Computational Footprinting Methods for Next-Generation Sequencing Experiments

Der Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen University vorgelegte Dissertation zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften

von

M.Sc. Eduardo Gade Gusmao

aus Recife, Brasilien

Abstract (English)

Transcriptional regulation orchestrates the proper temporal and spatial expression of genes. The identification of transcriptional regulatory elements, such as transcription factor binding sites (TFBSs), is crucial to understand regulatory networks driving cellular processes such as cell development and the onset of diseases.

The standard computational approach is to use sequence-based methods, which search over the genome's DNA for sequences representing the DNA binding affinity sequence of transcription factors (TFs). However, this approach is not able to predict active binding sites, i.e. binding sites that are being currently bound by TFs at a particular cell state. This happens as the sequence-based methods do not account for the fact that the chromatin dynamically changes its state between an open form (and accessible to TF binding) and closed (not accessible by TFs).

Advances in next-generation sequencing techniques have enabled the measurement of such open chromatin regions in a genome-wide manner with assays such as the chromatin immunoprecipitation followed by massive sequencing (ChIP-seq) and DNase I digestion followed by massive sequencing (DNase-seq). Current research has proven that such open chromatin genome-wide assays improve sequence-based detection of active TFBSs. The rationale is to restrict the sequence-based search of binding sites to genomic regions where these assays indicate the chromatin is open and accessible for TF binding, in a cell-specific manner.

We propose the first computational framework which integrates both DNase-seq and ChIP-seq data to perform predictions of active TFBSs. We have previously observed that there is a distinctive pattern at active TFBSs regarding both DNase-seq and ChIP-seq data. Our framework treats these data using signal normalization strategies and searches for these distinctive patterns, the so-called "footprints", by segmenting the genome using hidden Markov models (HMMs). Given that, our framework – termed HINT (HMM-based identification of TF footprints) – is categorized as a "computational footprinting method".

We evaluate our computational footprinting method by comparing the footprint predictions to experimentally verified active TFBSs. Our evaluation approach creates statistics which enables the comparison between our method and competing computational footprinting methods. Our comparative experiment is the most complete so far, with a total of 14 computational footprinting methods and 233 TFs evaluated.

Furthermore, we successfully applied our computational footprinting method HINT in two different biological studies to identify regulatory elements involved in specific biological conditions. HINT has proven to be a useful computational framework in biological studies involving regulatory genomics.

Abstrakt (Deutsch)

Die Transkriptionsregulation beschreibt die zeitliche und räumliche Expreßion der Gene. Mit Hilfe der Identifikation von transregulatorischen Elementen, wie beispielsweise Transkriptionsfaktorbindestellen, können regulatorische Netzwerke beßer verstanden werden. Regulatorische Netzwerke beschreiben zelluläre Prozeße wie zum Beispiel die Zellentwicklung und das Entstehen von Krankheiten.

Beim herkömmlichen rechnergestützten Ansatz zur Identifikation von Transkriptionsfaktorbindestellen wird auf Sequenzierungsmethoden zurückgegriffen, um die DNA des Genoms nach Sequenzen mit unterschiedlichen Bindungsneigungen zu Transkriptionsfaktoren (TF) zu durchsuchen. Mit diesem Ansatz ist es jedoch nicht möglich aktive Bindestellen vorherzusagen. Eine aktive Bindestelle ist beispielsweise dann gegeben, wenn an der DNA-Sequenz ein TF bindet. Dieser auf Sequenzierungstechniken beruhende Ansatz nimmt keinen Bezug darauf, daß der Zustand des Chromatins dynamisch zwischen offen (so daß ein TF binden kann) und geschloßen (so daß kein TF binden kann) wechseln kann.

Mit Sequenzierungsmethoden der nächsten Generation (next generation sequencing) kann offenes Chromatin genomweit identifiziert werden. Beispiele hierfür sind die Kombination von Chromatin ImmunoPrecipitation (ChIP-seq) oder DNase I Verarbeitung (DNase-seq) mit der Sequenzierungstechnik. Aktuelle Studien haben belegt, daß die Verwendung von ChIP-seq und DNase-seq zur Bestimmung von offenem Chromatin einen positiven Einfluß auf die Identifikation von aktiven TFBS haben. Dabei wird die Suche nach charakteristischen DNA-Sequenzen auf die Bereiche eingeschränkt, an denen das Chromatin offen ist und die TF somit in einer zellspezifischen Art binden können.

Wir führen zum ersten Mal in dieser Arbeit ein rechnergestütztes Rahmenwerk ein, das DNase-seq und ChIP-seq Daten kombiniert, um aktive TFBS vorherzusagen. Wir haben beobachtet, daß es bei aktiven TFBS ein ausgeprägtes Muster in DNase-seq und ChIP-seq Daten gibt. Unser Rahmenwerk führt zunächst eine Normalisierung des Signals aus und sucht dann in den Daten nach diesen Mustern, den sogenannten Fußabdrücken. Dabei wird das Genom mit einem Hidden Markov Modell segmentiert. Unsere Methode mit dem Namen HINT (HMM-basierte Identifikation von TF Fußabdrücken) ist als „rechnergestützte Fußabdruck Methode“ klaßifiziert.

In unserer Evaluierungßtudie haben wir die vorhergesagten Fußabdrücke von HINT mit bereits validierten Fußabdrücken verglichen. Dabei haben wir Statistiken erzeugt, um unsere Methode mit anderen zu vergleichen. Unsere Experimente sind mit insgesamt 14 verglichenen Methoden und 233 TF die umfangreichsten.

Zudem haben wir HINT erfolgreich bei zwei biologischen Studien angewandt, um regulatorische Elemente, die bei bestimmten biologischen Bedingungen vorkommen, zu identifizieren. HINT ist ein nützliches rechnergestütztes Rahmenwerk für biologische Studien in der regulatorischen Genomik.

Acknowledgements

First, I would like to thank my supervisor Dr. Ivan G. Costa for all the support provided during the course of my Ph.D. studies. Dr. Ivan G. Costa was essential for the success of the work presented in this document.

Furthermore, I thank Prof. Dr. Thomas Berlage, Prof. Dr. Martin Zenke and Prof. Dr. Stefan Decker for agreeing to be the referees of my thesis examination procedure, as well as Prof. Dr. Martin Grohe and Prof. Dr. Bernhard Rumpe for completing the examination committee.

Moreover, I thank Prof. Dr. Martin Zenke, Dr. Christoph Dieterich, Prof. Dr. Marcilio C. P. de Souto, Dr. Qiong Lin, Dr. Kristin Seré, Prof. Dr. Argyris Papantonis, Prof. Dr. Bernhard Lüscher, Prof. Dr. Wolfgang Wagner and Prof. Dr. Steffen Koschmieder for allowing to apply my skills in our fruitful collaborations.

Also, I would like to thank my colleagues, who helped me on some analyses, proofreading and productive discussions: Dr. Manuel Allhoff, Dr. Sonja Hänzelmann, Joseph Kuo, Ahmad Badar, Fabio Ticconi, Dr. Pablo A. Jaskowiak, Dr. Marcelo R. P. Ferreira, Dr. Juliana F. Pires, Dr. André C. A. do Nascimento and Barna Zajzon.

Finally, I thank my family and friends for all the support given during my stay in Aachen. In special, I would like to thank Christiani G. Gusmão, Herta G. Gusmão, Ingrid G. Gusmão, Stephanie G. Gusmão and Stefan Touet.

Selbstständigkeitserklärung

Ich versichere hiermit an Eides statt, daß ich die vorliegende Doktorarbeit selbstständig und ohne unzuläßige fremde Hilfe erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen. Ich habe die Grundsätze zur Sicherung guter wissenschaftlicher Praxis der RWTH Aachen zur Kenntnis genommen und eingehalten.

Aachen, September 3, 2016 

(Eduardo Gade Gusmao)

Publications

As required by §5(3) of the

> *Promotionsordnung für die Fakultät für Mathematik, Informatik und Naturwissenschaften der Rheinisch-Westfälischen Technischen Hochschule Aachen vom 27.09.2010 in der Fassung der zweiten Ordnung zur änderung der Promotionsordnung vom 30.06.2014 (veröffentlicht als Gesamtfassung)*,

a declaration of results that are published by the author as well as particular contributions to co-authored publications follows.

I am the main author of the following publications, which are co-authored with Dr. Ivan G. Costa and our collaborators Prof. Dr. Christoph Dieterich, Prof. Dr. Martin Zenke and Dr. Manuel Allhoff. All the results, with exception of the filter-based computational footprinting results (co-authored by Ahmad Badar), presented in this thesis were published in the following publications (in chronological order):

> **Gusmao EG**, Dieterich C, Zenke M, Costa IG (2014). Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. *Bioinformatics*, 30(22):3143–3151.

> **Gusmao EG**, Allhoff M, Zenke M, Costa IG (2016). Analysis of computational footprinting methods for DNase sequencing experiments. *Nature Methods*, 13(4):303–309.

Furthermore, some parts of the following publications, co-authored by me, were used in Section 5.6 as case studies in this thesis:

> Lin Q, Chauvistré H, Costa IG, **Gusmao EG**, Mitzka S, Häenzelmann S, Baying B, Klisch T, Moriggl R, Hennuy B, Smeets H, Hoffmann K, Benes V, Seré K, Zenke M (2015). Epigenetic and transcriptional architecture of dendritic cell development. *Nucleic Acids Research*, 43(20):9680–9693.

> Kolovos P, Georgomanolis T, Nikolic M, Koeferle A, Larkin JD, Feuerborn A, van Ijcken WF, **Gusmao EG**, Costa IG, Cook PR, Grosveld FG, Papantonis A (2016). Enhancer hijacking reveals a multimodal role of NF-κB during the immediate-early inflammatory response. *(in review)*.

Finally, the following publications did not directly contribute to this thesis, but shaped a general understanding of next-generation sequencing analysis, transcription factor motif analysis and machine learning. They were published during my Ph.D. studies in RWTH University Aachen.

Gusmao EG, de Souto MCP (2014). Issues on Sampling Negative Examples for Predicting Prokaryotic Promoters. *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2014)*, 494–501.

Ullius A, Lüscher-Firzlaff J, Costa IG, Walsemann G, Forst AH, **Gusmao EG**, Kapelle K, Kleine H, Kremmer K, Vervoorts J, Lüscher B (2014). The interaction of MYC with the trithorax protein ASH2L promotes gene transcription by regulating H3K27 modification. *Nucleic Acids Research*, 42(11):6901–6920.

Hänzelmann S, Beier F, **Gusmao EG**, Koch CM, Hummel S, Charapitsa I, Joussen S, Benes V, Brümmendorf TH, Reid G, Costa IG, Wagner W (2015). Replicative senescence is associated with nuclear reorganization and with DNA methylation at specific transcription factor binding sites. *Clinical Epigenetics*, 7(1):19.

Schemionek M, Herrmann O, Reher MM, Chatain N, Schubert C, Costa IG, Hänzelmann S, **Gusmao EG**, Kintsler S, Braunschweig T, Hamilton A, Helgason GV, Copland M, Schwab A, Müller-Tidow C, Li S, Holyoake TL, Brümmendorf TH, Koschmieder S (2015). MTSS1 is a critical epigenetically regulated tumor suppressor in CML. *Leukemia*, 30(4):823–832.

I implemented HINT and built the benchmarking dataset. All figures/tables in this thesis are authored by me and exceptions were explicitly stated in their respective captions. I performed all experiments in this study with the aid of Dr. Ivan G. Costa, Prof. Dr. Martin Zenke, Dr. Manuel Allhoff, Ahmad Badar, Dr. Sonja Hänzelmann and Joseph Kuo. All chapters of this thesis have been written by me. Ivan G. Costa provided support in all stages of this research including thesis manuscript writing. Mostly out of habit and to honor the fact that research is rarely an entirely solitary process I will, in this thesis, rely on the use of the first-person plural pronoun “we” in the text, as a nosism.

Glossary

AUC	Area under the ROC curve.
AUPR	Area under the PR curve.
bp	Base pair.
BWA	Burrows-Wheeler aligner software.
cDC	Classical dendritic cell.
CDP	Common dendritic cell progenitor.
CENTRIMO	Local motif enrichment analysis.
ChIP-seq	Chromatin immunoprecipitation followed by massive sequencing.
DC	Dendritic cell.
DH	Double-hit DNase-seq protocol.
DHS	DNase hypersensitivity site.
DNA	Deoxyribonucleic acid.
DNase-seq	DNase I digestion followed by massive sequencing.
DREME	Discriminative regular expression motif elicitation.
ENCODE	Encyclopedia of DNA Elements.
FC	Gene expression fold change.
FIMO	Find individual motif occurrences software.
FLR	Footprint likelihood ratio.
FP-Exp	Correlation of KS test statistic and gene expression FC.
FPR	False positive rate.
FS	Footprint score.
GEO	Gene expression omnibus.
HINT	HMM-based identification of TF footprints.
HMM	Hidden Markov model.
IQR	Interquartile region.
IUPAC	International union of pure and applied chemistry.
KS	Kolmogorov-Smirnov.
MACS	Model-based analysis for ChIP-Seq software.
MPBS	Motif-predicted binding site.
MPP	Multipotent progenitor cell.
NGS	Next-generation sequencing.
NK	Naked (deproteinized) DNA DNase-seq protocol.
OBS	Observed *vs* bias signal.
pDC	Plasmacytoid dendritic cell.
PFM	Position frequency matrix.
PIQ	Protein interaction quantification.
PR	Precision-recall.
PWM	Position weight matrix.
RGT	Regulatory genomics toolbox.
RNA	Ribonucleic acid.
ROC	Receiver operating characteristics.
SH	Single-hit DNase-seq protocol.
SRA	Sequence read archive.
TC	Tag count.
TF	Transcription factor.
TFBS	Transcription factor binding site.
TSS	Transcription start site.

List of Figures

List of Tables

Contents

CHAPTER 1

Introduction

1.1 Motivation

Gene Regulation and Transcription Factor Binding Sites

Every living organism is composed of multiple different cells. These cells contain genetic material encoded in the form of deoxyribonucleic acid (DNA) molecules, also known as genome. The genome can be represented as a categorical vector $\mathbf{g} = \langle g_1, \cdots, g_n \rangle$, where $g_i \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ represents the nucleotide at genomic position i. Certain regions within the genome encode the so-called genes. Genes can be read by specialized proteins to produce other proteins. This protein-producing cycle is the key mechanism for maintenance of life.

A couple of years ago, it was believed that, in possession of the complete genomic sequence $\mathbf{g}$ for a given organism, it would be possible to exactly determine its phenotype and disease susceptibility. However, after the analysis of the first genomes, it was clear that the simple determination of an organism's DNA nucleotide sequence is not enough to explain the great diversity of biological processes. Such processes are governed by a complex chain of events called "gene regulation". Gene regulation includes a wide range of mechanisms that happen inside a cell in which genes are turned "on" (i.e. they are expressed) and "off" (i.e. they are not expressed) dynamically. Depending on which genes are "on" or "off", the cell specializes in different functionalities (Alberts et al., 2007).

In the so-called post-genomic era, attention is turning to the understanding of how protein-coding genes (about $25,000$ in humans) and their products are regulated (Maston et al., 2006). These regulatory mechanisms drive the correct execution of biological processes and require a set of carefully orchestrated steps that depend on the correct spatial and temporal expression of genes (Maston et al., 2006). On the other hand, the deregulation of gene expression, i.e. errors regarding the regulatory steps, is often linked to diseases (ENCODE Project Consortium, 2012).

To understand the molecular mechanisms that dictate the cell's expression patterns, it is important to identify the regulatory elements involved in these activities. One of the most important regulatory players are transcription factors (TFs) – proteins that bind on the DNA enhancing or repressing the expression of genes. These proteins bind to particular genomic regions called transcription factor binding sites (TFBSs) (Maston et al., 2006). TFBSs may be active if they are currently being bound by a TF or inactive, if they are not currently being bound by a TF.

Importance of the Identification of All Active TFBSs of a Cell

The identification of all active TFBSs of a cell is a very important task, since they are the key players on regulatory mechanisms. By identifying active TFBSs we can develop regulatory networks, which encode the interplay between different genes to control specific cell functions. Such a task leads to the understanding of cellular mechanisms and the particular deregulatory steps which leads to disease.

There are a great number of successful experimental studies that benefited from the proper identification of active TFBSs. For instance, studies were able to: (1) unravel cellular mechanisms (Lin et al., 2015; Tsankov et al., 2015); (2) unravel disease mechanisms (Schaub et al., 2012; Vernot et al., 2012;

Charos et al., 2012); (3) understand the function of different regions in the genome (Yip et al., 2012; Whitfield et al., 2012; Natarajan et al., 2012) and (4) understand other cellular regulatory elements such as long noncoding ribonucleic acids (lncRNAs) (Tilgner et al., 2012; Bánfai et al., 2012).

In summary, the identification of active transcription factor binding sites is important because of its broad impact on many other cellular processes. Given the importance of the proper identification of cell-specific active TFBSs, our research focuses on performing such a task by applying computational methods to biological experimental data.

Computational Detection of Active TFBSs Must Consider the Chromatin Dynamics

Historically, the first computational approach to identify TFBSs was based solely on the DNA sequence (Stormo, 2000). Each TF has a particular DNA sequence affinity, i.e. they tend to bind to specific DNA sequences. The computational sequence-based methods search the genome $\mathbf{g}$ for DNA substrings that correspond to the affinity sequence of target TFs. However, although computational sequence-based methods are able to detect TFBSs, they are not able to tell whether these sites are active or inactive (Pique-Regi et al., 2011). This happens because such computational approach does not consider the fact that only a few regions in the genome are accessible for TFs to bind. These regions are called "open chromatin regions". The number of open chromatin regions and their location vary between different cell types and ultimately dictates which genes are accessible and being expressed (ENCODE Project Consortium, 2012).

Recent advances in biological techniques (Shendure and Ji, 2008) have enabled the creation of experimental methods to identify these open chromatin regions (ENCODE Project Consortium, 2012). We will explore two of these so-called "open chromatin next-generation sequencing (NGS) techniques": the chromatin immunoprecipitation followed by NGS – termed ChIP-seq (Johnson et al., 2007); and the DNase I cleavage followed by NGS – termed DNase-seq (Crawford et al., 2004; Sabo et al., 2004b). These techniques generate time series-like signals which span the entire genome and indicate open chromatin regions. These signals can be viewed as a numeric vector $\mathbf{x} = \langle x_1, \cdots, x_n \rangle$ where high data points $x_i \in \mathbb{N}^0$ indicate open chromatin regions. Moreover, certain patterns in the signals generated by DNase-seq and ChIP-seq are indicative of active TFBSs. Therefore, we can apply computational methods to process the DNase-seq and ChIP-seq signals and to identify these patterns. By doing so, we can detect active TFBSs considering the open chromatin information.

Computational Detection of Active TFBSs Using DNase-seq and ChIP-seq

The DNA is found wrapped in proteins called histones. There are a number of post-translational modifications on these histones which are indicative of open chromatin regions, such as the so-called H3K4me1 and H3K4me3. By performing a histone modification ChIP-seq experiment we are able to identify cell-specific open chromatin regions. Furthermore, the DNase-seq data also provides a robust map of open chromatin regions with a very high spatial resolution. By combining these two experimental data, we observe very characteristic patterns indicating the active binding of TFs in the genome (Figure 1.1). This pattern is commonly referred to as TF "footprints". A TF footprint is defined as a region likely to be represent an active TFBS (Boyle et al., 2011; Gusmao et al., 2012).

The experiments presented in this thesis focus on the computational treatment of DNase-seq and histone modification ChIP-seq data to perform computational predictions of active transcription binding sites. Such prediction is performed by searching the distinctive patterns, i.e. footprints, that the DNase-seq and histone modification ChIP-seq signals exhibit around active TFBSs. We use the traditional term "computational footprinting methods" for computational methods that searches for footprints using open chromatin data, such as DNase-seq and histone modification ChIP-seq. The computational footprinting framework presented in this thesis can be used in multiple different biological experiments to understand the regulation of genes.

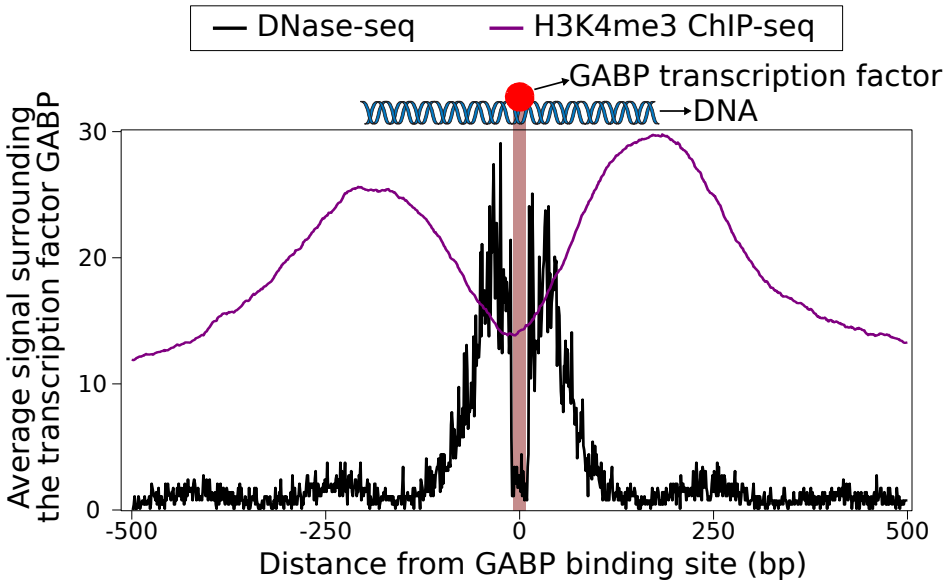

Figure 1.1: Distinctive pattern (footprint) of DNase-seq and ChIP-seq on active TFBSs. Average DNase-seq and histone modification H3K4me3 ChIP-seq signals surrounding the known (biologically verified) TF GABP active binding sites. Active TFBSs happen at depletions between two peaks of the DNase-seq signal (marked in red). Furthermore, these DNase-seq peaks, which determines an open chromatin region, happen at the depletion between two peaks of active histone modification marks. This distinctive pattern of signal depletion between two peaks is called a TF footprint. *Source: Gusmao et al. (2012)* (modified to fit thesis format and/or clarify key points).

1.2 Thesis Overview

In this thesis, we: (1) present a novel computational framework that uses DNase-seq and histone modification ChIP-seq data to detect active TFBSs, (2) evaluate the predictions made by our method using experimentally verified active TFBSs and (3) use our predictions in real biological scenarios to make inferences about the regulatory circuitry of particular cells. Figure 1.2 presents an overview of this thesis. In the following paragraphs we describe the Figure 1.2 in more detail.

Computational Footprinting Framework

For a particular cell type A we obtain DNase-seq and histone modification ChIP-seq data available in repositories such as the ENCODE Project Consortium (2012) (Figure 1.2a). We process these data using computational methods (Figure 1.2b) to generate a normalized DNase-seq signal $\mathbf{x}_A$ and normalized histone modification ChIP-seq signal $\mathbf{y}_A$ (Figure 1.2c). Then, we apply a computational footprinting method Θ on $\mathbf{x}_A$ and $\mathbf{y}_A$ (Figure 1.2d) generating a set of genomic regions (intervals) $S_A = \{s_{A1}, \cdots, s_{Am}\}$, where each genomic interval $s_{Ai} = [u, v]$ represent a predicted footprint, which is likely to be associated to an active TFBS (Figure 1.2e).

Evaluation of Predicted Footprints

The predicted footprints S_A are compared to experimentally verified active TFBSs (Figure 1.2e) to create statistics which evaluate how close our predictions are to the true active TFBSs (Figure 1.2f-g). These evaluation statistics (Figure 1.2g) are also used to compare our computational footprinting framework to competing methodologies.

Application to Real Biological Scenarios

Furthermore, the predicted footprints S_A are used in combination with downstream computational methods (Figure 1.2i) to generate real-scenario biological knowledge about the regulatory circuitry of the cell type A (Figure 1.2j).

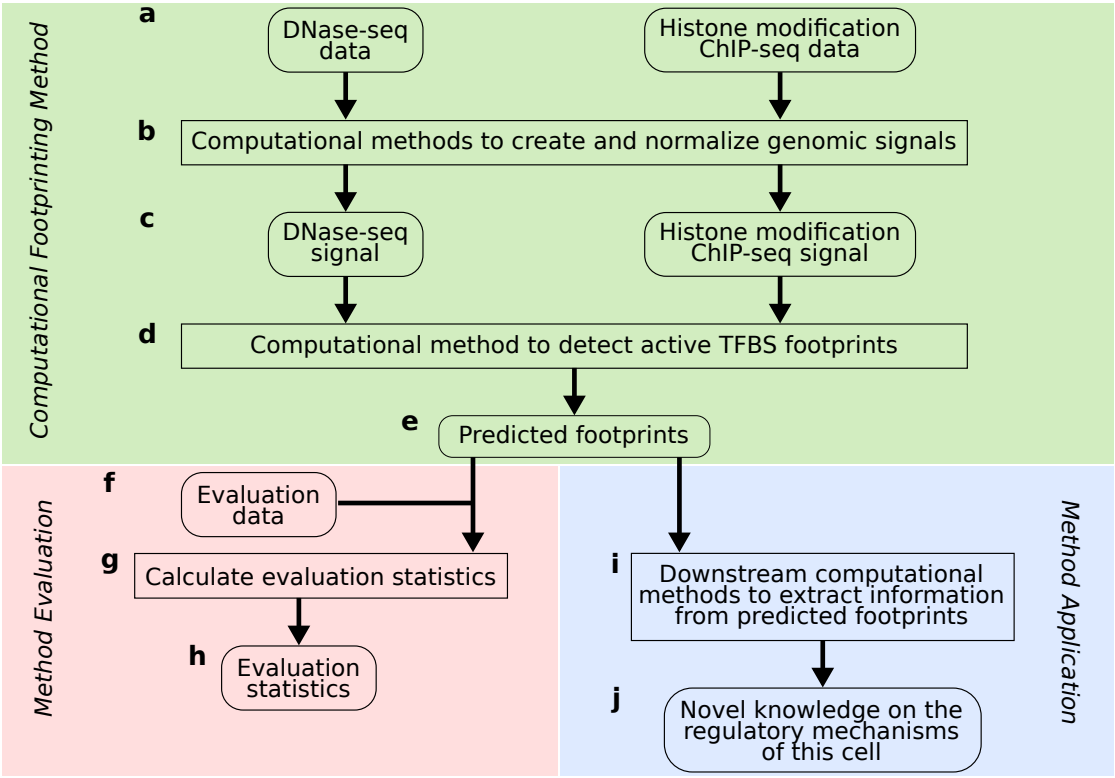

Figure 1.2: Thesis overview. This figure depicts the proposed thesis' workflow. Boxes with round-shaped edges represent data and square-shaped edges represent computational methods.

