In this way each region was described with a single value for each one of the PWM (see table 2.2).

  Since the dataset of PWMs gathered from the literature was highly redundant, motifs were grouped according to their DNA-binding domain. A straightforward approach to group motifs would be cluster them based on sequence similarity. Nevertheless, a

familial binding profile ignores the flanking positions of PWMs that are not aligned but which may be important in discriminating false positives. A recent paper (Oh et al., 2012) suggested an alternative approach, i.e. to consider all the redundant PWMs to search binding sites and then to summarize the information of single TFs at the level of their structural family. In line with this, PWMs were grouped according to their classification in families and subfamilies in TFClass (Wingender et al., 2013). A total number of 83 families and 263 subfamilies were considered. The lowest FIMO p-value among those obtained for the PWMs in a given family or subfamily was chosen as representative for each one of them. This approach also gives the advantage of reducing the initial number of features to be included in the supervised feature selection.
Furthermore, the sum of families and subfamilies showing at least a significant occurrence for one PWM was used as a proxy for cooperative binding at the region;

- The sum of C+G and the individual k-mers (with k equal to 2 or 4) counts were calculated;

- Repetitive elements in the mm9 genome were retrieved from the RepeatMasker (Smit et al., 1996) track of the UCSC genome browser (Fujita et al., 2011). A BED file for each class of repetitive elements was generated and overlapped with the regions of interest;

- The three-dimensional DNA shape features (Rohs et al., 2009) were predicted using the local sequence context in the 10 bp in the ETS core motif and for additional 15 bp on each side (see figure 2.1). The features included MGW (Minor Grove Width), Roll, propeller twist (ProT) and helix twist (HelT). Roll refers to the angle of deflection of two planer base pairs perpendicularly to the direction of the hydrogen bonds between two adjacent base pairs. Propeller twist indicates the angle of roll of one base relative to the other within the same hydrogen bond. Helix twist refers to the rotation of one base pair with respect to a neighboring one (Sinden, 1994). Measurements were obtained through all-atom Montecarlo simulations as recently described (Gordân et al., 2013).

- Nucleosome theoretical occupancy was calculated using a published algorithm (Kaplan et al., 2008). Calculations were performed using a sliding window of 147 bp. The average value among all the sliding windows was used as a proxy for the region.

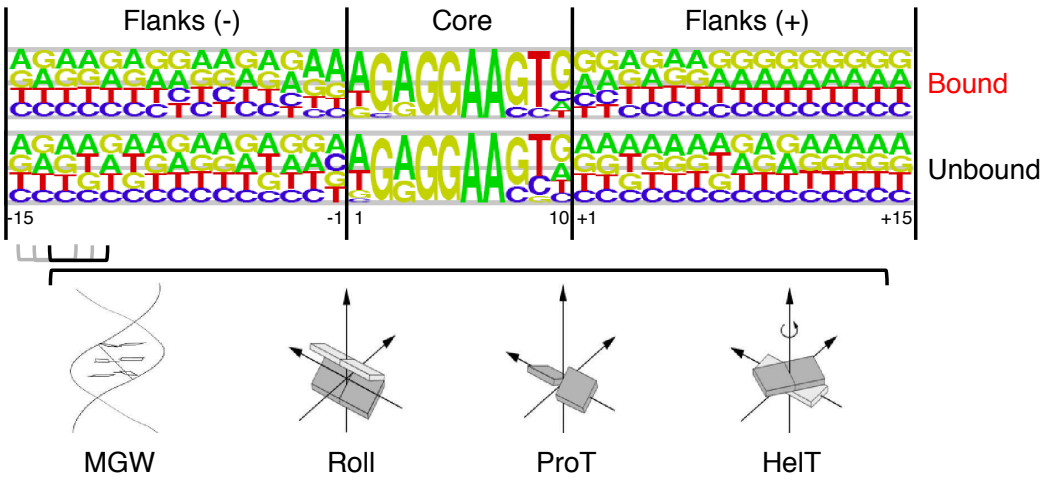

FIGURE 2.1: Each feature of DNA shape is predicted at each nucleotide using a window sliding the local surrounding sequence (ETS core +/- 15 flanking nucleotides). The sequence logo for the Pu.1-bound as well as unbound sequences is shown. Depicted as such, the two groups seems to be very similar. Nevertheless, the DNA shape measurements can capture properties of the single binding sites that are not apparent in a classical PWM representation.

2.5.1 Position Weight Matrices

Position-specific weight matrices are broadly used models for representing DNA-protein affinities. They are built upon collections of experimentally determined binding sites of a given DNA-binding protein (Wasserman and Sandelin, 2004). One of the main features of PWMs is that each nucleotide in independent of the nearby positions. This simplification has been demonstrated to be adequate in most of the cases (Weirauch et al., 2013) and improvement using higher order models are noticeable only for a few documented cases (Jolma et al., 2013, Weirauch et al., 2013).

Figure 2.2 gives an overview on the derivation and usage of a PWM. In summary, a position weight matrix (PWM) is constructed by dividing the probabilities at each nucleotide (see equation 2.1) by the expected background probabilities (see equation 2.2). These values are then log-transformed, so that probabilities can be sum up to obtain a quantitative PWM score (see equation 2.3 and figure 2.2). For large collections of binding sites, the scores are proportional to DNA-binding energies of the corresponding TF.

$$p(b,i) = \frac{f_{b,i} + s(b)}{N + \sum_{b' \in [A,C,G,T]} s(b')} \tag{2.1}$$

p(b,i) is the corrected probability of base b in position i

$f_{b,i}$ represents the counts of base b in position i

N is the number of sites in the matrix

s(b) is the pseudocount function

$$W_{b,i} = log2\frac{p(b,i)}{p(b)} \tag{2.2}$$

p(b) is the background probability of base b

p(b,i) = corrected probability of base b in position i

$W_{b,i}$ is the PWM vaule of base b in position i

$$S = \sum_{i=1}^{w} W_{li,i} \tag{2.3}$$

S is the PWM score of a sequence

l_i represents the nucleotide in position i in an input sequence

w equals the width of the PWM

2.6 Supervised learning using Support Vector Machines

Support vector machines (SVMs, Cortes and Vapnik, 1995) are supervised learning models used to discover patterns useful for classification and regression analysis (Drucker et al., 1997). Given a dataset of training examples, each belonging to one and only one category, a SVM training algorithm builds a model that can be used to assign new examples to a category. SVMs are mainly used for binary classification, even though implementations for multi-class classification are available (Chang and Lin, 2011).

More formally, a SVM finds a hyperplane or set of hyperplanes in a high-dimensional space able to separate the training examples belonging to different categories. This in turn can be used for classification or regression of new examples, by mapping them into this very same space. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class, by maximizing the so-called margin. If there exists no hyperplane that can perfectly split examples from different categories, the *Soft Margin* strategy is used to find a hyperplane that divides the examples as cleanly as possible. Formally, SVMs are linear classifiers. Nevertheless, they can efficiently perform non-linear

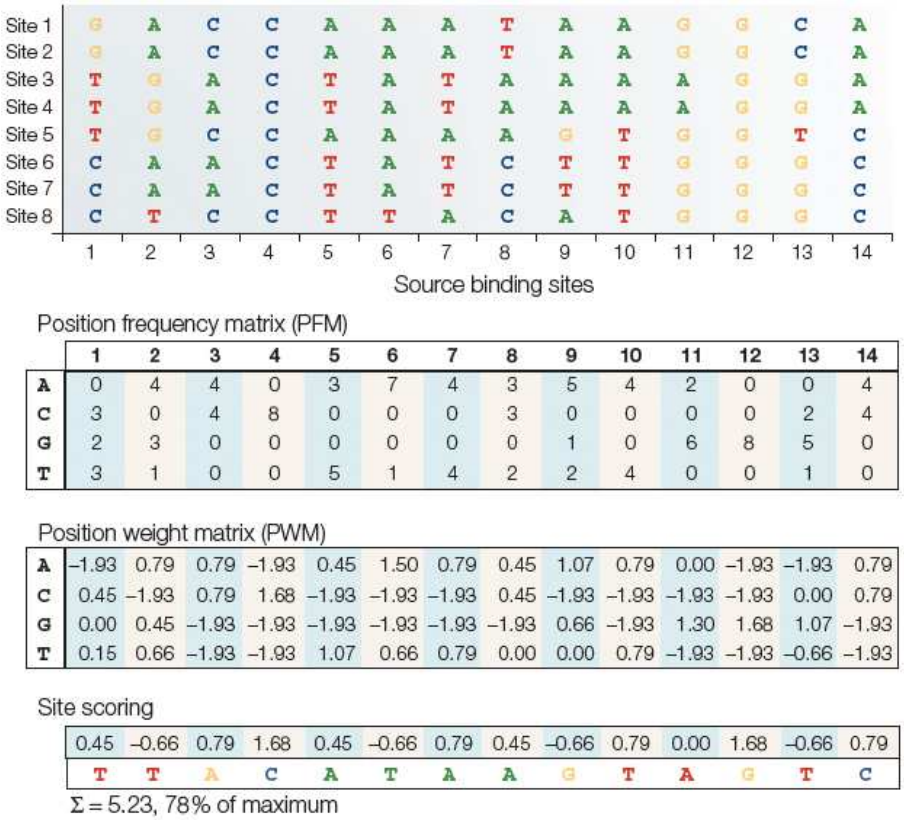

FIGURE 2.2: A set of experimentally validated MEF2-binding sites is represented in the upper panel. A matrix that contains the number of observed nucleotides at each position is created (Position Frequency Matrix). These frequencies are converted to log-transformed normalized values (Position Weight Matrix, see equations 2.1 and 2.2). Using a PWM model, a quantitative score for any DNA sequence can be generated by summing the values that correspond to the observed nucleotide at each position (Site Scoring). Adapted from Wasserman and Sandelin, 2004.

classification using what is called the *kernel trick*. This *trick* implies that SVMs are still performing linear classification, but input data points are mapped into high-dimensional feature spaces by a so-called *kernel* function.

SVMs were applied to classify Pu.1-bound regions showing a *canonical* binding site (see paragraph 2.4) from unbound sites. Considering 41,472 Pu.1-bound regions, the same number of regions was randomly chosen among the unbound sites. LibSVM (Chang and Lin, 2011) was used to train and test two-class SVMs.

Given the large initial amount of features (n=995, see table 2.3 and paragraph 2.5 for more details), a feature selection procedure (Guyon and Elisseeff, 2003) to identify the smallest set with the highest predictive power was devised. The use of fewer variables should not only result in an increase in accuracy but also in a simpler model, which allows a better biological understanding and interpretation of the results.

Group	# Features
PWMs	686
k-mers	147
DNA shape	146
Repeats	15
Kaplan et al., 2008	1

TABLE 2.3: For each one of the categories considered, the number of features is listed in the table. PWMs are collapsed on the families and subfamilies of TFs. k-mers encompass G+C and words with k=2 and k=4.

The procedure followed the steps shown in figure 3.6. Using 20% of the total instances, ten forward features selection were run randomizing training and validation datasets (50% each). The features selected in at least one out of the ten randomizations were then pooled and used to train the machine on the entire 20% and test on the remaining 80%. Training and test datasets were also randomized ten times. For each round of randomization, uninformative features, namely those showing no variance across the examples, were discarded. Features were then scaled properly (range 0-1) and ranked according to the value of absolute Spearman's rank correlation coefficient calculated among the values and the class of the training examples. Only those with showing a value >= 0.04 were retained (threshold estimated by elbow method). Forward selection consisted in adding features one by one (according to the described ranking) and keeping only those whose inclusion improved the accuracy on the validation set of at least 0.1%. This entire routine was wrapped into Python and R code.
A grid search was performed in order to choose the set of parameters giving the best performances on the validation set. In practice, for each round of feature selection, an exhaustive search through a manually specified subset of parameters was performed, and the set of parameters with the highest improvement of performance was retained. SVM with no *kernel* (linear SVM) or with radial basis function (RBF) *kernel* were tested. In both cases, parameter C was set to {0.01, 0.1, 1, 101, 100, 1000}. In case of RBF, parameter g was set to {0.0001, 0.001, 0.01, 0.1, 1, 10}. All the possible combinations were tested.
Performances were assessed using three indexes (bound are the positive dataset, unbound are the negative one, TP = true positive, FP = false positive, TN = true negative, FN = false negative):

- Overall accuracy, defining the fraction of instances correctly predicted, calculated as (TP+TN) / (TP+FP+TN+FN);

- Sensitivity, calculated as TP / (TP + FN), high values indicate the machine is very good in recognizing positive (bound) examples;
- Positive Predictive Value (PPV), calculated as TP / (TP + FP), high values correspond to a higher number of negative examples (unbound) predicted as positive (bound).

2.7 Mnase-seq data analysis

Paired-end 101 nt long reads were quality filtered and mapped to the mouse genome (mm9, NCBI Build 37) using Bowtie v0.12.7 (Langmead et al., 2009). The following parameters were used: -v 3 -m 1 -S -I 0 -X 250. In this way, the paired-end fragments with a unique match to the genome and showing three or fewer mismatches were retained. Duplicated fragments, which are likely to arise from selective PCR amplification, were discarded (see table 2.4 for statistics). Namely, given multiple fragments with both ends mapping to the same genomic coordinates, all fragments but one were discarded. Wiggle files (Fujita et al., 2011) at single bp resolution were generated with BedTools (Quinlan and Hall, 2010). In order to extract nucleosomal positions from this population-averaged profile PeakSplitter (Salmon-Divon et al., 2010) was run genome-wide on the wiggle file (with options -c 5 -f -v 0.7). For each one of the resulting regions the total number of fragments spanning the putative nucleosome dyad (namely the coordinate with the highest number of overlapping fragments) was calculated. This figure was used as proxy for occupancy. The dispersion of the midpoints of these fragments around the putative dyad (measured as standard deviation) was instead used as proxy for positioning. These calculations were performed by a custom C++ script.

Paired-end fragments for ESCs, NPCs and MEFs (Teif et al., 2012), aligned to the mm9 reference genome, were downloaded from GEO (Barrett et al., 2013). Alignments were processed as described in the previous paragraph. Final numbers of sequencing reads are summarized in table 2.4. Unless specified differently, all the heatmaps, the cumulative distributions and the nucleosome density plots have been computed using a 10 bp binning and the midpoint of each sequenced fragment as a proxy for the nucleosome dyad (hereafter referred to as midpoint analysis). Considering the heatmaps, the counts exceeding the 95th percentile of the overall distribution were set to the value of the 95th percentile. These counts were then normalized in the range 0-1, separately for each region.

Sample	# reads
BMDMs (rep. 1)	216,882,672
BMDMs (rep. 2)	250,307,803
BMDMs (rep. 3)	216,686,317
BMDMs (rep. 4)	148,756,092
BMDMs (empty, rep. 1)	212,660,520
BMDMs (shPu.1, rep. 1)	220,848,451
BMDMs (empty, rep. 2)	168,269,709
BMDMs (shPu.1, rep. 2)	181,527,195
in vitro	225,822,132
ESCs	443,856,962
NPCs	263,014,972
MEFs	399,506,104

TABLE 2.4: For each sample, the number of high-quality, uniquely aligned and properly paired reads (after filtering for PCR duplicates) is provided.

In order to sort the regions based on the size of the nucleosome-depleted region (NDR) at their center, the following strategy was applied. The number of nucleosome midpoints falling into the central 300 bp (+/- 150 bp) of each region was calculated. These numbers were used as a proxy for the overall occupancy of the area (lower the number, higher the depletion).

2.8 Support Vector Regressors

Support Vector Regressors (SVRs, Drucker et al., 1997) are a variant of SVMs that can be applied to address regression problems. It was used here to assess the fraction of variability in the nucleosomal occupancy pattern at Pu.1-bound and unbound sites in cells where Pu.1 is not expressed or in *in vitro* chromatin reconstitution experiments. SVRs were fed with the same features selected by the SVM. The theoretical nucleosomes occupancy (Kaplan et al., 2008) was excluded and a SVR was in parallel trained and tested with this feature alone.
Nucleosome occupancy at bound and unbound sites was evaluated by the log2-transformed number of fragments spanning the center of each region (corresponding to the Pu.1 ChIP-seq summit for the bound and to the GGAA core in case of the unbound). These numbers were calculated for the ESCs, NPCs, MEFs and the *in vitro* datasets.
The entire dataset of bound and unbound sites was split into 90% training and 10% test. The following procedure was run using the set of features coming from each one of the ten randomizations of the training and test datasets (see section 2.6) and separately for

each condition (ESCs, NPCs, MEFs and *in vitro*). Features were scaled to range 0-1. The training dataset was used to fit the experimentally determined nucleosome counts according to sequence features. The model obtained was then used to predict the nucleosome counts over the test dataset. Performances were evaluated through the coefficient of determination (R^2), calculated as the squared Pearson correlation coefficient among the predicted and the observed counts. This coefficient can be interpreted as the percentage of variation in the data that is explained by the model (i.e. the variation in the nucleosome occupancy that is explained by the features in the sequence). As mentioned, an independent SVR was fed with the theoretical nucleosomes occupancy alone, and its performances compared to those obtained by the model trained on all the remaining features.

The SVR implementation in the R package e1071 (Dimitriadou et al., 2008) with RBF kernel was used.

2.9 Data analysis upon Pu.1 depletion

BMDMs were infected with a retroviral vector either containing a short hairpin targeting the mRNA of Renilla (hereafter referred to as *empty vector*) or Pu.1 (hereafter referred to as shPu.1, see section 2.12.1 for details on the experimental procedure). ChIP-seq data from both samples were analyzed for enrichment versus the input DNA as described in section 2.1. All those peaks identified in the empty (using a p-value threshold of 1e-10) were retained only if also present in the untreated Pu.1 sample obtained in "wild-type" conditions (see section 2.1). Among them, those showing a significant enrichment for Pu.1 when compared to the shPu.1 (p-value <= 1e-10) were considered as Pu.1 sites whose occupancy was decreased by the depletion.

In order to get a more quantitative picture of the effect of the Pu.1 depletion, the entire dataset of peaks was sorted based on the ratio of the reads in the *empty* versus the shPu.1. Reads were counted in a window of 200 bp around the Pu.1 summit. After adding a pseudocount of 1 and normalizing for sequencing depth, ratios were calculated and used to split the dataset into quartiles (the 1st quartile corresponds to lower ratios, namely peaks that are not affected by the depletion, while the 4th quartile encompasses those peaks with the lowest occupancy in the Pu.1-depleted cells compared to the control). Bulk differences in nucleosomal occupancy

at these sites were evaluated summing up the nucleosomal fragments whose midpoint mapped into the 160 bp centered on the peak summit (the area that would ideally be occupied by a nucleosome if Pu.1 is not bound). The difference among the resulting distributions was tested using a Wilcoxon signed-rank test (which is a paired, non parametric test).

2.10 Chromatin-bound RNA-seq analysis

Chromatin-bound RNA-seq data from BMDMs were obtained from the literature (Bhatt et al., 2012). Quantitative estimation of the abundance of the transcripts (FPKM) was calculated using Cufflinks 2.0.2 (Trapnell et al., 2012) with options -N -u. Ensembl genes (Flicek et al., 2012) were used to guide assembly of the transcriptome.

2.11 Statistics and plots

All plots were drawn and statistics were performed using R.

2.12 Experimental procedures

The experiments described in the next paragraphs have been performed by Marta Simonatto, Silvia Bonifacio and Serena Ghisletti.

2.12.1 Cell culture, retroviral infection and ChIP

Macrophage cultures from bone marrows of C57/BL6 mice (Harlan) were generated as described (De Santa et al., 2007). The hairpin used in this study to deplete Pu.1 was selected among five designed using a publicly available software (http://katahdin.mssm.edu/siRNA). The shPU.1 sequence was cloned in a modified version of TtRMPVIR inducible retroviral vector (Genbank HQ456318) in which the puromycin resistance gene was inserted. The empty vector (containing an sh-Renilla sequence) was used as control.

At day zero bone marrow cells were isolated and 4e6 cells/plate were seeded in 10 cm dishes in TET-free BM medium. Cells were infected twice in two consecutive days after plating

using supernatants from transfected Phoenix-ECO packaging cells. Puromycin selection (3 μg/ml) was added on day 3. At day 5, shPU.1 expression was induced for 48 hours using doxycycline (0.5 μg/ml).

ChIP was carried out starting from 5e6-8e6 cells, using a previously described protocol (Ghisletti et al., 2010). ChIP DNA was prepared for HiSeq 2000 sequencing following standard Illumina protocols.

2.12.2 *In vitro* nucleosome assembly

Naked genomic DNA was purified from mouse macrophages by three consecutive phenol/chloroform extractions. DNA was sonicated to obtain fragments smaller than 2 kb, and fragments ranging from 600 to 2,000 bp were purified with Solid-Phase Reversible Immobilization (SPRI) beads (Agencourt AMPure XP, Beckman Coulter). DNA was combined with recombinant histones (EpiMark^{TM} Nucleosome Assembly Kit, NEB E5350) to generate nucleosomes by salt dialysis (Luger et al., 1999). DNA molecules were considered as multiple of 150 bp nucleosome-assembling units. Assembly reaction was performed mixing octamers and nucleosome-assembling units in a molar ratio 1:2, such that DNA was not limiting and octamer would assemble according to the sequence preference.

2.12.3 MNase digestion

MNase digestion was performed starting from 8e6-12e6 cells. Cell pellets were resuspended in a 15 mM NaCl, 15 mM Tris-HCl [pH 7.6], 60 mM KCl, 2 mM EDTA, 0.5 mM EGTA, 0.3 M sucrose buffer (0.5 mM PMSF, 1 mM DTT, 0.2 mM spermine, 1 mM spermidine) buffer and lysed upon addition of 0.4% NP40. Nuclei were washed with a 15 mM NaCl, 15 mM Tris-HCl [pH 7.6], 60 mM KCl, 0.3 M sucrose buffer (0.5 mM PMSF, 1 mM DTT, 0.2 mM spermine, 1 mM spermidine). Digestion was performed with 1.3 units of MNase (Roche 10107921001) in a 20 mM Tris-HCl [pH 7.6], 5 mM $CaCl_2$ digestion buffer, for 100 minutes at 37 °C. Nucleosomal DNA was isolated by diluting nucleosomes in digestion solution to a final concentration of 5 mM $MgCl_2$, 5 mM $CaCl_2$, 70 mM KCl and 10 mM HEPES [pH 7.9]. Digestion with 5 units of MNase was carried out at 37 °C and stopped after 100 minutes by adding EDTA to a final concentration of 50mM. DNA was purified from octamer proteins

with Qiagen PCR purification kit. Purified DNA was then run in a 1% agarose gel and the mononucleosomal band cut and purified first with Millipore DNA Gel Extraction Kit and then with Qiagen PCR purification kit. Digestion conditions were adjusted to obtain a mixture of DNA fragments constituted by 80% of mono-nucleosomes and 20% of di-nucleosomes or higher molecular weight forms. Mononucleosome-sized DNA fragments were isolated from agarose gels and subjected to 100 bp paired-end sequencing using the Illumina HiSeq 2000 platform.

2.12.4 *In vitro* ChIP

In vitro nucleosomes were partially digested with MNase (5U for 2 minutes in the digestion buffer described above) to obtain mainly di- and tri-nucleosomes and to eliminate any residual unwrapped DNA. They were then incubated with macrophage-derived nuclear extracts. Nuclear extracts were prepared from 2e7 cells. Cells were first lysed with hypotonic buffer (10 mM Tris-HCl, 1 mM KCl, 1.5 mM $MgCl_2$), then nuclei were lysed with a high-salt buffer (50 mM Tris-HCl, 200 mM NaCl, 10% glycerol, 0.2% NP40) and diluted 1:2 with a dilution buffer (10 mM Tris-HCl, 2 mM EDTA). Nuclear extracts were subjected twice (2 hours and overnight) to immunodepletion with 8 μg of Pu.1 antibody or normal rabbit IgG. Incubation of nuclear extracts and *in vitro* nucleosomes was performed at 4 °C for 2 hours, then 5 μg of anti-Pu.1 antibody were added for 1 hour and DNA-protein complexes recovered with G protein-coupled magnetic beads. Beads were washed 6 times with wash buffer (30 mM Tris-HCl, 200 mM NaCl, 10% glycerol, 0.1% NP40, 1 mM EDTA) and twice with TE. DNA was eluted in TE-2% SDS. DNA was then purified by Qiaquick PCR purification kit and quantified with PicoGreen (Invitrogen). ChIP DNA was prepared for HiSeq 2000 sequencing following standard Illumina protocols.

Chapter 3

Results

3.1 The *theoretical* cistrome of Pu.1

Pu.1 contacts DNA through a 10-nt long core (consensus sequence: AGAGGAAGTG, figure 3.1) that has been experimentally determined by an *in vitro* microwell-based assay (Wei et al., 2010). For brevity, we term this core sequence *canonical* Pu.1 binding site.