1.3 Contributions

The main contribution of this work is the development of a novel computational framework to treat data generated with the DNase-seq and ChIP-seq technologies and detect footprints (i.e. putative active TFBSs) based on these data. Our contributions are summarized as follows.

- **Novel computational footprinting method:** We devised a novel computational footprinting method based on hidden Markov models (HMMs). Our novel methodology is the first to successfully combine DNase-seq and histone modification ChIP-seq to predict footprints, i.e. putative active TFBSs. Such method was shown to provide robust active TFBS predictions on the basis of an extensive evaluation process.
- **Novel signal treatment strategy:** Novel DNase-seq and histone modification ChIP-seq signal treatment approaches were developed and formalized. Such treatment framework has shown to be robust and applicable to a wide range of different datasets.
- **DNase-seq experimental bias correction:** We created an approach to correct for known artifacts on DNase-seq data. Our experiments have shown the efficiency of such correction on bias mitigation.

- **Novel evaluation approach of computational footprinting methods:** Until now, computational footprinting methods have been evaluated using the "TF ChIP-seq approach". However, biases were pointed in such evaluation scheme (Yardımcı et al., 2014). Therefore, we develop a novel computational footprinting method evaluation approach based on gene expression.
- **Comprehensive computational footprinting method comparison:** We performed a comprehensive comparison including: (1) our novel HMM-based approach; (2) nine state-of-the-art computational footprinting methods and (3) four baseline approaches. Our comparative experiment is the most complete so far, with a total of 14 computational footprinting methods and 233 TFs evaluated.
- **Analysis of relevant features on computational footprinting:** A number of empirical analyses were performed. These analyses evaluated relevant features for the computational prediction of active TFBSs such as: method's parameter selection, experimental bias correction, optimal footprint scoring strategy and TF binding residence time.
- **Case studies:** We successfully applied our computational footprinting method in two different studies to identify regulatory elements involved in specific biological conditions.

1.4 Document Structure

In Chapter 2 we introduce all the concepts needed for the understanding of our work. We define the current challenges on computational identification of active TFBSs and provide a comprehensive literature review on computational footprinting methods.

In Chapter 3 we formalize our approach to address the detection of active TFBSs. We describe the treatment of the input DNase-seq and ChIP-seq data and the novel approach to detect active TFBSs based on HMMs. Furthermore, in Chapter 4 we describe the full experiment design of this project. We present the data used in our work, the execution of our computational footprinting approach and method evaluation strategies.

In Chapter 5 we present the results of our experiments, which encompasses: the analyses on relevant computational footprinting features, a comprehensive comparison of computational footprinting methods and case studies in which our methodology was successfully applied to real biological scenarios. In Chapter 6 we discuss all results presented in this thesis, highlighting all the key findings. Furthermore, we discuss future research opportunities. Further supplementary information and results can be found in the Appendix A.

CHAPTER 2

Background

In this chapter we provide background information required for the understanding of this thesis. First, we introduce the necessary biological concepts (Section 2.1). Next, we present the biological experimental techniques which are the main sources of data used in our analyses (Section 2.2). Then, we introduce the problem which we are going to address in this thesis – the computational prediction of active transcription factor binding sites (TFBSs; Section 2.3). Subsequently, we discuss the state-of-the-art computational solutions to address this problem – the computational footprinting methods (Section 2.4). Finally, we close this chapter with concluding remarks on the definitions made in this chapter and a brief overview on our strategy to solve the problem of predicting active TFBSs (Section 2.5).

2.1 Gene Regulation

In this thesis we focus on the biological field of gene regulation. Such research area focuses on the understanding of the cellular mechanisms behind the temporal and spatial expression of different genes on different cellular conditions. We start this section by describing the basic concepts of molecular biology (Section 2.1.1). Then, we describe the main biological processes regarding gene regulation (Section 2.1.2). Finally, we discuss the role of chromatin dynamics on such regulatory processes (Section 2.1.3). The concepts presented in this section are based on Alberts et al. (2007) and Lodish et al. (2007).

2.1.1. Basic Concepts of Molecular Biology

In this thesis we focus on two important macromolecules which are found inside cells: proteins (composed of amino acids) and nucleic acids (composed of nucleotides). Proteins assume many roles: catalysis of chemical reactions (enzymes), metabolite processing, cell signaling, regulation of the production of more proteins, structural function and others. Given such great variety of roles, one might regard these macromolecules as fundamental for the maintenance of living organisms. There are two types of nucleic acids: the deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). The main function of the DNA is to store the hereditary information of the organism. It is based on such information that new proteins are generated. The RNA have many important functions; however we will not focus on this molecule in this work.

The DNA molecule is formed by a double helix of paired nucleotide chains, each of which composed of the nucleotide types: adenine (A), cytosine (C), guanine (G) and thymine (T). Each nucleotide is composed of a sugar (deoxyribose), a phosphate group and a nitrogenous base (which determines the nucleotide type). Within each DNA strand of the double helix, nucleotides are connected through phosphodiester bonds (strong covalent bonds). Between each DNA strand, nucleotides are paired and connected through hydrogen bonds (weaker than covalent bonds). Cytosines always pair with guanines and adenines always pair with thymines. Because nucleotides are paired between the double helix structure, it is common to refer to nucleotides as base pairs (bp). Figure 2.1 depicts a graphical representation of the DNA molecule.

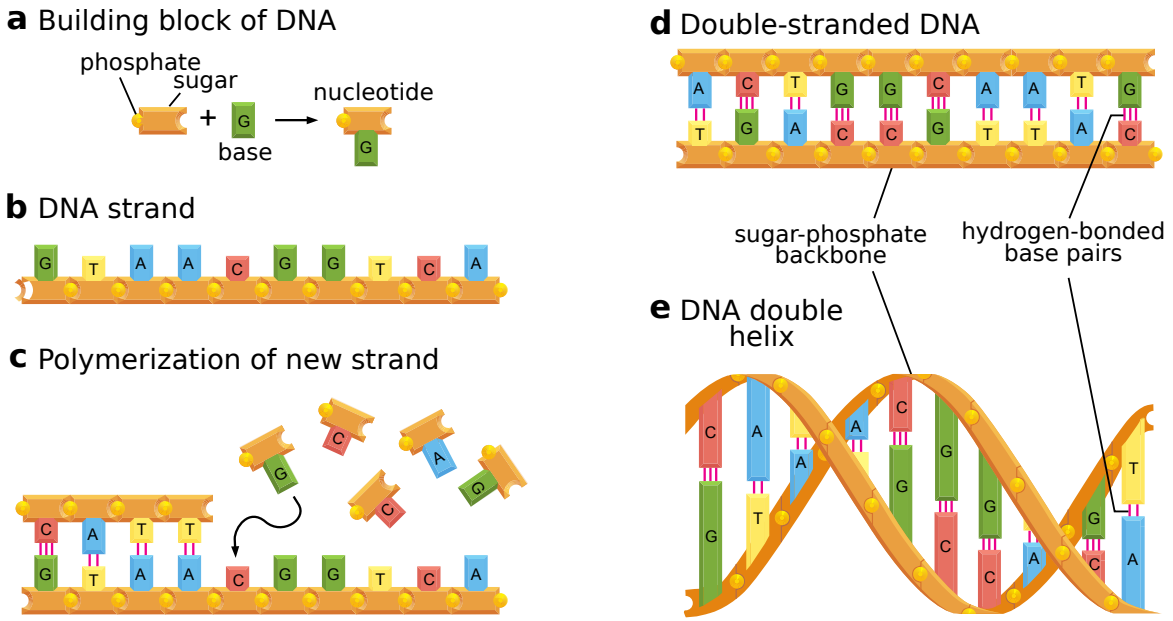

Figure 2.1: DNA structure. (**a**) Representation of the DNA's building block – the nucleotide. (**b**) Multiple nucleotides from all possible types (A, C, G and T) form a single strand of DNA, which in humans are as long as ~250,000,000 nucleotides. (**c**) The DNA generally occurs as a double strand. The biological process of polymerization allows the addition of nucleotides to a single strand, forming the DNA double strand. Cytosines (C) always pair with guanines (G) connected through three hydrogen bonds (pink lines); and adenines (A) always pair with thymines (T) through two hydrogen bonds. (**d**) Linear scheme of the double DNA strands. (**e**) Double helix structure of the double-stranded DNA molecule. This is the general structure in which the DNA occurs in nature. *Source: Alberts et al. (2007)* (modified to fit thesis format and/or clarify key points).

Proteins are chemical compounds with high molecular weight formed by a variable-length chain of amino acids. The amino acids that forms the proteins are composed of a central carbon atom which binds to a hydrogen, a carboxyl group, an amine group and a side chain. The side chain may be of various types and dictates the type of the amino acid. There are 20 amino acid types commonly found at proteins. The specific order of each amino acid type in a protein determines its three-dimensional structure. It is well-known that the protein's function is directly related to its structure. The simple substitution of one amino acid in the proteic chain is sufficient to modify the protein three-dimensional conformation leading to a reduced functional capability or total dysfunction. Figure 2.2 shows the different levels of protein structural conformation.

The process in which proteins are created based on the information encoded in the cell's DNA is called the "central dogma of molecular biology". Here, the key parts of this process, which aid in the understanding of this work, are presented. These key parts are: (1) the initiation, (2) the transcription and (3) the translation.

During the initiation phase, a number of proteins called transcription factors (TFs) bind in the DNA and recruit another protein called RNA polymerase (Figure 2.3a). The DNA region in which these TFs and RNA polymerase bind to start transcription is called promoter. Then, in the transcription phase, the RNA polymerase scans the DNA and creates an RNA molecule, based on the information encoded in the DNA (Figure 2.3b). The part of the DNA which is transcribed by the RNA polymerase is called gene. Finally, in the translation phase, the newly-generated RNA migrates outside the cell's nucleus and a protein called ribosome scans the RNA and creates a new protein molecule based on the information encoded in the RNA (Figure 2.3c). The rate in which the transcription occurs for a particular gene is called the gene's expression.

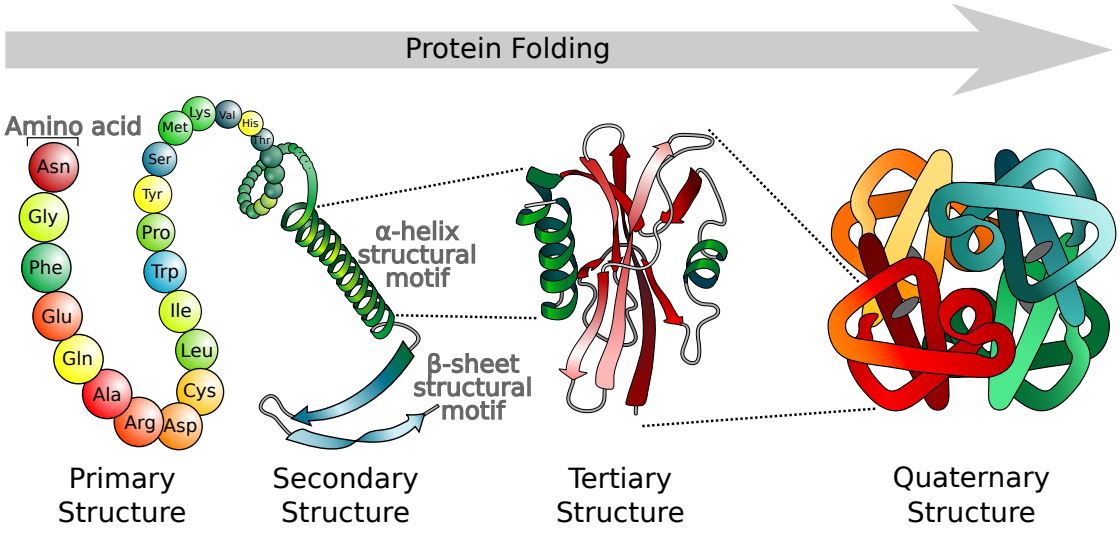

Figure 2.2: Protein structure. Proteins are formed by building blocks termed amino acids. The chain of amino acids that forms a specific protein is the protein's primary structure. Protein secondary structures, such as α-helixes or β-sheets, are formed through the natural folding of the amino acid chain given their physicochemical properties. The tertiary structure is composed of a number of secondary structures forming a stable protein unit. Finally, the quaternary structure represents the aggregation of multiple protein units to form fully functional protein. *Source: Mariana R. Villarreal* (modified to fit thesis format and/or clarify key points).

It is important to mention that, although only one of the DNA strands is read during the transcription process, both strands contain information necessary to produce RNA. Another important issue is the orientation of these two DNA strands. Each strand has two extremities: one corresponding to a hydroxyl group attached to the $3'$ carbon atom of the sugar; and the other corresponding to the phosphate group attached to the $5'$ carbon atom of the sugar. For this reason, processes involving the sliding of proteins in DNA have two orientations: forward ($5' \rightarrow 3'$) and reverse ($3' \rightarrow 5'$). The different strands in the DNA helix are attached to each other in opposite (anti-parallel) orientations. The transcription always occurs in the forward orientation.

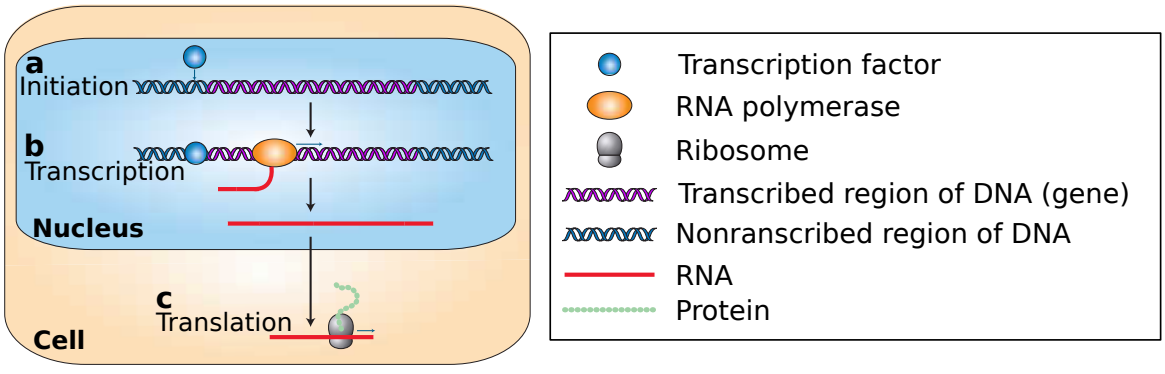

Figure 2.3: Central dogma of molecular biology. Depiction of the main steps of the central dogma of molecular biology necessary to create a protein molecule from the information encoded within the DNA molecule. Here we show the steps: (**a**) initiation, (**b**) transcription and (**c**) translation. *Source: Lodish et al. (2007)* (modified to fit thesis format and/or clarify key points).

2.1.2. Gene Regulation with Transcription Factors

The transcription initiation was previously described as the step in which the RNA polymerase binds to the promoter region in order to start the process of transcription. Nevertheless, there are many factors that contribute to the expression of particular genes in particular types/stages/conditions of a cell. We call "gene regulation" the wide range of mechanisms that are used by cells to increase or decrease the production of specific gene products. Gene regulation may happen in different stages of the central dogma. However, most part of the regulatory events happens at the transcription initiation level. A major role of the regulation at this level is played by proteins termed TFs, which use their physicochemical properties to direct the intensity level in which gene products are created. The TFs bind to DNA regions called TFBSs which are close (promoter region; approximately $< 1{,}000$ bp from the transcription start site) or far (distal regulatory regions; generally up to $1{,}000{,}000$ bp) from the gene. Different TFs may bind to different TFBSs to increase or decrease the expression of genes. Figure 2.4 shows a graphical representation of a basic regulatory landscape of a gene. The number of regulatory elements vary between genes; however the construct of TF and their DNA binding sites are generally present.

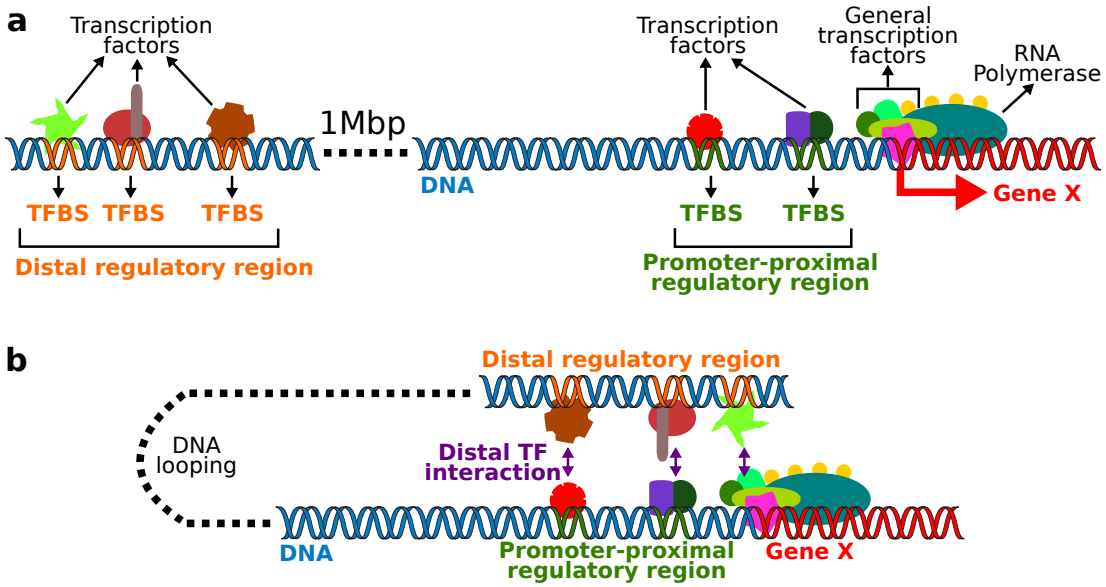

Figure 2.4: Basic regulatory landscape of a gene. (**a**) Schematic representation of a typical gene regulatory region with proximal and distal regulatory regions, composed of transcription factor binding sites (TFBSs), which are regions in the DNA being bound by proteins called transcription factors (TFs). The promoter typically spans less than 1 Kbp and is composed of: (1) a core promoter – where the transcriptional machinery is being bound and (2) promoter-proximal regulatory region – where TFs bind to increase/decrease gene expression. The distal regulatory regions are located up to 1 Mbp from the promoter. Among others, they are categorized as: (1) enhancers – where TFs bind to increase gene expression and (2) silencers – usually decreasing or completely silencing expression. (**b**) These distal elements may contact the core promoter or proximal promoter through a mechanism that involves looping out the intervening DNA. *Based on Lodish et al. (2007).*

The TFs contain a specific part (formally called "domains") within their structure, termed active site, which enables them to bind to the DNA. There is a relatively short number of smaller structural variants (which compose the final protein structure) in comparison to the number of different protein types. Some of these structural variants, including the ones containing active sites, are repeated between different protein. These DNA-binding protein domains usually have affinities towards specific

DNA sequences. These affinity sequences are termed "DNA motifs". Figure 2.5 shows four DNA-binding protein domains and examples of proteins that contain such domains and their respective DNA binding affinity motifs.

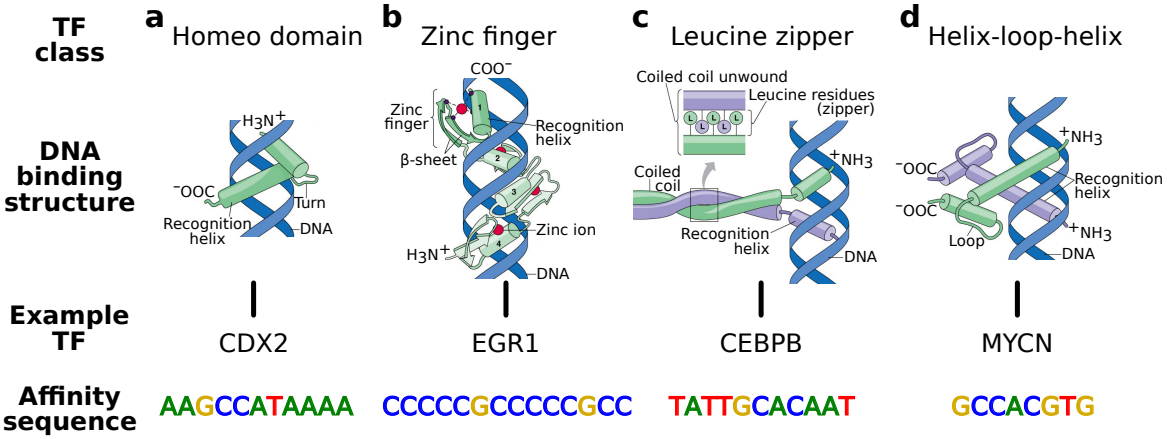

Figure 2.5: Different protein-DNA binding types. We show graphical representations of different protein-DNA binding types (top) and examples of proteins with such binding domain type and their DNA sequence binding affinity information (bottom). We show four DNA-binding TF classes: (**a**) Homeo domain (also known as helix-turn-helix), (**b**) zinc finger, (**c**) leucine zipper and (**d**) helix-loop-helix. This motivates the idea that, although there are many DNA-binding proteins, they usually have a few DNA-binding domains. It is important to mention that, although we show only one affinity DNA sequence for each example, proteins have flexibility within some parts of their motif to bind different nucleotides. *Source: Alberts et al. (2007)* (modified to fit thesis format and/or clarify key points).

2.1.3. Chromatin

The DNA is not isolated in the cell nucleus. Instead, it is found wrapped in proteic complexes, which are associated to the compaction of the DNA. The most important protein complex is formed by four pairs of histones named H2A, H2B, H3 and H4. The unit composed of the DNA wrapped in approximately 1.65 turns (~147 bp) around the histone complex is called nucleosome. From this lower level structure (nucleosome) the DNA structure is compacted in many levels. Such DNA+protein structure is termed chromatin. This compaction organization is depicted in Figure 2.6. Briefly, the chromatin can be found in a very condensed structure which does not allow transcription initiation (termed heterochromatin, or simply "closed chromatin"); or in a decondensed form, allowing transcription initiation and gene expression (termed euchromatin, or simply "open chromatin").

Different parts of the genome are open or closed at different times, allowing a specific set of genes to be expressed under different cell conditions. This is one of the main mechanisms behind the fact that we observe such a high number of different cells, each of which expressing a different set of genes, given that they all share the same underlying genomic information encoded in the DNA. Figure 2.7 shows a graphical example of two cells at different stages of commitment. Although the genomic region depicted is the same for these two cells, one present a closed chromatin structure, while the other present an open chromatin structure. The closed chromatin observed for the long-term hematopoietic stem cell (Figure 2.7a) does not allow the gene ATF3 to be transcribed, while the open chromatin structure present in the monocyte cell (Figure 2.7b) does allow the expression of ATF3 gene, since the TFs and transcription machinery are able to access that region and start the transcription process.

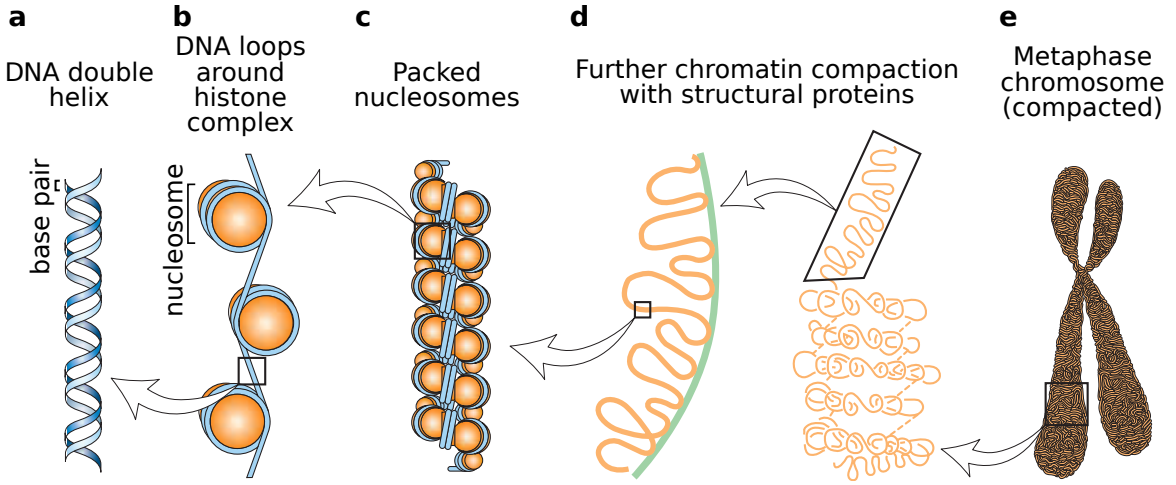

Figure 2.6: Chromatin conformation. The nuclear DNA have many compaction levels. (**a**) The lowest chromatin level corresponds to the DNA double helix. (**b**) The DNA double helix loops around the histone complexes forming the nucleosomes. (**c**) With additional structural proteins, the nucleosomes are packed in a structure termed 30-nm fiber. (**d**) The 30-nm fiber is further compacted in many compaction levels with the aid of further structural proteins. (**e**) The higher degree of chromatin compaction is represented in the cell's metaphase. *Source: Lodish et al. (2007)* (modified to fit thesis format and/or clarify key points).

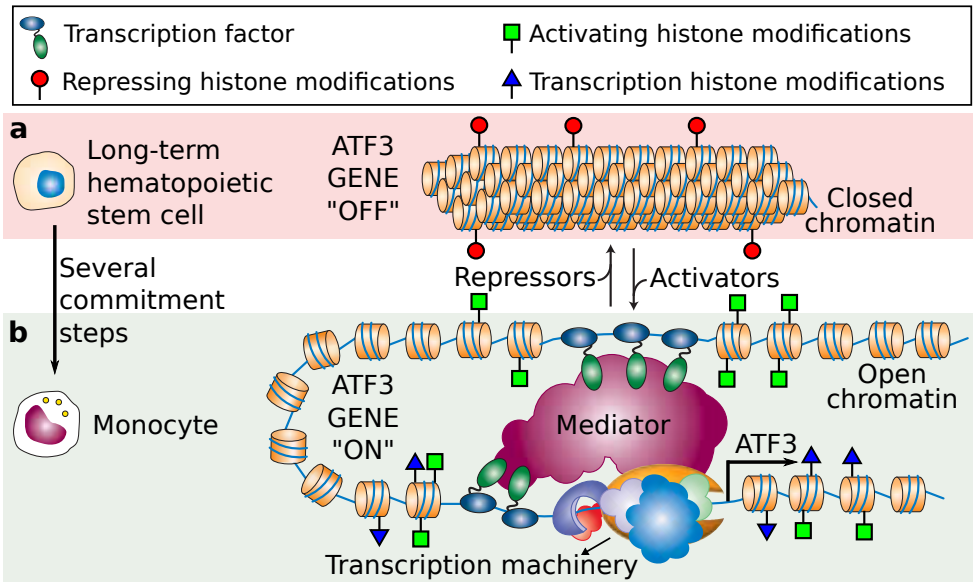

Figure 2.7: Open *vs* closed chromatin. Transription factor binding sites are cell-condition specific. This means that the same genomic locus might be accessible in a particular cell condition and not accessible in other condition. Such chromatin dynamics, which are modulated by genetic and epigenetic factors such as histone modifications, are responsible for the specialization of cells in performing the particular tasks required by the tissues they are located. This figure shows a graphical example of a locus (ATF3 gene) which is (**a**) closed in long-term hematopoietic stem cells (**b**) open in monocytes. Monocytes are a product of multiple specialization steps in hematopoietic stem cells. In this figure we show histone modifications which marks open/closed chromatin and proximal/distal regulatory regions. *Source: Lodish et al. (2007)* (modified to fit thesis format and/or clarify key points).

One of the main mechanisms associated to the chromatin switch between closed and open states is the post-translational histone modification. The histone proteins' N-terminal usually protrudes from the nucleosome and is termed histone tail. Histone tails can undergo post-translational chemical modifications at specific amino acids. Such modifications include the methylation (addition of a methyl group; labeled "me") and the acetylation (addition of an acetyl group; labeled "ac"). These modifications have a specific nomenclature dictated by: histone type, amino acid type, amino acid position within the histone tail and modification type. For instance, "H3K4me1" refers to the monomethylation (me1) of the lysine (K) in the fourth position of the tail of histone H3. Some histone modifications, such as H3K4me3, make the DNA more accessible to the binding of TFs; while others, such as H3K27me3, make the DNA less accessible to the binding of TFs. Figure 2.8 displays the different effects, on the chromatin structure, of modifications in lysines in the tail of histone H3.

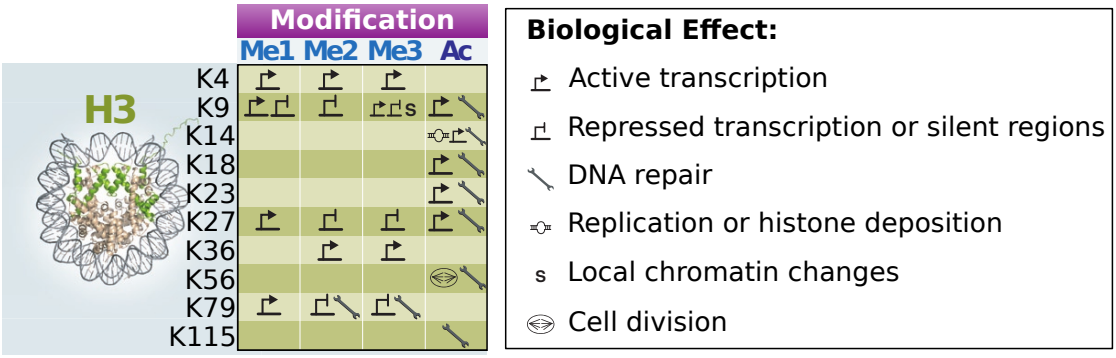

Figure 2.8: Main histone modifications on lysines of histone H3. The tail of the histone H3 undergo post-translation modifications which are associated to chromatin remodeling. Different modifications at different locations of the H3's tail have contrasting effects such as transcription initiation or repression. *Source: Lall (2007)* (modified to fit thesis format and/or clarify key points).

2.2 Next-Generation Sequencing Methods

Recently, novel DNA sequencing platforms have enabled the sequencing of a very large number of DNA fragments (up to a few billions) on one single assay with a significant decrease in cost and complexity (Hayden, 2014). However, although these techniques are able to sequence a very large number of DNA fragments per single execution; these fragments are small (usually up to hundreds of bp). We call these novel sequencing platforms "next-generation sequencing" (NGS) techniques (Shendure and Ji, 2008). Since the development of the first NGS technologies (Tucker et al., 2009), they have been constantly improving. We refer to Rusk (2010) for a full discussion on NGS technologies.

The emergence of NGS and its constant technological improvements have enabled the revisiting of traditional biological assays to investigate regulatory elements (described in Section 2.1.2) using the cell-specific chromatin dynamics context (described in Section 2.1.3). On revisiting such methods, their protocols could be adapted in order to fit the NGS technologies, which enables them to be performed in a genome-wide manner. Such large-scale analysis has potential to reveal the high-dimensional relationships between regulatory elements. NGS-based assays have enabled multiple current research progress, which unraveled the regulatory mechanisms linked to conditions, such as cell differentiation or the onset of diseases, of multiple cells (ENCODE Project Consortium, 2012; Neph et al., 2012; Thurman et al., 2012).

In this section we describe the following techniques: (1) Chromatin immunoprecipitation followed by NGS (ChIP-seq; Section 2.2.1) and (2) DNase I footprinting followed by NGS (DNase-seq; Sec-

tion 2.2.2). ChIP-seq combines chromatin immunoprecipitation with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins such as TFs or histones containing a particular modification. The DNase-seq method is used to identify regions of open chromatin, i.e. DNA regions accessible to the binding of TFs.

2.2.1. ChIP-seq

The ChIP-seq technique consists on retrieving target DNA-bound proteins and further sequencing of the DNA fragments retrieved using NGS techniques (Johnson et al., 2007). These target proteins can be, for instance, TFs or histones with a particular post-translational modification. This allows the genome-wide identification of the genomic regions in which a target protein is bound within a single experimental execution. When applied to a target TF, the ChIP-seq experiment allows us to identify the TFBSs. When applied to histones with particular post-translational modification, the ChIP-seq experiment allows us to identify the genomic regions in which these modified histones occur, and therefore make inferences on that region's particular chromatin structure.

The ChIP-seq protocol starts by isolating the nuclei of cells and breaking them in order to access the genomic material (chromatin). The isolated genomic material is cross-linked in order to preserve all protein-DNA binding events. Next, the cross-linked chromatin is sheared into approximately 200 bp DNA fragments. Afterwards, the chromatin lysate is treated with an antibody that targets a particular protein of interest. The solution is then immunoprecipitated. In this procedure, we retrieve only the sheared chromatin fragments that contains the protein of interest. The immunoprecipitated solution is separated and washed in order to keep only the DNA fragments. Then, these DNA fragments are sequenced using an NGS technique. It is important to mention that only the beginning of the retrieved DNA fragments are sequenced (50–100 bp) by NGS techniques. Such process is depicted in Figure 2.9a–c.

The sequenced DNA fragments (termed "reads") are mapped back into the reference genome using string alignment algorithms (Figure 2.9d), which are developed specially for mapping short DNA reads (length of 50–100 bp) into a big reference genome (human genome length is ~3.1 billion bp). Such demanding computational problem is considered a solved problem and there are many available algorithms such as Bowtie 2 (Langmead and Salzberg, 2012) or the Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2009). Given these aligned reads, we can generate a genomic signal by calculating the overlap between these reads at every genomic coordinate, i.e. every bp of the genome (Figure 2.9e–f). Nevertheless, since only the first 50–100 bp of the fragments are sequenced, they need to be extended to reflect the real length of the immunoprecipitated fragments (approximately 200 bp). This extension step reflects the fact that the protein is bound to virtually any location within the immunoprecipitated DNA fragment.

Finally, we can identify the binding locations of the target protein by evaluating the genomic regions with more reads mapped than expected by chance (often referred to as "enriched regions"). As shown in Figure 2.9f such regions with more ChIP-seq mapped reads than expected by chance can be seen as "peaks" in the signal generated by counting the number of mapped ChIP-seq reads in each genomic position (Figure 2.9g). The identification of significant peaks in ChIP-seq mapped reads is also a computational problem which was solved with the development of genomic peak-calling algorithms such as the model-based analysis for ChIP-seq (MACS) (Zhang et al., 2008). Since the ChIP-seq signal has a low resolution, i.e. it is smoothed given the fact that we have to extend the aligned reads, the target protein is considered to be likely bound anywhere within the called peaks.

2.2.2. DNase-seq

The DNase-seq technique consists in the observation of the DNA digestion by a certain cleavage agent able to break the DNA molecule (Crawford et al., 2004; Sabo et al., 2004b). The cleavage agent

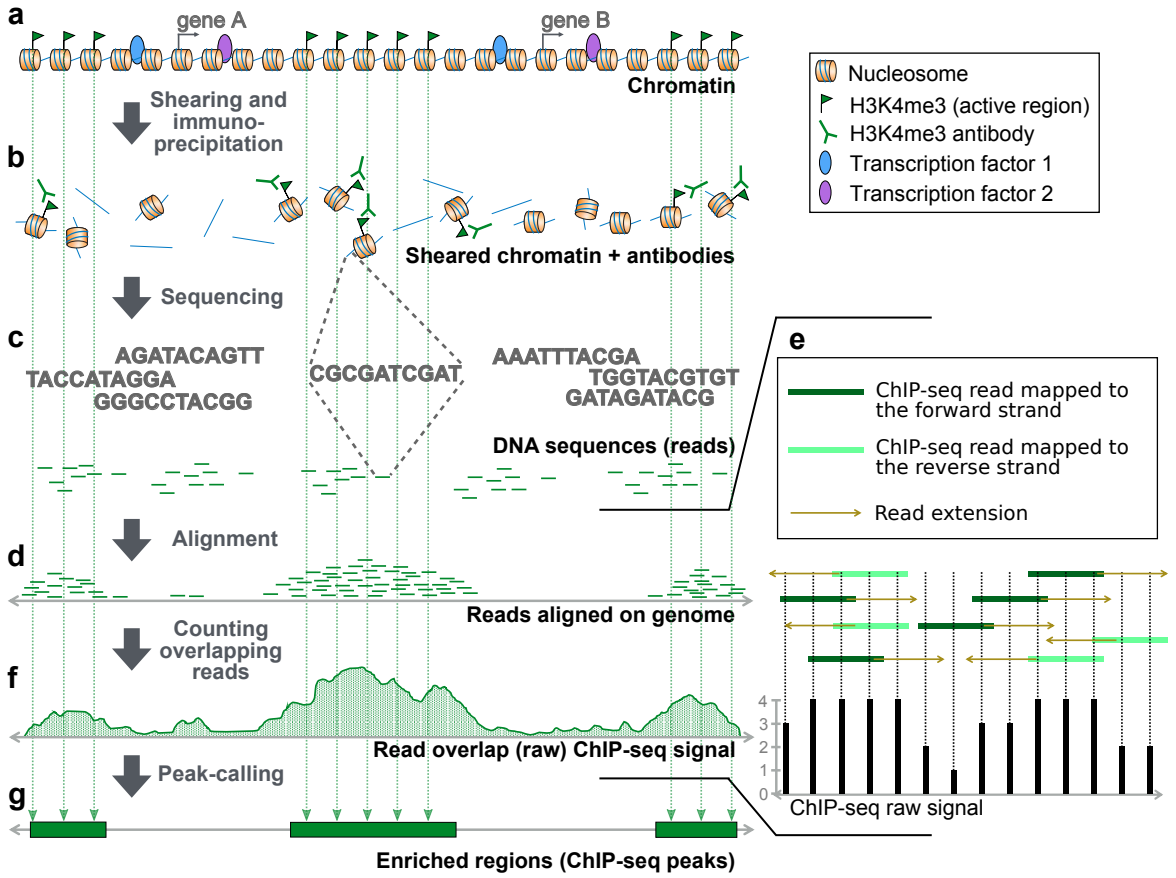

Figure 2.9: Chromatin immunoprecipitation sequencing experimental technique (ChIP-seq). (**a**) The protocol starts by obtaining the chromatin from multiple cells. Such chromatin is cross-linked to preserve all protein-DNA interactions. (**b**) The cross-linked chromatin is sheared and treated with antibodies that target a particular protein of interest. The antibodies bind the target-protein and the chromatin fragments bound by these antibodies are recovered (immunoprecipitated). (**c**) The DNA fragments are sequenced using NGS-based techniques. Only the first 50–100 bp of the ~200 bp fragments are sequenced. (**d**) The sequenced DNA fragments (reads) are mapped back into the reference genome using an alignment algorithm. (**e**) A genomic signal is generated by counting the overlap of the extended mapped reads. The reads are extended to match their original immunoprecipitated fragment length of ~200 bp. (**f**) The resulting genomic signal is enriched, i.e. present peaks, in the locations where the target protein is likely bound in the genome. (**g**) We can search for regions enriched with the ChIP-seq signal, which are the putative regions in which the target-protein is binding in the genome, using a peak-calling algorithm. *Based on protocol description in Johnson et al. (2007).*

used in this method is the enzyme deoxyribonuclease I (DNase I). The rationale of this method is that the DNase I enzyme can cleave the DNA in regions where it is accessible (i.e. open chromatin). Furthermore, within open chromatin regions, the DNase I enzyme cleaves the DNA only at protein-free regions, leaving "footprint marks" that can be traced back as DNA-bound (active) TFBSs.

The DNase-seq protocol starts by isolating the nuclei of cells and breaking them in order to access the genomic material (chromatin). The isolated genomic material is treated with optimal concentrations of DNase I, which cleaves the chromatin at random accessible positions. These accessible positions are the chromatin regions in which the DNA is open (i.e. not fully wrapped around the histone complexes) and protein-free (i.e. not being bound by proteins such as TFs). Such cleaved DNA fragments are isolated and sequenced using the same algorithms as described for the ChIP-seq procedure (Section 2.2.1). Such process is depicted in Figure 2.10a–e.

Then, a genomic signal is created by counting the number of overlapping reads at every genomic position. In the DNase-seq case, we only count the first base pair in the start ($5'$ position) of the reads, since that is the position in which the DNase I enzyme has cleaved the DNA and indicates an open chromatin region (Figure 2.10f–g). The resulting genomic signal represents a nucleotide-resolution map of the open chromatin positions within the whole genome.

Finally, we can detect the genomic regions with more reads mapped than expected by chance, often referred to as "DNase hypersensitivity sites" (DHSs; Figure 2.10h). DHSs are detected by using algorithms specially designed for such purpose such as the F-seq (Boyle et al., 2008). Each DHS is composed of several DNase-seq signal peaks. Note that the depletions within two of these nucleotide-resolution peaks are indicative of a region wherein the DNase I enzyme could not access because there was a protein binding in that region. The DNase-seq signal depletion between two DNase-seq peaks is called a "footprint" (Figure 2.10h). The identification of footprints gives us a genome-wide map of putative active TFBSs.

There are two different protocols to perform the DNase-seq experiment, termed "single-hit" and "double-hit". There are a few experimental differences but the most important is the fragment size selection and isolation. While the single-hit DNase-seq protocol selects for larger cleaved fragments representing the extremities of the DHSs, the double-hit protocol selects for shorter cleaved fragments within the DHSs. Nevertheless, the resulting genomic signal and all post-read alignment steps are the same between these two approaches.

It is important to point the differences between DNase-seq and ChIP-seq. In the DNase-seq method, we determine the binding of any protein in the region being analyzed, without knowing which protein is binding; however in ChIP-seq we only determine the binding of a particular target protein with a known antibody in the region of interest. Furthermore, while the DNase-seq can provide the precise protein binding location, the ChIP-seq tells us an approximated region for the binding of the target protein, since the protein binds virtually any *locus* within the ~200 bp immunoprecipitated fragments. The selection of the technique to use depends mainly on the experimental design and should consider these important details.

2.3 Computational Prediction of Active Transcription Factor Binding Sites

The identification of active TFBSs is a very important task, since they are the key players on most regulatory mechanisms. The detection of such regulatory elements have enabled significant advances in the understanding of many biological mechanisms such as cell differentiation (Lin et al., 2015; Tsankov et al., 2015) and the onset of diseases (Schaub et al., 2012; Vernot et al., 2012; Charos et al., 2012). In this thesis, we are going to address the computational prediction of active TFBSs.

The standard computational approach is the use of sequence-based methods, which search over the genome's DNA for sequences representing the DNA binding affinity sequence of TFs (Figure 2.5) (Stormo, 2000). However, this approach is not able to predict active binding sites, i.e. binding sites that are being currently bound by TFs at a particular cell state (Section 2.3.1). To introduce the cellular context, we can use the ChIP-seq for TFs (Section 2.3.2). However, success of ChIP-seq assays depends on the existence of a good antibody against the TFs of interest and on the availability of large numbers of cells. These two conditions are not always met in particular for primary cells. Furthermore, ChIP-seq is an expensive technique. Therefore, experiments involving ChIP-seq are restricted to the analysis of a small selection of TFs and cell types (Kim et al., 2008; Ouyang et al., 2009) or require the effort of large consortia (ENCODE Project Consortium, 2012). A solution to sequence-based and TF ChIP-seq-based methods' limitation is to explore the fact that an open chromatin structure is crucial and a prerequisite to the active binding of a TF on the DNA (Arvey et al.,

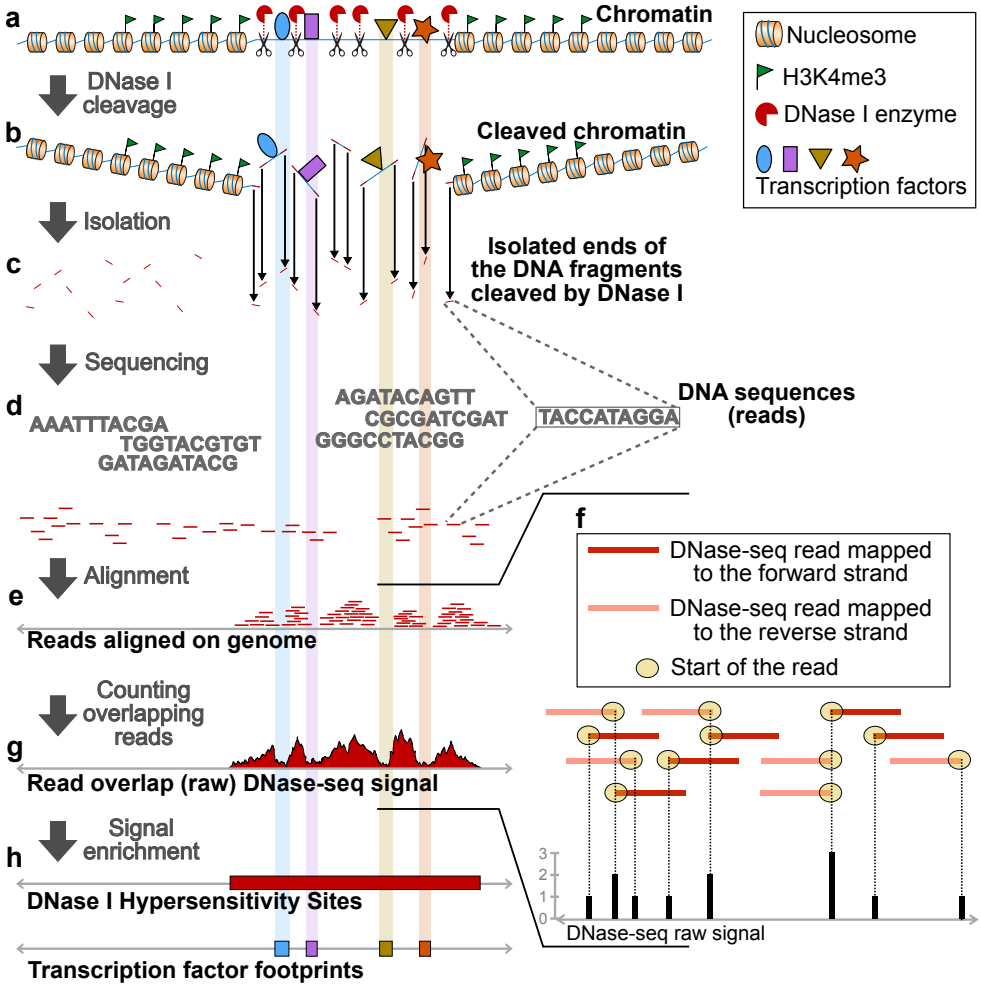

Figure 2.10: DNase I sequencing experimental technique (DNase-seq). (**a**) The protocol starts by obtaining the chromatin from multiple cells. (**b**) The chromatin is treated with optimal levels of DNase I enzyme, which cleaves the chromatin in accessible sites. These accessible sites are the ones in which the chromatin is open and is not being bound by proteins (as TFs or histones). (**c**) After the DNase I digestion we have multiple fragments of sheared chromatin. The extremities of these fragments are isolated. (**d**) The isolated DNA fragments' extremities are sequenced using NGS-based techniques. (**e**) The aligned DNA sequences are mapped back into the reference genome. (**f**) A signal is generated by counting the overlap of the mapped DNA sequences. Since we are interested in the positions in which the DNase I enzyme has cleaved the DNA, only the first bp (i.e. $5'$ bp) of the aligned read is count during the creation of the overlap signal. (**g**) The resulting read overlap signal exhibits peaks in the open chromatin (accessible to DNase I) regions. (**h**) Regions with high concentration of DNase-seq reads are termed DNase hypersensitivity sites (DHSs). Each of these regions are characterized by a number of depletions between peaks, which represent TF footprints. *Based on protocol description in Crawford et al. (2004) and Sabo et al. (2004b).*

2012). This solution is used by the computational chromatin-based methods (Section 2.3.3).

2.3.1. Sequence-Based Methods & Limitations

The standard computational approach to detect TFBSs is the use of sequence-based methods. These methods search the genome for the TF's DNA binding affinity motif (Stormo, 2000). Therefore, computational sequence-based methods to detect TFBSs can be viewed as string searching algo-

rithms. The most common sequence-based approach is called motif matching. This algorithm scans the genome and scores every contiguous DNA sequence using a matrix representation of the TF's motif termed position frequency matrix (more details in Section 4.2.1). The TFBSs predicted using computational sequence-based methods are called motif-predicted binding sites (MPBSs).

Computational sequence-based methods, such as the motif matching, have a low computational complexity, which makes their genome-wide application easy (Mathelier and Wasserman, 2013). However, while the genome is a large sequence of nucleotides (human genome length is ~3.1 billion bp), the TF's binding affinity sequences are small (usually between 5–20 bp) and degenerate (only a fraction of the motif is highly conserved). Therefore, it is hard to fine-tune the sensitivity at the expense of the specificity (Stormo, 2000). Furthermore, this technique has a major disadvantage: it is unable to identify active binding sites, i.e. binding sites that are actually being bound by proteins at a specific cellular condition (Boyle et al., 2011). This happens because computational sequence-based methods rely solely on the DNA sequence affinity of proteins. However, the DNA sequence is the same between different cells for a particular organism; independent of cell type, cellular condition, life stage, stimuli response, and others. The key characteristic which allows different cells with the same genetic material to express a different set of proteins is the chromatin structure. Therefore, information regarding the chromatin structure is a prerequisite to identify active (cell type-specific) TFBSs (Arvey et al., 2012; Thurman et al., 2012).

In practice, the fact that computational sequence-based approaches are unable to identify active binding sites is expressed as a very high number of false positive sequence-based predictions, representing the set of TFBSs not being accessed in a particular cellular condition. Figure 2.11 exemplifies this issue. We applied the motif matching tool "find individual motif occurrences" (FIMO) (Grant et al., 2011) in a genomic region using 520 TFs affinity representations with a conservative threshold to accept binding site hits. The result shows more than 3,000 MPBSs on a 3,000 bp region, which absolutely does not correspond to any possible biological regulatory model.

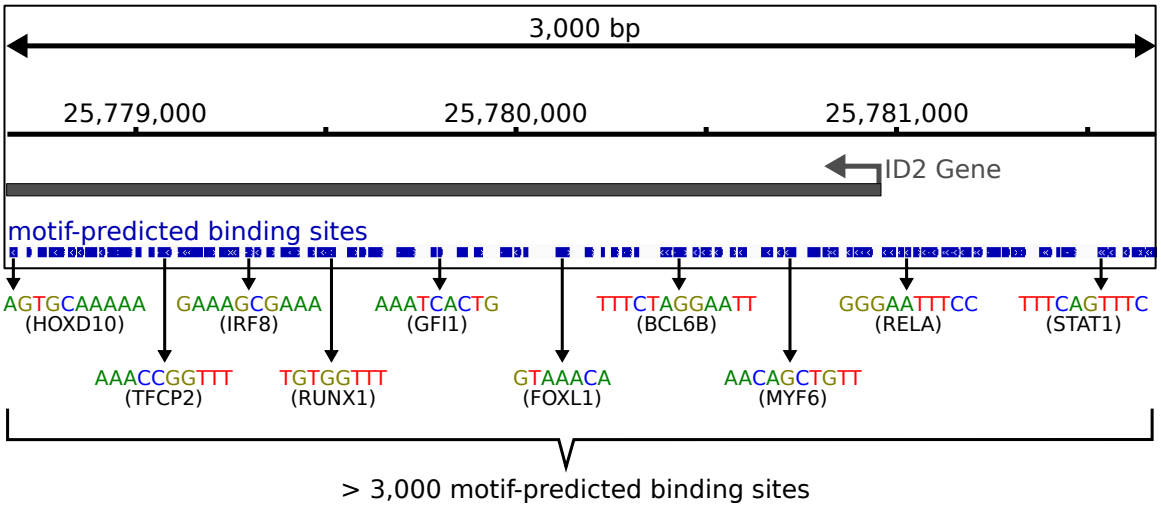

Figure 2.11: Main problem of computational sequence-based methods. The motif matching tool FIMO (Grant et al., 2011) was applied in a 3,000 bp genomic region using 520 TF DNA sequence binding affinity models and resulted in more than 3,000 MPBSs. Such biologically impossible scenario exemplifies the fact that computational sequence-based approaches (such as motif matching) are, alone, unable to identify active binding sites.

2.3.2. ChIP-seq for Transcription Factors

In order to make predictions of active TFBSs, we must use experimental data that give information on the chromatin dynamics and provide the required cell-specificity (Arvey et al., 2012; Thurman et al., 2012). Such information is obtained with the biological assays discussed in Section 2.2.

One way to obtain active TFBSs is to perform ChIP-seq on the target TFs of interest. Such assay provides all TFBSs for such target TFs with a very good accuracy and reasonable resolution.

However, using ChIP-seq to detect active TFBSs has a few limitations. First, the ChIP-seq relies on the quality of the antibody used on the immunoprecipitation step. There are many TFs in which the antibodies do not work properly or do not work at all. Second, if the experimental design relies on the identification of a small number of TFs (< 5), then ChIP-seq for TFs might be a good experimental choice; however, if one is interested in a higher number of TFs, the number of TF ChIP-seq assays makes the study very expensive and time consuming (Boyle et al., 2011; Pique-Regi et al., 2011).

2.3.3. Chromatin-Based Methods

Given the limitations of ChIP-seq for TFs, a solution is to use experimental open chromatin data to narrow the search of active TFBSs. As previously discussed, histone modifications *loci*, which are obtained using ChIP-seq, mark regions in which the chromatin is open, and therefore accessible to TFs (Park, 2009). Furthermore, the DNase-seq experimental assay provides open chromatin regions with a very high spatial specificity (Boyle et al., 2008).