As already pointed out, given the size of a typical mammalian genome (billions of bp),

FIGURE 3.1: Position-specific sequence logo showing the *in vitro* determined binding preferences for Pu.1 (Wei et al., 2010). The relative frequency of each nucleotide at each position is shown.

the sites that are found to be bound by a TF *in vivo* are a small fraction compared to the hundreds of thousands of sequences matching high-affinity recognition sites for the very same TF (Pan et al., 2010).

We first estimated this number for the murine genome (mm9, NCBI Build 37). We searched for high-affinity Pu.1 *canonical* binding sites (see figure 3.1) using FIMO (Grant et al., 2011). We identified a total of 731,453 occurrences (p <= 1e-4), among which 112,830 show no uniquely mappable nucleotides in a window of 100 bp centered on the *canonical* site, resulting in a set of 618,623 sites that could be potentially identified as bound.

The mappability filter ensures that this global pattern is comparable to the *in vivo* binding

data collected (as identified by ChIP followed by multi parallel sequencing using a read as short as 36 bp).

3.2 The collective *cis*-regulatory repertoire bound by Pu.1 *in vivo*

Pu.1 is expressed only in the hematopoietic system and specifically in myeloid cells, B lymphocytes and early T lymphocytes. In order to define the largest set of sites that can be contacted by Pu.1 *in vivo*, every murine ChIP-seq dataset of sufficient quality available in the literature was gathered (see table 2.1) and analyzed as described in section 2.3, resulting in 96,685 sites contacted by Pu.1.

Among the 618,623 *canonical* binding sites that could be potentially identified as bound,

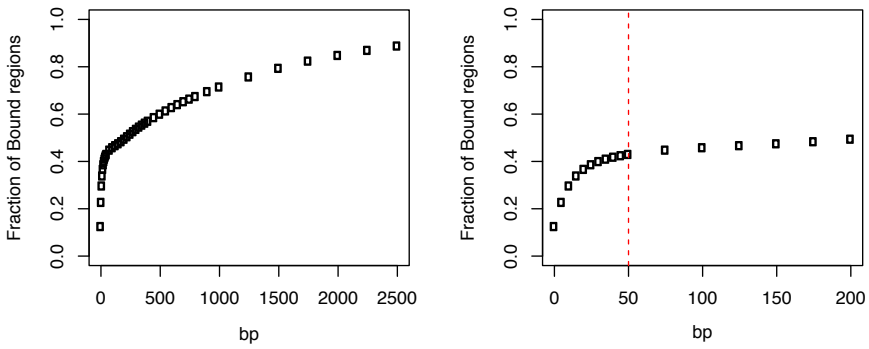

FIGURE 3.2: The fraction of Pu.1 bound regions overlapping a high-affinity *canonical* binding site (FIMO, p-value <= 1e-4) depends on the maximum distance considered between the experimentally determined summit and the *canonical* binding site. The right plot shows a magnification of the left one. The red dashed vertical line indicates the threshold chosen to split the dataset in bound with *canonical* binding site (bound/w) and without (bound/wo).

41,472 were found within 50 bp from the peak summit of the previously defined 96,685 Pu.1-bound regions (see figure 3.2). This means that 571,738 (93.2%) of the sites are never contacted by Pu.1 *in vivo*. Even assuming that part of them may be bound in conditions that were not recapitulated in the experiments that generated the datasets we collected, the vast majority of them is likely to be never engaged *in vivo*. On the other hand, 41,472 accounts for 42.9% of the total (see figure 3.3). This means that Pu.1 binding through lower affinity or composite sites and tethering interactions together account for more than 50%. Interestingly, 22.7% of the total bound/wo (Pu.1-bound regions showing no *canonical* binding site) regions

versus 14.9% in bound/w ones (Pu.1-bound regions at *canonical* binding site) resides within 2.5 kbp of RefSeq TSSs (p <= 0.01 in a Chi-squared test), which might indicate a relative enrichment of lower affinity sites or tethering interactions accounting for Pu.1 binding at TSSs.

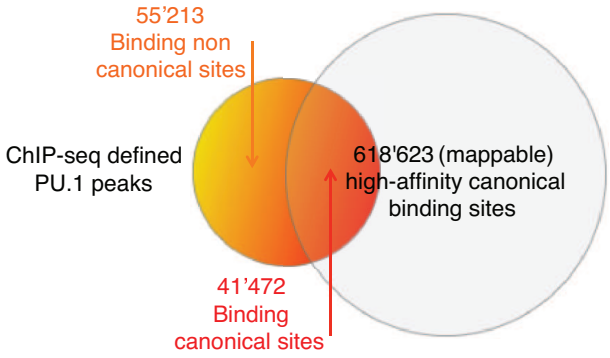

FIGURE 3.3: Venn diagram showing the overlap between Pu.1 peaks identified in ChIP-seq experiments from multiple cell types and computationally identified genomic Pu.1 sites.

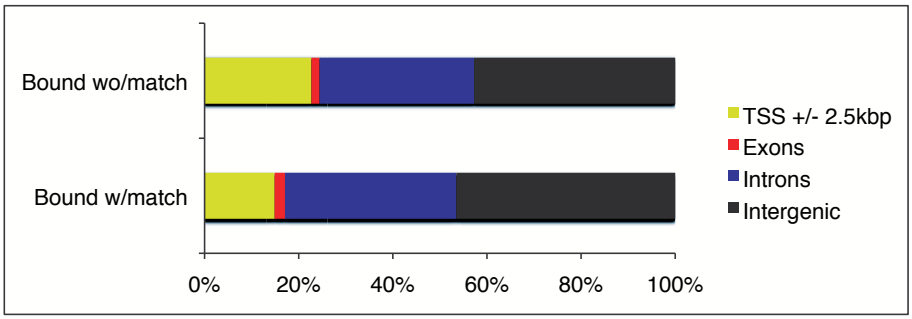

FIGURE 3.4: Bound/wo show a significanlty different distribution compared to bound/w (p <= 0.01 in a Chi-squared test). This can be attributed to a different representation of RefSeq TSS-proximal sites, which accounts for 22.7% of the total bound/wo regions versus 14.9% bound/w ones.

3.3 Discrimination of engaged and non-engaged Pu.1-binding sites *in vivo*

Starting from these data, we then asked if the hundreds of thousands of high-affinity recognition sites for Pu.1 showing no engagement in any of the cell types tested can be recognized from the engaged ones using only the information from the surrounding sequence. To this purpose, the 41,472 engaged sequences showing a *canonical* binding site were compared to the same number of unbound regions. Since the latter outnumber the former, 41,472 regions

were randomly chosen. While regions in both groups show a *canonical* binding site (FIMO p-value <= 1e-4), we asked if the affinity (according to the PWM used) for the sites was different. As shown in figure 3.5, this is indeed the case. Bound/w sites show statistically significant higher affinities than unbound ones (Mann-Whitney test, p = 2.31e-294).

Given this significant difference in affinity, we trained a Support Vector Machine (SVM,

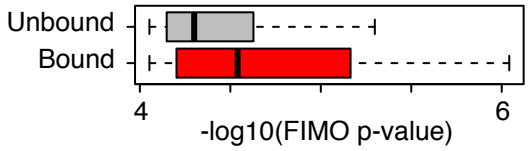

FIGURE 3.5: Higher affinities correspond to lower p-values, which correspond to higher -log10(p). In bulk, bound/w sites show higher affinities than unbound ones (Mann-Whitney test, p = 2.31e-294).

Cortes and Vapnik, 1995) using 20% of the examples and tested it on the remaining 80%. Pu.1 sequence preferences alone are poorly predictive of binding, resulting in an average accuracy of 58.5% (see figure 3.7, the mean value is shown, which is extremely stable over random initializations of the training and test datasets). We then measured features in the surrounding non-coding sequence, aimed at increasing this accuracy value. We collected a total of 995 sequence features assessed in 300 bp windows aligned to the summit of the ChIP-seq peaks in the case of bound regions, and to the invariant GGAA core of the Pu.1 binding site in the case of the unbound ones (see section 2.5 for a detailed description of the features and how they were extracted). We gathered 1,808 models (PWMs) from the literature describing known binding preferences for TFs. In order to avoid redundancies, the PWMs were grouped by TF family and subfamily. The described scoring procedure for the PWMs was repeated also for a more narrow window of 100 nucleotides. Among other features we included i) k-mers with k = 2 and k = 4, ii) C+G content, iii) the average theoretical nucleosome occupancy of the region calculated with a published algorithm (Kaplan et al., 2008), iv) the overlap with known classes of repetitive elements, and v) the three-dimensional (3D) DNA shape predicted for the 10 bp in the ETS core motif and for additional 15 bp on each side. DNA shape depends on sequence only and directly impacts on protein-DNA recognition (Rohs et al., 2009). The inclusion of these features is able to capture the fact that degenerated binding motifs can form very similar 3D shapes and conversely sequences with comparable affinity may display functionally relevant topological differences. In fact, DNA shape was recently shown to improve the prediction of engaged bHLH TF binding sites in *S. Cerevisiae* (Gordân et al., 2013).

Given this large amount of features, we devised a feature selection procedure (Guyon and

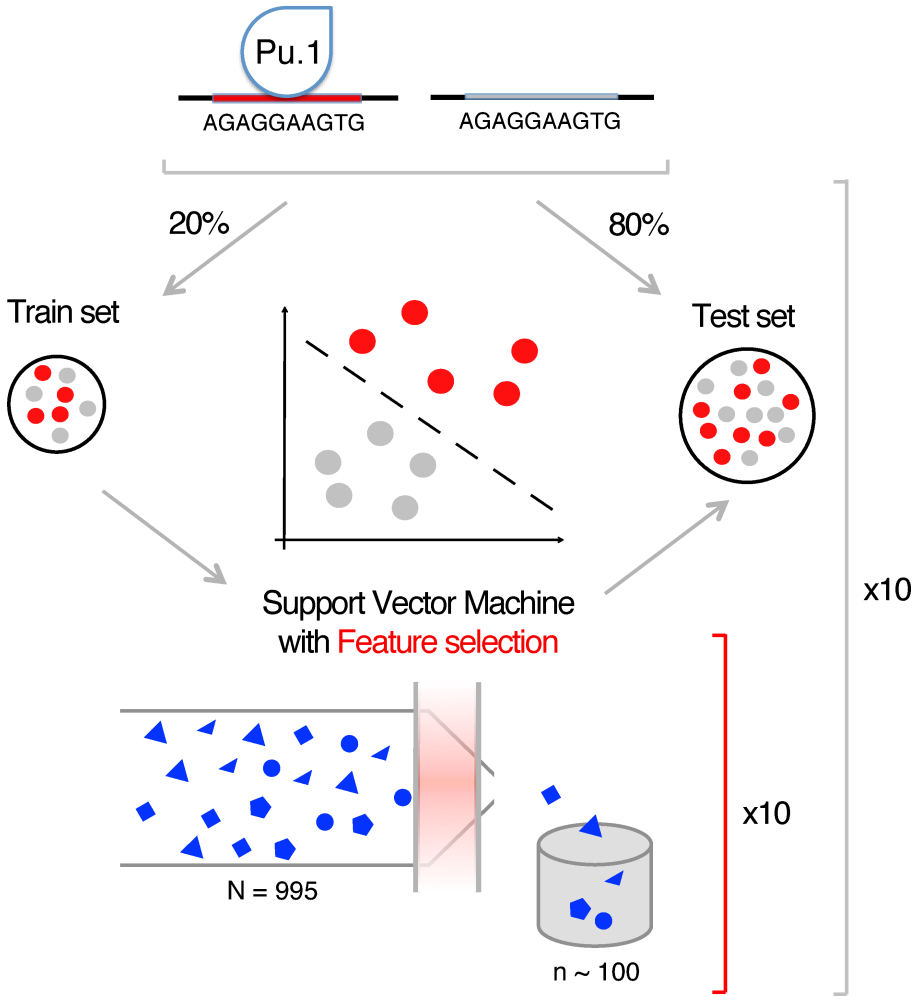

FIGURE 3.6: The entire set of sequences was split into training (20%) and test (80%) datasets. 50% of the training dataset was kept as a validation set. Forward feature selection was performed ten times and the selected features pooled. Final training and testing were performed over this pool of features. Features were pre-ranked according to the value of absolute value of the Spearman's rank correlation coefficient calculated among the values and the class of the training examples. Only those with a value >= 0.04 were retained. According to this ranking, forward selection was performed by adding features one by one and keeping only those whose inclusion improved the accuracy on the validation set of at least 0.1%. To estimate the robustness in the accuracy of the predictions and the reproducibility of the set of selected features, the approach was reiterated ten times on different permutations of the training and the test datasets.

Elisseeff, 2003) aimed at identifying the smallest set with the highest predictive power. The use of fewer variables should result not only in a more performing but also in a simpler model, which allows a better biological understanding and interpretation of the results. The entire dataset of sequences was split into training (20%) and test (80%). 50% of the training dataset was kept as a validation set. Forward feature selection was performed ten times and the selected features pooled. Performances on training and test sets were evaluated over this pool of features. To estimate the robustness of the accuracy achieved in the prediction

and the reproducibility of the set of selected features, the approach was reiterated ten times, reinitializing the training and test datasets (see figure 3.6).

Starting with the entire set of 995 features and through feature selection, we achieved an

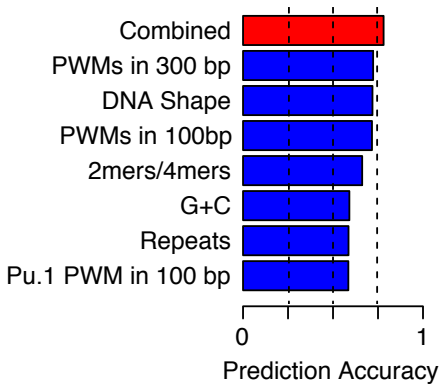

FIGURE 3.7: The average accuracy for the test datasets over ten training-test randomizations are shown. Standard deviations are negligible (see table 3.1) and not shown.

average accuracy of 78%(see figure 3.7, red bar). We then analyzed the contribution of individual groups of features to the prediction. Theoretical nucleosome occupancy and C+G content were found to have similar performances (accuracy of 59-60%). C+G content has in fact been reported to be a simple proxy of nucleosome occupancy (Tillo and Hughes, 2009). Interestingly, while 2-mers and 4-mers were more predictive (66.2%) than C+G only, a small number of DNA shape features alone achieved an average accuracy of 71.9%, slightly less than considering PWMs in a 300 nt window (72.2%). In the end, none of the single groups of features achieved the performance of the combination (see figure 3.7). When the feature

	Training			Testing		
Run	ACC	SEN	PPV	ACC	SEN	PPV
1	0.7888	0.7907	0.7878	0.7785	0.788	0.7733
2	0.7911	0.8002	0.7858	0.7834	0.7919	0.7787
3	0.7902	0.7904	0.7901	0.7821	0.7856	0.7801
4	0.7931	0.7996	0.7894	0.7825	0.785	0.7811
5	0.7814	0.7926	0.7753	0.7757	0.7826	0.772
6	0.7752	0.7854	0.7696	0.771	0.7829	0.7647
7	0.7859	0.7928	0.782	0.7816	0.7899	0.777
8	0.7969	0.8034	0.7931	0.7817	0.7883	0.778
9	0.7922	0.8007	0.7873	0.7798	0.7914	0.7735
10	0.7885	0.7932	0.7858	0.7858	0.7941	0.7811

TABLE 3.1: Accuracy (ACC), sensitivity (SEN) and positive predictive value (PPV) obtained from training and test datasets are shown.

selection routine was allowed to select between linear SVM or using the RBF (radial basis

function) as kernel and an exhaustive search for parameters was performed (grid search), RBF kernel was systematically preferred over the linear SVM. Nevertheless, while performances on the validation set increased, those on the test dataset dropped to values lower than those obtained using the linear SVM.

Considering the linear kernel and the ten training-test datasets randomizations performed,

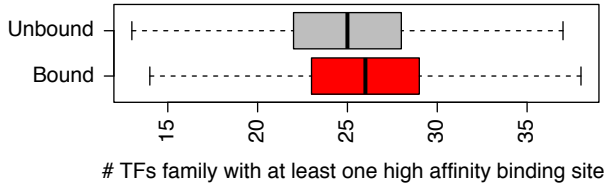

FIGURE 3.8: The sum of different families of PWMs with a putative binding site in a 300 bp region centered on the Pu.1-bound sites is significantly higher than that measured at the unbound ones (p-value = 1.54e-11, Mann-Whitney test).

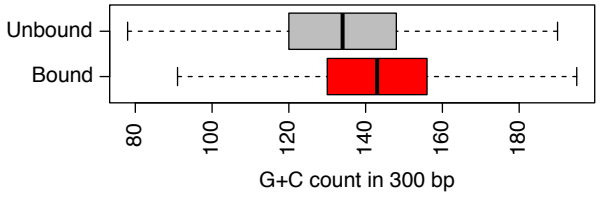

FIGURE 3.9: Bound sites show a significantly higher C+G content than unbound sites (p-value <= 1e-300, Mann-Whitney test).

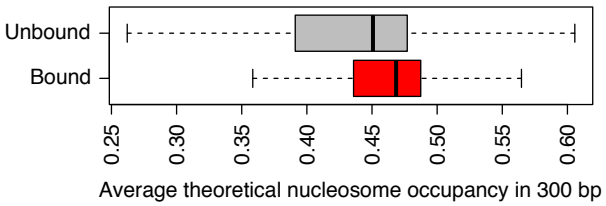

FIGURE 3.10: Bound sites show a significantly higher theoretical nucleosome occupancy than unbound sites (p-value <= 1e-300, Mann-Whitney test).

PWMs representing ETS-family binding preferences were systematically selected. On the contrary, the sum of PWMs families (used as a proxy for cooperative binding at the regions) was included only in 5 out of 10 runs. Nevertheless, the sum of distinct families of PWMs showing a putative binding site around Pu.1-bound sites is significantly higher (see figure 3.8) than that measured at the unbound ones (p-value=1.54e-11, Mann-Whitney test). Except for one case, C+G content was systematically selected along with the theoretical nucleosome occupancy. In fact, considering either of the two, bound sites show significantly higher

Category	Feature	# times (out of 10)
DNA shape	MGW flank -1	10
DNA shape	Roll core 10	10
DNA shape	Roll flank -1	10
DNA shape	MGW core 1	10
DNA shape	MGW core 7	10
DNA shape	Roll core 6	10
DNA shape	ProT flank -2	9
DNA shape	HelT core 3	9
DNA shape	MGW core 10	9
DNA shape	ProT flank +1	8
2mers/4mers	GC	10
2mers/4mers	CC	10
2mers/4mers	CG	10
2mers/4mers	AT	10
2mers/4mers	AAAT	9
2mers/4mers	TA	9
2mers/4mers	AG	9
2mers/4mers	AGGT	8
2mers/4mers	GATA	8
2mers/4mers	AGTG	7
PWMs	Fos-related factors (Family, +/- 50bp)	9
PWMs	B-ATF-related factors (Subfamily, +/- 50bp)	9
PWMs	Interferon regulatory factors (Family, +/- 50bp)	9
PWMs	Interferon regulatory factors (Subfamily, +/- 150bp)	8
PWMs	Runt-related factors (Family, +/- 50bp)	8
PWMs	Jun-related factors (Family, +/- 50bp)	7
PWMs	Jun factors (Subfamily +/- 50bp)	7
PWMs	CTCF-like factors (Subfamily, +/- 150bp)	6
PWMs	CTCF-like factors (Subfamily, +/- 50bp)	6
PWMs	Runt-related factors (Subfamily, +/- 150bp)	5
Repeats	LINE	10
Repeats	LTR	8

TABLE 3.2: Among those selected at least in 5 out of 10, the top ten selected features during multiple initialization of training-test datasets for each category are shown.

values than unbound ones (see figures 3.9 and 3.10). This is in line with observations at p53-contacted sites in *H. sapiens* (Nili et al., 2010). Considering the broader groups of features (i.e. PWMs, DNA shape, k-mers and overlap with repetitive elements), a summary of the features selected is given in table 3.2.

In the combined model, DNA shape features of the ETS core but also at -2, -1 and +1 flanking nucleotides were systematically selected (see table 3.2). Among the k-mers, those systematically selected are mostly reflecting a different C+G composition of bound and unbound sites. Nevertheless, they carry more information than C+G alone, by pushing the accuracy

to 66.2% compared to the 59% achieved by the C+G content only.

The families of TFs (PWMs group) that are more frequently selected are Jun/Fos and

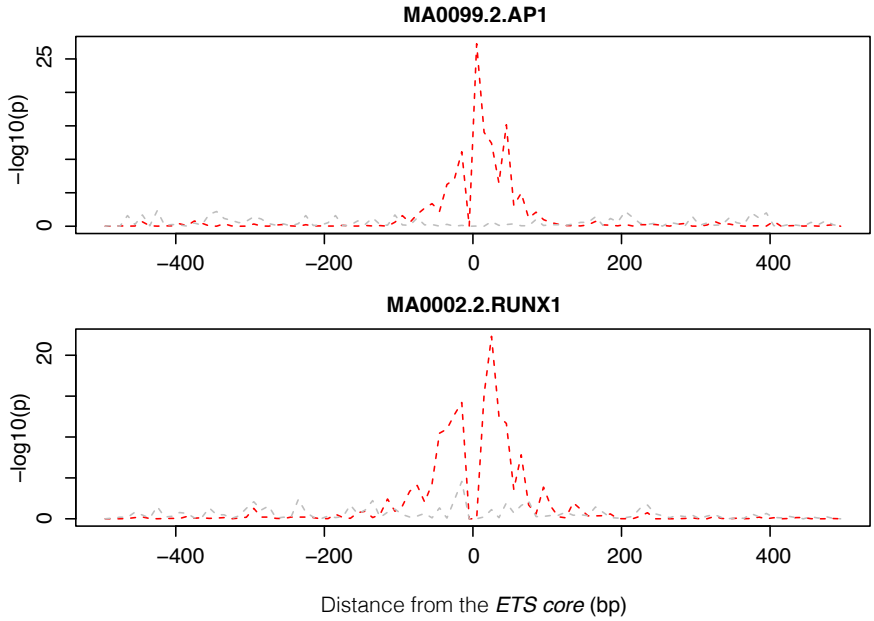

FIGURE 3.11: Runx1 and AP-1 sites (FIMO, p-value <= 1e-4) were annotated around the ETS cores. The cores have been aligned according to the GGAA in order to be able to spot existing spatial constraints. For bound (red) and unbound (grey) the number of Runx1 and AP-1 sites was summarized using a 10 bp binning. The probability for the observed number of sites to occur by chance was calculated for each bin (binomial probability estimated as the average bin frequency in 1,000 bp surrounding the ETS ore). Compared to p-values in the unbound (which show an almost completely flat distribution in the 1,000 bp window) the bound sites show a strong enrichment for both Runx1 and AP-1 sites in a narrow area on both sides of the ETS site.

ATF-like factors (AP-1 is a heterodimeric TF composed of proteins belonging to these families), Runt family members (like Pu.1, Runx1 is another essential transcription factor in hematopoiesis), IRF-like factors and CTCF. In line with this, we searched for Runx1 and AP-1 sites around the ETS cores of the unbound and bound sites. Both Runx1 and AP-1 sites show a strong enrichment in a narrow area (about 200 bp) around the bound ETS sites compared to unbound (see figure 3.11).

We also found of great interest that LINE and LTR repetitive elements were frequently selected by the machine learning approach. Using the whole set of unbound, we confirmed that 17.4% of them overlap LINEs, compared to only 6.9% of the bound/w sites (p <= 1e-300 in a Chi-squared test). LTRs also showed a highly significant difference as well (p <= 1e-300 in a Chi-squared test) but the gap among the two groups was smaller, with 17.1% of the unbound regions overlapping them, compared to a 12.1% for the bound/w sites. An intriguing

hypothesis suggests that these transposable elements elements containing the Pu.1 consensus site might be representative of a reservoir of elements ready to rewire the mammalian *cis*-regulatory repertoire (de Souza et al., 2013).

3.4 Nucleosomal organization at Pu.1 bound and unbound sites

Among the features that were systematically selected by the SVM, we found the C+G content and the theoretical nucleosome occupancy of primary interest. These data might indicate that DNA sequence drives higher nucleosomal occupancy at engaged (TF-bound) *cis*-regulatory regions, compared to unbound sites. To test this hypothesis, we looked at nucleosomal organization at bound and unbound sites in macrophages and in unrelated cell types in which Pu.1 is not expressed. Since we generated Mnase-seq data only in macrophages, in order to get a cleaner picture we considered only those Pu.1 sites that are specifically bound in BMDMs.