There is a distinctive pattern surrounding active TFBSs that is observed in genomic signals from DNase-seq and histone modification ChIP-seq. We refer to this pattern as the "grammar of active TFBSs" (Gusmao et al., 2012, 2014). This grammar, depicted in Figure 2.12, shows that TFBSs happen at depletions between two peaks of the DNase-seq signal (DNase footprints). These regions with multiple DNase-seq peaks, which comprise a DHS, occur within a depletion between two peaks of activating histone modifications (histone footprints) (Boyle et al., 2011; Gusmao et al., 2014).

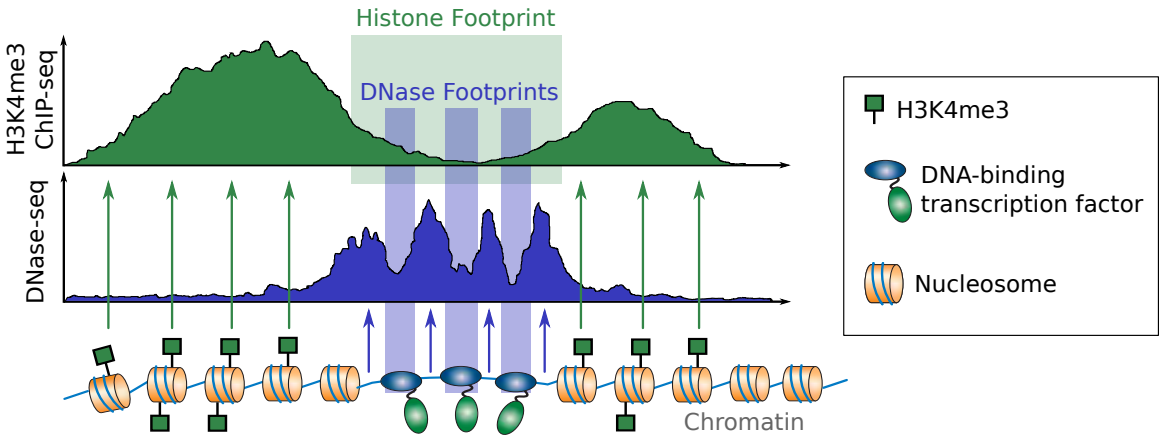

Figure 2.12: Grammar of active TFBSs. In this figure we show examples of histone modification ChIP-seq and DNase-seq signals for a genomic region (top) and a graphical representation of the chromatin landscape within this region (bottom). We and others observed that there is a clear pattern regarding these signals and the TFBSs. TFBSs happen at depletions between two peaks of the DNase-seq signal (DNase footprints). Furthermore, these DNase-seq peaks, which determines an open chromatin region, happen at the depletion between two peaks of active histone modification marks (histone footprints).

Such patterns can be used in order to make better predictions of active binding sites, when compared to purely sequence-based methods (Pique-Regi et al., 2011; Cuellar-Partida et al., 2012). Fur-

thermore, a complete map of all putative active TFBSs, i.e. footprints, for a particular cell type can be obtained with a few assays (in the example of Figure 2.12, two assays: ChIP-seq for H3K4me3 and DNase-seq). However, the detection of footprints with DNase-seq and histone modification ChIP-seq require special computational frameworks. Such computational frameworks which processes open chromatin NGS-based data gained popularity over the last years and are used to address the problem of active TFBS prediction. Such chromatin-based computational methods that use open chromatin data to predict active binding sites are called "computational footprinting methods" and are the main subject of this thesis.

2.4 Computational Footprinting Methods

In this section we present the state-of-the-art computational methods, which use the grammar of active TFBSs to perform predictions of active TFBSs – the computational footprinting methods (Section 2.4.1). We define the different types of computational footprinting methods (Section 2.4.2) and describe how they have been evaluated in the literature (Section 4.2). Next, we show the current challenges on the identification of active TFBSs using computational footprinting approaches (Section 2.4.4). Finally, we close this section with a comprehensive literature review on published computational footprinting methods (Section 2.4.5).

2.4.1. Method Definition

In this thesis we focus on computational footprinting methods to address the problem of active TFBS prediction. We formalize the concept of computational footprinting method as follows:

Computational Footprinting Method: *A computational framework to analyze open chromatin (NGS-based) data and create a genome-wide map of active TFBSs.*

The term "computational framework" refers to a set of methods and algorithms used to process the open chromatin data and perform the prediction of putative active TFBSs. Such computational framework has to be capable of executing within a reasonable amount of time with massive genome-wide data. Therefore, effort has to be done on applying efficient data structures and algorithms with minimal computational complexity. The output of computational footprinting methods consist of multiple genomic regions, each of which starts and ends at particular genomic coordinates, which represents the putative active binding sites. Furthermore, the predicted footprints should be as close as possible, in terms of genomic position and predicted region's width, to the real TFBSs. In other words, the method should have a high spatial specificity.

The Footprint Score (FS) Method

The simplest computational footprinting approach, termed "the footprint score (FS)" (Neph et al., 2012), consists on sliding a window across the genome and evaluating the ratio between the number of reads (from a particular open chromatin experiment such as DNase-seq) inside the window and inside the flanking regions of the window (Figure 2.13). More formally, let $\mathbf{x} = \langle x_1, \cdots, x_n \rangle$ be a genomic signal in which $x_i \in \mathbb{N}^0$ represents the number of DNase-seq mapped reads starting at the genomic position i within a genome with total length n. The FS for a particular window represented as a genomic region $r_i = [u, v]$, which is an interval from the genomic coordinate u to v (including both), is calculated as

$$\mathrm{FS}_{r_i} = \left(\frac{n^C_{r_i} + 1}{n^R_{r_i} + 1} + \frac{n^C_{r_i} + 1}{n^L_{r_i} + 1} \right), \tag{2.1}$$

where $n^C_{r_i}$, $n^L_{r_i}$ and $n^R_{r_i}$ are, respectively, the number of reads within, in the left and right flanking regions of the genomic region $r_i = [u, v]$. This is written as

$$n^C_{r_i} = \sum_{j=u}^{v} x_j, \qquad n^R_{r_i} = \sum_{j=v}^{2v-u} x_j, \qquad n^L_{r_i} = \sum_{j=2u-v}^{u} x_j. \tag{2.2}$$

As depicted in Figure 2.13, a negative $-\log \mathrm{FS}_{r_i}$ indicates more reads (i.e. DNase I cleavage hits) in the core of the window r_i than in its flanking regions. On the other hand, a positive $-\log \mathrm{FS}_{r_i}$ indicates more reads in the window's flanking region in comparison with the core. As a consequence of the grammar of active TFBSs, we regard the regions with $-\log \mathrm{FS}_{r_i} > t$, in which t is a cutoff threshold, as the predicted footprints. This simple approach which relies on evaluating a score given a sliding window exemplifies the rationale behind computational footprinting methods. However, the problem with such window-based approach is that the length of the footprint is unknown and varies significantly between different proteins (4–50 bp) (Neph et al., 2012). Furthermore, the length of the flanking regions, which in the FS method is the same as the length of the binding site, is also unknown and highly heterogeneous even with regard to the same TF (Sung et al., 2014).

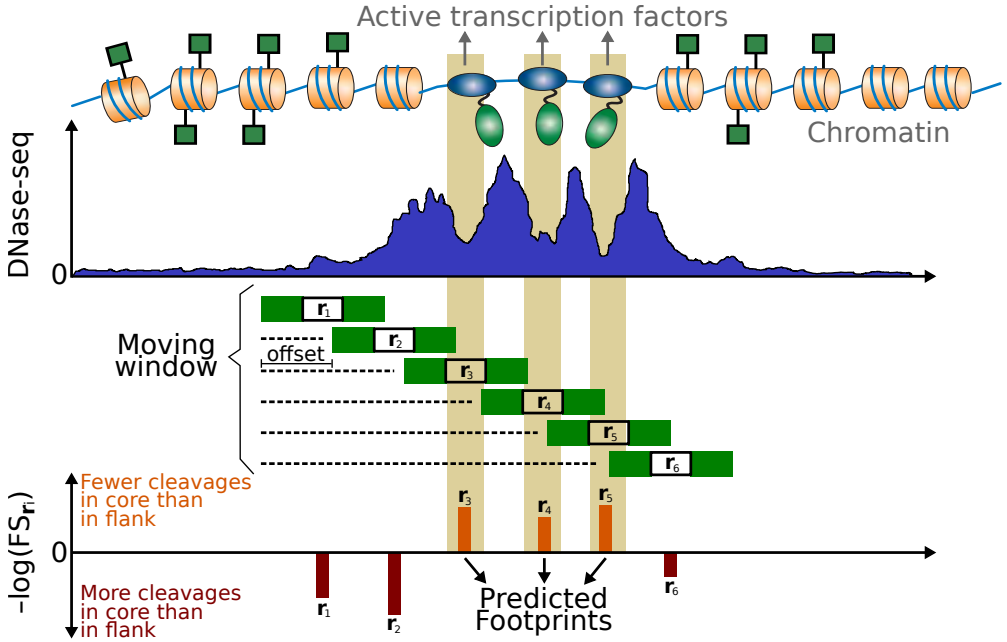

Figure 2.13: Simple example of computational footprinting. Depiction of a simple computational footprinting method which consists on sliding a window (denoted as r_i) composed of a core part (white box) and flanking regions (green box). The window, with a particular size, slides on the genome given a specific offset and corresponds to a predictive score. In the case of the FS shown in Equation 2.1, high positive $-\log \mathrm{FS}_{r_i}$ values indicate footprints, which are putative active TFBS predictions.

2.4.2. Types of Computational Footprinting Methods

Computational footprinting methods are broadly categorized as: (1) segmentation methods and (2) site-centric methods (Figure 2.14). The rationale behind segmentation methods is to scan the genome and use open chromatin data to segment the genome in open and closed chromatin regions. By measuring the different levels of histone modification ChIP-seq and/or DNase-seq signal intensity, the segmentation methods are able to identify patterns which correspond to the grammar of active TFBSs; consequently, detecting footprints in a genome-wide manner. This approach generates a map

of putative active binding sites without specifying the TFs that are binding these regions. Although further processing is necessary in order to identify which particular TF bind to these putative binding sites; the advantage of such approach is that binding sites of unknown TFs can be detected. Figure 2.14a shows an example of the segmentation computational footprinting approach. The example shown in this figure is similar to the FS shown in Figure 2.13.

On the other hand, site-centric methods (Figure 2.14b) start with putative binding sites obtained by using, for instance, a sequence-based prediction method such as motif matching (Section 2.3.1). Then, open chromatin experimental data around these *a priori* predictions are gathered and classified, generally using unsupervised machine learning methods. This approach leads to footprints for target TFs. The advantage of such approach is that we already know which TFs are binding to the predicted footprints. However, the disadvantage is that it depends on the *a priori* TF evidence (such as the DNA binding affinity motif), which is not always available. Consequently, site-centric techniques are only able to identify binding sites from well-known TFs.

Moreover, the computational complexity of the site-centric approach is larger than that of segmentation-based methods. To assess the complexity in terms of the big-$\mathcal{O}$ notation, let n be the length of the genome, w be the length of a window in which footprints are being searched, m be the number of TFs in which we are interested in analyzing and h be the number of MPBSs found by applying a motif-matching algorithm. The segmentation approach requires the sliding of a window of length w with offset of a fraction of w (usually $\frac{1}{3}$ in the case of the FS method) in the genome of size n. Therefore, the segmentation approach is $\mathcal{O}(n+w)$. The site-centric approach first requires the application of the motif matching, which has complexity $\mathcal{O}(nw)$. Then, it performs a classification algorithm in each MPBS, which has complexity of at least $\mathcal{O}(wh)$. Finally, the site-centric approach performs this operation for each one of the m TFs. Therefore, it has complexity of $\mathcal{O}(m(nw+wh)) = \mathcal{O}(m((n+h)w))$. In other words, the segmentation approach requires one execution per genome to provide footprint predictions for all putative TFBSs; the site-centric approach requires one genome-wide execution per TF. In practice, if one is interested on a genome-wide exploratory analysis (~500–1000 TFs), the execution of site-centric methods requires a significant amount of computational time.

2.4.3. Evaluation of Computational Footprinting Methods

There is no well-defined gold standard for the evaluation of footprinting methods. All works so far have used ChIP-seq of TFs in conjunction with MPBSs as ground truth (Pique-Regi et al., 2011; Cuellar-Partida et al., 2012). Such method provides a straightforward scenario for the evaluation of computational footprinting methods. The idea behind such an evaluation approach is that the TF ChIP-seq provides the cell-specificity and the MPBSs provides a countable structure which is used to calculate statistics. In the following we define such procedure.

In the so-called "ChIP-seq evaluation approach", MPBSs with ChIP-seq evidence (which can be, for instance, MPBSs close to TF ChIP-seq peak summits) are considered "true" TFBSs. On the other hand, MPBSs without ChIP-seq evidence are considered "false" TFBSs. Every TFBS prediction (i.e. footprint) that overlaps a true TFBS is considered a correct prediction (true positive – TP) and every prediction that overlaps a false TFBS is considered an incorrect prediction (false positive – FP). Therefore, true negatives (TN) and false negatives (FN) are, respectively, false and true TFBSs without overlapping predictions. This is depicted on Figure 2.15a.

The contingency table (TPs, FPs, TNs and FNs) enables the creation of receiver operating characteristic (ROC) curves, which describe the sensitivity increase as we decrease the specificity of the method (Figure 2.15b). The area under the ROC curve (AUC) is a good metric to evaluate the overall performance of computational footprinting methods. Furthermore, the contingency table also enables the evaluation of the area under the precision-recall (PR) curve (AUPR; Figure 2.15c). This metric is indicated for problems with imbalanced datasets (distinct number of positive and negative

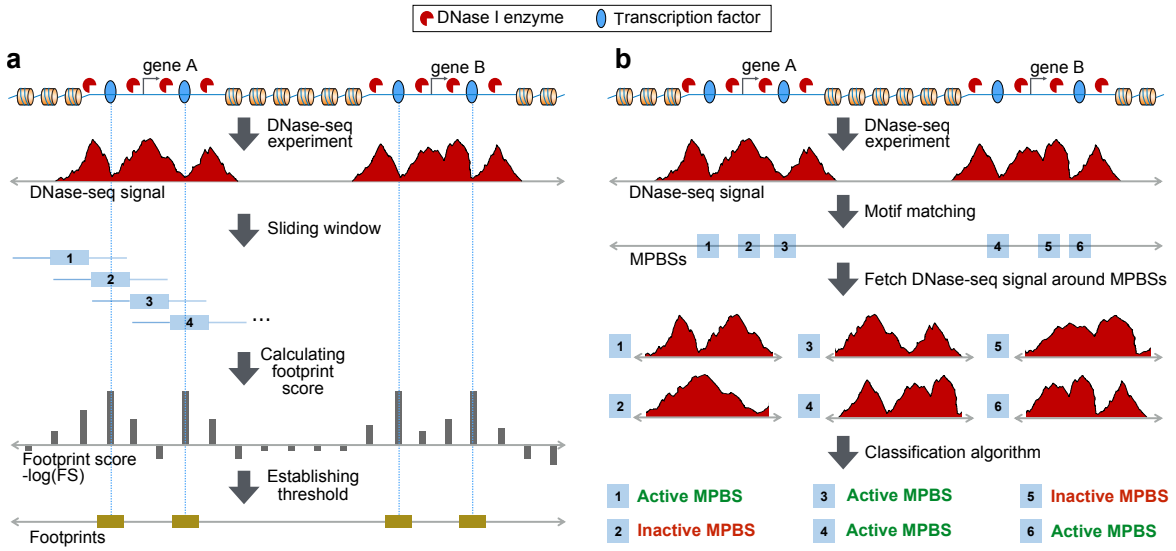

Figure 2.14: Segmentation *vs* site-centric computational footprinting methods. Example of each computational footprinting approach using DNase-seq data. Both approaches try to detect the grammar of active TFBSs. (**a**) The segmentation computational footprinting approach "reads" the DNase-seq signal as a time-series and searches for footprints based on the detection of patterns corresponding to the grammar of active TFBSs. In this case, the FS method is used as an example. The output of the segmentation approach corresponds to the genomic regions recognized as footprints. (**b**) The site-centric approach starts by detecting MPBSs using sequence-based algorithms such as motif matching. Then, the DNase-seq signal around all MPBSs is used to classify them as active/inactive binding sites. Such classification method distinguishes MPBSs that match the grammar of active TFBSs (active MPBSs) from the ones that do not (inactive MPBSs). In this case, the active MPBSs represent the footprint predictions.

examples) (Davis and Goadrich, 2006; Fawcett, 2006).

2.4.4. Current Challenges

There are a number of challenges in the development of computational footprinting methods for the identification of active TFBSs. Here we discuss current challenges which were not fully addressed by previous research studies. These challenges are the main motivations for the work presented in this thesis.

Integration of Multiple Data Sources

None of the computational footprinting methods published so far took advantage of the full-resolution spatial profile of all open chromatin NGS-based genomic signals. The current integrative approaches usually take only the full spatial signal for only one input data type. For instance, although a computational footprinting method called Centipede (Pique-Regi et al., 2011) uses the full-resolution profile for the DNase-seq signal, it only uses averaged versions of the other signal types, such as histone modification ChIP-seq. Studies that considered some sort of integrative approach (Pique-Regi et al., 2011; Cuellar-Partida et al., 2012; Sherwood et al., 2014; Kähärä and Lähdesmäki, 2015) argue that the high degree of variation between different open chromatin NGS-based genomic signals makes their integration difficult and prone to overfitting. Furthermore, a number of computational footprinting methods used smoothed versions of the open chromatin data (Cuellar-Partida et al., 2012; Sherwood et al., 2014; Kähärä and Lähdesmäki, 2015). Therefore, the full spatial profiles from the

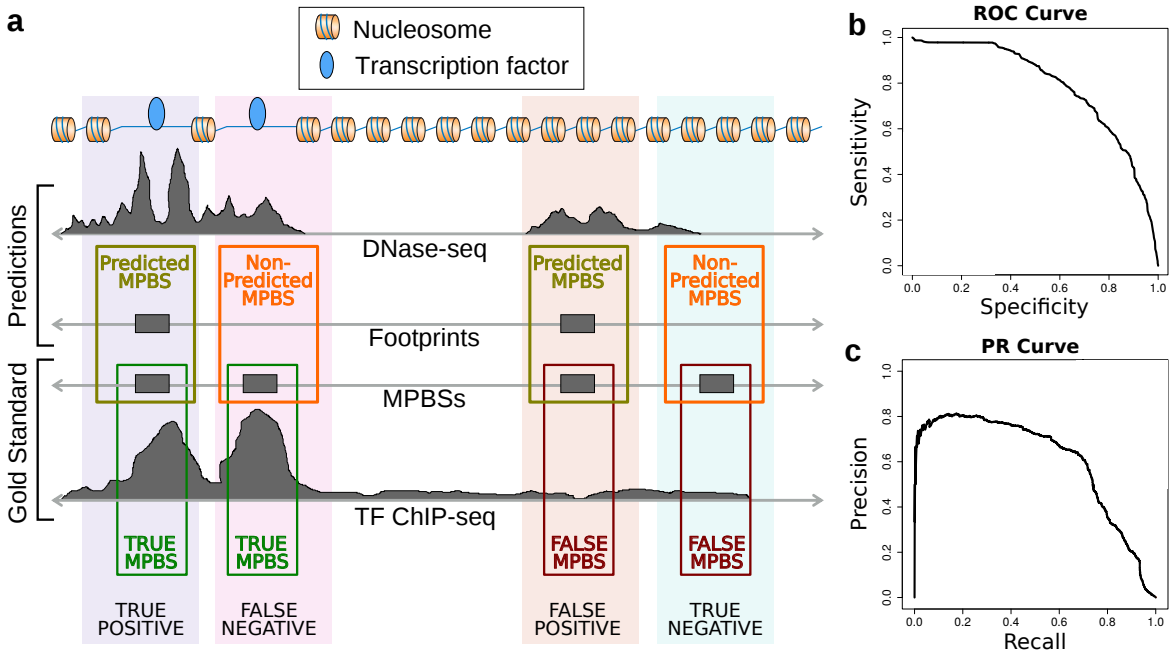

Figure 2.15: Evaluation of computational footprinting. (**a**) In the ChIP-seq evaluation scheme, MPBSs for a particular TF are considered true if they are within a ChIP-seq enriched region for that particular TF. Otherwise, they are considered false MPBSs. By overlapping this information with footprint predictions from a particular computational footprinting method we are able to generate a contingency table with true positives, true negatives, false positives and false negatives. By ranking the MPBSs based on the overlapping footprint's quality, we are able to create: (**b**) receiver operating characteristic (ROC) curves and (**c**) precision-recall (PR) curves. These statistics can be used to rank computational footprinting methods and evaluate their overall performance on identifying each TF individually.

integration of multiple sources of open chromatin data, such as DNase-seq and various histone modification ChIP-seq, has not been fully explored.

Treatment of Intrinsic Experimental Bias on Open Chromatin Data

It is known that open chromatin NGS-based genomic signals, such as DNase-seq and histone modification ChIP-seq, are noisy and intrinsically complex (Meyer and Liu, 2014). Most methods use smoothing techniques to handle such complexity (Pique-Regi et al., 2011; Cuellar-Partida et al., 2012; Sherwood et al., 2014; Kähärä and Lähdesmäki, 2015). Although a few attempts have been made to use these signals in their maximum possible resolution (Boyle et al., 2011; Sung et al., 2014), not much attention was given to data processing techniques such as signal normalization.

Furthermore, open chromatin NGS-based genomic signals are affected by multiple artifacts stemming from either the biological protocol or the computational pre-processing steps. These artifacts were summarized recently by Meyer and Liu (2014). He et al. (2014) showed that the DNase-seq sequence cleavage bias around TFBSs strongly affects the performance of the FS method, in a TF-specific manner. Such bias stems from the fact that the DNase I enzyme, used in the DNase-seq experimental technique, binds more often to certain DNA sequences than others. Since TFs also have a sequence binding preference (as depicted in Figure 2.5); the DNase-seq sequence cleavage bias might induce artificial footprints. This issue is observed, for the FS method, as a significant correlation between the amount of DNase-seq sequence cleavage bias surrounding each TF's binding sites and their computational footprinting prediction accuracies (Figure 2.16a).

He et al. (2014) also indicated several TFs, such as nuclear receptors, in which the DNase-seq footprint pattern resembles their DNase-seq sequence cleavage bias estimate. For instance, the average DNase-seq signal around binding sites of the transcription AR (androgen receptor) exhibits very similar patterns when a comparison is made between DNase-seq data from a cell type in which AR is known to be active (Figure 2.16b) and a cell type in which AR is known to be inactive (Figure 2.16c).

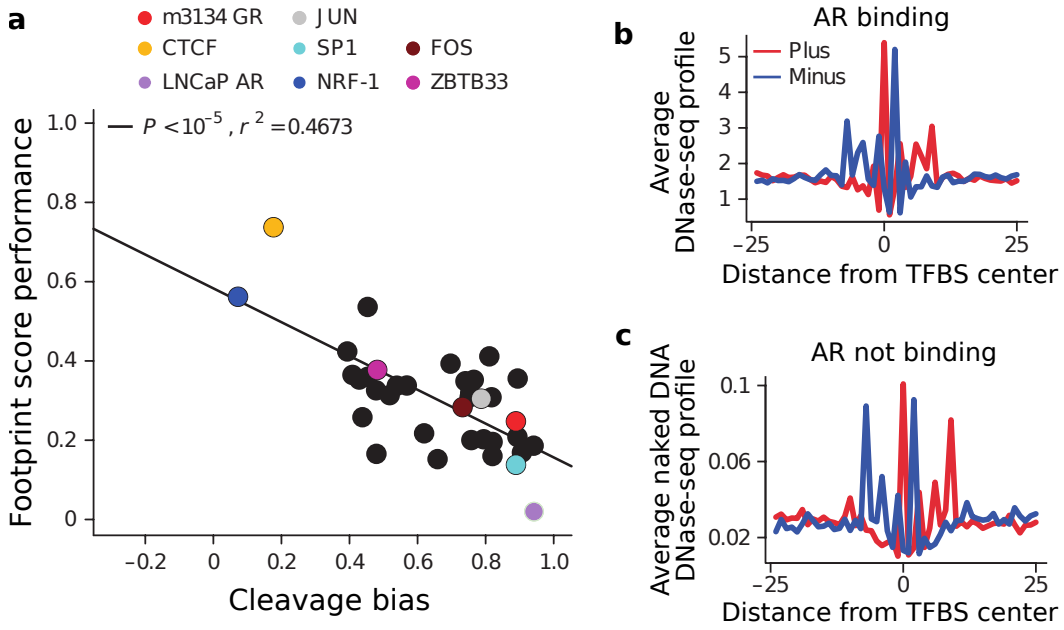

Figure 2.16: Impact of DNase-seq sequence cleavage bias on computational footprinting. (**a**) This graph shows the amount of DNase-seq sequence cleavage bias (*x*-axis) *vs* the performance (AUC from the ChIP-seq evaluation approach) of the FS footprinting method. We clearly observe that there is a strong negative correlation between these two variables. (**b**) Average DNase-seq signal profile around ChIP-seq peaks of the AR TF in a cell type in which it is known that the AR TF is being expressed (i.e. AR is actively binding). (**c**) Average DNase-seq signal profile around the same regions as in (b); however in this case the DNase-seq is from a naked DNA experiment, where all proteins were removed from the DNA. In this case AR is not binding the DNA, but a footprint pattern can still be found. *Source: He et al. (2014)* (modified to fit thesis format and/or clarify key points).

Lack of Benchmark Data for Method Evaluation

Except for a few studies (Sherwood et al., 2014; Yardımcı et al., 2014; Kähärä and Lähdesmäki, 2015), comparative analyses evaluating footprinting methods analyzed only a few (< 12) TFs. Also, a maximum of only four competing methods were evaluated using the same experiment design of a particular published study. In addition, the experiment design for evaluation of computational footprinting methods vary between different publications. Despite the importance of method evaluation (Nature Methods Editorial, 2015), there is a clear lack of benchmark data, evaluation standards and studies performing a comprehensive analysis of computational footprinting methods.

Furthermore, when we consider the few studies that performed comparative analyses so far, all of them have used the ChIP-seq evaluation scheme as described in Section 4.2. This evaluation requires TF ChIP-seq experiments to be carried out on the very same cells as the DNase-seq experiment and has a few caveats. First, TF ChIP-seq peaks are also observed in indirect binding events (Yardımcı et al., 2014), i.e. a binding site might have ChIP-seq evidence of TF A when in fact it is actually binding TF B which is interacting with A by an indirect binding event, such as DNA looping. Second,

they have a lower spatial resolution than DNase-seq. Consequently, false MPBSs might be regarded as true MPBSs by proximity to an active TFBS (Cuellar-Partida et al., 2012; Yardımcı et al., 2014). Therefore, it is important to devise a new evaluation strategy which does not rely on ChIP-seq data for independent evaluation of computational footprinting methods.