We first split the Pu.1-bound sites (the bound/w and the bound/wo separately) into TSS-

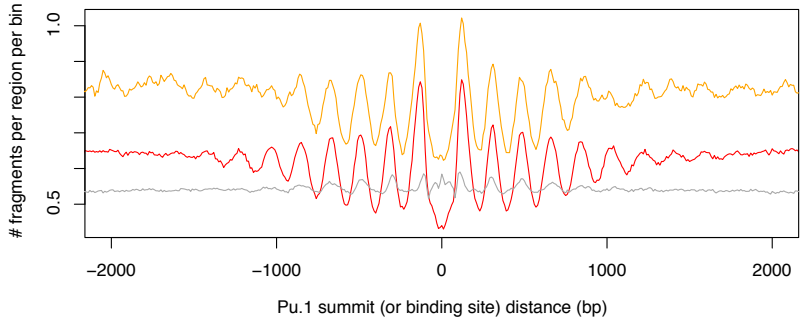

FIGURE 3.12: Cumulative distributions of nucleosome midpoints centered on the summit of the TSS-distal Pu.1-contacted regions in macrophages (bound/w and bound/wo shown respectively in red and orange) or on the GGAA of the computationally identified sites that are not bound *in vivo* (grey) (bin = 10 bp).

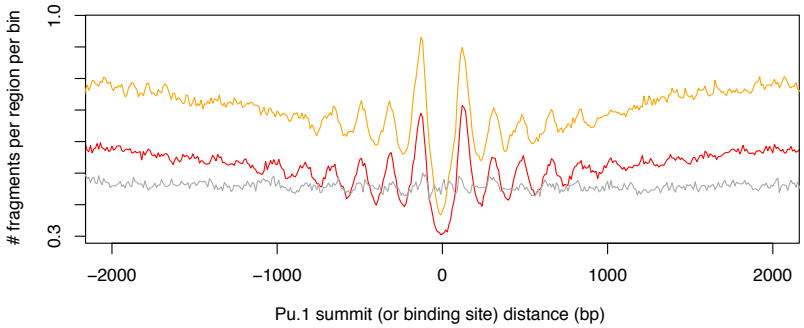

FIGURE 3.13: Same as in figure 3.12 but for the TSS-proximal sites.

proximal and TSS-distal sites. This was done in order to avoid any bias coming from the fact that TSSs are not as cell-type specific as TSS-distal enhancers. Considering BMDMs,

Pu.1 is able to induce the same pattern in both genomic contexts (a nucleosome-depleted area with nucleosome phased on either sides, see figures 3.12 and 3.13; see later in the thesis for a thorough investigation of nucleosomal patterns in macrophages).

Considering instead the situation in ESCs (in which Pu.1 is not expressed) the TSS-distal

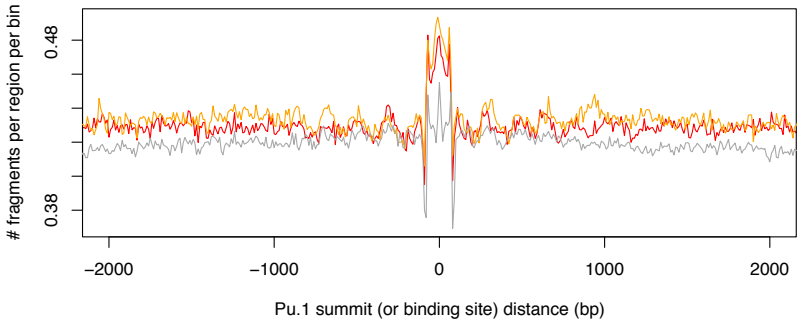

FIGURE 3.14: Cumulative distributions of the midpoints of the nucleosomal fragments centered on TSS-distal Pu.1 sites in ESCs. These are centered on the summit of the TSS-distal Pu.1-contacted regions in macrophages (bound/w and bound/wo shown respectively in red and orange) or on the GGAA of the computationally identified sites that are not bound *in vivo* (grey) (bin = 10 bp).

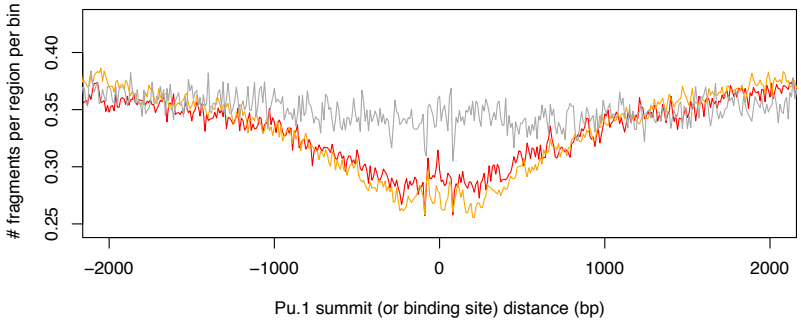

FIGURE 3.15: Same as in figure 3.14 but for TSS-proximal sites.

bound sites show increased nucleosome signal in a narrow area around the site contacted by Pu.1 *in vivo*, while the unbound sites do not. TSS-proximal sites instead show no clear increase in nucleosome occupancy over the Pu.1 sites. This might be due to the fact that these sites are active in this other system as well, being bound by a different combination of TFs, which is responsible for the nucleosome depletion (either compared to the surrounding regions or to the unbound sites).

The analysis at all Pu.1-bound sites was extended to nucleosomal data from neural precursors (NPCs) and mouse embryonic fibroblasts (MEFs) (Teif et al., 2012). These data were also aligned to the summit of Pu.1 peaks. Irrespective of the cell type considered, higher nucleosome occupancy extending for about a single nucleosome length and precisely overlapping

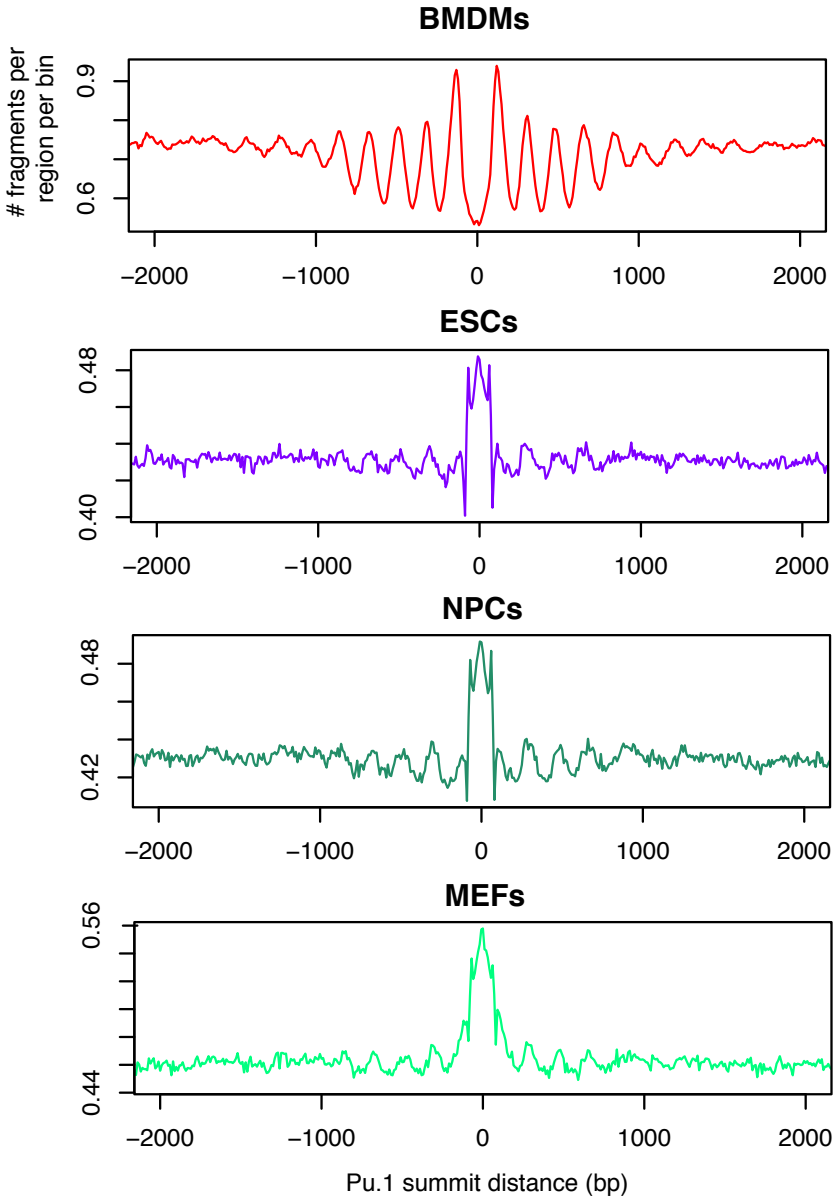

FIGURE 3.16: Cumulative distributions of the midpoints of the nucleosomal fragments centered on TSS-distal Pu.1 sites in macrophages and in unrelated cells that do not express Pu.1 (ESCs, NPCs, MEFs) (bin = 10 bp).

the macrophage Pu.1-bound, nucleosome-depleted regions was detected in case of TSS-distal sites (see figure 3.16). As already observed, TSS-proximal sites instead show no increase in nucleosome occupancy over the Pu.1 sites (see figure 3.17).

3.5 Pu.1-bound sites show spatial sequence constraints

The feature selection embedded in the SVM indicated that Pu.1-bound sites show a higher theoretical nucleosome occupancy (as well as C+G content) than unbound *canonical* Pu.1

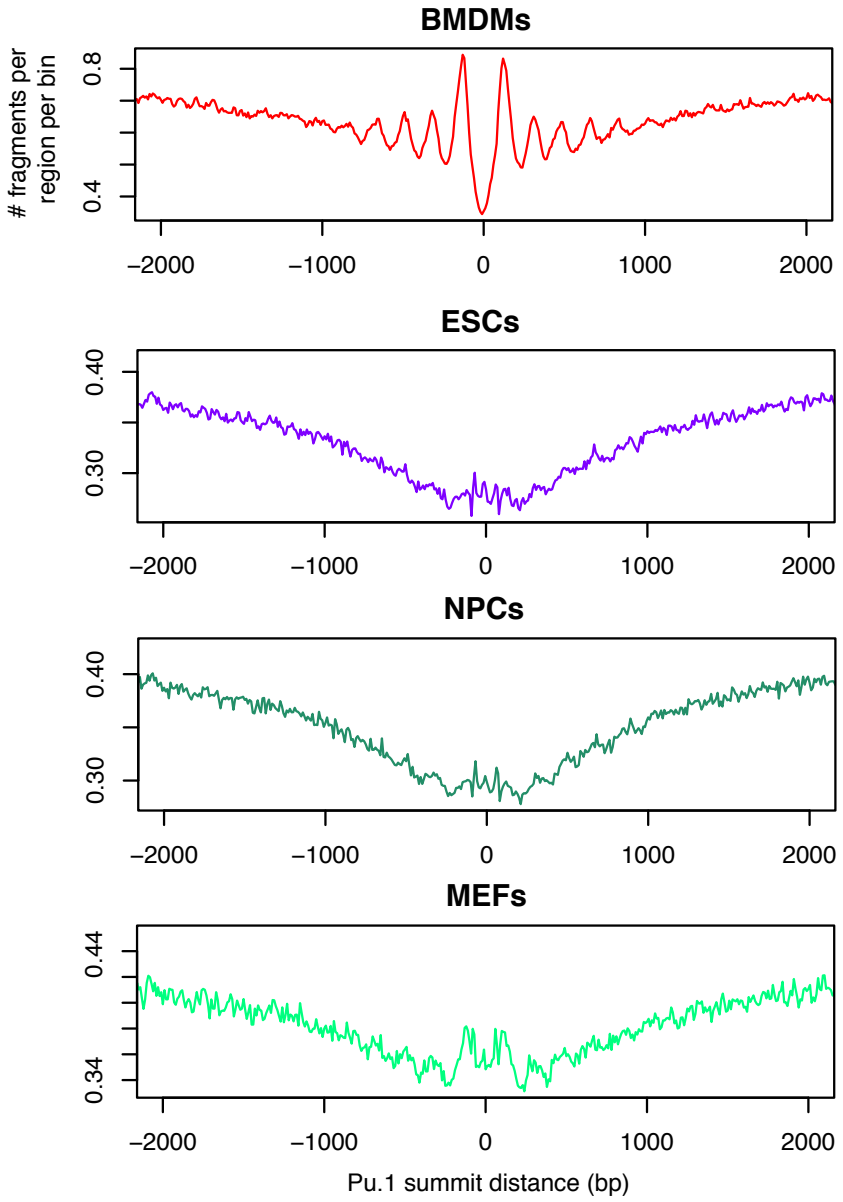

FIGURE 3.17: Cumulative distributions of the midpoints of the nucleosomal fragments centered on TSS-proximal Pu.1 sites in macrophages and in unrelated cells that do not express Pu.1 (ESCs, NPCs, MEFs) (bin = 10 bp).

sites. This sequence characteristic is reflected in the nucleosome patterns of the cell types in which Pu.1 is not expressed. This lead us to further investigate the features of the sequence along the engaged regions, centered on the summit of the regions bound by Pu.1 in BMDMs.

Cumulative distribution plots revealed features characteristic of nucleosome *container sites* (see figure 3.18, Valouev et al., 2011): an increase in the relative frequency of both AA dinucleotides (see figures 3.19 and 3.20) and AAAA polynucleotides (see figures 3.21 and 3.22) peaking at -100 and +100 positions relative to the summit of Pu.1 peaks (corresponding to repelling sequences) with an extended central core of G/C rich sequences that promote

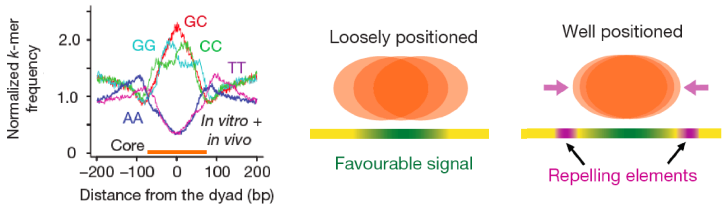

FIGURE 3.18: Schematic depiction of the *container site* positioning mechanism. The C/G-rich core (green) is known to favor nucleosome occupancy, but it is not able to precisely position the nucleosome. Flanking A/T-rich repelling elements (purple) add the ability to restrict the position of the nucleosome. Adapted from Valouev et al., 2011.

nucleosome occupancy (Tillo and Hughes, 2009) (note that the strong enrichment of CC/GG and AA/TT dinucleotides at the anchor point is enhanced by the central invariant nucleotides of the Pu.1 site, AGAGGAAGTG).

Any attempt of including these features in the SVM (spatial counts of AA and AAAA in

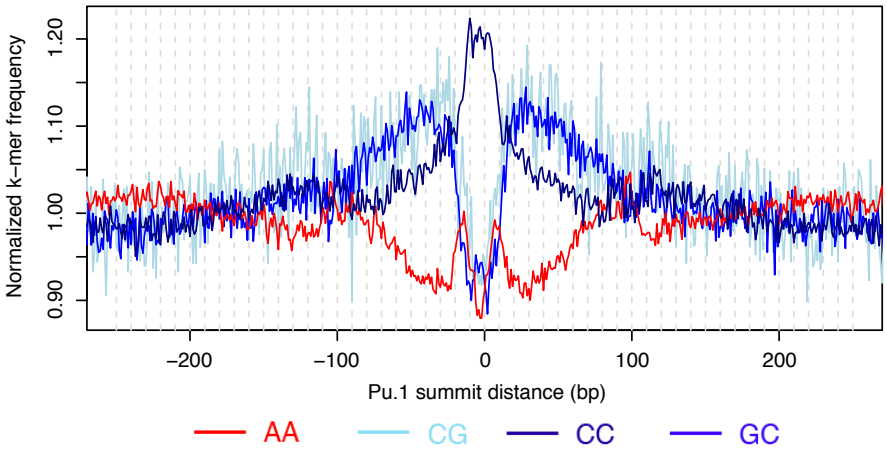

FIGURE 3.19: Frequency of AA and CG-rich dinucleotides around TSS-distal Pu.1 binding sites. CG-rich and AA normalized k-mer frequency is calculated as the relative frequency (the count of dinucleotides per bp per number of regions) divided by the average relative frequency in a larger region of +/- 500 bp.

coarse-grained bins or the ratio among the same counts in the central versus the side regions) did not show any increase in the performances. A possible explanation is that the *container site* is a feature characteristic only of a subset of the entire repertoire of engaged Pu.1-sites (see section 3.8, which reports evidences supporting this scenario).

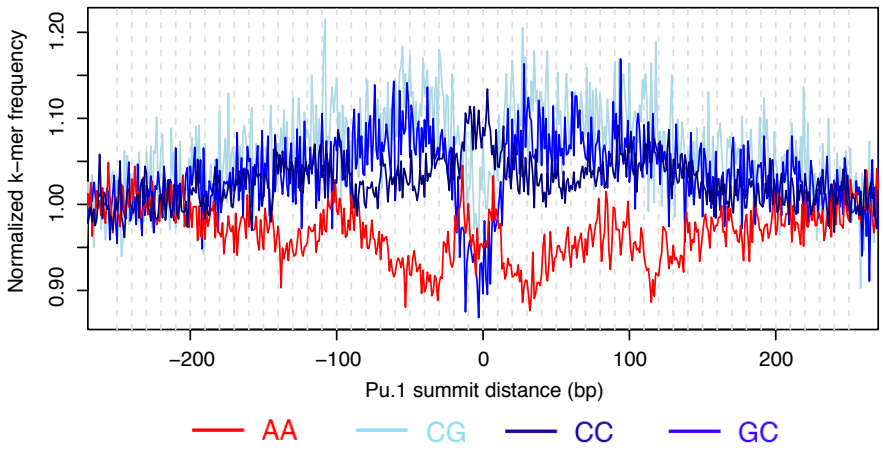

FIGURE 3.20: As described for figure 3.19, but for the TSS-proximal set of Pu.1.

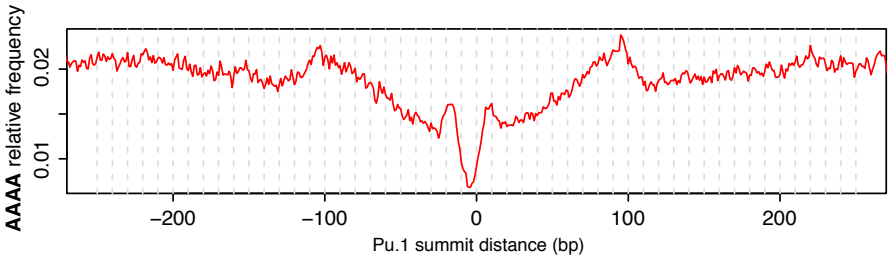

FIGURE 3.21: AAAA frequency at TSS-distal Pu.1 binding sites. AAAA relative frequency stands for the number of AAAA for each nucleotide position divided by the total number of regions considered.

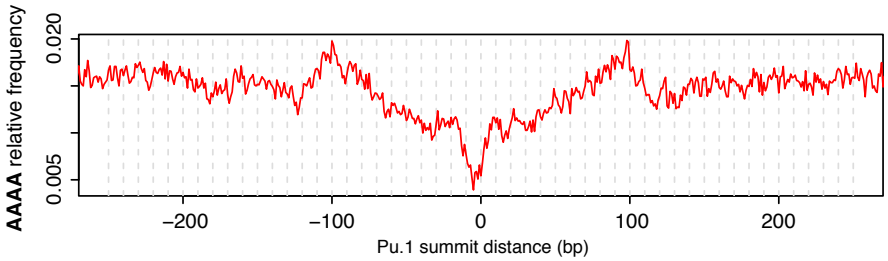

FIGURE 3.22: As in figure 3.21, but referring to the TSS-proximal set of Pu.1.

3.6 Nucleosomal organization at Pu.1 sites *in vitro*

In order to conclusively demonstrate the role of DNA sequence in controlling the basal nucleosomal landscape at Pu.1 sites, we assembled nucleosomes *in vitro* and analyzed them using the same pipeline. Naked genomic DNA extracted from mouse macrophages was sonicated and a smear from 600 to 2,000 bp fragments was purified and combined with recombinant histones to generate nucleosomes by salt dialysis (Luger et al., 1999). Assembly conditions in which DNA was not limiting were used to specifically focus on the effects of the primary sequence on nucleosome positioning (Luger et al., 1999, Valouev et al., 2011).

The cumulative distribution of nucleosome reads at TSS-distal as well as TSS-proximal Pu.1-bound sites in macrophages indicates that genomic sequence features are sufficient to generate a focused increase in nucleosomal density at both TSS-distal and proximal sites bound by Pu.1 in macrophages (see figure 3.23).

It is important to point out that, contrary to what we have observed *in vivo* (see figures

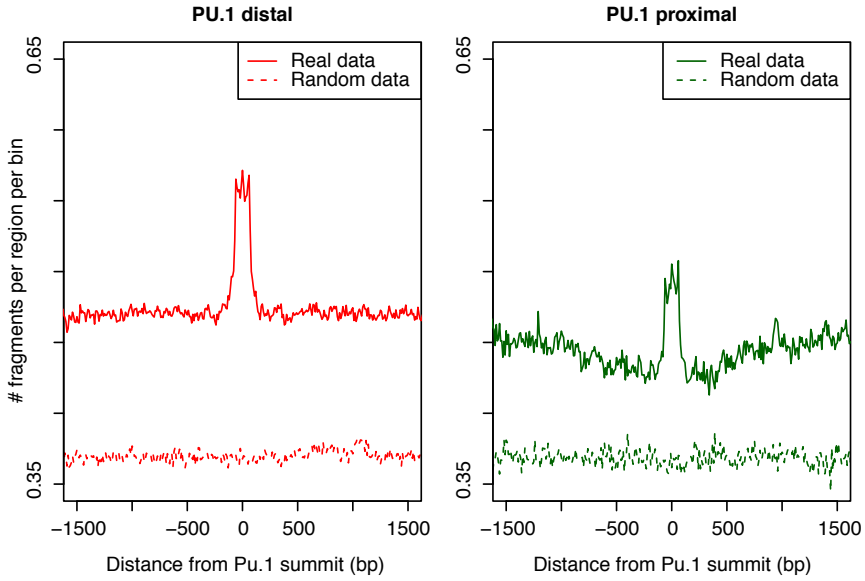

FIGURE 3.23: Cumulative midpoints distribution from *in vitro* assembled nucleosomes (bin = 10 bp).

3.15 and 3.17), the *in vitro* pattern at TSS-proximal sites also shows a local increase in nucleosomes positioned over the Pu.1-binding site.

We also looked for genome-wide evidences supporting the *container site*. The distributions of dinucleotides at strongly positioned nucleosomes in macrophages as well as *in vitro* are shown in figure 3.24. We first extracted all the nucleosomes supported by at least 10 sequenced fragments (but also less than 50 in order to avoid extreme outliers), sorted them by standard deviation of the fragments around the nucleosome dyad (smaller standard deviations correspond to better positioned nucleosomes), and used the top 100,000 for the analysis. Strongly positioned nucleosomes *in vitro* showed a higher frequency of AA dinucleotides rising at +/- 50 nt and peaking at +/- 100 nt from the dyad. The AA dinucleotides shoulders bracketed a central CC/GG-richer region extending for about 100 nt. Conversely, except for a rather narrow region around the dyad, the frequency of the same dinucleotides at 100,000 randomly picked nucleosomes was rather flat over the entire 500 nt considered. When the same procedure was applied to the nucleosomes in macrophages, a similar result was observed albeit

of lower magnitude. This is in line with a smaller contribution of the sequence determinants compared to other factors *in vivo*. These factors are often able to override the intrinsic occupancy and positioning dictated by the DNA sequence itself (Struhl and Segal, 2013). Taken together, these results confirmed that the *container site* observed in *H. sapiens* is a feature conserved in the genome of *M. musculus*.

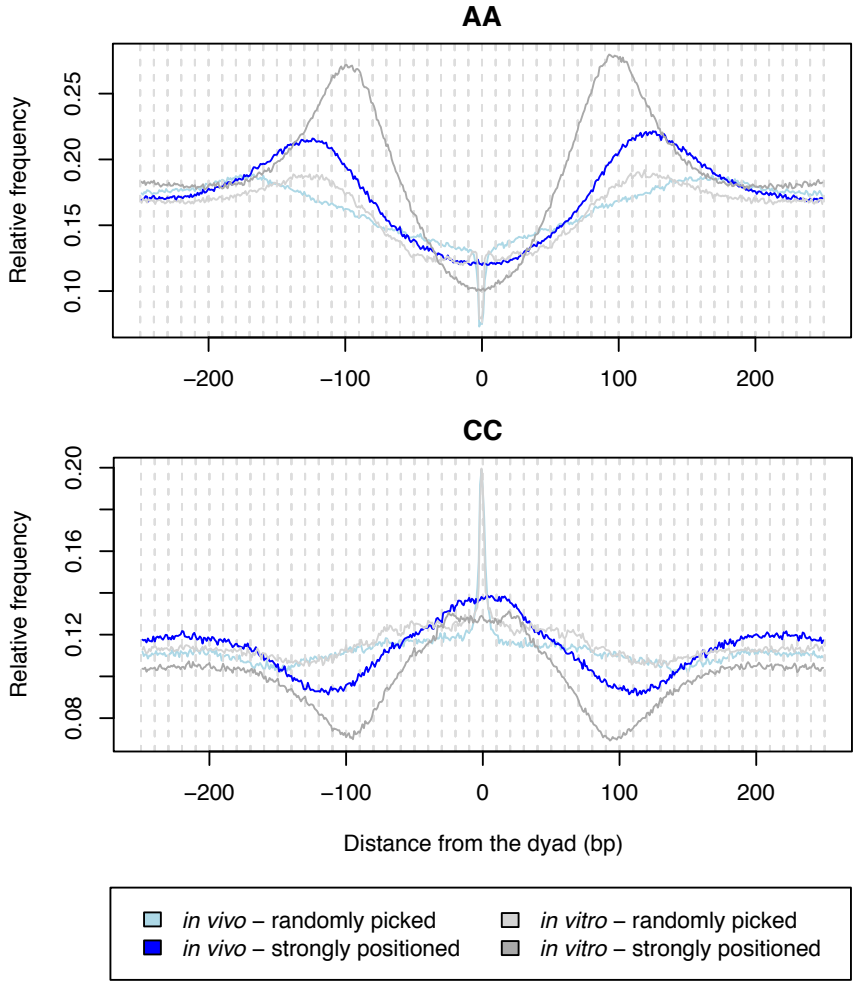

FIGURE 3.24: The top 100,000 positioned nucleosomes were extracted from the the *in vivo* (BMDMs) and *in vitro* reconstituted nucleosomes. The same number of nucleosomal positions was retrieved at random (independently for the *in vivo* and *in vitro* sets) and used as reference. The distributions of AA and CC dinucleotides are shown. For each nucleotide position, relative frequency stands for the number of regions showing that particular dinucleotide divided by the total number of regions considered (100,000 for every set).

3.7 Predicting nucleosome occupancy from features instructive for Pu.1 binding

The data presented so far suggest that nucleosomes may selectively occupy those Pu.1 sites that are contained within TF binding-competent *cis*-regulatory regions through different mechanisms. Since i) these nucleosomal patterns correlate with peculiar sequence features (namely a different C+G content and the presence of the *container site*) and ii) we were able to identify a limited subset of sequence features which is 78% accurate in predicting which Pu.1 canonical binding sites will be contacted *in vivo*, we then asked if we could use the same determinants to predict nucleosome occupancy in cells that do not express Pu.1.

The nucleosomal information at these sites was extracted from ESCs, NPCs, MEFs and

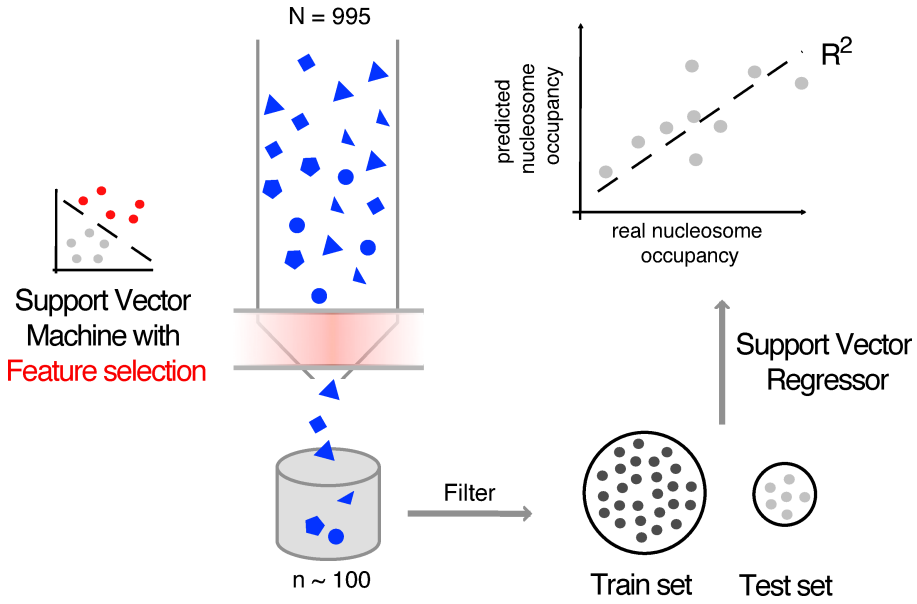

FIGURE 3.25: Operational scheme of the Support Vector Regressor used to predict nucleosome occupancy from the DNA sequence features predictive for Pu.1 binding. Filtering refers to the exclusion of the measurements of theoretical nucleosome occupancy from the features used.

in vitro patterns. The number of nucleosomal fragments spanning the center of each region (corresponding to the Pu.1 ChIP-seq summit for the bound or to the GGAA core in case of the unbound) was counted and the log2-transformed value used as a proxy for occupancy. The information for all the features except the theoretical nucleosomes occupancy (Kaplan et al., 2008) was used to feed a Support Vector Regressor (Drucker et al., 1997), which is a variant of SVM for regression. The entire dataset of bound and unbound sites was split into 90% training and 10% test. The training dataset was used to fit the experimentally determined

nucleosome counts in function of the features in the sequence. The model obtained was then used to predict the nucleosome counts over the test dataset (see schema in figure 3.25). Performance was evaluated through the coefficient of determination (R^2), calculated as the squared Pearson correlation coefficient among the predicted and the observed counts. Results

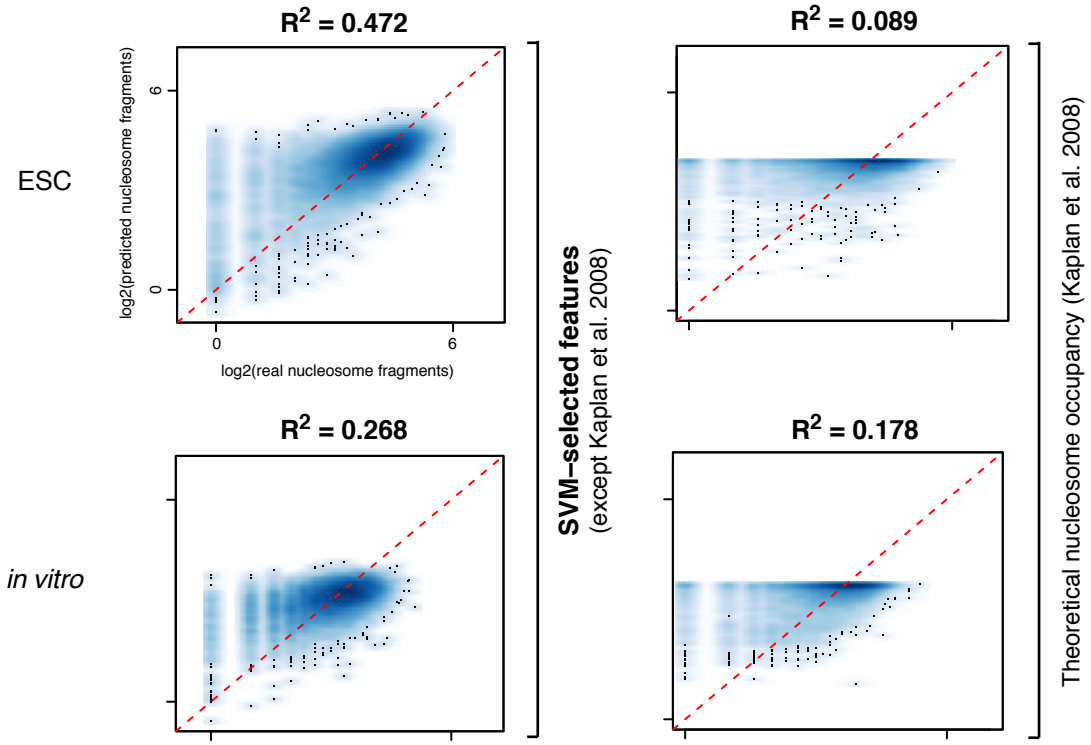

FIGURE 3.26: Smoothed scatterplots of the predicted values in function of the observed log2-transformed values of nucleosome occupancy in ESCs or *in vitro* over Pu.1 sites (using the set of features selected for one of the randomizations of the training-test SVM input datasets). The scatterplot on the right shows the results on the test dataset using only theoretical nucleosome occupancy. The plot on the left show instead the results using all the features selected except for it.

for a representative set of features are summarized in figure 3.26. Smoothed scatterplots show the predicted values in function of the observed values. The features discriminating Pu.1-bound from unbound sites explained 45% of the variability in the nucleosome occupancy pattern at these sites in ESC. Conversely, an SVR trained and tested using only the theoretical nucleosomes occupancy (Kaplan et al., 2008) explained less than 10% of the variability in the same data, which is in agreement with previously published data (Tillo et al., 2010). Interestingly, theoretical nucleosomes occupancy values (which are predicted by a model built upon yeast *in vitro* measurements) perform better on *in vitro* data compared to data from ESCs, and the SVM-selected features only slightly outperform it.

These results are robust when slightly different sets of features (corresponding to multiple

re-initialization of the SVM-based procedure used to predict Pu.1 binding) and different cell types are considered (see boxplots in figure 3.27).

Therefore, sequence determinants of Pu.1 binding also encode part of the information for

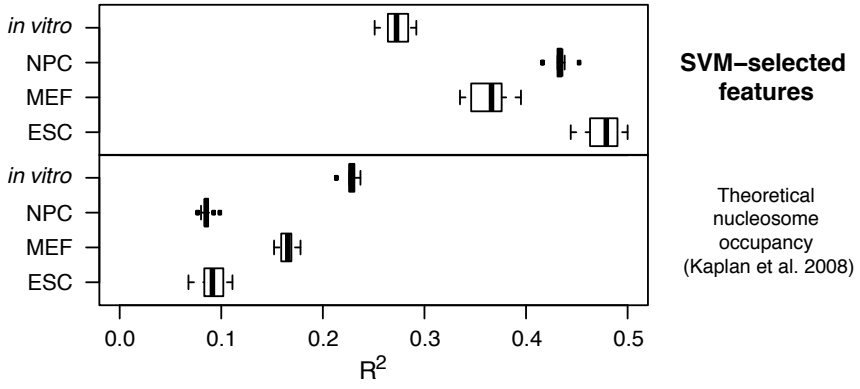

FIGURE 3.27: Boxplots showing the distribution of the (R^2) for the sets of features from the ten training-test datasets randomization of the SVM.

nucleosome affinity. It should be noticed that the results obtained are better (Kaplan et al., 2008, Tillo and Hughes, 2009) or in line (van der Heijden et al., 2012) with published models developed *ad hoc* to predict nucleosome occupancy from the local genomic sequence.

3.8 A detailed evaluation of chromatin organization at Pu.1 binding sites in BMDMs

These observations about the interplay of sequence determinants of Pu.1-binding and nucleosome affinity coupled to the unprecedented sequencing depth we reached in assaying the nucleosome pattern in BMDMs (see table 2.4) prompted us to a much detailed investigation of the nucleosomal patterns occurring at the Pu.1-binding sites (Ostuni et al., 2013). TSS-proximal and TSS-distal sites were defined upon annotation of the Pu.1-bound sites to Ensembl genes (Flicek et al., 2012), resulting in 17,401 (22.63% of the total) and 59,481 (77.37%) regions, respectively.

3.8.1 TSS-distal sites

As already shown, when TSS-distal Pu.1 peaks (corresponding to putative enhancers) were used as central anchoring points, we detected regular arrays of well-positioned nucleosomes (with up to seven nucleosomes on each side of the Pu.1-bound region, see figure 3.12). Since this cumulative distribution is not informative of the behavior of individual genomic regions, we generated a heatmap in which Pu.1 summit-centered nucleosome patterns were sorted based on the decreasing width of the central NDR (see figure 3.28). Regions at the bottom of the heatmap are characterized by narrow NDRs flanked on each side by one prominent nucleosome and then additional nucleosomes whose occupancy progressively diminish with increasing distance from the center. Conversely, regions at the top show broad NDRs that are less clearly demarcated because of the much lower degree of occupancy of the flanking nucleosomes. Pu.1-bound TSS-distal regions sorted by decreasing NDR width were then split into deciles and further analyzed. Although significantly different ($p = 7.89e\text{-}95$ in a Kruskal-Wallis test), Pu.1 occupancies were relatively similar in magnitude across all deciles, with slightly higher scores (score is equivalent to -10*log10(p-value) of the ChIP-seq enrichment over the input) only in the first decile. This is an indication that different degrees of Pu.1 occupancy are not a major determinant of the width of the NDR (see figure 3.29). Considering a larger (+/- 1.5 kbp) area centered on Pu.1, regions in the 1st decile (at the top of the heatmap, broader NDRs) are characterized by an overall lower nucleosome occupancy

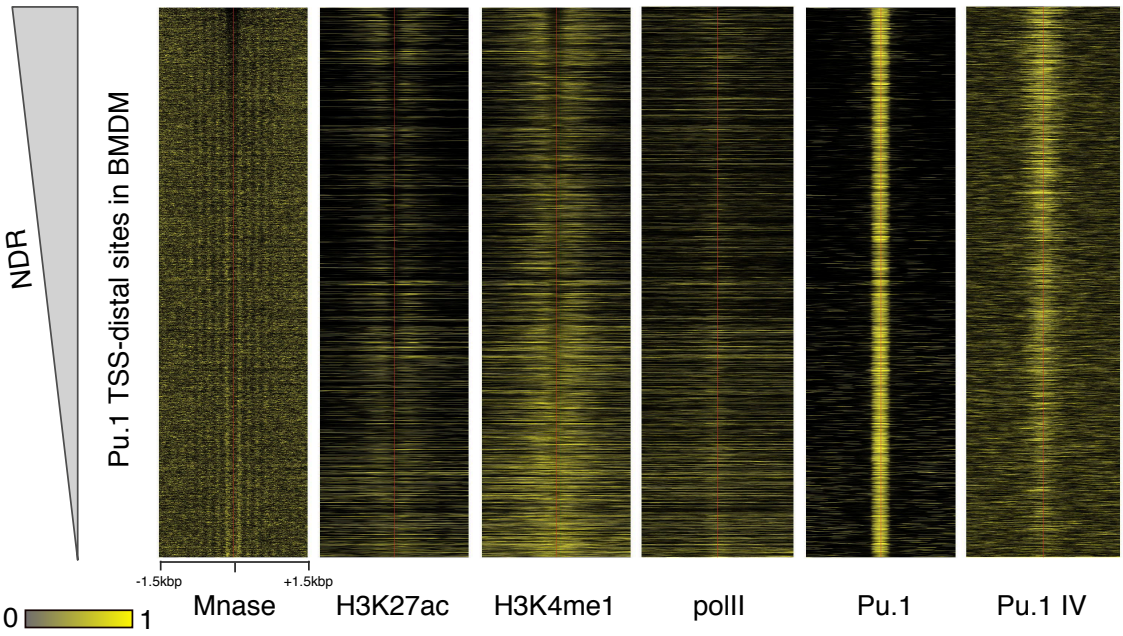

FIGURE 3.28: TSS-distal Pu.1-bound sites in macrophages were sorted according to the extent of the NDR. Nucleosome patterns (Mnase), hPTMs (H3K27ac, H3K4me1), polII and Pu.1 binding profiles (*in vivo* - Pu.1 - and *in vitro* - Pu.1 IV) are shown as heatmaps. Considering the nucleosome midpoints, the counts exceeding the 95th percentile of the overall distribution were set to the value of the 95th percentile. These counts were then normalized in the range 0-1, separately for each region. Considering the ChIP-seq data, the same procedure was applied except that the 0-1 normalization was applied to the entire set (this would emphasize absolute differences in the level of the modifications in the different deciles, while the region-wise normalization is better aimed at showing relative differences at the level of the single region, which is better suited to highlight nucleosome positions).

than those in the 10th (see figure 3.30). This suggests that different properties in terms of nucleosome organization extend beyond the centrally located regulatory region. Importantly, the two NDR-flanking nucleosomes (heretofore indicated as -1 and +1 nucleosomes) are prominent in the regions belonging to the 10th decile and almost absent in those in the 1st, contributing to the width of the NDR in this group. Therefore, although the nucleosome map shows a continuum of behaviors, qualitatively different classes of NDRs that surround Pu.1 peaks can be identified.

H3K27ac and H3K4me1 showed a peculiar bimodal behavior, with bulk signal decreasing in the lower deciles (broader NDRs) and increasing again in the upper deciles (more narrow NDRs). This trend (shown in figure 3.31) is mirrored by hPTMs enrichment that can be spot by statistical analysis of the ChIP-seq data (see table 3.3). Considering H3K4me1, the 1st decile shows almost 60% of overlap with H3K4me1 peaks, a figure that slightly decreases to 52% in the 4th decile and then increases gradually from the 4th to the 10th, up to 80%. A similar trend is observed for the H3K27ac and H3K4me3. Given the extremely different

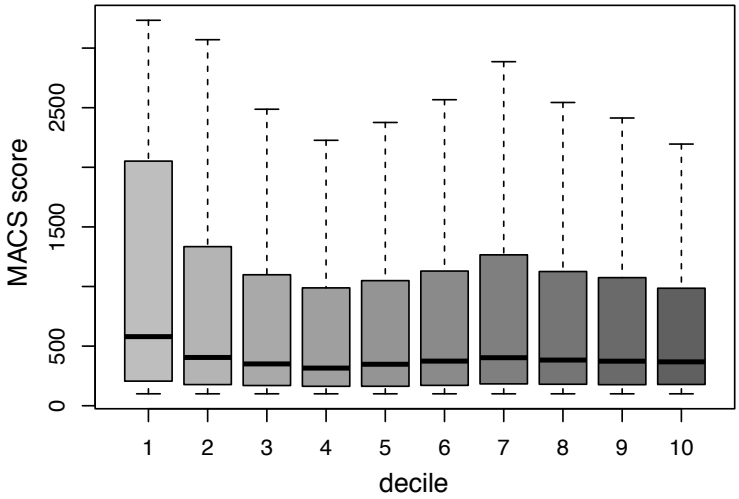

FIGURE 3.29: Pu.1 ChIP-seq score (according to MACS) of the peaks in different deciles are shown. Groups are significantly different (p = 7.89e-95 in a Kruskal-Wallis test) even though only the first decile (larger NDRs) displays a marked increase in ChIP-seq determined occupancies.

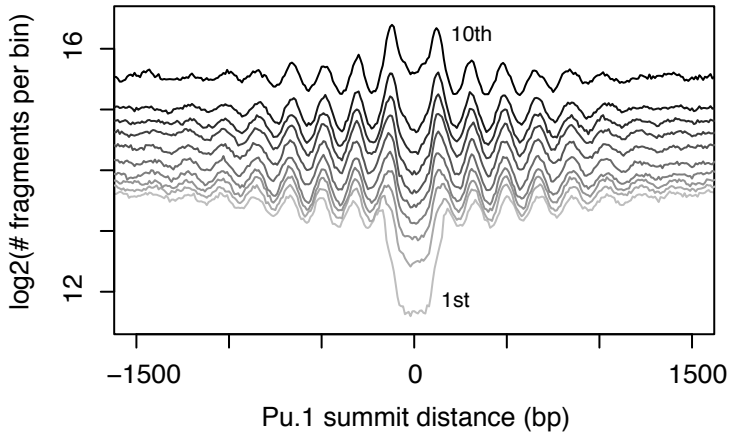

FIGURE 3.30: Bulk signals of the nucleosome midpoints in each one of the deciles defined according to the NDR width (see figure 3.28).

average level of nucleosome occupancy at the deciles, the hPTMs data were normalized according to it. This resulted in a smoother transition from higher to lower levels of H3K27ac and H3K4me1, going from the 10th to the 1st decile. Overall, the regions in the lower deciles show a higher density of nucleosomes, which have a lower probability of being modified. On the other hand, DNA in the upper deciles has a lower propensity to be found into nucleosomes, but these nucleosomes show a higher probability of being modified. Same as saying that the relative amount of modified histones is different, but the absolute amount is comparable among the two subsets. This results raise an important issue (that is not among the

aims of this thesis) and put into a completely different light the results from ChIP targeting hPTMs: normalization by input DNA alone or by using also the nucleosome occupancy of the area can lead to different interpretations.

As observed for the signals of hPTMs (before normalization by the average nucleosome occupancy of the area) polII accumulation shows a bimodal behavior. The bulk signal decreases in the lower deciles (broader NDRs) and increases again in the upper deciles (see figure 3.33), which is also in line with the statistical analysis of the enriched regions (see table 3.3).

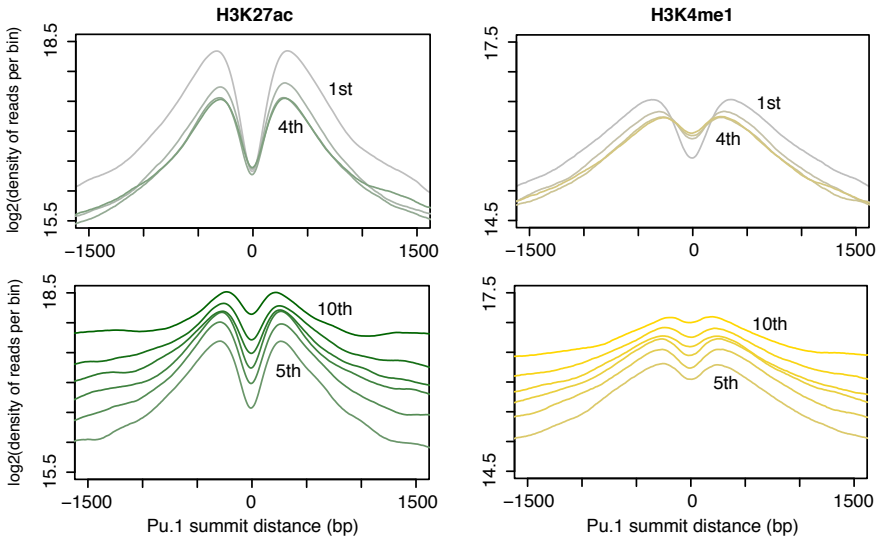

FIGURE 3.31: Bulk signals (density of reads) for the ChIP-seq of H3K27ac (green) and H3K4me1 (yellow) in each one of the deciles defined according to the NDR width (see figure 3.28) are shown.

Decile	H3K4me1	H3K4me3	H3K27ac	polII	IV Pu.1	*Canonical* site	CpGi
1	0.597	0.077	0.434	0.134	0.244	0.472	0.042
2	0.564	0.056	0.332	0.111	0.206	0.497	0.034
3	0.532	0.040	0.296	0.099	0.183	0.532	0.020
4	0.521	0.042	0.293	0.107	0.167	0.530	0.023
5	0.550	0.044	0.316	0.130	0.153	0.522	0.023
6	0.630	0.042	0.393	0.149	0.146	0.504	0.018
7	0.694	0.047	0.449	0.166	0.141	0.486	0.021
8	0.721	0.042	0.464	0.182	0.114	0.494	0.019
9	0.754	0.042	0.518	0.209	0.098	0.475	0.016
10	0.807	0.045	0.597	0.263	0.075	0.473	0.015

TABLE 3.3: Pu.1-bound TSS-distal regions in each decile were overlapped with enrichment peaks derived from ChIP-seq datasets (H3K4me1, H3K4me3, H3K27ac, polII and IV Pu.1, which stands for *in vitro* ChIP targeting Pu.1, see section 3.8.1.3 for details), *canonical* Pu.1-binding sites and CpGi (Illingworth et al., 2010).

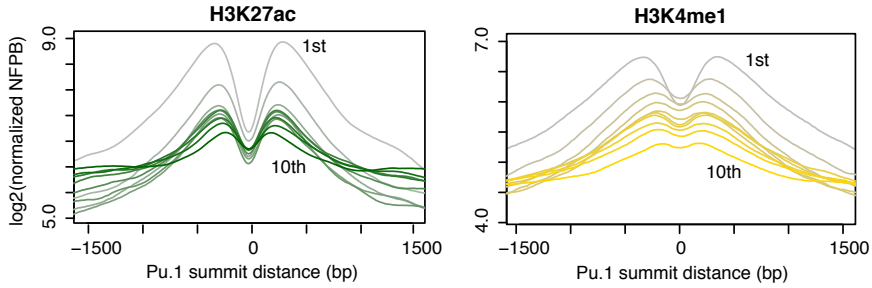

FIGURE 3.32: Bulk signals (density of reads) for the ChIP-seq of H3K27ac (green) and H3K4me1 (yellow) in each one of the deciles defined according to the NDR width (see figure 3.28) are shown. Compared to figure 3.31 the signal of each decile has been divided by the average nucleosome occupancy of the area. NFPB stands for *number of fragments per bin*.