Yardımcı et al. (2014) proposed the use of gene expression data to evaluate footprint predictions using the following idea: the higher the expression of a particular TF in cell type A in comparison with cell type B, the higher the quality of the footprint predictions for that particular TF in cell type A (in comparison with B). Nevertheless, this evaluation strategy, which can be used to complement the ChIP-seq evaluation strategy, has not been systematically explored.

Transcription Factor Binding Residence Time

The open chromatin NGS-based techniques described in Section 2.2 also have constraints intrinsic to the regulatory mechanism of the cell. These experimental limitations were not fully explored in the light of computational footprinting method's performance. The main limitation regards the residence time of TFs binding on the DNA. Sung et al. (2014) showed that short-lived TFs, i.e. TFs that have a low binding residence time, display a lower DNase-seq cleavage protection pattern, i.e. low number of DNase-seq mapped reads surrounding the footprint, when compared to a TF with higher binding residence time (Figure 2.17). Such fleeting TFs are harder to detect than other TFs with longer residence time since their protection pattern is less pronounced.

Nevertheless, a systematic evaluation on the extent of the negative impact on footprint prediction accuracy given the TF binding residence time issue has not been made. It is very important to measure such impact to determine the feasibility of computational footprinting for certain TFs and cell types. Furthermore, a TF-wise quantification of such impact would assist in determining the overall quality of each TF's footprint predictions.

2.4.5. Review on Computational Footprinting Methods

A number of computational footprinting methods have been proposed. These methods use different combinations of open chromatin NGS-based experimental data sources, different algorithms and target different experiment designs. Here, we discuss the main published methods, providing a comprehensive literature review on computational footprinting methods.

Hesselberth et al.

One of the first attempts to create a computational footprinting method for DNase-seq data was performed by Hesselberth et al. (2009). In their study, they performed the DNase-seq experiment in the *Saccharomyces cerevisiae* organism (yeast). They used a three-phase segmentation approach to detect footprints in the DNase-seq data. In the first phase, the authors consider every possible window that was contained within one of the specified target regions (DHSs) and compute a depletion score for each of these regions. The second phase consists of selecting high-scoring windows using a greedy algorithm, eliminating from consideration any window that overlapped a window with a higher score. Finally, in a third phase, the authors shuffle the input data independently within each target region and repeat the entire procedure, using the resulting scores to estimate quality scores. They introduced the FS (Equation 2.1) as a quality metric of footprints.

Within this systematic identification of DNase-seq footprints, Hesselberth et al. (2009) analyzed many features regarding such footprint predictions. They identified many known sequence motifs in these footprints, observing that collectively, 35.2% of the footprints with a false discovery rate of 0.05 overlapped a conserved factor binding site inferred from ChIP-seq data. Furthermore, they observed that the patterns of DNase I protection surrounding different TFs had different average shapes, i.e. the

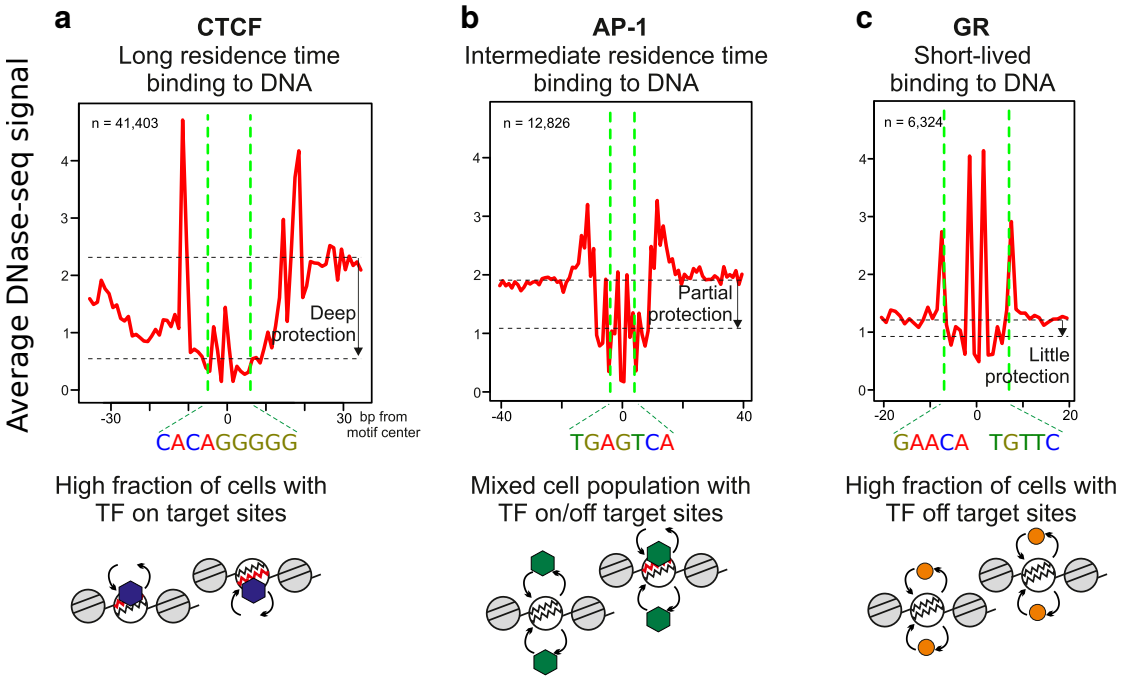

Figure 2.17: TF residence time. This graphs depicts the average DNase-seq signal around TFBSs of TFs with: (**a**) long residence time – CTCF; (**b**) intermediate residence time – AP-1 (C-JUN) and (**c**) short residence time – GR. Sung et al. (2014) suggests that the DNase-seq signal width of the protection against the DNase I enzyme might be correlated with the residence time of TFs in the DNA. *Source: Sung et al. (2014)* (modified to fit thesis format and/or clarify key points).

DNase-seq average signal varies depending on the binding type of TFs. Finally, they created a very consistent genome-wide map of TFBSs for the *Saccharomyces cerevisiae*, which led into insights on the chromatin architecture and gene expression of this organism.

Neph et al.

Neph et al. (2012) used a simplified version of the segmentation-based method originally proposed in Hesselberth et al. (2009). Their method consists on applying a sliding window to find genomic regions (6–40 bp) with low DNase-seq signal between regions (3–10 bp) with high DNase-seq signal (peak-dip-peak pattern). They performed their experiments on human DNase-seq data. They also use the FS to determine the most significant predictions. Their study amplified the analysis scale significantly, by detecting footprints for 41 diverse human cells with data from the encyclopedia of DNA elements (ENCODE) repository (ENCODE Project Consortium, 2012). Such a large-scale study was able to provide multiple new insights on computational footprinting. First, they found that genetic variants affecting allelic chromatin states are concentrated in footprints, and that these elements are preferentially sheltered from DNA methylation. Second, they showed that the average TF-wise patterns of DNase I digestion is correlated to the crystallographic topography of protein-DNA interfaces, indicating that TF structure has been evolutionarily imprinted on the genome. Finally, they performed an extensive "brute force" *de novo* motif finding algorithm and found 683 unique DNA sequence affinity motif structures, of which 394 (58%) matched distinct experimentally-verified motif models present in Jaspar (Mathelier et al., 2014), Uniprobe (Robasky and Bulyk, 2011) and Transfac (Matys et al., 2006) motif repositories.

Boyle et al.

Boyle et al. (2011) designed a segmentation computational footprinting approach, which is based on using hidden Markov models (HMMs) to predict the DNase-seq pattern described in Figure 2.12. Briefly, their HMM uses a normalized DNase-seq signal to find regions with depleted DNase I digestion (footprints) between two peaks of intense DNase I cleavage. As the DNase-seq profiles required a nucleotide-resolution signal, which is usually noisy, the authors used a Savitzky-Golay smoothing filter to reduce noise and to estimate the slope of the DNase-seq signal (Madden, 1978). Their HMM had five states, with specific states to identify the decrease/increase of DHS signals around the peak-dip-peak region. They also provided numerous insights into computational footprinting. First, they described cell-specific footprint patterns, which correlate significantly with gene expression fold change between different cells. Second, they described a conservation phenomenon which was not observed in the conservation study performed by Hesselberth et al. (2009). They find that for most TFs, there is a marked drop in conservation ~10 bp immediately flanking the footprint. Beyond this drop, conservation increases again before gradually decreasing to background levels, creating a "shoulder" in this signal. They also described in details the unique binding characteristics that the human insulator CTCF footprints displays. Finally, they used the STAMP (Mahony and Benos, 2007) method to detect putative TFs in footprints, which simply searches for TFs on known DNA sequence affinity information. Therefore, a *de novo* motif finding approach was not performed as in Neph et al. (2012).

Pique-Regi et al. (Centipede)

One of the most common footprinting approaches was created by Pique-Regi et al. (2011). Their strategy, termed Centipede, is a site-centric approach, which gathers experimental and genomic information around MPBSs. It then uses a Bayesian mixture model as an unsupervised classification tool to label each retrieved MPBS as either "bound" or "unbound". Their approach was the first to integrate multiple different experimental assays. The experimental data include DNase-seq and histone modification ChIP-seq. The DNase-seq data were used at its full spatial resolution (nucleotide-resolution), by obtaining raw DNase-seq signal surrounding a 200 bp window around each MPBS. However, only the average histone modification ChIP-seq signal was used. The genomic data include the scores from the computational sequence-based approach used to create the MPBSs, sequence conservation and distance to the nearest gene. They evaluated their approach on six TFs with ChIP-seq data available, using the ChIP-seq evaluation approach. Their reported average AUC for the six tested TFs were as high as 98.11%. However, they did not observe a gain in accuracy when using a model with both DNase-seq and histone modification ChIP-seq (median AUC $= 96.52\%$).

Cuellar-Partida et al.

Cuellar-Partida et al. (2012) proposed a site-centric method to include open chromatin NGS-based experimental data as priors for the motif matching procedure (Section 2.3.1). Their method uses a probabilistic classification approach inspired in Bayes decision theory to compute better log-posterior odds scores than the ones observed by purely using the DNA sequence binding affinity model. Before the computation of prior probabilities the DNase-seq or histone modification ChIP-seq signals are smoothed as follows. To create smoothed DNase-seq signals they calculated the number of reads aligning to a window of 150 bp, specified every 20 bp. Histone modification smoothed signal input data was specified in a 25 bp resolution. In every position, it was summed 1 if a mapped read fell within 0–200 bp from the 25 bp window and 0.25 if it occurred within 200–300 bp. After the computation of prior probabilities with smoothed DNase-seq signal, they perform the motif match using the program FIMO (Grant et al., 2011). They performed the first comparative study on computational

footprinting methods, where they compared their method with Centipede and with a much simpler approach termed "tag count" (TC). The TC approach, similar to the FS, consists on ranking the MPBSs using the number of DNase I cleavage hits within a window of length l around the MPBSs. More formally, let the interval $r_i = [u, v]$ be the i^{th} MPBS from a set R of MPBSs and $\mathbf{x} = \langle x_1, ..., x_n \rangle$ be the DNase-seq genomic signal from a genome of size n. The TC for $r_i = [u, v]$ is calculated as

$$\text{TC}_{r_i} = \sum_{j=\frac{u+v}{2}-\frac{l}{2}}^{\frac{u+v}{2}+\frac{l}{2}-1} x_j. \tag{2.3}$$

Their results showed that, for their approach, the DNase-seq method dramatically improved the sequence-based prediction of TFBSs. Furthermore, they find that adding the histone modifications H3K4me3 or H3K27ac to their DNase-seq model improved the accuracy slightly. The comparison showed that Centipede outperformed their method using the gold standard proposed in Pique-Regi et al. (2011). However, in this case, they found that using the simple TC approach would outperform both Centipede and their approach. They associate such results with the biases generated by such gold standard created on the basis of TF ChIP-seq data.

Piper et al. (Wellington)

Piper et al. (2013) devised a segmentation approach based on a Binomial test. For a given candidate footprint, it tests the hypothesis that there are more reads in the flanking regions than within the footprint. Following an observation that DNase-seq cuts of the double-hit protocol are strand-specific, Wellington only considers reads mapped to the upstream flanking region of the footprints. They evaluate their method and competing methods in a ChIP-seq-based gold standard created with 214 human TF ChIP-seq datasets. First, they showed that using such observed strand imbalance of reads increases the computational footprinting predictive power. Furthermore, their strategy outperformed the competing methods by Hesselberth et al. (2009), Neph et al. (2012) and Pique-Regi et al. (2011). Finally, this study performs a great contribution by creating a DNase-seq data processing package in the programming language Python termed pyDNase. Such package allows a user-friendly application of their methodology and also further DNase-seq data processing tools.

Sherwood et al. (PIQ)

Sherwood et al. (2014) developed a computational footprinting framework termed protein interaction quantification (PIQ). PIQ is a site-centric method, which uses Gaussian process to model and smooth the footprint profiles around candidate MPBSs (Sherwood et al., 2014). Active footprints are estimated with an expectation propagation algorithm. Finally, PIQ indicates the set of motifs which footprint signals are distinguishable from noise to reduce the set of candidate TFs. They compared their method with competing methods in a very large benchmarking dataset containing 303 TFs binding on K562 human cell type. Through the same evaluation procedure used in the aforementioned works (Pique-Regi et al., 2011; Cuellar-Partida et al., 2012; Piper et al., 2013), they measured a mean AUC of 0.93 for PIQ against 0.87 for Centipede (Pique-Regi et al., 2011) and 0.65 for Neph (Neph et al., 2012) approach.

Nevertheless, this study contains many further analyses which provide insights into computational footprinting. They analyzed the differentiation of mouse embryonic stem cells into pancreatic and intestinal endoderm cells and were able to identify and experimentally validate eight pioneer TF families that perform changes in the chromatin dynamics. One of the most interesting findings is that these pioneer TFs change the chromatin directionally. Besides the identification of pioneer TFs, they also detected "settler" TFs, which binds the DNA after the chromatin structure changes performed by the pioneer TFs.

Yardımcı et al. (FLR)

The issues on computational footprinting presented by He et al. (2014) (Section 2.4.4) were analyzed in Yardımcı et al. (2014). In this study, they proposed a site-centric method (termed FLR) based on a mixture of multinomial models to classify MPBSs as active/inactive in an unsupervised manner. The method uses an expectation maximization algorithm to find a mixture of two multinomial distributions, representing active (footprints) and inactive (background) MPBSs. The background model is initialized with either DNase-seq sequence cleavage bias frequencies or estimated *d*e novo. After successful estimation, MPBSs are scored with the log odds ratio for the footprint *vs* background model. The model takes DNase-seq cuts within a small window around the candidate profiles (25 bp up/downstream) as input. DNase-seq sequence cleavage bias is estimated for 6-mers based on the DNA sequences extracted within the same regions in which the cuts were retrieved. They showed that their method significantly outperformed the simple TC approach. Furthermore, they also criticize the TF ChIP-seq evaluation method on the basis that it is not able to identify indirect binding events. For that reason, they performed a simple analysis based on gene expression and observed that the footprints retrieved by their approach are significantly enriched on cell types where the tested TFs are being expressed.

Sung et al. (DNase2TF)

Sung et al. (2014) also performed a number of analysis that contributed to the discussion initiated in He et al. (2014). First, they developed a new segmentation computational footprinting approach with very simple premises, which is called DNase2TF. DNase2TF is based on the calculation of a binomial *z*-score based on the levels of DNase-seq depletion surrounding candidate footprints. At a second step, DNase2TF interactively merges close candidate footprints whenever they improve depletion scores. DNase2TF corrects for DNase-seq sequence cleavage bias using cleavage statistics for 2 or 4-mers. They reported that their method outperformed Hesselberth et al. (2009), Centipede (Pique-Regi et al., 2011) and Wellington (Piper et al., 2013). Furthermore, as He et al. (2014), they also raised the issue that some TF DNase-seq signatures resemble their cleavage bias. Moreover, they showed that one of the main problems with DNase-seq footprinting was related to the fact that some TFs have a very low residence time on DNA. Since they bind to the DNA in a short time period, the DNase-seq protocol is not able to produce a clear peak-dip-peak pattern.

Kähärä et al. (BinDNase)

Kähärä and Lähdesmäki (2015) developed a supervised site-centric method based on logistic regression to predict active/inactive MPBSs. The algorithm starts with nucleotide-resolution DNase-seq signal around the MPBSs (100 bps up/downstream) and selects discriminatory features using a backward greedy approach. As a supervised approach, the method requires positive and negative examples, which is obtained from TF ChIP-seq data. They showed that their approach does not present any significant gain in performance by modeling DNase-seq sequence cleavage bias. Furthermore, they present a discussion on the standardization of DNase-seq data pre-processing, showing that data on major repositories such as ENCODE are not always analyzed standardly. They state that their supervised approach outperforms unsupervised site-centric approaches such as Centipede (Pique-Regi et al., 2011) and PIQ (Sherwood et al., 2014). However, since their approach is supervised (i.e. needs TF ChIP-seq data for model training), BinDNase is simply a sanity check for the TF ChIP-seq data. Therefore, it has little use in real-case scenarios and shares the same issues regarding the usage of TF ChIP-seq (Section 2.3.2).

Overview of Computational Footprinting Methods

In this section we made a comprehensive discussion on state-of-the-art computational footprinting methods. A summary of the main computational footprinting methods and their features is presented in Table 2.1. In this table we list the main characteristics of these computational footprinting methods:

- **Type.** The type of computational footprinting method: site-centric (SC) or segmentation (SEG).
- **Algorithm.** The main algorithm that the method uses to perform footprint predictions.
- **Bias Correction.** Whether the method performs DNase-seq sequence cleavage bias correction or does not perform such correction. Such correction estimates the sequence cleavage bias for all DNA sequences of length k termed k-mers and use these bias estimates to correct footprint predictions.
- **Resolution/Smoothing.** Whether the method applies a smoothing technique in the input open chromatin data or if it uses the full base-pair (bp) resolution data.
- **Footprint Ranking.** The metric used to rank the footprint predictions. It is used as a quality metric to filter out lower-scored footprints.
- **Availability.** The availability of software tool or source code. The methods obtain a '+' if they are public available ('–' otherwise).
- **Usability.** Defines how complex it is to execute the method. The methods natively supporting standard genomic files and being executed with few commands (≤ 3) have '+' ('–' otherwise).
- **Others.** Other important additional information about the method.

Table 2.1: Overview of computational footprinting methods. *Source: Gusmao et al. (2016)* (modified to fit thesis format and/or clarify key points).

Name	Type	Algorithm	Bias Correction	Resolution/ Smoothing	Footprint Ranking	Availability	Usability	Others
BinDNase	SC	Logistic Regression	No	bp / Sliding Window	Probability	+	–	Require TF ChIP-seq for Training
Boyle	SEG	HMM	No	bp	None	–	–	
Centipede	SC	Bayesian Mix. Model	No	bp	Probability	+	–	Integrates Histone and Sequence Data
Cuellar	SC	Weighted Motif Match	no	Sliding Window	Sequence-based Score	+	–	
DNase2TF	SEG	Sliding Window	4-mer	bp	p-values	+	+	
FLR	SC	Mixture Model	6-mer	bp	Log-Odds	+	–	Bias Correction for Each TF
Neph	SEG	Sliding Window	no	bp	FS	–	–	
PIQ	SEG	GP/Expectation Propagation	No	bp / GP	Probability	+	+	Support Replicates, Time Series
Wellington	SEG	Sliding Window	No	bp	p-value	+	+	

2.5 Discussion

In this chapter we introduced the main concepts within the molecular biology field of gene regulation. Then, we defined the problem we are going to address in this thesis, which is to identify active TFBSs, i.e. DNA regions being bound by regulatory proteins at a particular cell state or condition. We have discussed that sequence-based computational approaches which takes advantage of the protein-DNA sequence binding affinity are not able to identify active sites, since the chromatin dynamics also needs to be considered. Nevertheless, we show that novel open chromatin assays such as DNase-seq and ChIP-seq capture such cell-specific chromatin dynamics. However, the magnitude and complexity of the data generated by these biological experimental assays call for robust computational frameworks. Finally, we discussed a particular type of computational framework for open chromatin data – the computational footprinting methods – which addresses the TFBS identification problem by processing such open chromatin data and searching for patterns that are indicative of active TFBSs. We performed a comprehensive literature review on the main computational footprinting methods and discussed the current challenges on this field.

In this thesis we investigate computational footprinting methods in detail. Among our goals are:

- The development of a computational footprinting method which takes advantage of the full (nucleotide-resolution) grammar of active TFBSs given by the DNase-seq and histone modification ChIP-seq data. Given the experimental flexibility of the footprints obtained using a segmentation-based approach, we are going to develop our method using the segmentation approach.

- The investigation of techniques to process and normalize the DNase-seq and histone modification ChIP-seq signals. Furthermore, we will analyze the correction of the DNase-seq signal for DNase-seq sequence cleavage bias and other experimental artifacts, as these issues were correlated with a decrease in accuracy for other computational footprinting methods.

- The comparison of multiple computational footprinting methods. We will investigate particularities associated to each method. Furthermore, we will analyze the correlation between the accuracy of computational footprinting methods and multiple biological genomic features. Moreover, we will also show the application our computational footprinting framework in real case scenarios.

- The development of an alternative evaluation approach to that using TF ChIP-seq, to avoid interpreting the results only in the light of a single evaluation methodology. Furthermore, an attempt to create a benchmark dataset will be made, in order to standardize method comparison within this field. Such benchmark dataset and the comparative method analyses will be performed on a large compendium comprising many TFs, TF ChIP-seq data and gene expression data.

- The investigation of the extent of the TF residence time's impact on footprint prediction performance. We plan to identify potential problematic TFs using only the input open chromatin data to assist in the biological interpretation of the computational footprinting methods' results.

CHAPTER 3

Methods

In this chapter we describe the computational footprinting framework we devised to address the problem of active transcription factor binding site (TFBS) identification. Here, we exclusively present and formalize our novel computational footprinting approach. Method parameterization and execution details of our and other methods will be made in the next chapter.

Our methodological framework is divided into two main parts:

- In the first part, we discuss the input data processing (Section 3.1). We use data from next-generation sequencing (NGS) open chromatin experiments which gives information regarding the chromatin structure, allowing an accurate search for patterns of active TFBSs. In this first part we describe all steps involved in the generation of DNase-seq and ChIP-seq signals from the aligned reads and treatment of these signals through several steps which aim at reducing bias and normalizing such genomic signals.
- In the second part, we define the novel method to identify footprints (i.e. putative active TFBSs) using the processed signals (Section 3.2). Such method is based on the probabilistic framework of hidden Markov models (HMMs).

Furthermore, we describe some details of the computational implementation of the methods described in this chapter (Section 3.3). Finally, we close this chapter with a few concluding remarks on the methodology choice and novelty of our approach (Section 3.4).

3.1 Input Signal Processing

As mentioned in Section 2.4.4, biological data from open chromatin NGS-based experiments, such DNase-seq and ChIP-seq are affected by biases and noises intrinsic to the experimental protocol. Therefore, a number of data processing steps are performed to assuage these biases and noises and also to prepare the data for our computational footprinting method.

The input data processing pipeline is schematically represented in Figure 3.1. In this thesis, we use DNase-seq and histone modification ChIP-seq data as the input for our computational footprinting framework. With such data we are able to identify patterns of active transcription factor binding by analyzing the regions in the genome which are both open and protected against DNase I digestion (Section 2.3.3). Our method receives as input the DNase-seq and histone modification ChIP-seq aligned reads (Figure 3.1a).

Given the aligned reads, we create genomic signals by counting the overlap between these reads at every genomic coordinate (base pair; bp). This step, which is exemplified in Figure 3.1b, is formally shown in Section 3.1.1. We refer to this signal as read overlap (raw) genomic signal. The read overlap DNase-seq genomic signal is affected by the DNase-seq sequence cleavage bias. Such bias stems from the fact that the DNase I enzyme prefers to bind to, and cleave, certain genomic sequences. Given that, we perform a DNase-seq sequence cleavage bias correction. The DNase-seq sequence cleavage bias correction, exemplified in Figure 3.1c, is thoroughly defined in Section 3.1.2. This step performs local corrections in the DNase-seq signal while preserving signal scale, magnitude and shape.

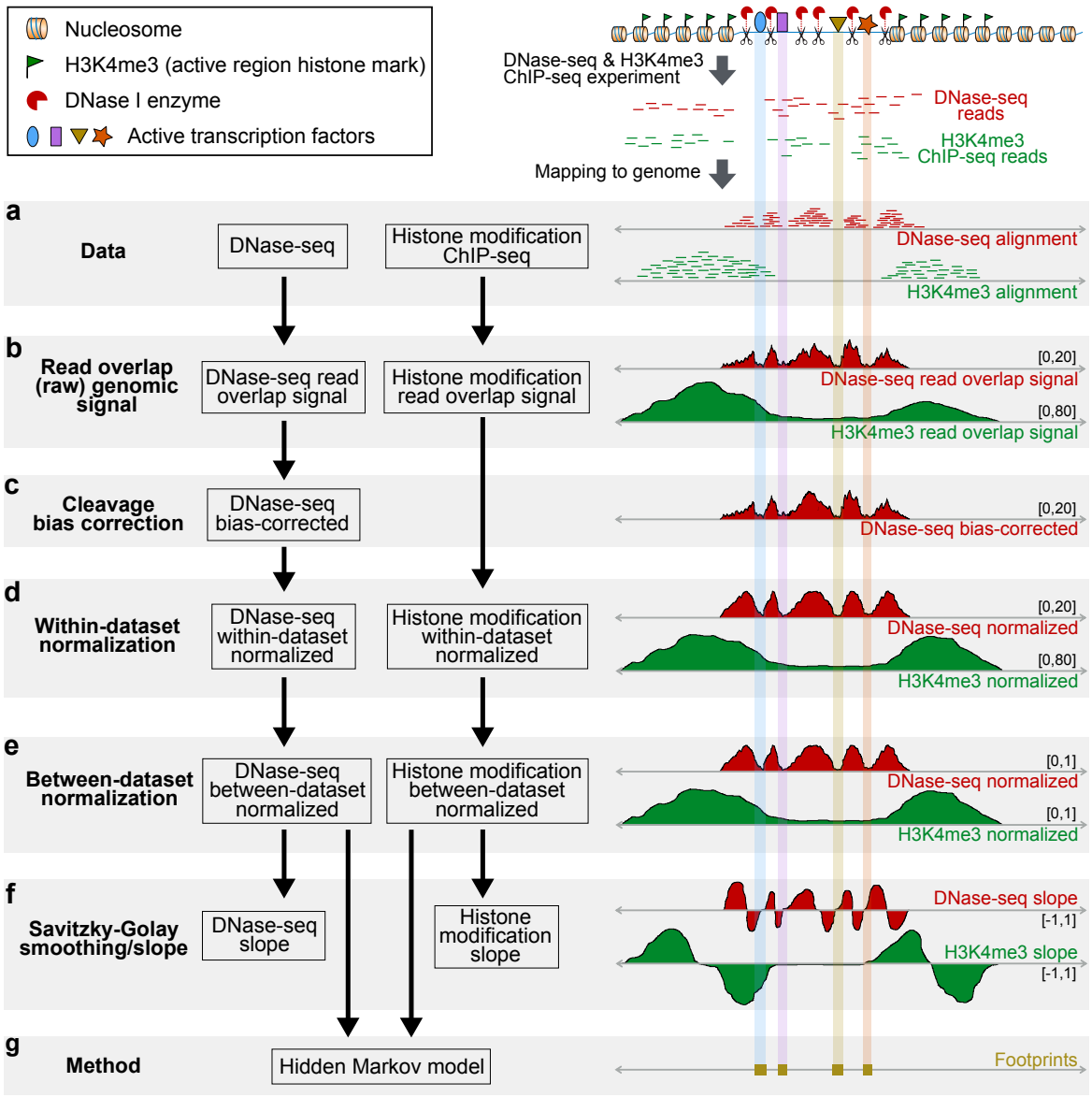

Figure 3.1: Overview of input signal processing framework. This figure provides a schematic representation of the input signal processing pipeline (left panels) and visual examples of the effects of these processing steps (right panels). The interval in each genomic signal represents the data scale. (**a**) We obtained DNase-seq (always represented in red) and histone modification data (always represented in green) as aligned reads. (**b**) A read overlap (raw) signal is generated by counting the number of overlapping aligned reads. (**c**) The DNase-seq signal needs to be corrected for sequence cleavage bias. (**d**) Then, both read overlap histone modification ChIP-seq and sequence cleavage bias-corrected DNase-seq signals are normalized using a within-dataset approach. This decreases the magnitude differences between the signal peaks within each dataset, while preserving the scale of the data. (**e**) Afterwards, both signals undergo a between-dataset normalization procedure to allow all datasets to be within the same scale of $[0, 1]$. (**f**) The slope of both signals are calculated using the Savitzky-Golay smoothing filter and differentiation methodology. (**g**) The DNase-seq and histone modification normalized and slope signals are the input for our computational footprinting method.