We then tried to assess if enhancers in distinct deciles have a different impact on the transcriptional rate of the neighboring genes. Even though computational assignment of TSS-distal regions to core promoters of target genes is a very inaccurate task (see section 1.10.3) each Pu.1-bound TSS-distal element was assigned to the nearest RefSeq gene with detectable mRNA in the macrophage (see section 2.10 for details). Distributions of FPKMs from the different deciles are significantly different (p = 2.13e-14 in a Kruskal-Wallis test), mainly due to an increase in the 9th and 10th deciles (those ones with higher nucleosome occupancy). Besides, we run GREAT (McLean et al., 2010), which reports enrichment for functional annotations of a dataset of non-coding genomic regions through a probabilistic assignment of each region to nearby genes. We restricted the analysis to the terms associated to biological processes in the Gene Ontology and used very stringent criteria (Bonferroni-corrected hypergeometric p-value <= 0.01, fold enrichment of at least 2). This resulted in very similar lists, irrespective of the decile analyzed. All of them were found enriched for terms related to Immune Response and Hematopoietic System Development, and show no decile-specific enriched terms.

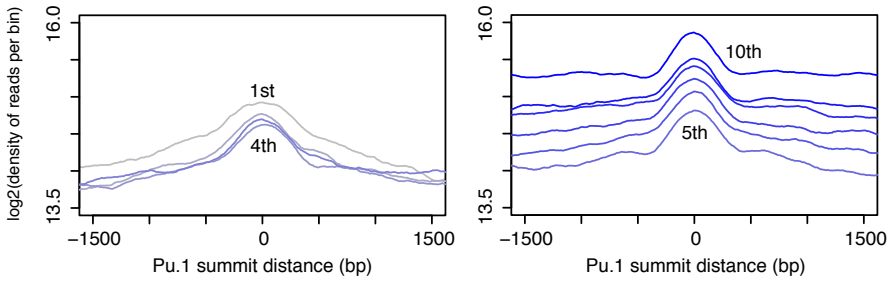

FIGURE 3.33: Bulk signals (density of reads) for the ChIP-seq of polII in each one of the deciles defined according to the NDR width (see figure 3.28) are shown.

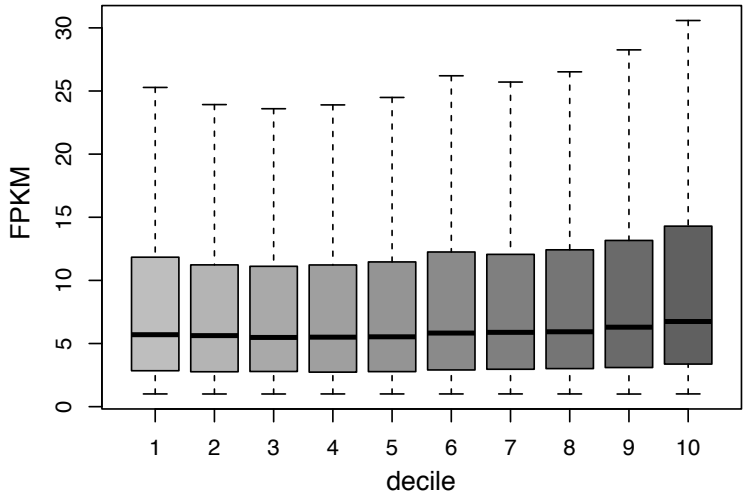

FIGURE 3.34: Each region was annotated to the nearest gene showing detectable mRNA in macrophages. The corresponding FPKMs were used to derive decile-specific distributions. These are significantly different (p = 2.13e-14 in a Kruskal-Wallis test) mainly due to an increase in the 9th-10th deciles.

3.8.1.1 Sequence determinants

We then investigated the composition of the DNA sequences belonging to each decile, considering a region of +/- 150 bp from the Pu.1 summit. We were particularly interested in understanding if the *container site* features observed in the bulk of Pu.1-contacted regions (see figure 3.19) were characteristic of one or more distinct deciles.

The overall C+G content showed a progressive increase from the 1st to the 10th decile (p

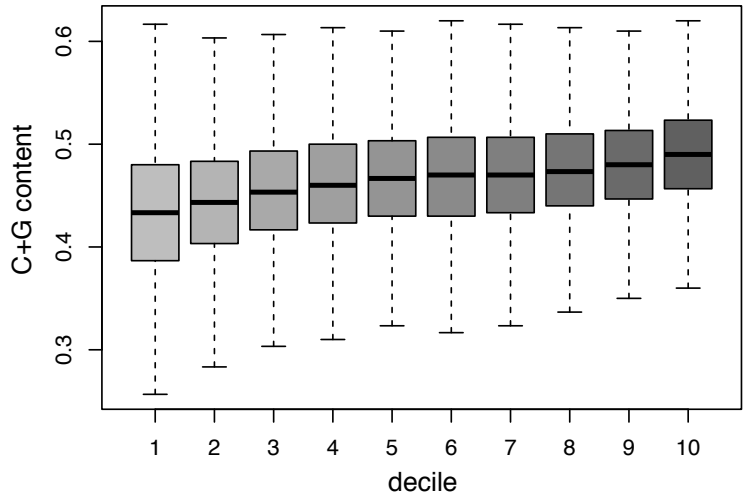

FIGURE 3.35: Considering a region of +/- 150 bp from the Pu.1 summit, C+G content increases with the nucleosome occupancy observed at the Pu.1-bound TSS-distal sites (p <= 1e-300 in a Kruskal-Wallis test).

<= 1e-300 in a Kruskal-Wallis test, see figure 3.35), which is consistent with the progressive increase in nucleosome occupancy. AA dinucleotides, which contribute to generate the repelling elements in *container sites* (Valouev et al., 2011), were more represented in the 1st decile with a peak at -100 and +100 positions (see figure 3.36). This relative enrichment of AA dinucleotides in the flanks may determine the strong depletion of the -1 and +1 nucleosomes in the first decile, a hypothesis that is directly addressed in the next paragraph. In a reciprocal manner, the 1st decile showed a relative depletion of GC and CC dinucleotides in the flanks.

Taken together, these data indicate qualitative and quantitative differences in sequence com-

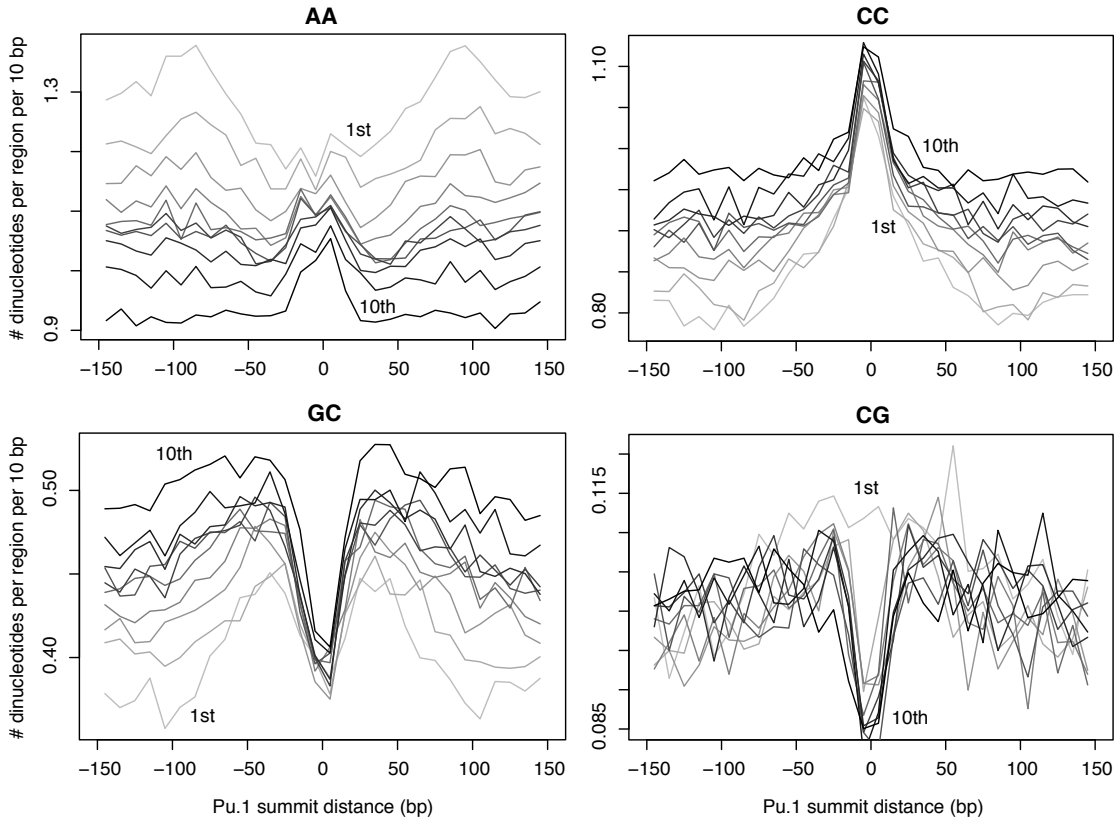

FIGURE 3.36: The positional content for AA, CC, GC and CG is shown as average frequency in the population (considering a bin of 10 bp) for each dinucleotide.

position across deciles and suggest that the interplay between Pu.1 and different underlying sequences may eventually determine the features of distinct classes of NDRs in macrophages. Besides, the *container site* is a feature characteristic only of a subset of Pu.1-bound TSS-distal sites, namely those showing broader NDRs in macrophages.

We then assessed the relative over-representation for binding sites of known TFs using Pscan (Zambelli et al., 2009). Basically, the sequences belonging to each deciles were compared to

the pool of sequences coming from all deciles. Considering some of the well-known families of TFs involved in myeloid differentiation and innate inflammatory response, we observed that while Runt, Maf, and Egr/Klf families are enriched in the upmost and lowest deciles, Irfs and Stats are enriched in the lowest and Nfkb in the upmost. Besides, the ATF-like matrices (including the AP-1 subunits) are evenly distributed among deciles.

3.8.1.2 Nucleosomal patterns in unrelated cell-types and *in vitro*

In order to directly determine the impact of sequence composition on nucleosomal organization at these *cis*-regulatory elements, we analyzed nucleosome occupancy in unrelated cell types that do not express Pu.1 (ESCs, NPCs, MEFs) and in *in vitro* reconstituted mouse chromatin.

In previous paragraphs we already observed that, considering unrelated cell types, higher

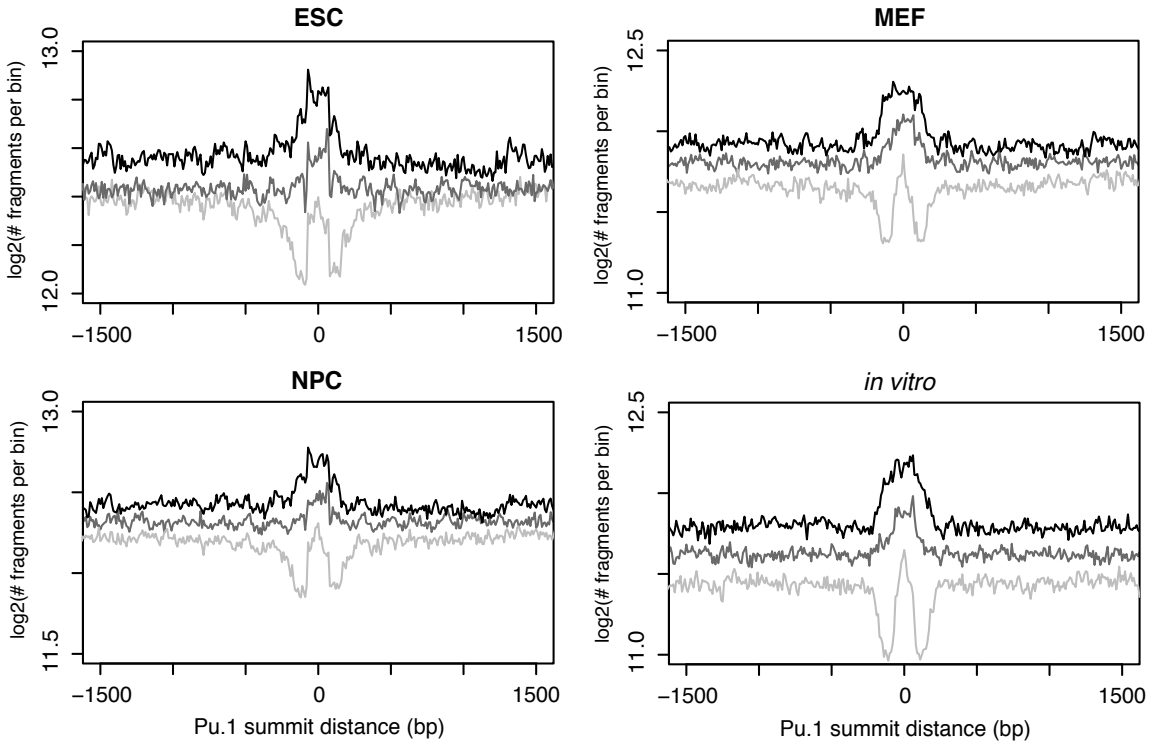

FIGURE 3.37: Cumulative nucleosome profile in cells other than macrophages and *in vitro*. Regions in the 1st (lightgrey), 5th (darkgrey) and 10th (black) deciles are shown.

nucleosome occupancy extending for about a single nucleosome length and overlapping the macrophage Pu.1-bound, nucleosome-depleted regions is detected in case of TSS-distal sites (see figure 3.16). Considering *in vitro* data instead, this holds true for TSS-distal as well as

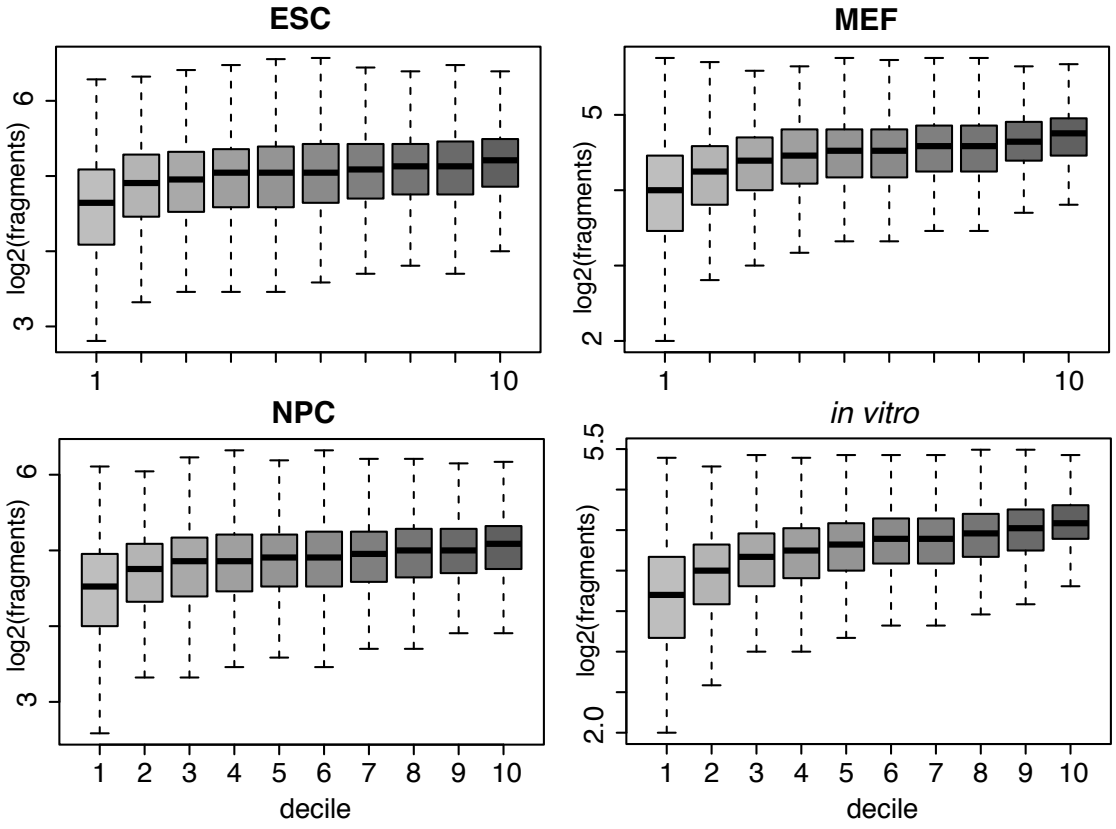

FIGURE 3.38: Considering a region of +/- 150 bp from the Pu.1 summit, the overall nucleosome occupancy in ESCs, NPCs, MEFs and *in vitro* reconstituted chromatin is shown (in all four situations, p <1e-300 in a Kruskal-Wallis test).

TSS-proximal sites (see figure 3.23). When data corresponding to the individual deciles were analyzed separately, distinct regulatory mechanisms impinging on nucleosome occupancy and positioning became apparent (see figure 3.37). Considering the first decile, the central NDR observed in macrophages showed a focused increase in nucleosomal density bracketed by two narrow areas of nucleosome depletion in all other cells. This behavior is entirely compatible with the enrichment in this decile of well-positioned nucleosomes controlled by *container sites* demarcated by AA-rich flanks, which in fact are mainly observed in the 1st decile (see figure 3.36). It is important to notice that these well-positioned nucleosomes occur in the context of the lower nucleosome occupancy characteristic of the 1st decile (see figure 3.38). At the opposite side of the range, the 10th decile was characterized by central nucleosomes with higher occupancy but much lower positioning (as indicated by the width of the signal on the x-axis), which occurred in regions with an overall higher occupancy (see figure 3.38).

Given the existence of these two peculiar categories at the edges of a more continuos distribution of Pu.1-bound sites, we decided to assess if Pu.1 shows a different capability to engage

binding using an *in vitro* reconstituted system.

3.8.1.3 *in vitro* ChIP against Pu.1

We devised an *in vitro* ChIP-seq approach in which *in vitro*-assembled nucleosomes were first digested with MNase and then incubated with macrophage-derived nuclear extracts in order to allow the formation of protein-DNA complexes. Pu.1-bound nucleosomes were immunoprecipitated and subject to HT-sequencing. An *in vitro* Pu.1 ChIP-seq performed with Pu.1-immunodepleted nuclear extracts was used as a reference. Depending on the stringency applied to the *in vivo* dataset of Pu.1-binding sites (the cell type a-specific Pu.1 cistrome, see section 3.2), between 26% and 40% of the Pu.1 binding events observed *in vivo* were recapitulated in the *in vitro* assay (see figure 3.39).

Considering the TSS-distal Pu.1-bound sites in macrophages dissected by NDRs-deciles (see

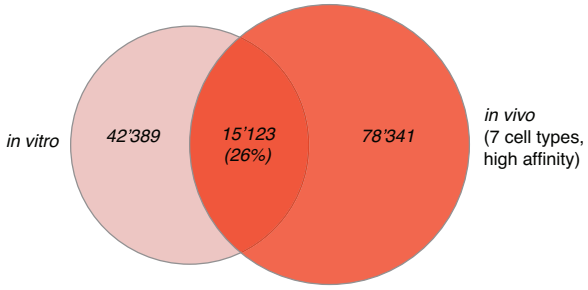

FIGURE 3.39: Venn diagram showing the overlap between *in vitro* and *in vivo* Pu.1-bound sites. *Low-affinity* sites defined in section 2.3.

table 3.3 and figure 3.28), those in the lower deciles (broader NDRs) are those that show the higher rate of binding *in vitro* (up to 25%). Conversely, those in the upper deciles (more narrow NDRs, higher nucleosome occupancy) show a lower rate of binding *in vitro* (down to 7%). Taken together, these results show that an overall high degree of occupancy along the regulatory sequence has a higher detrimental impact on Pu.1 binding than a well-positioned nucleosome precisely located on the Pu.1 binding site. This is compatible not only with the hypothesis that the positioned nucleosome observed in the lower deciles is preventing unspecific binding to the site, but also with a nucleosome-driven mechanism of binding site recognition.

These observations were mirrored by a more quantitative analysis, carried out using nucleosome fragments from the *in vitro* reconstituted chromatin which overlapped the Pu.1-binding

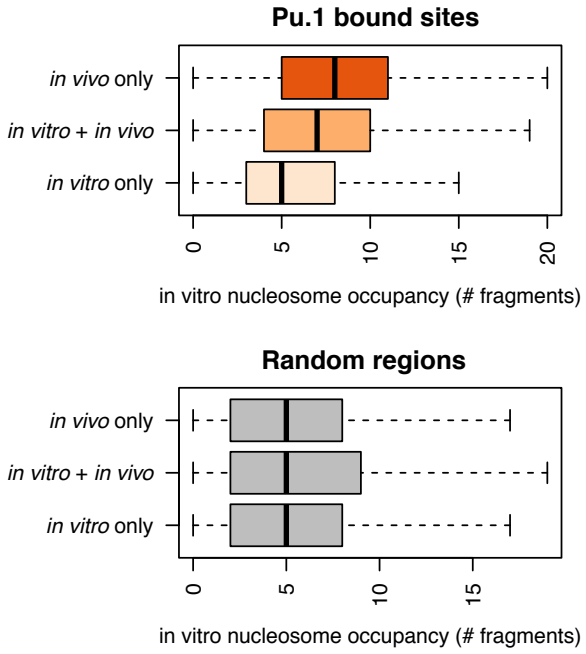

FIGURE 3.40: Boxplots showing the *in vitro* nucleosome occupancy at sites bound by Pu.1 either *in vivo*, *in vitro* or in both conditions. Nucleosome occupancy at random matched sequences is also shown.

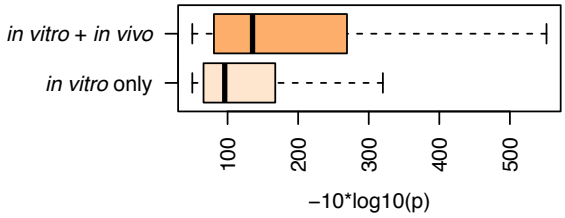

FIGURE 3.41: Pu.1 ChIP-seq scores (score is equivalent to -10*log10(p-value)) at sites bound by Pu.1 both *in vitro* and *in vivo* or *in vitro* only (p <= 1e-300 in a Mann-Whitney test).

sites in the three groups depicted in figure 3.39 (namely sites bound by Pu.1 either *in vivo*, *in vitro* or in both conditions). Those binding events occurring only *in vitro* were restricted to regions of low nucleosome occupancy (see figure 3.40) and show lower *in vitro* TF-occupancy (measured by ChIP-seq occupancy) compared to those occurring both *in vivo* and *in vitro* (see figure 3.41).

Interestingly, among the sites contacted by Pu.1 *in vivo*, those that are also found *in vitro* show a more extensive overlap with *canonical* Pu.1-binding sites (71.6%, see table 3.4) than expected by the overlap of the entire population (42.9%, see section 3.2). Our interpretation of this data is that cooperative binding is disfavored in the *in vitro* conditions (which are

	Canonical Pu.1 sites	TSS (Ensembl genes)
in vitro and *in vivo*	71.6%	18.5%
in vitro only	21.7%	7.7%
in vivo only	40.0%	30.8%

TABLE 3.4: Regions showing a statistically significant enrichment for Pu.1 in an *in vitro* reconstitution experiment were overlapped with the *canonical* Pu.1-binding sites and the TSSs of Ensembl genes.

characterized by non-physiological concentrations of partner TFs and absence of active mechanisms, i.e. ATP is not added to the reaction). In this context, Pu.1-binding is favored at *canonical* sites, at which Pu.1 can bind alone (compared to higher affinity sites, Pham et al., 2013 showed that lower affinity sites are buried in a sequence context enriched for binding sites of partner TFs). Considering the same datasets, those bound *in vitro* and *in vivo* show a less extensive overlap with TSSs of Ensembl genes compared to those binding events that are not recapitulated *in vitro*. Consistent with this - and as it will become apparent in the next paragraph - TSS-proximal Pu.1-bound sites in macrophages show more resistance to Pu.1-binding *in vitro*.

3.8.2 TSS-proximal sites

We then moved to the TSS-proximal set of Pu.1-bound sites in macrophages. The nucleosomal patterns at these sites showed a lower number of phased nucleosomes around the site compared to TSS-distal ones (see panels relative to BMDMs of figures 3.17 and 3.16). Overall, they also show a lower nucleosome occupancy in bulk (compare the y-axes of figures 3.17 and 3.16, panels relative to BMDMs).

As described for the TSS-distal sites, TSS-proximal ones were also split in deciles according to the width of the NDR overlapping the Pu.1-bound site. Considering a larger (+/- 1.5 kbp) area centered on Pu.1, still the regions in the 1st decile are characterized by an overall lower nucleosome occupancy than those in the 10th (see figure 3.42), but the differences span a shorter range of values (2-fold compared to the 4-fold showed by TSS-distal deciles). Considering instead the NDRs themselves, qualitatively different classes can be identified. Furthermore, comparing the extreme deciles, a larger difference was found in the bulk nucleosome occupancy of the NDRs, if compared to TSS-distal ones (5 orders of magnitude compared to 4 orders).

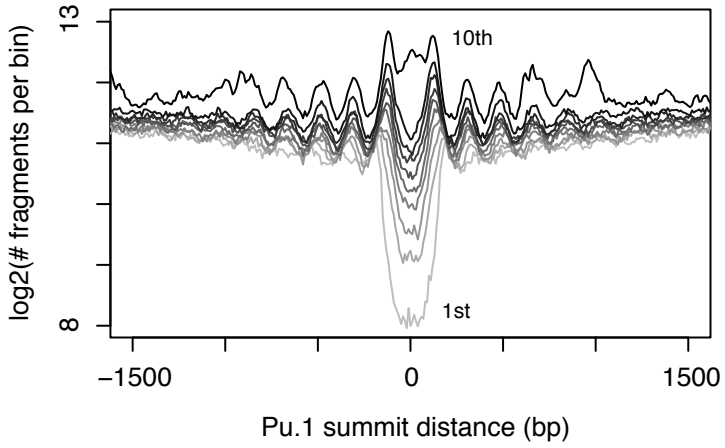

FIGURE 3.42: Bulk signals of the nucleosome midpoints in each one of the deciles defined according to the NDR width.

This could suggest that the regions in the lower deciles (broader NDRs) might have a higher probability to coincide with core promoters, while the upper deciles (more narrow NDRs, higher nucleosome occupancy) to be TSS-proximal enhancers. By directly testing this hypothesis, we found that this is indeed the case. Each region was assigned the nearest TSS and the decile-specific distributions were constructed upon the corresponding distances (see figure 3.43, p <= 1e-300 in a Kruskal-Wallis test).

Even prior to normalization for the average nucleosome occupancy of the area, H3K27ac and

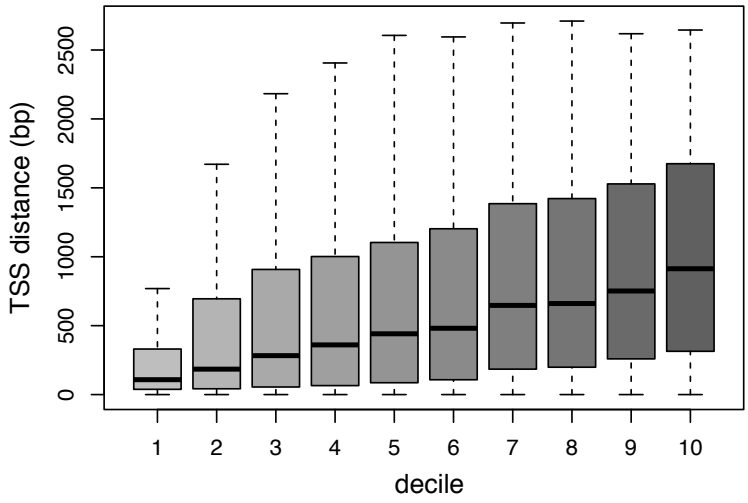

FIGURE 3.43: The distributions of the distances from the nearest TSSs for the regions in the different deciles defined according to the NDR width are shown (p <= 1e-300 in a Kruskal-Wallis test).

H3K4me3 patterns did not show the peculiar bimodal behavior observed for the H3K27ac and H3K4me1 at TSS-distal sites (a decrease in signal from the 1st to the 4th deciles and an increase from the 5th to the 10th deciles, as shown in figure 3.31). According to this,

also the overlap with statistically significant enriched regions showed a smoother transition from high to low percentages (see table 3.5). Considering H3K4me3, the 1st decile shows almost 90% of overlap with H3K4me3 peaks, a figure that decreases almost linearly down to 24% in the 10th decile. The same trend, but coming down to higher absolute values, is observed for the H3K27ac. On the contrary, H3K4me1 show an inversion of this trend, in which around 30% of the regions in the 1st decile overlap H3K4me1, a figure increasing up to 60-70% in the upper deciles. The reversing of the K4me1/K4me3 ratio from the 1st to the 10th decile is in line with a strong relative enrichment for core promoters in the 1st decile and of TSS-proximal enhancers in the 10th one (consistent to what is shown in figure 3.43).

In this case, normalizing by average level of nucleosome occupancy at the deciles did not

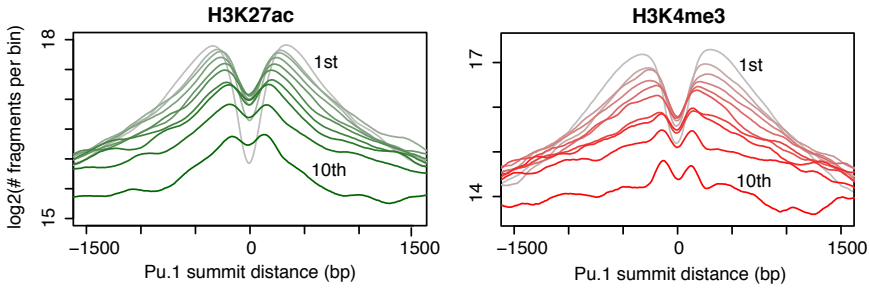

FIGURE 3.44: Bulk signals (density of reads) for the ChIP-seq of H3K27ac (green) and H3K4me3 (red) in each one of the deciles defined according to the NDR width are shown.

change the picture as dramatically as for the TSS-distal sites. Overall, the observations made for the TSS-distal sites hold for the TSS-proximal ones as well. The regions in the lower deciles have a higher density of nucleosomes, with a lower probability of being modified. On the other hand, DNA in the upper deciles show a lower propensity to be found into nucleosomes, but with a higher probability of being modified.

The overlaps with CpG islands (Illingworth et al., 2010) also mirrored the results obtained for the H3K4me3 peaks. Interestingly, the overlaps with the *canonical* Pu.1-binding sites showed particularly low values at broad NDRs (down to 30%). This is compatible with a scenario in which Pu.1 is able to bind at core promoters mainly through cooperative interactions or tethering mechanisms. As already mentioned, a recent paper (Pham et al., 2013) showed that lower affinity Pu.1-binding sites are found in sequences enriched for binding sites of putative partner TFs.

As observed for the signals of hPTMs, polII accumulation shows a smooth transition from lower to upper deciles (see figure 3.46). This is also in line with the statistical analysis of the

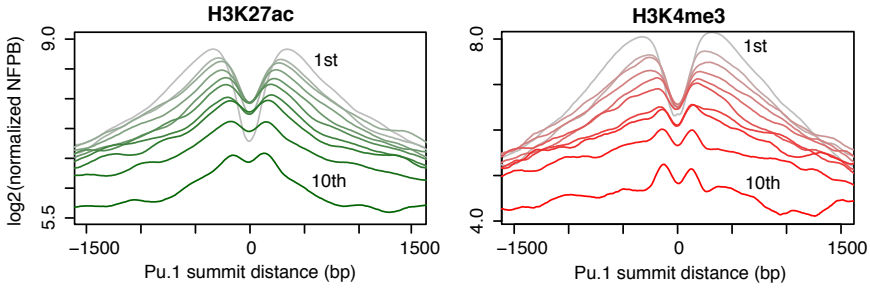

FIGURE 3.45: Bulk signals (density of reads) for the ChIP-seq of H3K27ac (green) and H3K4me3 (red) in each one of the deciles defined according to the NDR width are shown. Compared to figure 3.44 the signal of each decile has been divided by the average nucleosome occupancy of the area. NFPB stands for *number of fragments per bin*.

Decile	H3K4me1	H3K4me3	H3K27ac	polII	IV Pu.1	Canonical site	CpGi
1	0.302	0.893	0.939	0.748	0.059	0.305	0.787
2	0.459	0.802	0.902	0.679	0.093	0.364	0.649
3	0.521	0.766	0.881	0.674	0.078	0.375	0.606
4	0.574	0.703	0.826	0.622	0.091	0.367	0.547
5	0.615	0.675	0.820	0.608	0.072	0.405	0.501
6	0.623	0.610	0.752	0.552	0.082	0.390	0.443
7	0.676	0.509	0.715	0.496	0.084	0.416	0.367
8	0.701	0.496	0.678	0.489	0.064	0.420	0.332
9	0.675	0.408	0.576	0.413	0.071	0.468	0.268
10	0.617	0.236	0.445	0.316	0.074	0.542	0.153

TABLE 3.5: Pu.1-bound TSS-proximal regions in each decile were overlapped with enrichment peaks derived from ChIP-seq datasets (H3K4me1, H3K4me3, H3K27ac, polII and IV Pu.1, which stands for *in vitro* ChIP targeting Pu.1, see section 3.8.1.3 for details), *canonical* Pu.1-binding sites and CpGi (Illingworth et al., 2010).

enriched regions, showing an almost linear decrease of the measured overlap from 75% down to 32% (see table 3.3).

Each region was then annotated to the nearest gene showing detectable mRNA in macrophages

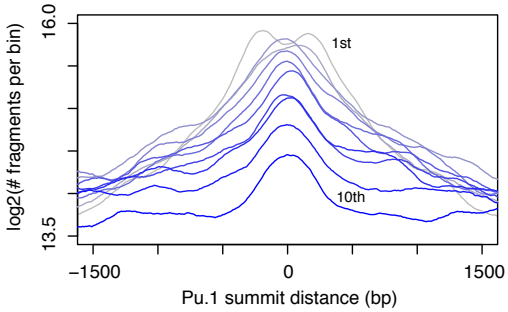

FIGURE 3.46: Bulk signals (density of reads) for the ChIP-seq of polII in each one of the deciles defined according to the NDR width are shown.

(see section 2.10 for details). The corresponding FPKMs were used to build decile-specific

distributions. According to boxplots in figure 3.47, groups are only slightly significantly different ($p = 0.00758$ in a Kruskal-Wallis test). This is due to a minor decrease in the FPKMs belonging to the 10th decile (which are also, compared to regions in the other deciles, those TSS-proximal enhancers that are further away from the nearest TSS, see figure 3.43).

We then evaluated the nucleosome organization in ESCs and *in vitro* at the very same TSS-proximal sites. The bulk signal (see figure 3.48) did recapitulate the positioned nucleosome over the site in ESCs and *in vitro*.

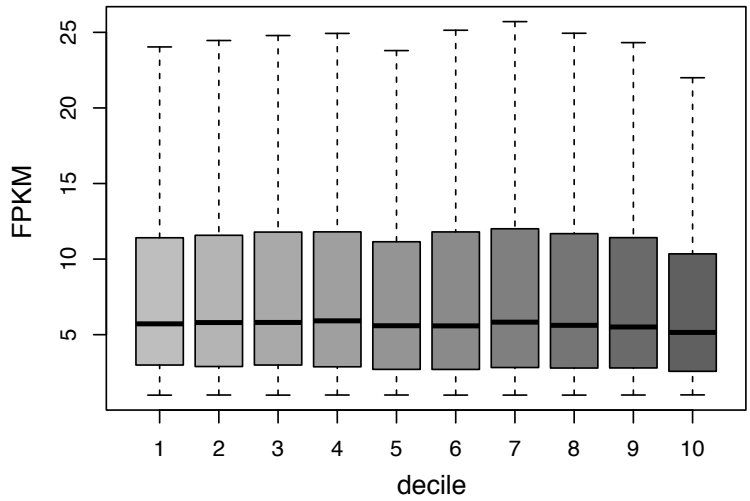

FIGURE 3.47: Each region was annotated to the nearest gene showing detectable mRNA in macrophages. The corresponding FPKMs were used to derive decile-specific distributions. These are slightly significantly different ($p = 0.00758$ in a Kruskal-Wallis test) mainly due to a decrease in the FPKMs belonging to the 10th decile.

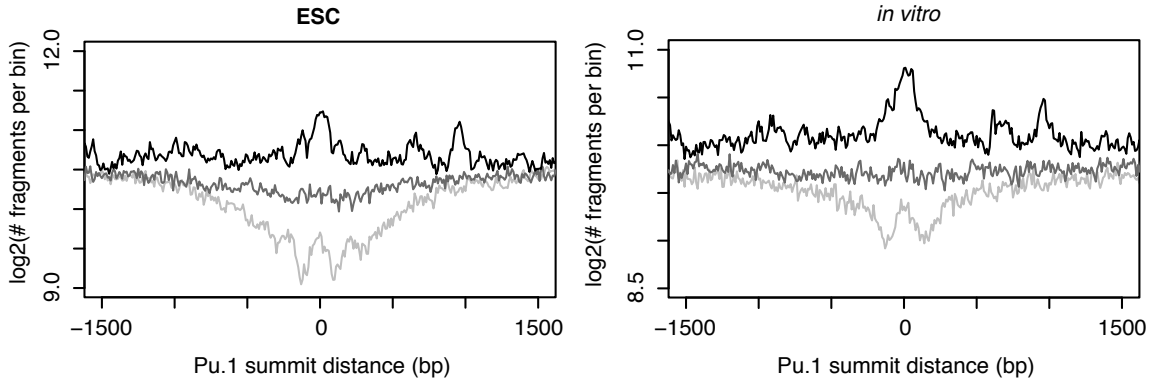

FIGURE 3.48: Cumulative nucleosome profile in ESCs and *in vitro*. Regions in the 1st (light-grey), 5th (darkgrey) and 10th (black) deciles are shown.

3.9 Is Pu.1 required to maintain the nucleosomal organization at *cis*-regulatory elements in BMDMs?

Pu.1 is expressed from the very early stages of macrophage differentiation, where it supervises almost every regulatory event. Besides, it shows the capability to reprogram different cells (e.g. fibroblasts) to macrophage-like cells (see section 1.9.1). The data presented in this thesis support a scenario in which Pu.1 engages its recognition sites only at peculiar sequence contexts, which also correspond to precise nucleosomal conformations. Besides, we showed that Pu.1 is able to contact thousands of its *in vivo* binding sites over *in vitro* reconstituted chromatin, which is an indication of the ability of Pu.1 to invade some chromatin environments even in the absence of ATP-dependent chromatin remodelers. Although a formal demonstration of this pioneering activity is still lacking, all these evidences point to a role for Pu.1 in defining and maintaining the precise nucleosomal conformations at its binding sites.

To directly address the role of Pu.1 in counteracting DNA sequence-driven nucleosome occu-

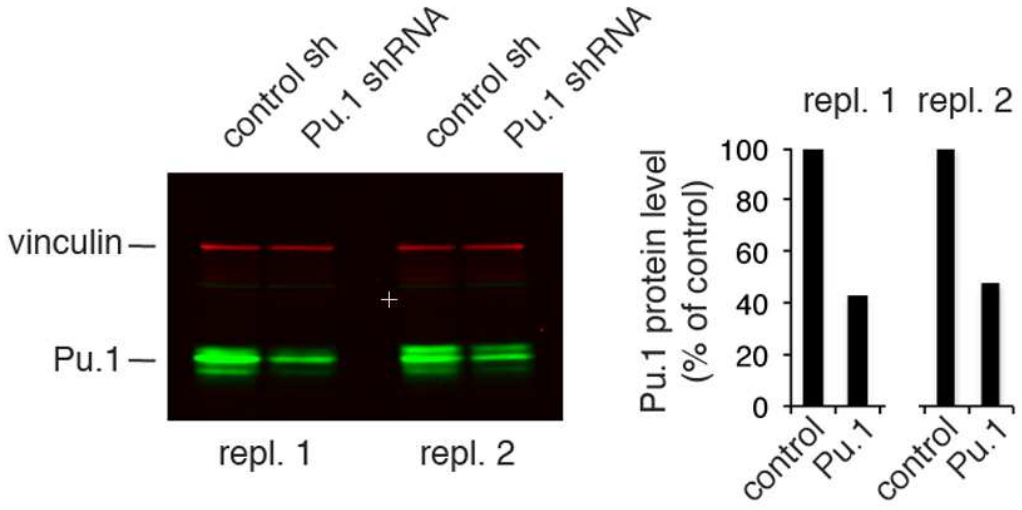

FIGURE 3.49: Acute depletion of Pu.1 in terminally differentiated macrophages using a retrovirally encoded Tet-regulated shRNA. Data from two biological replicates are shown. Vinculin was used as loading control.

pancy and therefore in maintaining nucleosome depletion and accessibility of the underlying regulatory regions in macrophages, we generated a retroviral vector for inducible, doxycycline-regulated expression of an shRNA targeting Pu.1. Bone marrow-derived cells (that proliferate and differentiate in macrophages in M-CSF-containing medium) were infected at day 1 and 2 after plating, selected in puromycin and then induced to express the Pu.1 shRNA at day 5. 48h after shRNA induction we reproducibly obtained around 60% depletion of Pu.1 protein

levels (see figure 3.49). It must be notice that a complete depletion of Pu.1 is not compatible with macrophage survival. Because of the residual amount of the protein not all genomic regions were equally stripped of Pu.1. Therefore we carried out a Pu.1 ChIP-seq to classify regulatory regions based on the level of reduction of Pu.1 binding and in parallel samples of cells we analyzed nucleosome profiles by MNase-Seq. Cells infected with a control retroviral vector (*empty vector*) were used as a reference.

In the experimental setting of the retroviral infection, only a restricted fraction of Pu.1-

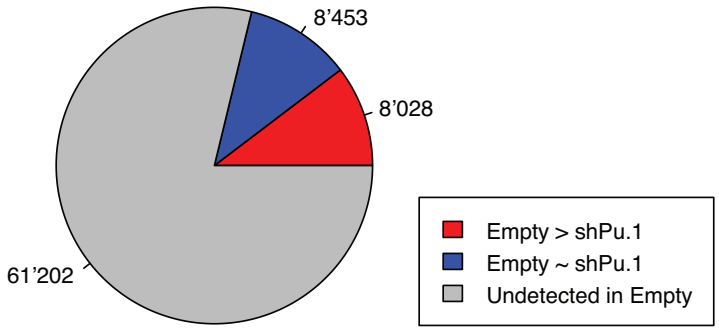

FIGURE 3.50: The pie shows the fraction of Pu.1-bound sites in macrophages that are not affected (grey) and those found in the *empty vector* experiment (blue and red). Among them, nearly half are found to be significantly diminished by the acute depletion of Pu.1 (red).

bound sites in macrophage could be recapitulated (21.2%, see figure 3.50). Among them, nearly half are found to be significantly diminished by the acute depletion of Pu.1. In order to quantify the effect of the lower concentration of Pu.1 on the local nucleosomal organization, the TSS-distal Pu.1 peaks identified by ChIP-seq were divided in quartiles based on the ratio of the Pu.1 signal in Pu.1-depleted vs control cells (the fourth quartile corresponding to peaks showing the stronger reduction in Pu.1 binding, see figure 3.51).

A quantitatively strong and statistically significant increase in nucleosomal reads at TSS-distal Pu.1 regions was detected in both replicates, particularly in the fourth quartile (see figure 3.52 and reported p-values).

Overall, these data indicate that Pu.1 is essential in maintaining the nucleosome depletion at its binding sites, but it seems much less important in affecting the phasing of the nearby nucleosomes (see Discussion).

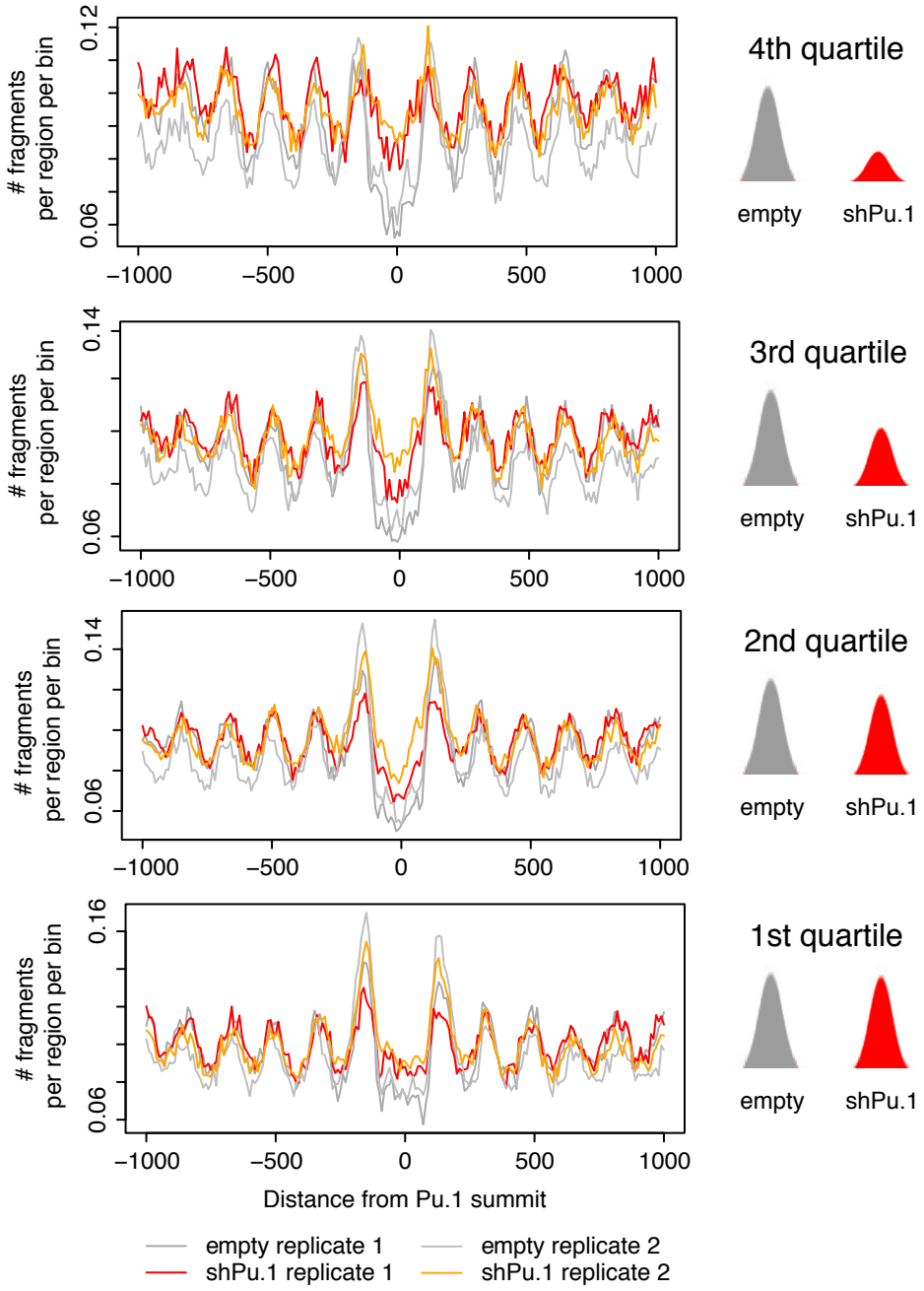

FIGURE 3.51: Pu.1 peaks were divided in quartiles based on the degree of signal reduction in Pu.1-depleted vs. control cells. The 4th quartile corresponds to Pu.1 peaks with the higher reduction in binding occupancy in depleted cells. Distributions of the midpoints of the nucleosome fragments were centered on the summit of Pu.1 peaks. MNase-seq data from two different biological replicates were independently analyzed.

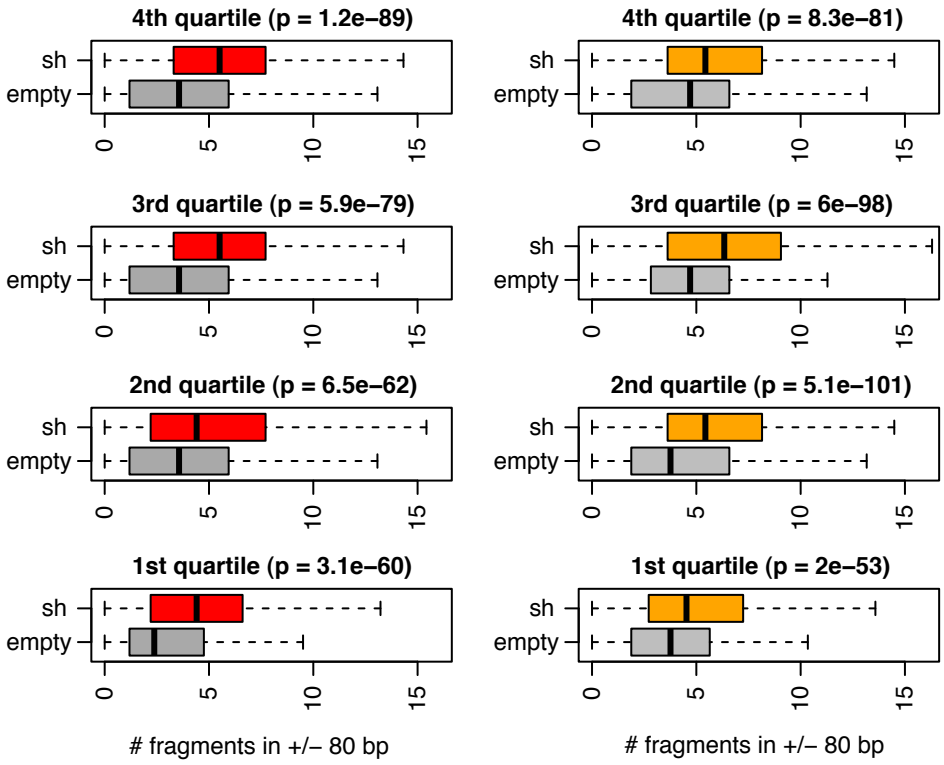

FIGURE 3.52: Midpoints of nucleosomal fragments in +/- 80 bp surrounding Pu.1 peaks were quantified and shown as boxplots, separately for the two replicates (red and orange). Wilcoxon signed-rank test was used to assess the statistical significance of the different nucleosomal occupancy of the corresponding region upon Pu.1 depletion.