Both bias-corrected DNase-seq and read overlap histone modification ChIP-seq signals have different magnitude throughout the genome, i.e. the height of the peaks within each of these signals vary greatly in distinct genomic regions. In order to normalize these signals while preserving the signal scale and shape, we perform a within-dataset (local) normalization procedure (Section 3.1.3). The effect of such normalization procedure, as seen on Figure 3.1d, is the decrease in peak height variability between different genomic regions. Furthermore, since our goal is to integrate both DNase-seq and histone modification ChIP-seq signal, we perform a between-dataset (global) normalization approach (Section 3.1.4). As seen on Figure 3.1e, such normalization approach brings these two different signals to the same data range, i.e. the data is fit into the interval $[0,1]$, without losing their underlying shape. After such normalization procedures, both signals will present less variability within and between datasets without enhancing background noise and without changes in the shape and duration of the signal's peaks.

Our computational footprinting method uses the increase and decrease of the signals. Therefore, we also apply a method called Savitzky-Golay smoothing filter and differentiation to calculate the slope of the signal based on a certain window of data points (Section 3.1.5). We observe in Figure 3.1f that such slope signal assumes high positive values when there is an increase in the genomic signal and a low negative value when there is a decrease in the genomic signal. The normalized and slope versions of the DNase-seq and histone modification ChIP-seq signals correspond to our computational footprinting method's input, as will be described in Section 3.2.

3.1.1. Read Overlap Signal

NGS experiments, such as DNase-seq and ChIP-seq, provide multiple reads, i.e., short deoxyribonucleic acid (DNA) sequences that are aligned into the genome. Here we formally define the genome as a vector

$$\mathbf{g} = \langle g_1, \cdots, g_n \rangle, \tag{3.1}$$

where n equals the number of base pairs (coordinates) in the genome and each $g_i \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ represents a nucleotide. As described in Section 2.1.1 the DNA has two strands, which we refer to as the forward and reverse strand. Throughout this thesis consider $\mathbf{g}$ as the forward strand. Strand differentiation will be mentioned only when this issue is important. Moreover, the reverse strand can be inferred from the forward strand since each nucleotide pairs with a specific matching nucleotide. We denote as $\mathbf{g}[u..v]$ a substring of $\mathbf{g}$ from the genomic coordinate u to v for all $u \leq v$, including both within the interval. Therefore, $\mathbf{g}[u..v]$ has total length $u - v + 1$.

Furthermore, we refer to the term "genomic region" to denote an interval from a particular genomic coordinate u to another genomic coordinate v. The genomic regions, as the genomic DNA substrings, have both initial (u) and final (v) positions within the interval and $u \leq v$ for all intervals, which have length $u - v + 1$. A "genomic region set" is a collection of genomic regions, which are be represented as $R = \{r_1, \cdots, r_m\}$.

We represent the reads obtained from any open chromatin NGS-based experiment, which are aligned into a genome $\mathbf{g}$, as a genomic regions set. Let $R = \{r_1, \cdots, r_m\}$ be the set of m genomic regions representing the reads from a particular NGS experiment aligned in $\mathbf{g}$. In this case, each $r_i = [u, v, s]$ represents a triple, where u is the coordinate in $\mathbf{g}$ where the aligned read starts, v is the coordinate in $\mathbf{g}$ where the aligned read ends and $s \in \{+, -\}$ corresponds to the DNA strand in which the read was aligned to ($+$ represents the forward strand, while $-$ represents the reverse strand).

With such a set R of genomic regions representing the aligned reads we are able to create a genomic signal $\mathbf{x}$, defined as a vector

$$\mathbf{x} = \langle x_1, \cdots, x_n \rangle, \tag{3.2}$$

by evaluating the overlap between the aligned reads R. Each $x_i \in \mathbb{N}^0$ represents the number of reads in R that overlapped at the genomic position i.

However, as previously discussed in Section 2.2, only the first base pairs of the DNA fragments obtained from the biological experiments are sequenced by NGS techniques. We are interested in evaluating the overlap of different aligned genomic regions (representing the aligned reads) for different biological experiments (DNase-seq and ChIP-seq). Consequently, we first define a mapping function, which maps a particular read interval to a genomic region based on an extension parameter η. Such function is written as

$$f^{\text{ext}}(r_i, \eta) = f^{\text{ext}}([u, v, s], \eta) = \begin{cases} [u, u+\eta] & \text{if } s = \text{+} \\ [v-\eta, v] & \text{else.} \end{cases} \tag{3.3}$$

With the extension function, we are able to define the overlap signal ($\mathbf{x}$) as

$$x_i = \sum_{r_j \in R} \mathbf{1}\left(i \in f^{\text{ext}}(r_j, \eta)\right), \tag{3.4}$$

where $\mathbf{1}(\cdot)$ is an indicator function, which returns 1 if its parameter proposition is true or 0 otherwise.

The extension parameter used for the DNase-seq is $\eta = 1$ bp, since we are interested in the regions in which the DNase I enzyme nicked the DNA, i.e. the start of each read. The extension parameter used for ChIP-seq experiments is $\eta = 200$ bp. Such read size matches the average length of the DNA fragments retrieved during the chromatin immunoprecipitation procedure.

3.1.2. DNase-seq Sequence Cleavage Bias

DNase-seq data was found to be affected by the DNase-seq sequence cleavage bias (He et al., 2014; Meyer and Liu, 2014). This happens because the DNase I enzyme has an intrinsic preference to bind to (and cleave) certain DNA sequences. In this section we describe our approach to estimate the DNase-seq sequence cleavage bias and to correct the DNase-seq signal for such bias.

Estimation of DNase-seq Sequence Cleavage Bias

The estimation of DNase-seq sequence cleavage bias is performed based on DNA sequence words of length k (k-mers). Since we want to capture the DNase-seq sequence cleavage bias within particular regions enriched with DNase I activity, such bias estimation is performed in a set of genomic regions of interest $H = \{h_1, \cdots, h_m\}$. Our approach consists on measuring, within these genomic regions of interest: (1) the observed DNase I cleavage score for a k-mer $\mathbf{w}$, which corresponds to the number of DNase-seq cleavage hits centered on $\mathbf{w}$; and (2) the background DNase-seq cleavage score, which is defined by the total number of times $\mathbf{w}$ occurs. Then, the bias estimation is computed as the ratio between the observed and background cleavage scores. Such estimation is performed for all possible k-mers within the DNA alphabet $\{\text{A}, \text{C}, \text{G}, \text{T}\}$.

The process of estimation and correction of DNase-seq sequence cleavage bias is strand-specific, which means that we will consider the DNA sequences and signal generated separately for each DNA strand. However, for simplicity of notation, we will not explicitly denote strandedness in the equations.

For each possible k-mer $\mathbf{w}$, which is a string of length k constructed with symbols from the DNA alphabet $\{\text{A}, \text{C}, \text{G}, \text{T}\}$, the observed cleavage score $o_{\mathbf{w}}$ is calculated, for a set of genomic regions of interest $H = \{h_1, \cdots, h_m\}$, as

$$o_{\mathbf{w}} = 1 + \sum_{i=1}^{m} \sum_{j \in h_i} x_j \mathbf{1}\left(\mathbf{g}[j - \frac{k}{2} .. j + \frac{k}{2}] = \mathbf{w}\right). \tag{3.5}$$

Similarly, the background cleavage score $h_{\mathbf{w}}$ is calculated as

$$h_{\mathbf{w}} = 1 + \sum_{i=1}^{m} \sum_{j \in h_i} \mathbf{1}\left(\mathbf{g}[j - \frac{k}{2}..j + \frac{k}{2}] = \mathbf{w}\right). \tag{3.6}$$

Finally, the estimated cleavage bias b_i for a genomic position $k+1 \leq i \leq m-k+1$, given that $\mathbf{w} = \mathbf{g}[i - \frac{k}{2}..i + \frac{k}{2}]$, is calculated as

$$b_i = \frac{o_{\mathbf{w}}}{h_{\mathbf{w}}}. \tag{3.7}$$

The estimated genomic bias signal point b_i represents how many times the k-mer sequence $\mathbf{g}[i - \frac{k}{2}..i + \frac{k}{2} + 1]$ was cleaved by the DNase I enzyme in comparison to its total occurrence in the set of regions of interest H.

Correction of DNase-seq Sequence Cleavage Bias

The DNase-seq sequence cleavage bias correction is performed on smoothed versions of both read overlap DNase-seq ($\mathbf{x}$) and bias score $\mathbf{b}$ signals. The rationale is that we want to avoid dramatic signal changes generated within nucleotide-resolution bias signals.

First, we create a smoothed DNase-seq signal $\hat{\mathbf{x}}$ using a 50 bp window, which is written as

$$\hat{x}_i = \frac{x_j}{\sum_{j=i-25}^{i+24} x_j}. \tag{3.8}$$

Then, we create a smoothed bias score signal $\hat{\mathbf{b}}$ using the same 50 bp window as for the smoothed DNase-seq signal, which is denoted as

$$\hat{b}_i = \frac{b_j}{\sum_{j=i-25}^{i+24} b_j}. \tag{3.9}$$

With $\hat{\mathbf{x}}$ and $\hat{\mathbf{b}}$, we are able to calculate a signal of bias-correction factors $\mathbf{c}$ as

$$c_i = \hat{x}_i \hat{b}_i. \tag{3.10}$$

The pre-processed bias-corrected DNase-seq genomic signal ($\hat{\mathbf{x}}^{\text{bc}}$) is obtained by applying

$$\hat{x}_i^{\text{bc}} = \log(x_i + 1) - \log(c_i + 1). \tag{3.11}$$

The pre-processed bias-corrected DNase-seq signal generated by Equation 3.11 may include negative values. Since a few posterior statistical analyses required a signal consisting only of positive values, we have shifted the entire signal by adding the global (genomic) minimum value. The global minimum value ζ in the pre-processed bias-corrected DNase-seq signal is denoted as

$$\zeta = \min_{i=1,\cdots,n} \hat{x}_i^{\text{bc}}. \tag{3.12}$$

The final DNase-seq bias-corrected signal $\mathbf{x}^{\text{bc}}$ is calculated by summing the pre-processed bias-corrected DNase-seq genomic signal ($\hat{\mathbf{x}}^{\text{bc}}$) and the absolute global minimum value (ζ). Such summation is simply defined as

$$x_i^{\text{bc}} = \hat{x}_i^{\text{bc}} + |\zeta|. \tag{3.13}$$

where $|\cdot|$ represents the absolute value of a number.

3.1.3. Within-Dataset Normalization

The next pre-processing step is applied on both DNase-seq sequence bias-corrected signal and read overlap histone modification ChIP-seq signal. From this point further, DNA strand information is not relevant anymore because we disregard the underlying DNA sequence. We will denote both these signals (bias-corrected DNase-seq and read overlap histone modification ChIP-seq) here as $\mathbf{x}$ for simplicity. This procedure is applied separately on each genomic signal. The within-dataset normalization step aims to reduce the intrinsic variability present within DNase-seq or ChIP-seq data. Such variability arise from the multiple biological and computational protocol steps.

First, the genome is partitioned into a set of non-overlapping bins $Y = \{y_1, \cdots, y_m\}$, where each y_l represents the interval $[((l-1)\cdot\iota)+1, l\cdot\iota]$ for a particular interval-length parameter ι. Furthermore, we also create a genome partition of overlapping bins $Z = \{z_1, \cdots, z_m\}$, where each z_l represents the interval y_l extended by $\iota/2$ on both sides.

We are able to create a within-signal normalized signal by dividing the signal by non-zero signal averages (Boyle et al., 2011) inside the proposed bins. For a given genomic signal entry x_i at genomic coordinate i, such that $i \in y_l$, we apply

$$x_i^{\text{norm1}} = \frac{x_i}{\sum\limits_{j \in z_l} x_j \mathbf{1}(x_j > 0) \Big/ \sum\limits_{j \in z_l} \mathbf{1}(x_j > 0)}. \tag{3.14}$$

3.1.4. Between-Dataset Normalization

After the within-dataset normalization, we perform a between-dataset normalization procedure to force values inside the interval $[0,1]$ by fitting the within-dataset normalized signals into a logistic function (Hon et al., 2009).

Let $Y = \{y_1, \cdots, y_m\}$ and $Z = \{z_1, \cdots, z_m\}$ be non-overlapping and overlapping genomic partitions, respectively, as described in Section 3.1.3. For a given genomic signal entry x_i at genomic coordinate i, such that $i \in y_l$, we apply

$$x_i^{\text{norm2}} = \frac{1}{1 + e^{-(x_i^{\text{norm1}} - \varsigma_{z_l}^t)/\sigma_{z_l}}}, \tag{3.15}$$

where $\varsigma_{z_l}^t$ is the t^{th} percentile of the signal data points within the interval z_l and σ_{z_l} is the standard deviation of the signal data points within the interval z_l, given by

$$\sigma_{z_l} = \sqrt{\frac{\sum_{j \in z_l} \left(x_j^{\text{norm1}} - \mu_{z_l}\right)^2}{2\iota}}, \tag{3.16}$$

where μ_{z_l} is the mean of the signal data points within the interval z_l, given by

$$\mu_{z_l} = \sum_{j \in z_l} \frac{x_j^{\text{norm1}}}{2\iota}. \tag{3.17}$$

After the application of the within-dataset and between-dataset normalization procedures (for both DNase-seq and histone modification ChIP-seq), we consider the output as our "normalized" signal. For simplicity, we will denote such signal as $\mathbf{x}^{\text{norm}}$.

3.1.5. Savitzky-Golay Smoothing and Slope

Our computational footprinting method uses an additional signal, which indicates upward and downward trends in the normalized genomic signals. We will use the slope of the normalized signal to

assess such information (Boyle et al., 2011). In order to estimate the slope of the genomic signals we apply a Savitzky-Golay smoothing filter followed by differentiation (Madden, 1978; Luo et al., 2005).

The Savitzky-Golay smoothing filter and differentiation method consists of fitting the data into a polynomial, performing a convolution (based on a specific window length τ) with a vector containing Savitzky-Golay coefficients (Madden, 1978).

Let an odd-number τ be a specific window length in which the smoothing is going to be performed. The Savitzky-Golay convolution is expressed as

$$x_i^{\text{slope}} = \sum_{j=-\hat{\tau}}^{\hat{\tau}} c_{j+\hat{\tau}} x_{i+j}^{\text{norm}}, \tag{3.18}$$

where $\hat{\tau} = \frac{\tau-1}{2}$ and $\mathbf{c}$ is the vector of Savitzky-Golay coefficients.

The derivation of the Savitzky-Golay coefficients is performed using an analytic solution that enables the smoothing and differentiation within the same convolution depicted in Equation 3.18. First, a polynomial will be fitted by linear least squares to a set of τ adjacent data points. These are the same data points within the window of the convolution represented in Equation 3.18. Then, let $\mathbf{z}$ be a variable which represents the index of the equally-spaced convolution, i.e. $\mathbf{z} = \{z_1, z_2, \cdots, z_\tau\} = \{-\hat{\tau}, \cdots, 0, \cdots, \hat{\tau}\}$. The fitted polynomial of degree τ is described as

$$Y = c_0 + c_1 z + c_2 z^2 + c_\tau z^\tau. \tag{3.19}$$

The coefficients c_i are obtained by solving the linear square's normal equations

$$\mathbf{c} = (\mathbf{J}^\intercal \mathbf{J})^{-1} \mathbf{J}^\intercal \hat{\mathbf{x}}, \tag{3.20}$$

where $\hat{\mathbf{x}}$ is the vector of signals within the current convolution window of length τ (Equation 3.18), and the i^{th} row of the Jacobian matrix $\mathbf{J}$, denoted as

$$\frac{\partial \hat{\mathbf{x}}}{\partial \mathbf{c}}, \tag{3.21}$$

has values $\langle 1, z_i, z_i^2, \cdots, z_i^\tau \rangle$.

For the full linear squares derivation of the Savitzky-Golay coefficients and more details on the effects of such smoothing filter, please refer to Luo et al. (2005).

3.2 Computational Footprinting with Hidden Markov Models

In order to detect footprints in genomic signals of the DNase-seq and histone modification ChIP-seq experiments we need a technique which is able to segment the genome from a multidimensional input. The grammar of active TFBSs shows a clear sequential pattern with regard to the intensities of the DNase-seq and histone modification signals. However, the length of each of these pattern's segment, i.e. length of the background regions (with no detectable signals), the length of histone modification or DNase-seq peaks and the duration of the footprints are diverse. Furthermore, segmented regions might present a similar level of the signals. For instance, both the background genomic regions and footprint genomic regions, which should definitely be separated by the computational footprinting segmentation method, have the same signal landscape, i.e. low (close to zero) signals of both DNase-seq and histone modification ChIP-seq signals. The difference between these regions is that the footprint happen within two peaks of DNase-seq signals, which happen within two peaks of active histone modification signals. Given these remarks, an obvious choice for such a segmentation task

are hidden Markov models (HMMs). HMM is a computational technique based on Markov stochastic processes. Such computational model is defined thoroughly in Section 3.2.1.

After defining the HMMs, we proceed to discuss how this probabilistic model is used to segment the genome in the context of the identification of active TFBSs. Figure 3.2 shows a schematic pipeline of our computational footprinting framework using HMMs. First, we define a number of different model topologies based on the grammar of active TFBSs and on remarks made by recent studies on the heterogeneity of such grammar (Section 3.2.2). The different HMM topologies take different input signals, which can be normalized and/or slope versions of the DNase-seq (Figure 3.2a) and histone modification ChIP-seq (Figure 3.2b) signals. The HMM topologies used in this thesis are stated in Figure 3.2c. Next, we define how the model is trained in a supervised manner, using annotation of known TFBSs and a maximum-likelihood probability approach (Section 3.2.3; Figure 3.2d). Finally, the DNase-seq and histone modification ChIP-seq data are used as input for the trained HMMs to make predictions of active TFBSs. To accomplish such a task, we use the Viterbi algorithm (Section 3.2.4; Figure 3.2e).

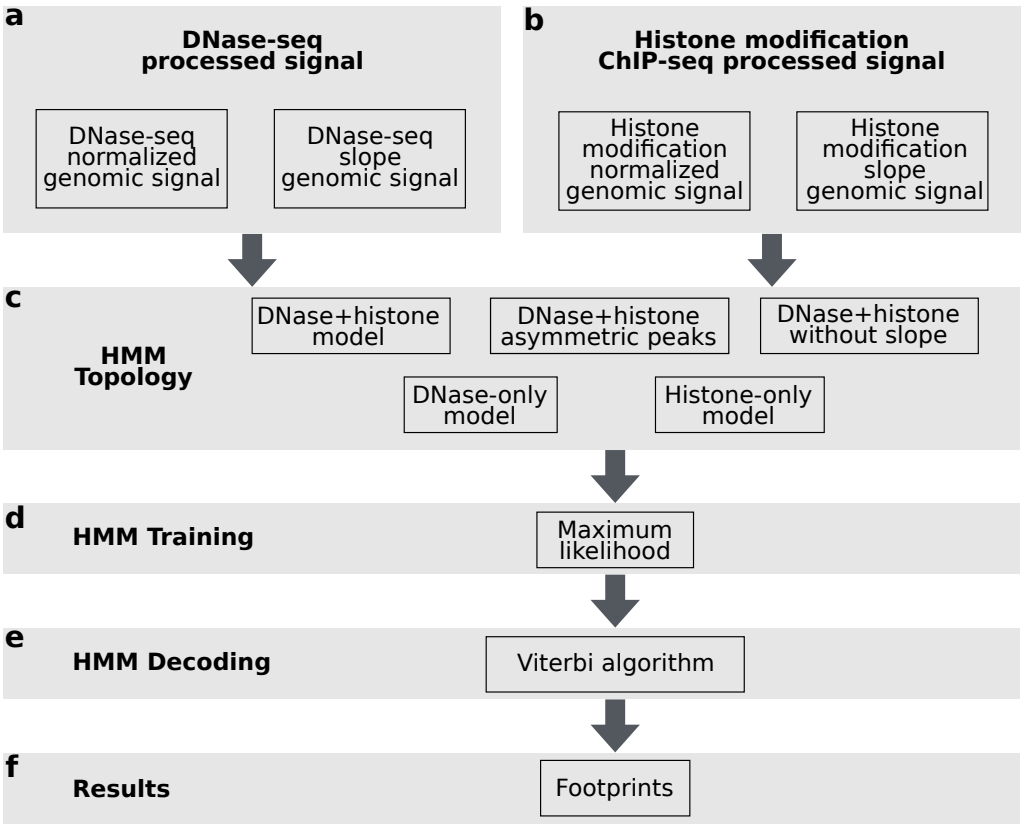

Figure 3.2: Computational footprinting framework. Graphical representation of the computational footprinting method pipeline. (**a**,**b**) Our computational footprinting method receives as input normalized and/or slope signals of DNase-seq and/or histone modification ChIP-seq. (**c**) Different HMM topologies were used. Such HMM topologies take different types of input data. (**d**) All HMMs are trained using the supervised maximum likelihood method. (**e**) We use the Viterbi algorithm to apply the HMM in the genomic signal and predict footprints. (**f**) The final footprints represent our predictions of putative active TFBSs.

3.2.1. Multivariate Continuous HMM

Markov chains are probabilistic models composed of a collection of states and transitions between these states (Rabiner, 1989). These transitions correspond to the probability of changing between states. The HMMs follow the same baseline idea, however they also contain within their model an unknown sequence of states associated to each input symbol (Rabiner, 1989; Durbin et al., 1998). In this section we formalize the concept of HMMs.

First, we define the input data for our HMM as a matrix

$$\mathbf{X} = \{x_{ij}\}^{d \times n} \tag{3.22}$$

of d observed multivariate continuous genomic signals, each of which has length n. For a given multivariate observation $\langle \mathbf{x_{\cdot 1}}, \cdots, \mathbf{x_{\cdot t}}, \cdots, \mathbf{x_{\cdot n}} \rangle$ from $\mathbf{X}$, we have a corresponding hidden sequence path $\mathbf{q} = \langle q_1, \cdots, q_t, \cdots, q_n \rangle$, where $q_t \in W = \{1, \cdots, w\}$ represents the state emitting the vector $\mathbf{x}_{\cdot t}$ at the t^{th} genomic position and w is the total number of states given a particular HMM topology.

HMMs have two independence assumptions (Rabiner, 1989). The first assumption is that the probability to reach state t depends only on the previous state $t-1$

$$p(q_t | q_1, \cdots, q_{t-1}) = p(q_t | q_{t-1}), \tag{3.23}$$

and the second assumption dictates that the probability density function of emitting an input vector $\mathbf{x_{\cdot t}}$ observed at state t, depends only on this current state

$$p(\mathbf{x_{\cdot t}} | q_1, \cdots, q_t) = p(\mathbf{x_{\cdot t}} | q_t). \tag{3.24}$$

Given the formalism previously defined, there are three general problems which can be addressed directly through computationally efficient implementations of HMMs (Durbin et al., 1998). Let Θ be the parameters of a HMM, we state these problems as:

Problem 1 Estimate the HMM parameters Θ in order to maximize $p(\mathbf{X}|\Theta)$.

Problem 2 Given an observed multivariate input $\mathbf{X}$ and an HMM represented by the parameters Θ, find the sequence of hidden states $\mathbf{q}$ which best explains the input given the HMM, i.e. that maximizes $p(\mathbf{X}, \mathbf{q}|\Theta)$.

Problem 3 Given an observed multivariate input $\mathbf{X}$ and an HMM represented by the parameters Θ, compute the probability of the input sequence given the HMM $p(\mathbf{X}|\Theta)$.

The first problem regards the HMM parameter estimation, i.e. model training. This problem will be addressed in Section 3.2.3. The second and third problems represent our genomic segmentation methodology using the HMM states in order to predict active binding sites. These problems will be explored in Section 3.2.4. For a more thorough discussion, including proof of theorems, we refer to Rabiner (1989); Durbin et al. (1998); Mitchell (1997); Bishop (2006); Duda et al. (2000).

The multivariate continuous HMM used to address the aforementioned problems is defined, in terms of its parameters, as

$$\Theta = \{\mathbf{A}, \mathbf{E}, \mathbf{s}\}. \tag{3.25}$$

The parameter $\mathbf{A}$ represents the matrix which contains the probabilities of transitioning between the states of the HMM. We formalize this as

$$\mathbf{A} = \{a_{uv}\}^{w \times w}, \tag{3.26}$$

where a_{uv} represents the probability of transition from state u to v, which is

$$a_{uv} = p(q_t = v | q_{t-1} = u). \tag{3.27}$$

The parameter $\mathbf{E}$ represents the vector of probability density functions which represent the emissions of symbols by the HMM. More formally,

$$\mathbf{E} = \langle e_1(\mathbf{x}), \cdots, e_w(\mathbf{x}) \rangle, \tag{3.28}$$

where each state u has a probability $e_u(\mathbf{x})$ of emitting the vector symbol $\mathbf{x}$. Such probability density function is represented by

$$e_u(\mathbf{x}) = p(\mathbf{x}_{\mathbf{t}} | q_t = u). \tag{3.29}$$

The parameter $\mathbf{s}$ represents the initial state transition probabilities. This is represented as a vector of probabilities

$$\mathbf{s} = \langle s_1, \cdots, s_w \rangle, \tag{3.30}$$

where each s_i represents the probability of starting in a particular state.

3.2.2. HMM Topology

We refer to HMM topology as the number of states w and the predefined possible transitions between these states ($a_{uv} > 0$). The mathematical modeling of a problem with HMMs require the knowledge of the problem in order to be able to create a meaningful HMM topology. We implemented a number of different HMM topologies, depicted in Figures 3.3–3.7. It is important to mention that all HMM states from all topologies have transitions to itself, which were omitted in all figures for simplicity. In this section we will define these topologies and discuss the rationale behind each topology choice. To enhance clarity, the HMM states will also be represented with labels using the `UPPERCASE COURIER` font. Also, the HMM topologies' names will be represented using the SMALL CAPITALS font.

DNASE + HISTONE MODEL

The DNASE + HISTONE MODEL (Figure 3.3) represents the main topology from our method. It combines both DNase-seq and histone modification ChIP-seq in an HMM structure devised to recognize the grammar of active TFBSs described in Section 2.3.3. The idea behind this topology is that we are going to model the depletion between two peaks of DNase-seq using DNase-specific states and we model the open chromatin region within the depletion between two peaks of histone modification ChIP-seq using histone-specific states. We shall refer to this as the ORIGINAL DNASE + HISTONE MODEL.

In this topology, the input matrix $\mathbf{X}$ (Equation 3.22) consists on the normalized and slope versions of the DNase-seq and histone modification ChIP-seq signals. This input matrix can be represented as a vector of input signal vectors

$$\mathbf{X} = \langle \quad \mathbf{x}^{\text{norm}}_{\text{dnase}}, \quad \mathbf{x}^{\text{slope}}_{\text{dnase}}, \quad \mathbf{x}^{\text{norm}}_{\text{histone}}, \quad \mathbf{x}^{\text{slope}}_{\text{histone}} \quad \rangle. \tag{3.31}$$

The probability density function used for the emission probabilities ($\mathbf{E}$) correspond to a multivariate Gaussian (normal) density function with full covariance matrix. This is described as

$$\begin{aligned} p(\mathbf{x}_{\mathbf{t}} | q_t = u) &= p(\mathbf{x}_{\mathbf{t}} | \boldsymbol{\mu}^u, \boldsymbol{\Sigma}^u) \\ &= \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}^u|}} e^{-\frac{1}{2}(\mathbf{x}_{\mathbf{t}} - \boldsymbol{\mu}^u)^T (\boldsymbol{\Sigma}^u)^{-1} (\mathbf{x}_{\mathbf{t}} - \boldsymbol{\mu}^u)}, \end{aligned} \tag{3.32}$$

where $\boldsymbol{\mu}^u$ and $\boldsymbol{\Sigma}^u$ are, respectively, the d-dimensional mean vector and full covariance matrix of the

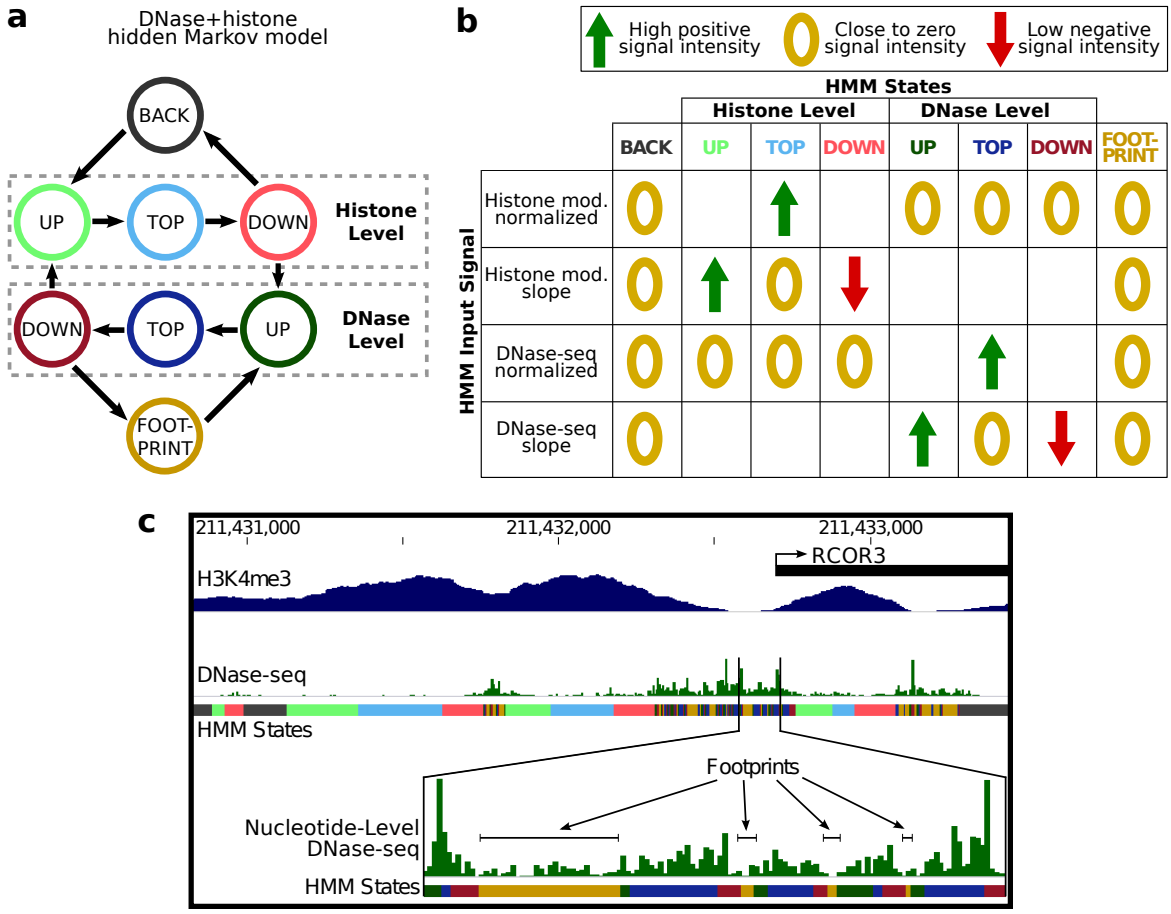

Figure 3.3: DNase + Histone model HMM topology and genomic segmentation. (**a**) DNase + Histone model HMM topology. Each circle represents a labeled HMM state. Each arrow represents an allowed transition between states. Self-transitions exist in all states and were omitted for simplicity. (**b**) Summary of the normalized and slope versions of the DNase-seq and histone modification signals' intensities at each state of the DNase + Histone model. The blank cells within this table correspond to variable signal intensity between different input data and, although important for the final HMM decoding, are not prerequisite for the HMM's recognition of the grammar of TFBSs. (**c**) DNase-seq and H3K4me3 (ChIP-seq) signals around the promoter region of the RCOR3 gene. This region was annotated using the DNase + Histone model. The color code of the annotation matches the color code of the DNase + Histone model representation. We are able to observe several putative footprint predictions of varied sizes. *Source: Gusmao et al. (2014)* (modified to fit thesis format and/or clarify key points).

emission probability density function at state u.

Figure 3.3a shows a graphical representation of the DNase + Histone model. The first state (BACK) corresponds to the "background" regions with low concentration of DNase-seq and histone modification ChIP-seq signals. The histone level states represent a peak in the histone modification ChIP-seq signal, recognizing an increase in the histone modification ChIP-seq signal based on high positive $x^{\text{slope}}_{\text{histone}}$ values (UP), summit regions with $x^{\text{slope}}_{\text{histone}}$ values close to zero and high $x^{\text{norm}}_{\text{histone}}$ values (TOP) and a decrease based on negative values of the $x^{\text{slope}}_{\text{histone}}$ signal (DOWN). From the histone level DOWN state, the model can either return to BACK (isolated histone modification peaks without further DNase hypersensitivity sites) or continue to the DNase level UP state. The DNase level states are equivalent to the histone level states, with the exception that the $x^{\text{norm}}_{\text{dnase}}$ and $x^{\text{slope}}_{\text{dnase}}$ signals are being recognized instead. From the DNase level DOWN state, the model decides between returning to a region

of higher histone modification ChIP-seq signals (histone level UP state) and visiting the FOOTPRINT state, which represents the dip between two peaks of intense DNase I cleavage. The regions of the genome where the HMM has recognized as FOOTPRINT are the ones reported by our method as the predicted footprints.

In Figure 3.3b we provide a full representation of the signal intensity levels observed in each state. This diagram summarizes the aforementioned discussion of the observed signal intensities at each model state. The different input signals have a clear sequential pattern when considered in combination. Such pattern is captured by our HMM's emission distribution full covariance matrix ($\mathbf{\Sigma}^u \; \forall \; u \in W$).

Figure 3.3c shows an example of a genomic region annotated by the DNASE + HISTONE MODEL. In this example, we are able to visualize the difference in resolution between the histone modification ChIP-seq and DNase-seq signals. The HMM is able to segment the genome and capture these resolution differences. This can be seen by the different time intervals in which the HMM remained at each particular state between DNase level and histone level states.

DNASE + HISTONE ASYMMETRIC PEAKS MODEL

The DNASE + HISTONE ASYMMETRIC PEAKS MODEL (Figure 3.4) is an extension of the ORIGINAL DNASE + HISTONE MODEL to account for the histone modification signal asymmetry, i.e. the fact that some open chromatin regions have very low signals of active histone modifications on either its downstream (left peak of the grammar of active TFBSs) or upstream (right peak of the grammar of active TFBSs) regions (Kundaje et al., 2012). For such, two additional transitions were added (shown in red in Figure 3.4) in order to allow the DNase level states to be visited when there are no histone modification peaks before or after DNase hypersensitivity sites.

In this topology, the input matrix $\mathbf{X}$ is the same as depicted for the ORIGINAL DNASE + HISTONE HMM (Equation 3.31), i.e. the normalized and slope versions of the DNase-seq and histone modification ChIP-seq signals.

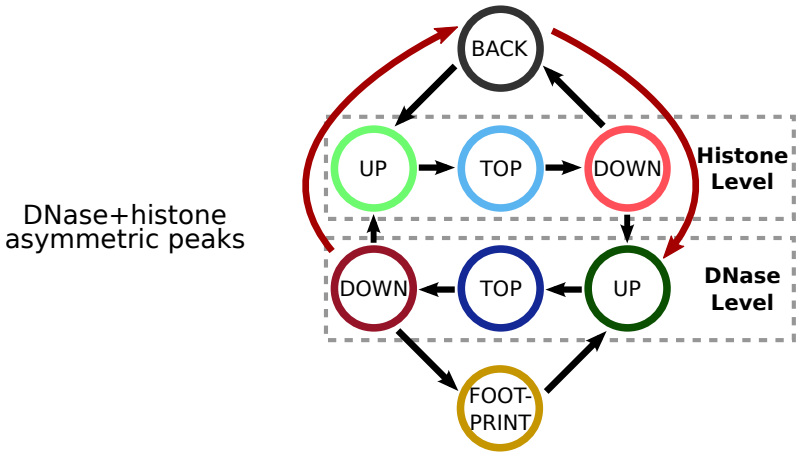

Figure 3.4: DNASE + HISTONE ASYMMETRIC PEAKS MODEL topology. Each circle represents a labeled HMM state. Each arrow represents an allowed transition between states. The red arrows represent the transitions added from the ORIGINAL DNASE + HISTONE MODEL. Self-transitions exist in all states and were omitted for simplicity. *Source: Gusmao et al. (2014)* (modified to fit thesis format and/or clarify key points).

DNase + Histone Without Slope Model

The DNase + histone without slope model (Figure 3.5) is a simplification of the original DNase + Histone model. The simplification consists on removing the slope signals and performing footprint predictions using only the normalized data. In the DNase + histone without slope model, the `UP`, `TOP` and `DOWN` states from the original DNase + histone HMM are compressed into one state – `HIGH` – which recognizes high levels of DNase-seq signal (DNase level state) or high levels of histone modification signal (histone level state).

In this topology, the HMM needs only the normalized signal and becomes bivariate (DNase-seq and histone modifications normalized signals). The input matrix $\mathbf{X}$ can be represented as a vector of input signal vectors

$$\mathbf{X} = \langle \quad \mathbf{x}^{\text{norm}}_{\text{dnase}}, \quad \mathbf{x}^{\text{norm}}_{\text{histone}} \quad \rangle. \tag{3.33}$$

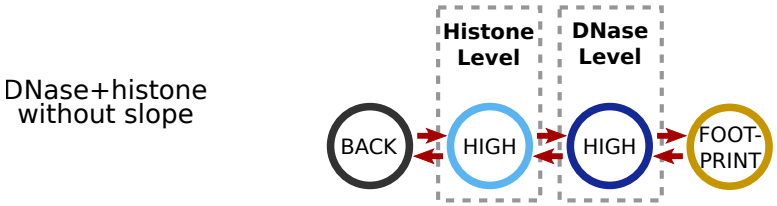

Figure 3.5: DNase + histone without slope model topology. Each circle represents a labeled HMM state. Each arrow represents an allowed transition between states. The red arrows represent the transitions added from the original DNase + Histone model. Self-transitions exist in all states and were omitted for simplicity. *Source: Gusmao et al. (2014)* (modified to fit thesis format and/or clarify key points).

DNase-only Model

The DNase-only model (Figure 3.6) represents an alternative model to the DNase + Histone models. Such model uses only DNase-seq signal and the following modifications were performed in comparison to the original DNase + Histone HMM. The histone level states were removed and additional transitions were added: (1) from the DNase level `DOWN` state to the `BACK` state and (2) from the `BACK` state to the DNase level `UP` state.

In this topology, the input matrix $\mathbf{X}$ can be represented as a vector of DNase-seq input signal vectors

$$\mathbf{X} = \langle \quad \mathbf{x}^{\text{norm}}_{\text{dnase}}, \quad \mathbf{x}^{\text{slope}}_{\text{dnase}} \quad \rangle. \tag{3.34}$$

Histone-only Model

The histone-only model (Figure 3.7) represents an alternative model to the DNase + Histone models. Such model uses only histone modification ChIP-seq signal. The changes in comparison to the original DNase + histone HMM are exactly the same as for the DNase-only model; however, instead of removing the histone level states, the DNase level states are removed and additional transitions are created: (1) from the histone level `DOWN` state to the `FOOTPRINT` state and (2) from the `FOOTPRINT` state to the histone level `UP` state.

In this topology, the input matrix $\mathbf{X}$ can be represented as a vector of histone modification ChIP-seq input signal vectors

$$\mathbf{X} = \langle \quad \mathbf{x}^{\text{norm}}_{\text{histone}}, \quad \mathbf{x}^{\text{slope}}_{\text{histone}} \quad \rangle. \tag{3.35}$$

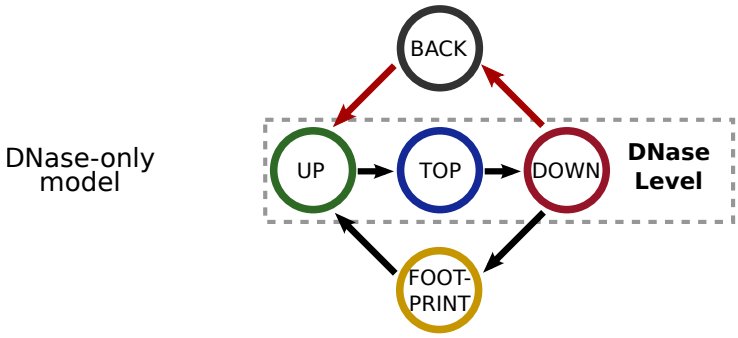

Figure 3.6: DNASE-ONLY MODEL topology. Each circle represents a labeled HMM state. Each arrow represents an allowed transition between states. The red arrows represent the transitions added from the ORIGINAL DNASE + HISTONE MODEL. Self-transitions exist in all states and were omitted for simplicity. *Source: Gusmao et al. (2014)* (modified to fit thesis format and/or clarify key points).

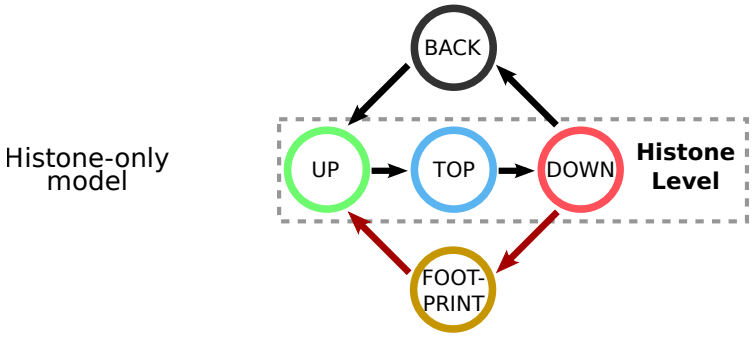

Figure 3.7: HISTONE-ONLY MODEL topology. Each circle represents a labeled HMM state. Each arrow represents an allowed transition between states. The red arrows represent the transitions added from the ORIGINAL DNASE + HISTONE MODEL. Self-transitions exist in all states and were omitted for simplicity. *Source: Gusmao et al. (2014)* (modified to fit thesis format and/or clarify key points).

Unlike all aforementioned HMM topologies, the HISTONE-ONLY MODEL generates broad footprint predictions. This stems from the fact that the histone modification ChIP-seq signals have a lower resolution than the DNase-seq signals. The HISTONE-ONLY MODEL footprint predictions resemble, in terms of length, DNase hypersensitivity sites. The resolution difference between footprints predicted by models that use DNase-seq and models that use histone modification ChIP-seq only can be visualized in Figure 2.12.

3.2.3. HMM Training

We estimate the HMM parameters in a supervised manner, using the maximum likelihood method. For a given annotation sequence of the HMM states $\mathbf{q} = \langle q_1, \cdots, q_t, \cdots, q_n \rangle$ and sample data $\mathbf{X}$, the transition probabilities are estimated as

$$a_{uv} = \frac{\hat{a}_{uv}}{\sum_{j=1}^{w} \hat{a}_{uj}}, \tag{3.36}$$

where $\hat{a}_{uv}$ represents the number of transitions from state u to state v observed in the annotated training data, formally defined as

$$\hat{a}_{uv} = \sum_{i=1}^{n-1} \mathbf{1}(q_i = u, q_{i+1} = v). \tag{3.37}$$

To calculate the emission probability density functions we need to estimate the Gaussian's mean (μ_i^u) and covariance matrix (σ_{ik}^u) for every state u and input signal types i and k. This is performed, using the maximum likelihood method, as

$$\mu_i^u = \frac{\sum_{j=1}^{n} x_{ij} \mathbf{1}(q_j = u)}{\sum_{j=1}^{n} \mathbf{1}(q_j = u)}, \tag{3.38}$$

where μ_i^u is the Gaussian's mean at state u for the signal i and

$$\sigma_{ik}^u = \frac{\sum_{j=1}^{n} (x_{ij} - \mu_i^u)^T (x_{kj} - \mu_k^u) \mathbf{1}(q_j = u)}{\sum_{j=1}^{n} \mathbf{1}(q_j = u) - 1}. \tag{3.39}$$

where σ_{ik}^u is the Gaussian's variance at state u between signals i and k.

As we define the HMM to start at the `BACK` state (the first HMM state), the initial transition vector $\mathbf{s}$ was manually set with the following probabilities

$$\begin{array}{l} s_1 = 1 \\ s_t = 0 \quad \forall \quad t \neq 1 \end{array}. \tag{3.40}$$

3.2.4. HMM Decoding

Given HMMs with topologies described in Section 3.2.2 and parameters estimated as described in Section 3.2.3 we are able to perform the prediction of footprints. This prediction is performed using a well-known HMM decoding technique termed Viterbi "algorithm" (Rabiner, 1989), which addresses the **Problem 2** defined in Section 3.2.1. Briefly, it computes the sequence of hidden states $\mathbf{q}$ that maximizes $p(\mathbf{X}, \mathbf{q}|\Theta)$. Then, given the computed sequence of hidden states we are able to identify the ones which corresponds to the `FOOTPRINT` state. In this section we formalize the Viterbi algorithm applied in the context of computational footprinting.

We are interested on identifying the most probable path $\mathbf{q}^*$ given the input $\mathbf{X}$ on an HMM Θ. In formal terms, we are interested in evaluating the following equation

$$\mathbf{q}^* = \arg\max_{\mathbf{q}} p(\mathbf{X}, \mathbf{q}|\Theta). \tag{3.41}$$

The solution to the equation 3.41 can be found in an exaustive way by evaluating $p(\mathbf{X}, \mathbf{q}|\Theta)$ for all w^n possible instances of the n-length vector $\mathbf{q}$, in which each element assumes one of the w HMM states. It is clear, however, that the complexity of such approach, in terms of the big-$\mathcal{O}$ notation is $\mathcal{O}(w^n)$, i.e. it grows exponentially given the input vector with length n. Fortunately, it is possible to solve the equation 3.41 using a dynamic programming algorithm which relies on the HMM independence claims described by equations 3.23 and 3.24 with a polynomial complexity $\mathcal{O}(n \times w^2)$ using the Viterbi algorithm. The Viterbi algorithm is formalized, in the context of our multivariate HMM, in the following.

Let $v_u(t)$ be a Viterbi variable, which corresponds to the probability of the most probable path of the input subset $\langle \mathbf{x}_{\cdot\mathbf{1}}, \cdots, \mathbf{x}_{\cdot\mathbf{t}} \rangle$ ending at state u. Assuming knowledge of $v_u(t)$, we are able to calculate the probability for the path subset $\langle \mathbf{x}_{\cdot\mathbf{1}}, \cdots, \mathbf{x}_{\cdot\mathbf{t+1}} \rangle$ using the HMM independence claims as

$$v_v(t+1) = e_v(\mathbf{x}_{\cdot\mathbf{t+1}}) \max_u v_u(t) a_{uv} \tag{3.42}$$

Let out HMM decoding start at a figurative initial time 0. We define the initial Viterbi variables for all HMM states as our initial HMM probabilities (equation 3.30) as

$$v_u(0) = s_u. \tag{3.43}$$

From this initial time we are able to calculate the Viterbi variables for all the following input time points using equation 3.42. Furthermore, we dynamically construct a "pointer" vector $\boldsymbol{\phi}$ in which we add the most probable states in each iteration of Viterbi variable calculations for the following input time points. The algorithm is fully described as follows. In the following algorithm we denote as ε an additional figurative last state of our path $\mathbf{q}$ in order to formally describe the algorithm termination.

Viterbi Algorithm

1. Initialization:
  1.1. $\nu_u(0) = s_u$
2. Iteration $(t = 1, \cdots, n)$**:**
  2.1. $\nu_v(t) = e_v(\mathbf{x}_\mathbf{t}) \max_u \nu_u(t-1) a_{uv}$
  2.2. $\phi_v(t) = \arg\max_u \nu_u(t-1) a_{uv}$
3. Termination:
  3.1. $p(\mathbf{X}, \mathbf{q}^*) = \max_u \nu_u(n) a_{u\varepsilon}$
  3.2. $q_n^* = \arg\max_u \nu_u(n) a_{u\varepsilon}$
4. Reassembly $(t = n, \cdots, 1)$**:**
  4.1. $q_{t-1}^* = \phi_{q_t^*}(t)$

The footprint predictions are defined as the set of genomic intervals F in which contiguous predicted hidden states $q_t = \texttt{FOOTPRINT}$. This can be written as

$$F = \{f_i = [m,n] : q_t = \texttt{FOOTPRINT} \quad \forall \quad m \leq t \leq n \quad \text{and} \quad q_{m-1}, q_{n+1} \neq \texttt{FOOTPRINT}\}. \quad (3.44)$$

3.3 Implementation

We implemented our signal processing methodology and our HMM-based computational footprinting framework as a Python command line tool. Our method is called HINT – HMM-based Identification of TF Footprints – and will be referenced as such throughout this thesis. Such command line tool implements all the steps described in this chapter.

HINT is part of the regulatory genomics toolbox (RGT), which is a computational framework composed of a Python package and/or command line tools to handle genomic signals such as DNase-seq and ChIP-seq. The HINT tool was first released in August 2014 and is available under the terms of the GNU General Public License v3 (GPL v3). HINT python package dependencies are summarized in Table 3.1.

The minimal input data required for HINT are BAM files, which is the standard file format for aligned reads for either DNase-seq or histone modification ChIP-seq. Additionally, the user may input a reference genome in order to perform the DNase-seq sequence cleavage bias correction (Section 3.1.2). The tool outputs a BED file, which is the standard format for genomic regions (intervals). Such output BED file corresponds to the predicted footprints.

HINT was tested on Python 2.7, Numpy 1.4.0, Scipy 0.7, Scikit-learn 0.14, Pysam 0.7.5, HMM-learn 0.0.1. We used a local Linux Ubuntu 15.04 LTS x86 64-bit machine running with 8 Intel Core i7-4770 CPU at 3.40GHz and 32 GB RAM. Furthermore, we ran HINT on an HPC cluster mainly based on Intel Xeon-based 8– to 128–way SMP 64-bit nodes with Scientific Linux release 6.6 (Carbon).

Table 3.1: HINT tool python package dependencies.

Package	Version	Website
Numpy	$\geq$ 1.4.0	`http://www.numpy.org/`
Scipy	$\geq$ 0.7.0	`http://www.scipy.org/`
Scikit-learn	$\geq$ 0.14	`http://scikit-learn.org/`
HMMlearn	$\geq$ 0.1.1	`https://github.com/hmmlearn/hmmlearn/`
Pysam	$\geq$ 0.7.5	`https://github.com/pysam-developers/pysam`

For more information on HINT implementation please see:

`http://www.regulatory-genomics.org/hint/`

3.4 Discussion

In this chapter we described our computational footprinting framework. In the first part (Section 3.1), we process the DNase-seq and histone modification ChIP-seq signals, as summarized in Figure 3.1. In the second part (Section 3.2), we described our HMM-based approach (HINT), as summarized in Figure 3.2. Our computational footprinting framework introduced new concepts to solve the identification of active TFBS problem:

- We created a novel DNase-seq and histone modification ChIP-seq signal processing framework that corrects for DNase-seq sequence cleavage bias and normalizes the signal considering both within- and between-dataset signal variability. Furthermore, we applied the Savitzky-Golay smoothing filter to obtain the slope of the genomic signal.
- We devised HMMs to segment the genome and search for the grammar of active TFBSs as shown in Section 2.3.3. This novel approach is the first to integrate the full spatial profiles of both DNase-seq and histone modification ChIP-seq signal.
- Our model is also flexible enough to consider only DNase-seq or histone modification ChIP-seq data separately. Allowing for experimental flexibility.
- From a methodological perspective, HMMs are a favorable method choice. Window-based segmentation methods (Hesselberth et al., 2009; Neph et al., 2012; Piper et al., 2013) has a high dependency on footprint size definition. They rely on an extensive search using multiple window size extensions. Such methods do not take advantage of the HMM's decoding algorithms, which are able to model the length of the footprints using a probabilistic framework. Furthermore, site-centric approaches (Pique-Regi et al., 2011; Cuellar-Partida et al., 2012; Sherwood et al., 2014; Yardımcı et al., 2014) do require a much higher preparation time, running time, depend highly on the model's parameters and do not necessarily fit our main goal, which is to provide a map of all putative active TFBSs for a particular cell type.