Chapter 4

Discussion

The mechanisms by which a TF engages only a specific (and small) fraction of all its genomic consensus sites are still not well-understood. This thesis tackled this very general question and found out that the answer is intrinsically related to the the interplay existing between TF-binding and nucleosome-mediated occlusion of the functional DNA sequences they recognize, which is at the hearth of regulated gene expression. In order to do that, we exploited the availability of extensive high-throughput binding data for a single TF, Pu.1, and coupled it to novel computational analyses, machine learning approaches and the ability to generate nucleosomal patterns *in vivo* and *in vitro* at an unprecedented resolution. The main results can be outlined as follows:

- we trained a SVM-based machine learning approach able to discriminate with 78% accuracy those randomly occurring Pu.1 sites that do not show binding competence from those that are contacted *in vivo* and are therefore potentially functional;
- a feature selection approach was embedded in the training of the SVM, allowing the identification of some of the molecular determinants of binding competence; among them we found i) differences in the ETS core and immediately flanking sequences, especially at the level of the imposed local DNA shape; ii) binding preferences for partner TFs; iii) higher C+G content and theoretical nucleosome occupancy in case of the engaged sites compared to those that are never contacted *in vivo*;

- using the same sequence determinants, we trained a regression model that explains up to 45% of the variability observed in the nucleosome occupancy in cells where Pu.1 is not expressed. This is equal or better than what can be achieved by published *ad hoc* models (Kaplan et al., 2008, van der Heijden et al., 2012);

- we thoroughly investigated both the sequence and the nucleosomal organization at Pu.1-bound sites in macrophages, in cell types where Pu.1 is not expressed, and in *in vitro* reconstituted chromatin. We found out that i) Pu.1-binding generates NDRs flanked by arrays of positioned nucleosomes; ii) the regions bearing Pu.1 sites that can be engaged *in vivo* show an intrinsic higher nucleosome affinity compared to its flanks, as measured by *in vitro* reconstitution experiments, iii) which is also observed in cell types that are not expressing Pu.1; iv) Pu.1-binding occurs in regions with different sequence composition, which is mirroring different nucleosomal configurations (both in the presence and in the absence of Pu.1); in particular, two extreme patterns have been identified. In the former, a region of lower nucleosome occupancy is either occupied by Pu.1 or by a well-positioned nucleosome. In the latter, regions of broad higher occupancy extend on both sides of a centrally located, prominent but less well-positioned nucleosome. These nucleosomal patterns show distinctive underlying DNA sequence features. While in the former case a *container site* is observed, the latter shows a significantly higher C+G content. Importantly, when analyzing Pu.1 recruitment to *in vitro* assembled chromatin, the second configuration showed a stronger inhibition of Pu.1 binding, suggesting that chromatin remodelers may be selectively required only at sequences characterized by an extended high nucleosomal occupancy.

These findings point to a basic organizational principle of mammalian *cis*-regulatory sequences: TF-engagement at its consensus sites and nucleosome occupancy are coordinately controlled by overlapping sequence features. This model also suggests that co-evolution of these features may be crucial to maintain cell-type specific enhancer activation. At least in principle, mutations that change the sequence of a TF binding site or its flanks may have no deleterious effects on binding affinity but may impact on the ability of the same sequence to promote nucleosome occupancy. This would result in the uncoupling of TF-engagement from the sequence determinants important for the proper nucleosomal configuration of the

region and would impair the functional properties of the regulatory region (e.g. by allowing unrestricted TF access to the underlying sequence). We hypothesize that these sites are subject to a *joint* selective pressure for the concurring maintenance of these overlapping features. Disentangling the evolutionary conservation of these single features is a task we have not addressed yet.

The higher nucleosome affinity encoded at the engaged sites compared to the non-engaged ones is compatible with two non-exclusive mechanisms, namely nucleosome-driven TF-binding and nucleosome-mediated masking of the TF binding sites. While the latter might ensure that enhancers exert their action only in the presence of the proper lineage-determining TFs, the former might be relevant to display the binding sites to TFs. We have identified two subsets of Pu.1-engaged sites showing extremely different sequence content and nucleosomal configurations, which are valuable to contextualize these two mechanisms.

Those sites showing lower occupancy in general, but also a well-positioned nucleosome right over the Pu.1 site, are those more likely to be contacted by Pu.1 in *in vitro* reconstitution experiments. This might be a consequence of the lower occupancy, or the result of a nucleosome-assisted mechanism of binding site selection (as recently suggested for the Progesterone Receptor in Ballaré et al., 2012, in its inactive configuration the regulatory element might be constrained such that it can display the Pu.1 site for binding); this mechanism is only one side of the coin. Considering the general lower occupancy of the genomic region, the nucleosome over the Pu.1 site could also serve to prevent aberrant ectopic activation by broadly expressed TFs, in those tissues where the lineage-determining TF is not present.

On the other hand, those sites showing a higher nucleosome occupancy are more clear candidates in preventing stable Pu.1 binding unless the proper ATP-dependent chromatin remodelers are available; in fact, these sites cannot be bound in experiments of *in vitro* chromatin reconstitution in the absence of ATP. In line with this, it has been shown that the Glucocorticoid Receptor (GR) binding to naked DNA proceeds over an extended period of 5-7 min, while it is rapidly recruited to chromatin over brief periods of 30 seconds. GR is transiently trapped and released at dense arrays of nucleosomes, at which it is able to direct the action of chromatin remodelers, which in turn are necessary to stabilize its binding (Nagaich et al., 2004). We can envision that a similar mechanism is in place at these Pu.1-engaged regions. Their higher nucleosome occupancy would represent an unsurmountable barrier for any TF

(avoiding unwanted ectopic regulatory activity), unless the proper sequence is recognized by a pioneer TF that is in turn able to recruit the remodeling machinery.

Although a formal demonstration of this pioneering activity is still lacking, we showed that Pu.1 is able to contact thousands of its *in vivo* binding sites also when chromatin is reconstituted *in vitro*. This is an indication of the competence of Pu.1 to invade some chromatin environments even in the absence of ATP-dependent chromatin remodelers.

It is also interesting to notice that moving across deciles there is not only a difference in the occupancy of the NDR but also a delta in the occupancy of a larger genomic area (around 1.5-fold difference from 1st to 10th decile *in vitro*, a difference that is even amplified in macrophages). This could suggest that some information about the higher order chromatin structure is encoded in the genomic sequence itself. Nevertheless, we cannot exclude that this is in part due to an experimental bias (as a result of an under-representation of polynucleosomes, which can lead to an underestimate more compact chromatin configuration).

We also assessed the impact of the acute depletion of Pu.1 during macrophage differentiation. These experiments indicated that Pu.1 is essential in maintaining the nucleosome depletion at its binding sites, but it seems much less important in affecting the phasing of the nearby nucleosomes. This could be an indication that the maintenance of the NDR depends on the nuclear Pu.1 concentration, while the phasing of the nearby nucleosome does not. This might be further validated comparing the nucleosomal patterns at those Pu.1-bound sites that are invariably contacted in cell types showing a broad range of Pu.1 concentration (in the B- and T- cells lineages, or in progenitor cells). It is important to stress that Pu.1 is essential to macrophage differentiation, so a complete depletion is not feasible under physiological conditions. In this context, introducing Pu.1 in a non-physiological setting might be a valid surrogate to understand how *de novo* deposition of Pu.1 affects the existing nucleosomal configurations (e.g. over-expressing Pu.1 in fibroblasts).

We can envision a number of experiments aimed at corroborating the pioneering activity of Pu.1 and at increasing our understanding about the interplay between Pu.1 and the nucleosomal context of its binding sites:

- we hypothesize that those Pu.1-bound sites showing a higher nucleosome occupancy cannot be bound in experiments of *in vitro* chromatin reconstitution in the absence of

ATP; the *in vitro* chromatin reconstitution experiment with ATP, followed by a ChIP against Pu.1 must be performed in order to disentangle this hypothesis;

- higher affinity for nucleoprotein templates and longer residence times on chromatin is a *bona fide* criteria to distinguish pioneer from non-pioneer TFs. Longer residence times (less nuclear mobility) testify the ability of pioneers to scan the chromatin fiber for their targets. FRAP experiments have been performed to measure nuclear mobility of FoxA1 (Sekiya et al., 2009). Performing FRAP for Pu.1 and compare its nuclear mobility to those of other TFs involved in myeloid differentiation and inflammatory response, would represent a further positive indication about its pioneering activity;

- the impact of the relative positioning between the nucleosome dyad and its recognition sequence over the ability of Pu.1 to access and engage this site remains an open question. This could be tackled by performing EMSA with synthetic oligos. The *601 sequence* (Lowary and Widom, 1998) is a reliable standard positioning signal for *in vitro* studies. This could be engineered moving the Pu.1 site to different positions relative to the nucleosome dyad, and differences in binding-site recognition assessed;

- even though we reached an unprecedented sequencing depth for a nucleosomal pattern in a single cell type, still the number of fragments describing each nucleosome in the population is too low to define their positions with high confidence. Since the fraction of the genome which is of primary interest to study *cis*-regulation in a single cell type is relatively small compared to the size of the genome, higher resolutions can be achieved by target enrichment (TE) strategies. Nevertheless, considering the high range of G+C that must be covered and the extensive overlap with repetitive elements (Tewhey et al., 2009), standard TE strategies will be inappropriate. Instead, it has been recently shown that locus-specific enrichment of mononucleosomal DNA using hybridization to BACs increased the coverage up to 500 fold, compared to previous genome-wide sequencing efforts (Yigit et al., 2013).

It is also important to stress that the performances achieved by the SVM are in line with the state-of-the-art of the field. Discrimination of engaged TF-binding sites in mammalian genomes using sequence information has been successfully tackled in two recent papers

(Yáñez-Cuna et al., 2012, Arvey et al., 2012). While both studies addressed cell type specificity of binding for different TFs and co-regulators (e.g. p300) using SVMs, only Arvey et al., 2012 compared TF binding events occurring *in vivo* to a negative set (namely nearby regions). The results obtained are very good (AUC higher than 0.9) but are calculated only on the best 1,000 ChIP-seq peaks, which represent a very small minority of the entire regulatory repertoire of the majority of mammalian TFs. Arvey et al., 2012 used the surrounding sequence (200 bp away) as negative set, which might be challenging in the sense that the surrounding sequence of a regulatory element might have similar C+G content. Nevertheless, their strategy did not address the real question, namely trying to disentangle occurring binding events from similar recognition sequences that are not productively engaged *in vivo*. When applied to our data, this approach performed poorly compared to ours (AUC of 0.66 compared to an average of 0.86 achieved by our approach).

Considering that in the future we will be able to define smarter methods to capture and summarize the information content of DNA stretches of a few hundreds base pairs, we are aware that the performances achieved certainly represent a lower boundary. More importantly, we have so far used a pool of Pu.1-engaged sites that are coming from different cell-types within the hematopoietic compartment. This means that we are selecting for those features in the sequence that are more general but we are missing those that are context-specific (e.g. binding sites for Pu.1 partners either peculiar for the myeloid, the B- or the T- lineages). We plan to overcome these limitations by taking into account cell-type specific binding into more sophisticated predictors. At the same time, it cannot be excluded (on the contrary, we think this is probably the case) that a part of the variability could be only explained by epigenetic factors. Preliminary analysis using RRBS data (Bock et al., 2012, RRBS is a biased experimental technique so the information could be extracted only for 4.4% of the unbound sites and for 16.6% of the Pu.1-bound sites) suggested a significant difference in the cytosine methylation level of the unbound sites compared to those ones engaged in macrophages (AUC = 0.8). Further analyses are needed in order to disentangle the relative contribution of genetics and epigenetics. In fact, cytosine methylation might be either redundant with (and so explained by) the sequence features themselves or it might add further information (representing a memory of a previous developmental stage, an information that cannot be directly ascribed to the regulatory sequence itself).

Another open question in the field is whether TFBSs found in enhancers show differences in the local sequence context compared to those found in TSSs. At least from the results we obtained so far this does not seem to be the case. TSS-distal Pu.1-bound regions outnumber the TSS-proximal sites. Nevertheless, this does create a bias in the predictions (wrong predictions are equally distributed, $p = 0.36$ in a Chi-squared test). This is in line with the hypothesis that although the genomic context is different, the local sequence determinants for Pu.1 binding are very similar. Still, we cannot exclude the existence of subtle differences. This could be addressed in the future by training a machine specifically on the TSS-proximal sites and testing it on both the sets (and the other way round).

Another point we still have not addressed concerns our understanding of the orientation of the regularly spaced nucleosomes observed around Pu.1-binding sites. A recent paper (Kundaje et al., 2012) suggested that most of the TFs show an asymmetry, namely an array of organized nucleosomes only on one side. We think this is an important point to be investigated, but we also think that a statistic to understand, site by site, if this asymmetry is real or just a computational artifact, is missing. This is the only way to formally show that the clustering of the sites performed to understand if there is an orientation bias is not just highlighting spurious differences at the two sides of the TFBSs.

Bibliography

Dvir Aran and Asaf Hellman. DNA methylation of transcriptional enhancers and cancer predisposition. *Cell*, 154(1):11, 2013.

Michelle N Arbeitman, Eileen EM Furlong, Farhad Imam, Eric Johnson, Brian H Null, Bruce S Baker, Mark A Krasnow, Matthew P Scott, Ronald W Davis, and Kevin P White. Gene expression during the life cycle of *Drosophila melanogaster*. *Science*, 297(5590):2270–2275, 2002.

Yojiro Arinobu, Shin-ichi Mizuno, Yong Chong, Hirokazu Shigematsu, Tadafumi Iino, Hiromi Iwasaki, Thomas Graf, Robin Mayfield, Susan Chan, Philippe Kastner, et al. Reciprocal activation of GATA-1 and PU.1 marks initial specification of hematopoietic stem cells into myeloerythroid and myelolymphoid lineages. *Cell Stem Cell*, 1(4):416–427, 2007.

Cosmas D Arnold, Daniel Gerlach, Christoph Stelzer, Łukasz M Boryń, Martina Rath, and Alexander Stark. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, 339(6123):1074–1077, 2013.

Aaron Arvey, Phaedra Agius, William Stafford Noble, and Christina Leslie. Sequence and chromatin determinants of cell-type–specific transcription factor binding. *Genome research*, 22(9):1723–1734, 2012.

Jonathan Back, David Allman, Susan Chan, and Philippe Kastner. Visualizing PU.1 activity during hematopoiesis. *Experimental hematology*, 33(4):395–402, 2005.

Gwenael Badis, Michael F Berger, Anthony A Philippakis, Shaheynoor Talukder, Andrew R Gehrke, Savina A Jaeger, Esther T Chan, Genita Metzler, Anastasia Vedenko, Xiaoyu Chen, et al. Diversity and complexity in DNA recognition by transcription factors. *Science*, 324(5935):1720–1723, 2009.

Cecilia Ballaré, Giancarlo Castellano, Laura Gaveglia, Sonja Althammer, Juan González-Vallinas, Eduardo Eyras, Francois Le Dily, Roser Zaurin, Daniel Soronellas, Guillermo P Vicent, et al. Nucleosome-driven transcription factor binding and gene regulation. *Molecular cell*, 2012.

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

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

Michael F Berger, Gwenael Badis, Andrew R Gehrke, Shaheynoor Talukder, Anthony A Philippakis, Lourdes Pena-Castillo, Trevis M Alleyne, Sanie Mnaimneh, Olga B Botvinnik, Esther T Chan, et al. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell*, 133(7):1266–1276, 2008.

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

Dev M Bhatt, Amy Pandya-Jones, Ann-Jay Tong, Iros Barozzi, Michelle M Lissner, Gioacchino Natoli, Douglas L Black, and Stephen T Smale. Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. *Cell*, 150(2): 279–290, 2012.

Christoph Bock, Isabel Beerman, Wen-Hui Lien, Zachary D Smith, Hongcang Gu, Patrick Boyle, Andreas Gnirke, Elaine Fuchs, Derrick J Rossi, and Alexander Meissner. DNA methylation dynamics during *in vivo* differentiation of blood and skin stem cells. *Molecular cell*, 2012.

Kristin Brogaard, Liqun Xi, Ji-Ping Wang, and Jonathan Widom. A map of nucleosome positions in yeast at base-pair resolution. *Nature*, 486(7404):496–501, 2012.

Michael Bulger and Mark Groudine. Looping versus linking: toward a model for long-distance gene activation. *Genes & development*, 13(19):2465–2477, 1999.

Yi Cao, Zizhen Yao, Deepayan Sarkar, Michael Lawrence, Gilson J Sanchez, Maura H Parker, Kyle L MacQuarrie, Jerry Davison, Martin T Morgan, Walter L Ruzzo, et al. Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming. *Developmental cell*, 18(4):662–674, 2010.

Juan Manuel Caravaca, Greg Donahue, Justin S Becker, Ximiao He, Charles Vinson, and Kenneth S Zaret. Bookmarking by specific and nonspecific binding of foxa1 pioneer factor to mitotic chromosomes. *Genes & development*, 27(3):251–260, 2013.

Matteo Cesaroni, Davide Cittaro, Alessandro Brozzi, Pier Giuseppe Pelicci, and Lucilla Luzi. CARPET: a web-based package for the analysis of ChIP-chip and expression tiling data. *Bioinformatics*, 24(24):2918–2920, 2008.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

Varodom Charoensawan, Sarath Chandra Janga, Martha L Bulyk, M Madan Babu, and Sarah A Teichmann. DNA sequence preferences of transcriptional activators correlate more strongly than repressors with nucleosomes. *Molecular cell*, 47(2):183–192, 2012.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3): 273–297, 1995.

Antoine Coulon, Carson C Chow, Robert H Singer, and Daniel R Larson. Eukaryotic transcriptional dynamics: from single molecules to cell populations. *Nature Reviews Genetics*, 14(8):572–584, 2013.

Menno P Creyghton, Albert W Cheng, G Grant Welstead, Tristan Kooistra, Bryce W Carey, Eveline J Steine, Jacob Hanna, Michael A Lodato, Garrett M Frampton, Phillip A Sharp, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107(50):21931–21936, 2010.

Francesca De Santa, Maria Grazia Totaro, Elena Prosperini, Samuele Notarbartolo, Giuseppe Testa, and Gioacchino Natoli. The histone H3 lysine-27 demethylase Jmjd3 links inflammation to inhibition of polycomb-mediated gene silencing. *Cell*, 130(6):1083–1094, 2007.

Francesca De Santa, Iros Barozzi, Flore Mietton, Serena Ghisletti, Sara Polletti, Betsabeh Khoramian Tusi, Heiko Muller, Jiannis Ragoussis, Chia-Lin Wei, and Gioacchino Natoli. A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS biology*, 8(5):e1000384, 2010.

Flávio SJ de Souza, Lucía F Franchini, and Marcelo Rubinstein. Exaptation of transposable elements into novel cis-regulatory elements: is the evidence always strong? *Molecular biology and evolution*, 30(6):1239–1251, 2013.

Elzo de Wit and Wouter de Laat. A decade of 3C technologies: insights into nuclear organization. *Genes & development*, 26(1):11–24, 2012.

Rodney P DeKoter and Harinder Singh. Regulation of b lymphocyte and macrophage development by graded expression of pu. 1. *Science*, 288(5470):1439–1441, 2000.

Wulan Deng, Jongjoo Lee, Hongxin Wang, Jeff Miller, Andreas Reik, Philip D Gregory, Ann Dean, and Gerd A Blobel. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell*, 149(6):1233–1244, 2012.

Joseph L DeRisi, Vishwanath R Iyer, and Patrick O Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278(5338):680–686, 1997.

Thomas Derrien, Jordi Estellé, Santiago Marco Sola, David G Knowles, Emanuele Raineri, Roderic Guigó, and Paolo Ribeca. Fast computation and applications of genome mappability. *PLoS One*, 7(1):e30377, 2012.

Evgenia Dimitriadou, Kurt Hornik, Friedrich Leisch, David Meyer, and Andreas Weingessel. Misc functions of the department of statistics (e1071), TU Wien. *R package*, pages 1–5, 2008.

Dale Dorsett. Cohesin: genomic insights into controlling gene transcription and development. *Current opinion in genetics & development*, 21(2):199–206, 2011.

Harris Drucker, Chris JC Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. Support vector regression machines. *Advances in neural information processing systems*, pages 155–161, 1997.

Ian Dunham, Ewan Birney, Bryan R Lajoie, Amartya Sanyal, Xianjun Dong, Melissa Greven, Xinying Lin, Jie Wang, Troy W Whitfield, Jiali Zhuang, et al. An integrated encyclopedia of DNA elements in the human genome. 2012.

Ellis Englesberg, Joseph Irr, Joseph Power, and Nancy Lee. Positive control of enzyme synthesis by gene C in the L-arabinose system. *Journal of bacteriology*, 90(4):946–957, 1965.

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.

Jason Ernst, Pouya Kheradpour, Tarjei S Mikkelsen, Noam Shoresh, Lucas D Ward, Charles B Epstein, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael Coyne, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, 2011.

Ru Feng, Sabrina C Desbordes, Huafeng Xie, Ester Sanchez Tillo, Fiona Pixley, E Richard Stanley, and Thomas Graf. PU. 1 and C/EBPα/β convert fibroblasts into macrophage-like cells. *Proceedings of the National Academy of Sciences*, 105(16):6057–6062, 2008.

Romain Fenouil, Pierre Cauchy, Frederic Koch, Nicolas Descostes, Joaquin Zacarias Cabeza, Charlène Innocenti, Pierre Ferrier, Salvatore Spicuglia, Marta Gut, Ivo Gut, et al. CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. *Genome research*, 22(12):2399–2408, 2012.

Paul Flicek, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, et al. Ensembl 2012. *Nucleic acids research*, 40(D1):D84–D90, 2012.

Pauline A Fujita, Brooke Rhead, Ann S Zweig, Angie S Hinrichs, Donna Karolchik, Melissa S Cline, Mary Goldman, Galt P Barber, Hiram Clawson, Antonio Coelho, et al. The UCSC genome browser database: update 2011. *Nucleic acids research*, 39(suppl 1):D876–D882, 2011.

Daniel J Gaffney, Graham McVicker, Athma A Pai, Yvonne N Fondufe-Mittendorf, Noah Lewellen, Katelyn Michelini, Jonathan Widom, Yoav Gilad, and Jonathan K Pritchard.

Controls of nucleosome positioning in the human genome. *PLoS genetics*, 8(11):e1003036, 2012.

Kathryn E Gardner, C David Allis, and Brian D Strahl. Operating on chromatin, a colorful language where context matters. *Journal of molecular biology*, 409(1):36–46, 2011.

Gaetano Gargiulo, Samuel Levy, Gabriele Bucci, Mauro Romanenghi, Lorenzo Fornasari, Karen Y Beeson, Susanne M Goldberg, Matteo Cesaroni, Marco Ballarini, Fabio Santoro, et al. NA-seq: a discovery tool for the analysis of chromatin structure and dynamics during differentiation. *Developmental cell*, 16(3):466–481, 2009.

Serena Ghisletti, Iros Barozzi, Flore Mietton, Sara Polletti, Francesca De Santa, Elisa Venturini, Lorna Gregory, Lorne Lonie, Adeline Chew, Chia-Lin Wei, et al. Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. *Immunity*, 32(3):317–328, 2010.

Walter Gilbert and Benno Müller-Hill. The lac operator is DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 58(6):2415, 1967.

Paul G Giresi, Jonghwan Kim, Ryan M McDaniell, Vishwanath R Iyer, and Jason D Lieb. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome research*, 17(6):877–885, 2007.

Raluca Gordân, Ning Shen, Iris Dror, Tianyin Zhou, John Horton, Remo Rohs, and Martha L Bulyk. Genomic regions flanking E-box binding sites influence DNA binding specificity of bhlh transcription factors through DNA shape. *Cell reports*, 2013.

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

David S Gross and William T Garrard. Nuclease hypersensitive sites in chromatin. *Annual review of biochemistry*, 57(1):159–197, 1988.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

Outi Hallikas, Kimmo Palin, Natalia Sinjushina, Reetta Rautiainen, Juha Partanen, Esko Ukkonen, and Jussi Taipale. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell*, 124(1):47–59, 2006.

Emily E Hare, Brant K Peterson, Venky N Iyer, Rudolf Meier, and Michael B Eisen. Sepsid even-skipped enhancers are functionally conserved in *Drosophila* despite lack of sequence conservation. *PLoS genetics*, 4(6):e1000106, 2008.

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, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–112, 2009.

Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C Lin, Peter Laslo, Jason X Cheng, Cornelis Murre, Harinder Singh, and Christopher K Glass. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Molecular cell*, 38(4):576–589, 2010.

Joung-Woo Hong, David A Hendrix, and Michael S Levine. Shadow enhancers as a source of evolutionary novelty. *Science*, 321(5894):1314, 2008.

Robert S Illingworth, Ulrike Gruenewald-Schneider, Shaun Webb, Alastair RW Kerr, Keith D James, Daniel J Turner, Colin Smith, David J Harrison, Robert Andrews, and Adrian P Bird. Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS genetics*, 6(9):e1001134, 2010.

Ilya P Ioshikhes, Istvan Albert, Sara J Zanton, and B Franklin Pugh. Nucleosome positions predicted through comparative genomics. *Nature genetics*, 38(10):1210–1215, 2006.

François Jacob and Jacques Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology*, 3(3):318–356, 1961.

Cizhong Jiang and B Franklin Pugh. Nucleosome positioning and gene regulation: advances through genomics. *Nature Reviews Genetics*, 10(3):161–172, 2009.

Arttu Jolma, Teemu Kivioja, Jarkko Toivonen, Lu Cheng, Gonghong Wei, Martin Enge, Mikko Taipale, Juan M Vaquerizas, Jian Yan, Mikko J Sillanpää, et al. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome research*, 20(6):861–873, 2010.

Arttu Jolma, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, et al. DNA-binding specificities of human transcription factors. *Cell*, 152(1):327–339, 2013.

Michael H Kagey, Jamie J Newman, Steve Bilodeau, Ye Zhan, David A Orlando, Nynke L van Berkum, Christopher C Ebmeier, Jesse Goossens, Peter B Rahl, Stuart S Levine, et al. Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467 (7314):430–435, 2010.

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, et al. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, 458(7236): 362–366, 2008.

Tommy Kaplan, Xiao-Yong Li, Peter J Sabo, Sean Thomas, John A Stamatoyannopoulos, Mark D Biggin, and Michael B Eisen. Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early *Drosophila* development. *PLoS genetics*, 7(2):e1001290, 2011.

Tae-Kyung Kim, Martin Hemberg, Jesse M Gray, Allen M Costa, Daniel M Bear, Jing Wu, David A Harmin, Mike Laptewicz, Kellie Barbara-Haley, Scott Kuersten, et al. Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295):182–187, 2010.

Roger Kornberg. The location of nucleosomes in chromatin: specific or statistical? 1981.

Roger D Kornberg. Chromatin structure: a repeating unit of histones and DNA. *Science*, 184(4139):868–871, 1974.

Roger D Kornberg and Yahli Lorch. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell*, 98(3):285–294, 1999.

Tony Kouzarides. Chromatin modifications and their function. *Cell*, 128(4):693–705, 2007.

Ivan V Kulakovskiy, Yulia A Medvedeva, Ulf Schaefer, Artem S Kasianov, Ilya E Vorontsov, Vladimir B Bajic, and Vsevolod J Makeev. HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic acids research*, 41(D1):D195–D202, 2013.

Anshul Kundaje, Sofia Kyriazopoulou-Panagiotopoulou, Max Libbrecht, Cheryl L Smith, Debasish Raha, Elliott E Winters, Steven M Johnson, Michael Snyder, Serafim Batzoglou, and Arend Sidow. Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome research*, 22(9):1735–1747, 2012.

Hojoong Kwak, Nicholas J Fuda, Leighton J Core, and John T Lis. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science*, 339(6122):950–953, 2013.

Andrew T Kwon, Alice Yi Chou, David J Arenillas, and Wyeth W Wasserman. Validation of skeletal muscle cis-regulatory module predictions reveals nucleotide composition bias in functional enhancers. *PLoS computational biology*, 7(12):e1002256, 2011.

Michael TY Lam, Han Cho, Hanna P Lesch, David Gosselin, Sven Heinz, Yumiko Tanaka-Oishi, Christopher Benner, Minna U Kaikkonen, Aneeza S Kim, Mika Kosaka, et al. Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. *Nature*, 2013.

Ben Langmead, Cole Trapnell, Mihai Pop, Steven L Salzberg, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3): R25, 2009.

Toby Lawrence and Gioacchino Natoli. Transcriptional regulation of macrophage polarization: enabling diversity with identity. *Nature Reviews Immunology*, 11(11):750–761, 2011.

Martina I Lefterova, David J Steger, David Zhuo, Mohammed Qatanani, Shannon E Mullican, Geetu Tuteja, Elisabetta Manduchi, Gregory R Grant, and Mitchell A Lazar. Cell-specific determinants of peroxisome proliferator-activated receptor γ function in adipocytes and macrophages. *Molecular and cellular biology*, 30(9):2078–2089, 2010.

Boris Lenhard, Albin Sandelin, and Piero Carninci. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature Reviews Genetics*, 13(4): 233–245, 2012.

Mike Levine. Transcriptional enhancers in animal development and evolution. *Current biology*, 20(17):R754–R763, 2010.

Wenbo Li, Dimple Notani, Qi Ma, Bogdan Tanasa, Esperanza Nunez, Aaron Yun Chen, Daria Merkurjev, Jie Zhang, Kenneth Ohgi, Xiaoyuan Song, et al. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature*, 2013.

Zhaoyu Li, Paul Gadue, Kaifu Chen, Yang Jiao, Geetu Tuteja, Jonathan Schug, Wei Li, and Klaus H Kaestner. Foxa2 and H2A.Z mediate nucleosome depletion during embryonic stem cell differentiation. *Cell*, 151(7):1608–1616, 2012.

Louisa M Liberman and Angelike Stathopoulos. Design flexibility in *cis*-regulatory control of gene expression: synthetic and comparative evidence. *Developmental biology*, 327(2): 578–589, 2009.

Colin R Lickwar, Florian Mueller, Sean E Hanlon, James G McNally, and Jason D Lieb. Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function. *Nature*, 484(7393):251–255, 2012.

Ryan Lister, Mattia Pelizzola, Robert H Dowen, R David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *nature*, 462(7271): 315–322, 2009.

Ryan Lister, Mattia Pelizzola, Yasuyuki S Kida, R David Hawkins, Joseph R Nery, Gary Hon, Jessica Antosiewicz-Bourget, Ronan OMalley, Rosa Castanon, Sarit Klugman, et al. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature*, 471(7336):68–73, 2011.

PT Lowary and J Widom. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *Journal of molecular biology*, 276 (1):19–42, 1998.

Karolin Luger, Armin W Mäder, Robin K Richmond, David F Sargent, and Timothy J Richmond. Crystal structure of the nucleosome core particle at 2.8 å resolution. *Nature*, 389(6648):251–260, 1997.

Karolin Luger, Thomas J Rechsteiner, and Timothy J Richmond. Preparation of nucleosome core particle from recombinant histones. *Methods in enzymology*, 304:3–19, 1999.

Richard W Lusk and Michael B Eisen. Evolutionary mirages: selection on binding site composition creates the illusion of conserved grammars in *Drosophila* enhancers. *PLoS genetics*, 6(1):e1000829, 2010.

Luca Magnani, Jérôme Eeckhoute, and Mathieu Lupien. Pioneer factors: directing transcriptional regulators within the chromatin environment. *Trends in Genetics*, 27(11):465–474, 2011.

T Maniatis, JV Falvo, TH Kim, TK Kim, CH Lin, BS Parekh, and MG Wathelet. Structure and function of the interferon-β enhanceosome. In *Cold Spring Harbor symposia on quantitative biology*, volume 63, pages 609–620. Cold Spring Harbor Laboratory Press, 1998.

Glenn A Maston, Sara K Evans, and Michael R Green. Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, 7:29–59, 2006.

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, 2008.

Cory Y McLean, Dave Bristor, Michael Hiller, Shoa L Clarke, Bruce T Schaar, Craig B Lowe, Aaron M Wenger, and Gill Bejerano. GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology*, 28(5):495–501, 2010.

Menie Merika and Dimitris Thanos. Enhanceosomes. *Current opinion in genetics & development*, 11(2):205–208, 2001.

Tarjei S Mikkelsen, Manching Ku, David B Jaffe, Biju Issac, Erez Lieberman, Georgia Giannoukos, Pablo Alvarez, William Brockman, Tae-Kyung Kim, Richard P Koche, et al.

Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–560, 2007.

Alan C Mullen, David A Orlando, Jamie J Newman, Jakob Lovén, Roshan M Kumar, Steve Bilodeau, Jessica Reddy, Matthew G Guenther, Rodney P DeKoter, and Richard A Young. Master transcription factors determine cell-type specific responses to TGF-β signaling. *Cell*, 147(3):565–576, 2011.

Akhilesh K Nagaich, Dawn A Walker, Ron Wolford, and Gordon L Hager. Rapid periodic binding and displacement of the glucocorticoid receptor during chromatin remodeling. *Molecular cell*, 14(2):163–174, 2004.

Gioacchino Natoli. Maintaining cell identity through global control of genomic organization. *Immunity*, 33(1):12–24, 2010.

Hillary CM Nelson, John T Finch, Bonaventura F Luisi, and Aaron Klug. The structure of an oligo (dA)· oligo (dT) tract and its biological implications. *Nature*, 330(6145):221–226, 1987.

Shane Neph, Andrew B Stergachis, Alex Reynolds, Richard Sandstrom, Elhanan Borenstein, and John A Stamatoyannopoulos. Circuitry and dynamics of human transcription factor regulatory networks. *Cell*, 2012a.

Shane Neph, Jeff Vierstra, Andrew B Stergachis, Alex P Reynolds, Eric Haugen, Benjamin Vernot, Robert E Thurman, Sam John, Richard Sandstrom, Audra K Johnson, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489 (7414):83–90, 2012b.

Efrat Lidor Nili, Yair Field, Yaniv Lubling, Jonathan Widom, Moshe Oren, and Eran Segal. p53 binds preferentially to genomic regions with high DNA-encoded nucleosome occupancy. *Genome research*, 20(10):1361–1368, 2010.

Stephen L Nutt and Barbara L Kee. The transcriptional regulation of B cell lineage commitment. *Immunity*, 26(6):715–725, 2007.

Young Min Oh, Jong Kyoung Kim, Seungjin Choi, and Joo-Yeon Yoo. Identification of co-occurring transcription factor binding sites from DNA sequence using clustered position weight matrices. *Nucleic acids research*, 40(5):e38–e38, 2012.

Stuart H Orkin and Leonard I Zon. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell*, 132(4):631–644, 2008.

Valerio Orlando, Helen Strutt, and Renato Paro. Analysis of chromatin structure by in vivo formaldehyde cross-linking. *Methods*, 11(2):205–214, 1997.

Ulf Andersson Ørom, Thomas Derrien, Malte Beringer, Kiranmai Gumireddy, Alessandro Gardini, Giovanni Bussotti, Fan Lai, Matthias Zytnicki, Cedric Notredame, Qihong Huang, et al. Long noncoding RNAs with enhancer-like function in human cells. *Cell*, 143(1):46–58, 2010.

Renato Ostuni, Viviana Piccolo, Iros Barozzi, Sara Polletti, Alberto Termanini, Silvia Bonifacio, Alessia Curina, Elena Prosperini, Serena Ghisletti, and Gioacchino Natoli. Latent enhancers activated by stimulation in differentiated cells. *Cell*, 152(1):157–171, 2013.

Yongping Pan, Chung-Jung Tsai, Buyong Ma, and Ruth Nussinov. Mechanisms of transcription factor selectivity. *Trends in Genetics*, 26(2):75–83, 2010.

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):1170–1177, 2007.

Aleksandra Pekowska, Touati Benoukraf, Joaquin Zacarias-Cabeza, Mohamed Belhocine, Frederic Koch, Hélène Holota, Jean Imbert, Jean-Christophe Andrau, Pierre Ferrier, and Salvatore Spicuglia. H3K4 tri-methylation provides an epigenetic signature of active enhancers. *The EMBO journal*, 30(20):4198–4210, 2011.

Carlos-Filipe Pereira, Ihor R Lemischka, and Kateri Moore. Reprogramming cell fates: insights from combinatorial approaches. *Annals of the New York Academy of Sciences*, 1266 (1):7–17, 2012.

Thu-Hang Pham, Julia Minderjahn, Christian Schmidl, Helen Hoffmeister, Sandra Schmidhofer, Wei Chen, Gernot Längst, Christopher Benner, and Michael Rehli. Mechanisms of *in vivo* binding site selection of the hematopoietic master transcription factor PU. 1. *Nucleic acids research*, 2013.

Elodie Portales-Casamar, Supat Thongjuea, Andrew T Kwon, David Arenillas, Xiaobei Zhao, Eivind Valen, Dimas Yusuf, Boris Lenhard, Wyeth W Wasserman, and Albin Sandelin. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic acids research*, 38(suppl 1):D105–D110, 2010.

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

Oren Ram, Alon Goren, Ido Amit, Noam Shoresh, Nir Yosef, Jason Ernst, Manolis Kellis, Melissa Gymrek, Robbyn Issner, Michael Coyne, et al. Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. *Cell*, 147(7): 1628–1639, 2011.

Timothy Ravasi, Harukazu Suzuki, Carlo Vittorio Cannistraci, Shintaro Katayama, Vladimir B Bajic, Kai Tan, Altuna Akalin, Sebastian Schmeier, Mutsumi Kanamori-Katayama, Nicolas Bertin, et al. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, 140(5):744–752, 2010.

Ho Sung Rhee and B Franklin Pugh. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature*, 483(7389):295–301, 2012.

Remo Rohs, Sean M West, Alona Sosinsky, Peng Liu, Richard S Mann, and Barry Honig. The role of dna shape in protein-dna recognition. *Nature*, 461(7268):1248–1253, 2009.

Mali Salmon-Divon, Heidi Dvinge, Kairi Tammoja, and Paul Bertone. PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci. *BMC bioinformatics*, 11(1):415, 2010.

Robert M Samstein, Aaron Arvey, Steven Z Josefowicz, Xiao Peng, Alex Reynolds, Richard Sandstrom, Shane Neph, Peter Sabo, Jeong M Kim, Will Liao, et al. Foxp3 exploits a pre-existent enhancer landscape for regulatory T cell lineage specification. *Cell*, 151(1): 153–166, 2012.

Amartya Sanyal, Bryan R Lajoie, Gaurav Jain, and Job Dekker. The long-range interaction landscape of gene promoters. *Nature*, 489(7414):109–113, 2012.

Sandra C Satchwell, Horace R Drew, and Andrew A Travers. Sequence periodicities in chicken nucleosome core DNA. *Journal of molecular biology*, 191(4):659–675, 1986.

Eran Segal and Jonathan Widom. Poly (dA:dT) tracts: major determinants of nucleosome organization. *Current opinion in structural biology*, 19(1):65–71, 2009.

Eran Segal, Yvonne Fondufe-Mittendorf, Lingyi Chen, AnnChristine Thåström, Yair Field, Irene K Moore, Ji-Ping Z Wang, and Jonathan Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104):772–778, 2006.

Takashi Sekiya, Uma M Muthurajan, Karolin Luger, Alexei V Tulin, and Kenneth S Zaret. Nucleosome-binding affinity as a primary determinant of the nuclear mobility of the pioneer transcription factor foxA. *Genes & development*, 23(7):804–809, 2009.

Kate Senger, Grant W Armstrong, William J Rowell, Jennifer M Kwan, Michele Markstein, and Michael Levine. Immunity regulatory DNAs share common organizational features in *Drosophila*. *Molecular cell*, 13(1):19–32, 2004.

Richard R Sinden. *DNA structure and function*. Gulf Professional Publishing, 1994.

Stephen T Smale and James T Kadonaga. The RNA polymerase II core promoter. *Annual review of biochemistry*, 72(1):449–479, 2003.

Arian FA Smit, Robert Hubley, and Phil Green. RepeatMasker Open-3.0, 1996.

Zachary D Smith and Alexander Meissner. DNA methylation: roles in mammalian development. *Nature Reviews Genetics*, 2013.

Lingyun Song, Zhancheng Zhang, Linda L Grasfeder, Alan P Boyle, Paul G Giresi, Bum-Kyu Lee, Nathan C Sheffield, Stefan Gräf, Mikael Huss, Damian Keefe, et al. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome research*, 21(10):1757–1767, 2011.

François Spitz and Eileen EM Furlong. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*, 2012.

Kevin Struhl and Eran Segal. Determinants of nucleosome positioning. *Nature structural & molecular biology*, 20(3):267–273, 2013.

Vladimir B Teif, Yevhen Vainshtein, Maïwen Caudron-Herger, Jan-Philipp Mallm, Caroline Marth, Thomas Höfer, and Karsten Rippe. Genome-wide nucleosome positioning during embryonic stem cell development. *Nature structural & molecular biology*, 2012.

Ryan Tewhey, Masakazu Nakano, Xiaoyun Wang, Carlos Pabón-Peña, Barbara Novak, Angelica Giuffre, Eric Lin, Scott Happe, Doug N Roberts, Emily M LeProust, et al. Enrichment of sequencing targets from the human genome by solution hybridization. *Genome Biol*, 10 (10):R116, 2009.

Robert E Thurman, Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T Maurano, Eric Haugen, Nathan C Sheffield, Andrew B Stergachis, Hao Wang, Benjamin Vernot, et al. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, 2012.

Desiree Tillo and Timothy R Hughes. G+C content dominates intrinsic nucleosome occupancy. *BMC bioinformatics*, 10(1):442, 2009.

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.

Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, 7(3):562–578, 2012.

Anton Valouev, Steven M Johnson, Scott D Boyd, Cheryl L Smith, Andrew Z Fire, and Arend Sidow. Determinants of nucleosome organization in primary human cells. *Nature*, 474(7352):516–520, 2011.

Harm van Bakel, Kyle Tsui, Marinella Gebbia, Sanie Mnaimneh, Timothy R Hughes, and Corey Nislow. A compendium of nucleosome and transcript profiles reveals determinants of chromatin architecture and transcription. *PLoS genetics*, 9(5):e1003479, 2013.

Mariëtte PC van de Corput, Ernie de Boer, Tobias A Knoch, Wiggert A van Cappellen, Adrian Quintanilla, Leanna Ferrand, and Frank G Grosveld. Super-resolution imaging reveals three-dimensional folding dynamics of the β-globin locus upon gene activation. *Journal of Cell Science*, 125(19):4630–4639, 2012.

Thijn van der Heijden, Joke JFA van Vugt, Colin Logie, and John van Noort. Sequence-based prediction of single nucleosome positioning and genome-wide nucleosome occupancy. *Proceedings of the National Academy of Sciences*, 109(38):E2514–E2522, 2012.

Axel Visel, Edward M Rubin, and Len A Pennacchio. Genomic views of distant-acting enhancers. *Nature*, 461(7261):199–205, 2009.

Zhibin Wang, Chongzhi Zang, Kairong Cui, Dustin E Schones, Artem Barski, Weiqun Peng, and Keji Zhao. Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell*, 138(5):1019–1031, 2009.

Wyeth W Wasserman and Albin Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4):276–287, 2004.

Gong-Hong Wei, Gwenael Badis, Michael F Berger, Teemu Kivioja, Kimmo Palin, Martin Enge, Martin Bonke, Arttu Jolma, Markku Varjosalo, Andrew R Gehrke, et al. Genome-wide analysis of ETS-family DNA-binding *in vitro* and *in vivo*. *The EMBO journal*, 29 (13):2147–2160, 2010.

Assaf Weiner, Hsiuyi V Chen, Chih Long Liu, Ayelet Rahat, Avital Klien, Luis Soares, Mohanram Gudipati, Jenna Pfeffner, Aviv Regev, Stephen Buratowski, et al. Systematic dissection of roles for chromatin regulators in a yeast stress response. *PLoS biology*, 10(7): e1001369, 2012.

Matthew T Weirauch, Atina Cote, Raquel Norel, Matti Annala, Yue Zhao, Todd R Riley, Julio Saez-Rodriguez, Thomas Cokelaer, Anastasia Vedenko, Shaheynoor Talukder, et al. Evaluation of methods for modeling transcription factor sequence specificity. *Nature biotechnology*, 2013.

Malgorzata Wiench, Sam John, Songjoon Baek, Thomas A Johnson, Myong-Hee Sung, Thelma Escobar, Catherine A Simmons, Kenneth H Pearce, Simon C Biddie, Pete J Sabo, et al. DNA methylation status predicts cell type-specific enhancer activity. *The EMBO journal*, 30(15):3028–3039, 2011.

Edgar Wingender, Torsten Schoeps, and Jürgen Dönitz. TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic acids research*, 41(D1):D165–D170, 2013.

Deborah R Winter, Lingyun Song, Sayan Mukherjee, Terrence S Furey, and Gregory E Crawford. DNase-seq predicts regions of rotational nucleosome stability across diverse human cell types. *Genome research*, 2013.

Chunlei Wu, Ian MacLeod, and Andrew I Su. BioGPS and MyGene. info: organizing online, gene-centric information. *Nucleic acids research*, 41(D1):D561–D565, 2013.

Zeba Wunderlich and Leonid A Mirny. Different gene regulation strategies revealed by analysis of binding motifs. *Trends in genetics*, 25(10):434–440, 2009.

Jian Yan, Martin Enge, Thomas Whitington, Kashyap Dave, Jianping Liu, Inderpreet Sur, Bernhard Schmierer, Arttu Jolma, Teemu Kivioja, Minna Taipale, et al. Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell*, 154(4):801–813, 2013.

J Omar Yáñez-Cuna, Huy Q Dinh, Evgeny Z Kvon, Daria Shlyueva, and Alexander Stark. Uncovering cis-regulatory sequence requirements for context-specific transcription factor binding. *Genome research*, 22(10):2018–2030, 2012.

Chuhu Yang, Eugene Bolotin, Tao Jiang, Frances M Sladek, and Ernest Martinez. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene*, 389(1):52–65, 2007.

Kuangyu Yen, Vinesh Vinayachandran, Kiran Batta, R Thomas Koerber, and B Franklin Pugh. Genome-wide nucleosome specificity and directionality of chromatin remodelers. *Cell*, 149(7):1461–1473, 2012.

Erbay Yigit, Quanwei Zhang, Liqun Xi, Dan Grilley, Jonathan Widom, Ji-Ping Wang, Anjana Rao, and Matthew E Pipkin. High-resolution nucleosome mapping of targeted regions using BAC-based enrichment. *Nucleic acids research*, 41(7):e87–e87, 2013.

Guo-Cheng Yuan and Jun S Liu. Genomic sequence is highly predictive of local nucleosome depletion. *PLoS computational biology*, 4(1):e13, 2008.

Federico Zambelli, Graziano Pesole, and Giulio Pavesi. Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. *Nucleic acids research*, 37(suppl 2):W247–W252, 2009.

Gabriel E Zentner and Steven Henikoff. Regulation of nucleosome dynamics by histone modifications. *Nature structural & molecular biology*, 20(3):259–266, 2013.

Jingli A Zhang, Ali Mortazavi, Brian A Williams, Barbara J Wold, and Ellen V Rothenberg. Dynamic transformations of genome-wide epigenetic marking and transcriptional control establish T cell identity. *Cell*, 149(2):467–482, 2012.

Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, 9(9):R137, 2008.

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 & molecular biology*, 16(8):847–852, 2009.

Zhenhai Zhang, Christian J Wippo, Megha Wal, Elissa Ward, Philipp Korber, and B Franklin Pugh. A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome. *Science*, 332(6032):977–980, 2011.

Michael J Ziller, Hongcang Gu, Fabian Müller, Julie Donaghey, Linus T-Y Tsai, Oliver Kohlbacher, Philip L De Jager, Evan D Rosen, David A Bennett, Bradley E Bernstein, et al. Charting a dynamic DNA methylation landscape of the human genome. *Nature*, 500 (7463):477–481, 2013.

Robert P Zinzen, Charles Girardot, Julien Gagneur, Martina Braun, and Eileen EM Furlong. Combinatorial binding predicts spatio-temporal *cis*-regulatory activity. *Nature*, 462(7269): 65–70, 2009.