CHAPTER 4

Experiments

The main goal of this chapter is to present the experimental framework used in this thesis. In this chapter we are going to:

- Describe the execution of our novel computational footprinting approach HINT, which was formally presented in the previous chapter.
- Describe the execution of all competing computational footprinting methods.
- Formalize the computational footprinting method evaluation methodologies and metrics.
- Introduce common downstream analyses, which use footprint predictions to make inferences about the regulatory network of a cell.
- State all statistical methods employed in this thesis.

The experimental framework used in this thesis is divided in two parts: the execution and the evaluation of computational footprinting methods. First, we describe the execution of computational footprinting methods (Section 4.1). We provide the detailed experimental workflow of our computational footprinting method HINT. Furthermore, we describe the execution of all competing methods used in this thesis. In the second part, we define the methodology used to evaluate all computational footprinting methods (Section 4.2). We describe two different evaluation strategies: the traditional transcription factor (TF) ChIP-seq approach and the novel strategy we devised, which is based on gene expression.

In addition, we describe two common downstream analyses, which use footprint predictions: the TF enrichment analysis and the *de novo* motif finding. These analyses are described in Section 4.3. Moreover, all statistical methods used on the analyses presented in this thesis are described in Section 4.4. Finally, we close this chapter with a final discussion on our experimental framework (Section 4.5).

4.1 Execution of Computational Footprinting Methods

In this section we describe the experimental framework regarding the execution of our computational footprinting method – HINT – and the competing footprinting methods (Figure 4.1). First, we describe the data used as input for these methods: DNase-seq and histone modification ChIP-seq (Section 4.1.1). Then, we proceed by characterizing HINT's signal processing pipeline (Section 4.1.2) and HINT's execution (Section 4.1.3). Finally, we describe the execution of all competing computational footprinting methods (Section 4.1.4)

4.1.1. Data

Data-seq Data

DNase-seq aligned reads were obtained in ENCODE Project Consortium (2012). We obtained data generated from the single-hit (labeled as "SH") and double-hit (labeled as "DH") DNase-seq pro-

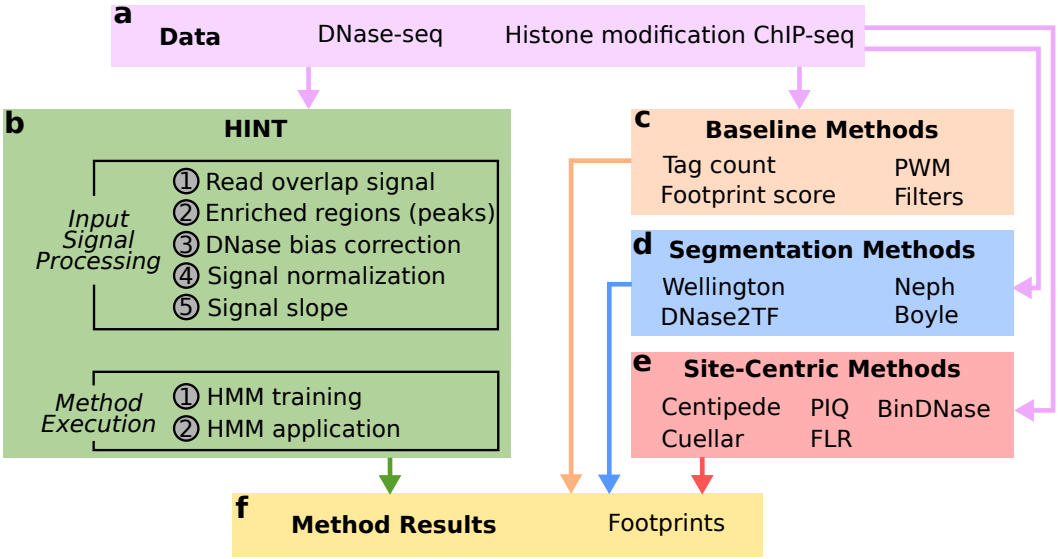

Figure 4.1: Experimental framework of the execution of the computational footprinting methods. (**a**) The computational footprinting methods take as input different combinations of DNase-seq and histone modification ChIP-seq data. (**b**) The computational footprinting method proposed in this thesis – HINT – is divided in: input signal processing and method execution. (**c**–**e**) Competing methods, categorized as (**c**) baseline, (**d**) segmentation and (**e**) site-centric, are executed to perform a comparative analysis with the performance of HINT. (**f**) All computational footprinting methods generate footprint predictions.

tocols. Furthermore, we obtained naked deoxyribonucleic acid (DNA) DNase-seq data (labeled as "NK"). The naked DNA DNase-seq data results from the application of the same DNase-seq protocol as shown for the single-hit and double-hit version (Section 2.2.2); however it is applied to a cell in which all active TFs have been removed from the DNA. A full description on all DNase-seq data used in this thesis can be found at Supplementary Table A.1.

Each DNase-seq dataset correspond to the DNase-seq experiment aligned reads for one cell type. We divided the cell types in which we obtained DNase-seq data into three categories: the `Comparative Dataset`, `Analysis Dataset` and `Full Dataset`. Below, we show the cell types associated to each category:

- `Comparative Dataset` – Data used in the method comparison analysis:
  - Single-hit DNase-seq protocol (SH): H1-hESC, K562 and GM12878.
- `Analysis Dataset` – Data used in the analysis of relevant computational footprinting features:
  - Single-hit DNase-seq protocol (SH): H1-hESC, HeLa-S3, HepG2, HUVEC, K562, GM12878, LNCaP and MCF-7.
  - Double-hit DNase-seq protocol (DH): H7-hESC, HepG2, HUVEC, K562 and m3134.
  - Naked DNA DNase-seq protocol (NK): MCF-7, K562 and IMR90.
- `Full Dataset` – All cell types from the encyclopedia of DNA elements (ENCODE) Tier 1 and Tier 2 experiments (SH, DH and NK) (ENCODE Project Consortium, 2012). These data were used to investigate relevant features regarding the DNase-seq sequence cleavage bias.

Histone Modification ChIP-seq Data

Histone modifications ChIP-seq aligned reads were also obtained in ENCODE Project Consortium (2012). We only obtained histone modification ChIP-seq data from cell types of the `Comparative Dataset`, i.e. H1-hESC, K562 and GM12878. For each cell type, it was obtained data regarding the activating histone modifications: H3K4me1, H3K4me3, H3K9ac, H3K27ac and H2A.Z. Additionally, to perform some analysis of relevant computational footprinting features we obtained data regarding the histone modifications H3K4me1 and H3K4me3 for cell types HeLa-S3 and HepG2. See Supplementary Table A.2 for a full histone modification ChIP-seq data description.

Genomic Information on Experimental Datasets

Both DNase-seq and histone modification ChIP-seq data are based on the human genome build 37 (hg19), except the DNase-seq for m3134 cell type, which is based on mouse genome build 37 (mm9). Chromosome Y was removed from all analyses. It is usual to remove chromosome Y from such types of analyses since it is present in male subjects only; and therefore introduces biases. The genomic sequences (DNA) for the human genome (hg19) and mouse genome (mm9) were also obtained in ENCODE Project Consortium (2012).

4.1.2. HINT Signal Processing

HINT takes as input DNase-seq and/or histone modification ChIP-seq data. Here we discuss the processing steps of such data from the obtained aligned reads to the final signal which is used by HINT to make the footprint predictions. For simplicity, we ignore the fact that signals and intervals are defined on distinct chromosomes.

Read Overlap Signal

We obtained the DNase-seq aligned reads and histone modification ChIP-seq aligned reads in ENCODE Project Consortium (2012) (Section 4.1.1). These datasets are already processed to remove known experimental and computational artifacts (ENCODE Project Consortium, 2012; Derrien et al., 2012; Ashoor et al., 2013; Diaz et al., 2012). We created the read overlap signal from both DNase-seq and histone modification ChIP-seq pre-processed aligned reads as described in Section 3.1.1.

DNase Hypersensitivity Sites and Histone Modification ChIP-seq Peaks

To optimize execution time we constraint the application of our method (HINT) only in genomic regions enriched with either DNase-seq and histone modification ChIP-seq signals. Enriched regions are genomic regions with more reads aligned than expected by chance, given a statistical model. Here, we describe how we obtained the enriched regions for DNase-seq data – called DNase hypersensitivity sites (DHSs) – and the enriched regions for histone modification ChIP-seq data – called histone modification ChIP-seq peaks.

DNase Hypersensitivity Sites (DHSs)

DHSs, i.e. regions enriched with DNase-seq data, are estimated based on the DNase-seq read overlap signal. The process consists on evaluating a smoothed DNase-seq signal and then finding regions with more aligned reads than expected by chance based on a p-value cutoff of 0.01 calculated based on a fitted Gamma distribution. The Gamma distribution was shown to outperform other models for DNase-seq data (Boyle et al., 2008).

For that, we used the F-seq software (Boyle et al., 2008), which was devised specially for DNase-seq data and has shown to provide accurate DHSs (Boyle et al., 2008, 2011). We run the F-seq software version 1.81 with the default parameters, except for the feature length option (set to 300) (Boyle et al., 2011). F-seq source code is found in `http://fureylab.web.unc.edu/software/fseq/`.

Histone Modification ChIP-seq Peaks

The histone modification ChIP-seq peaks were obtained by applying the "model-based analysis for ChIP-seq" (MACS) software (Zhang et al., 2008) version 2.0.9 to the histone modification read overlap signals. Such peak-calling software was devised specially for ChIP-seq data. We executed MACS software with the default parameters. However, we used two extra command options which are indicated in the case of histone modification ChIP-seq (--nomodel --nolambda). The MACS software source code is found at `http://liulab.dfci.harvard.edu/MACS/`.

DNase-seq Sequence Cleavage Bias Correction

The intrinsic DNase-seq sequence cleavage bias was previously shown to affect certain footprinting methods (He et al., 2014). In this study, we are going to explore the DNase-seq sequence cleavage bias correction using two approaches: (1) The "DHS sequence bias" considers the sequence bias estimates within DHSs of each DNase-seq experiment. This approach captures DNase I cleavage, read fragmentation and sequence complexity biases of DHSs of each DNase-seq experiment (He et al., 2014). The "naked DNA sequence bias" considers the sequence bias estimates within naked DNA DNase-seq experiments (Yardımcı et al., 2014). In this case, all DNA regions are open, therefore the sequence bias estimates will mainly capture the DNase I cleavage bias. Both strategies use the formalism described in Section 3.1.2.

In the DHS sequence bias correction approach, we use the same DNase-seq dataset in which the read overlap genomic signal was created, to estimate the DNase-seq sequence cleavage bias. Furthermore, the genomic regions of interest H from Section 3.1.2 correspond to the DHSs. In this case, each dataset (each DNase-seq experiment for a particular cell type) has their own DNase-seq sequence cleavage bias estimations and the correction is made on a cell-specific manner.

In the naked DNA sequence bias correction approach, we use a naked DNA DNase-seq experiment to estimate the sequence cleavage bias. Moreover, the estimation is made on the whole genome, since there are no significantly enriched signals in such dataset. Therefore, the set of genomic regions of interest correspond to a single region which encompasses the whole genome, i.e. $H = \{[1,n]\}$, where n is the total number of base pairs (bp) in the genome. In this case, each naked DNA DNase-seq dataset has their own sequence cleavage bias estimations. The sequence cleavage bias correction for a DNase-seq dataset is made using the naked DNA DNase-seq sequence cleavage bias estimates which correlated best with the DHS sequence bias estimates for that DNase-seq dataset.

Another methodological choice is the size of the k-mer sequence cleavage bias estimates. For both approaches (DHS and naked DNA sequence bias correction) we used 6-mers. As observed by He et al. (2014), a 6-mer bias model captures significantly more information than $k < 6$ models and the information gained with $k > 6$ models are not significant and does not justify the increase in computational complexity (since the number of estimates is exponential).

Signal Normalization

In possession of the DNase-seq sequence cleavage bias corrected signals and read overlap histone modification ChIP-seq signals we proceed to the signal normalization step, which consists on the treatment of these genomic signals to: (1) reduce the within-dataset variability (as described in Sec-

tion 3.1.3) and (2) reduce the variability between these different genomic signals (as described in Section 3.1.4).

Within-dataset Normalization

In the within-dataset normalization step we divide the genome in multiple bins, to estimate and correct the magnitude of the signal peaks, as shown in Section 3.1.3. The length of the bin, denoted by ι was set to $10,000$ bp. The reason for such a choice is that shorter regions would not capture enough signal and lose statistical power. On the other hand larger regions would not achieve the goal of correcting the magnitude of the peaks within the dataset range of signals (Gusmao et al., 2014).

Between-dataset Normalization

In the between-dataset normalization step, we fit the signal into a logistic function to force the values to be within the interval $[0, 1]$. In this step, we estimated the mean μ, standard deviation σ and the percentile ς using data from chromosome 1, which was removed from the evaluation strategy. Furthermore, we used the 98^{th} percentile. Such a choice was made by observing the amount of the genome which is enriched for both DNase-seq and histone modification ChIP-seq, which is, on average ~2% (Gusmao et al., 2014).

Signal Slope

Given the normalized DNase-seq and histone modification ChIP-seq signals, we proceed to calculate the signals' slope. The goal is to create the additional slope signal required by HINT (as described in Section 3.1.5).

In the application of the Savitzky-Golay technique to calculate the signals' slope, we used a 2^{nd}-order polynomial. Furthermore, the odd-valued window size τ for smoothing and estimation of the slope of the signal was set to 9 bp for the DNase-seq signal, as suggested by Boyle et al. (2011). For the histone modification ChIP-seq signal, such parameter was set to 201 bp, as it fits the read extension length (η) considered during the creation of the read overlap signal.

HINT Input Signal

After these processing steps we have four different input signals for our computational footprinting method HINT. Different HINT's hidden Markov model (HMM) topologies use different combinations of these four signals (the description of all topologies is found in Section 3.2.2). The four HINT's input signals are:

1. $\mathbf{x}_{\text{dnase}}^{\text{norm}}$ – the real-valued vector of normalized DNase-seq genomic signals.

2. $\mathbf{x}_{\text{dnase}}^{\text{slope}}$ – the real-valued vector of slope DNase-seq genomic signals.

3. $\mathbf{x}_{\text{histone}}^{\text{norm}}$ – the real-valued vector of normalized histone modification ChIP-seq genomic signals.

4. $\mathbf{x}_{\text{histone}}^{\text{slope}}$ – the real-valued vector of slope histone modification ChIP-seq genomic signals.

Figure 4.2 shows an example of two genomic regions with read overlap (count), normalized and slope signals for two distinct genomic regions. In this figure we are able to observe the effect of the within-dataset normalization strategy by looking at the different ranges of the histone modification base overlap count signal (in black) – [0,40] on the genomic region depicted in Figure 4.2a *vs* [0,80] on the genomic region depicted in Figure 4.2b in comparison to the range between these two regions of the normalized signal [0,1]. The normalization preserves the shape of the peaks and does not attenuate

undesired background noise signals. Furthermore, the effect of the between-dataset normalization approach is more straightforward, as all normalized signals will be within the range [0,1]. Moreover, this figure shows the difference in resolution between the different data sources. While ChIP-seq peaks are smoothed and spans on average 250 bp, the DNase-seq peaks can be as short as 5 bp.

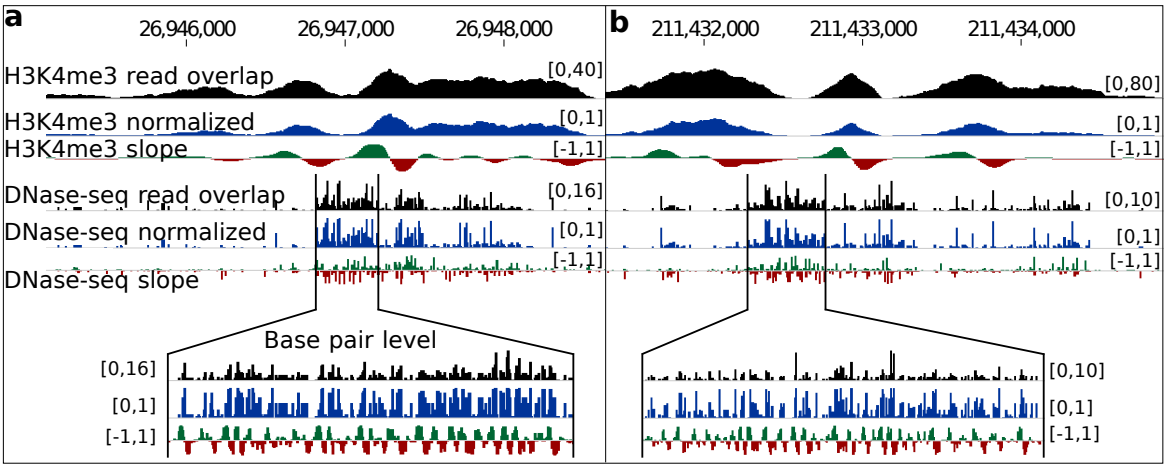

Figure 4.2: Genomic signal processing examples. Examples of histone modification H3K4me3 ChIP-seq and DNase-seq signals before treatment (read overlap; in black), normalized (in blue) and after Savitzky-Goaly smoothing and differentiation (slope; positive values in green and negative values in red). Each signal's range is displayed between square brackets next to each signal. In this figure we show examples for two different regions in human chromosome 1. *Source: Gusmao et al. (2014)* (modified to fit thesis format and/or clarify key points).

4.1.3. HINT Method Execution

Our computational footprinting method HINT segments the genome using normalized and slope versions of the DNase-seq and histone modification ChIP-seq signals. Such segmentation task is performed on the basis of the grammar of active transcription factor binding sites (TFBSs). As shown in Section 3.2.2 we have devised a number of different HMM topologies, which take different combinations of input signals and address particularities of the TFBS patterns on the open chromatin data. Here, we provide experimental details on how HINT is trained and applied on the genome.

HINT Training

We train the HMM models in a supervised manner. Briefly, a manual annotation is created for each cell type, histone modification and HMM topology (Figures 3.3– 3.7) based on the DNase-seq and histone modification ChIP-seq data.

We selected a 10,000 bp region (with genomic coordinates 211,428,000–211,438,000 in human chromosome 1) around the promoter region of the gene RCOR3 and performed a cell-specific manual annotation, in which each genomic position is assigned with a state from our HMM topology. This promoter-proximal regulatory region annotated with HMM states was used to train the models which use histone modifications H3K4me3, H3K9ac, H3K27ac and H2A.Z. As the histone modification H3K4me1 is known to be associated to distal regulatory regions, we have additionally annotated an enhancer region (with genomic coordinates 26,942,000–26,952,000 in human chromosome 1). The selection of these regions was made randomly, but we checked ENCODE Project Consortium (2012) tracks for evidence that the gene RCOR3 was expressed in all cell types analyzed and that the

enhancer region was far ($>$ 100 Kbp) from known genes and expressed regions but associated with the expression of the closest gene's transcription start site (TSS).

In order to help the annotation of the footprints, motif-predicted binding sites (MPBSs) obtained by applying motif matching with all position frequency matrices (PFMs) from Jaspar (Mathelier et al., 2014), Uniprobe (Robasky and Bulyk, 2011) and Transfac (Matys et al., 2006) PFM repositories were detected inside the training regions (the details on the identification of MPBSs can be found in Section 4.2.1). We consider "real" footprints all the DNase-seq signal depleted regions between two DNase-seq peaks that overlap a MPBS. For the HISTONE-ONLY MODEL, we considered as footprints all the DHSs within these regions. We trained five HMMs per cell type, one for each histone modification (H3K4me3, H3K9ac, H3K29ac and H2A.Z with the promoter-proximal regulatory region and H3K4me1 with the distal regulatory region). In the case of the DNASE-ONLY MODEL, only one HMM was trained for each cell type. Such training was performed in the promoter-proximal region. The regions used for training were excluded from all further analyses. In possession of the manually annotated regions for each cell type, histone modification and HMM topology, all HMMs were trained using the maximum-likelihood process described in Section 3.2.3.

Here, we show an example of a complete set of HMM parameters, regarding the ORIGINAL DNASE + HISTONE HMM topology (Figure 3.3) trained with DNase-seq + H3K4me3 using data from the H1-hESC cell. The Table 4.1 represents the transition matrix. Each number represents the probability of performing a transition from the HMM state depicted in the table's first row to the HMM state depicted in the table's first column. In this transition matrix we are able to observe that only the self-transitions and the transitions allowed by our model topology have a non-zero probability. The transition matrix is the structure that directly defines our HMM topology.

The Table 4.2 exhibits the emission distribution mean values. It contains the mean in which each signal type (represented in the columns) assumes at each state (represented in the rows). A closer look into these vectors of means for each state and signal shows the grammar of active TFBS in a numerical form. The states `BACK` and `FOOTPRINT` have low absolute means for all signal types. `UP`, `TOP` and `DOWN` states have, respectively, high positive, close to zero and low negative slope signals. Such emission parameters models the signal magnitude of the active TFBS grammar. We are able to model both magnitude and shape of the signals when we consider the transition probabilities and the mean component of the emission probability distributions.

Moreover, the Table 4.3 shows all covariance matrices from the emission distributions. The full covariance matrix is depicted for each state, in which rows and columns are sorted by the input signals: DNase-seq normalized, DNase-seq slope, H3K4me3 normalized and H3K4me3 slope. The covariance matrix component of the emission probability distributions reflect the relationship between our signals in our multivariate model. For instance, it is interesting to observe a negative value (-0.0053) at the DNase level `UP` state (UP(D) in the table) on the covariance matrix corresponding to the normalized DNase-seq *vs* the normalized histone modification ChIP-seq signal. Such data behavior is in line with the observed grammar of TFBSs, where the DNase-seq signal generally start to increase at the decrease of the histone modification ChIP-seq signal. Finally, when we consider all HMM parameters, we are able to model the magnitude, shape and relationship between the chromatin dynamics signals.

HINT Application

To reduce the dimensionality of the data, we used the DHS and histone modification ChIP-seq peaks (Section 4.1.2). In the DNASE + HISTONE HMM topologies, we have extended these enriched regions by $5,000$ bp on each side and merged the resulting regions. We apply the trained HMM models in these extended and merged regions. In the DNASE-ONLY MODEL and HISTONE-ONLY MODEL the extension process is the same, however, using only the DHSs and histone modification ChIP-seq peaks, respectively.

Table 4.1: Example of HMM transition matrix. Transition probabilities of the HMM trained with DNase + H3K4me3 using H1-hESC data. Transitions are specified from the states in the rows to the states in the columns. Histone level states are denoted with "(H)" and DNase level states with "(D)". The FOOTPRINT state is abbreviated as "FP". *Source: Gusmao et al. (2014)* (modified to fit thesis format and/or clarify key points).

	BACK	**UP (H)**	**TOP (H)**	**DOWN (H)**	**UP (D)**	**TOP (D)**	**DOWN (D)**	**FP**
BACK	0.9997	0.0003	0.0	0.0	0.0	0.0	0.0	0.0
UP (H)	0.0	0.9915	0.0085	0.0	0.0	0.0	0.0	0.0
TOP (H)	0.0	0.0	0.9901	0.0099	0.0	0.0	0.0	0.0
DOWN (H)	0.0057	0.0	0.0	0.9861	0.0082	0.0	0.0	0.0
UP (D)	0.0	0.0	0.0	0.0	0.6515	0.3485	0.0	0.0
TOP (D)	0.0	0.0	0.0	0.0	0.0	0.783	0.217	0.0
DOWN (D)	0.0	0.0339	0.0	0.0	0.0	0.0	0.577	0.3891
FP	0.0	0.0	0.0	0.0	0.0564	0.0	0.0	0.9436

Table 4.2: Example of HMM emission's mean vectors. Signals' mean values for each state of the HMM trained with DNase + H3K4me3 using H1-hESC data. Histone level states are denoted with "(H)" and DNase level states with "(D)". The FOOTPRINT state is abbreviated as "FP". *Source: Gusmao et al. (2014)* (modified to fit thesis format and/or clarify key points).

	DNase norm.	**DNase slope**	**Histone norm.**	**Histone slope**
BACK	0.0045	-0.0002	0.0441	0.0007
UP (H)	0.0501	0.0043	0.1983	0.2995
TOP (H)	0.0445	-0.0075	0.4693	0.0158
DOWN (H)	0.0636	0.0003	0.2309	-0.4237
UP (D)	0.1537	0.6343	0.0894	-0.0647
TOP (D)	0.4244	0.0059	0.1091	-0.0735
DOWN (D)	0.1578	-0.6562	0.0816	-0.0434
FP	0.0902	-0.0162	0.1009	-0.0436

Given the trained HMM models, we identify footprints using the Viterbi decoding algorithm, as described in Section 3.2.4. This process generates a set of footprints for every trained HMM model and every cell type. We observed that the small transition probabilities from an HMM state to the FOOTPRINT state and from the FOOTPRINT state to an HMM state often results in small delays on entering in the FOOTPRINT state and leaving such state slightly early. Therefore, we perform a small extension on the footprints. All resulting footprints are extended by 5 bp to each side. These footprints represent our predicted active TFBSs.

In HINT's HMM topologies that use histone modification data, we are able to create predictions using more than one histone modification. For that, we simply merge all the footprint predictions made using each histone modification individually. The use of combinations of histone modifications will be discussed with more details in the next chapter.

4.1.4. Execution of Competing Methods

In this section we present the full description of the parameterization and execution of all competing computational footprinting methods which were evaluated in this thesis. These methods are categorized as segmentation methods (Neph (Neph et al., 2012), Boyle (Boyle et al., 2011), Wellington (Piper et al., 2013) and DNase2TF (Sung et al., 2014)) and site-centric methods (Cen-

Table 4.3: Example of HMM emission's covariance matrices. Covariance matrices for each state of the HMM trained with DNase + H3K4me3 using H1-hESC data. Within each state's matrix, lines and rows are sorted by signal type as DNase-seq normalized, DNase-seq slope, H3K4me3 normalized and H3K4me3 slope. Histone level states are denoted with "(H)" and DNase level states with "(D)". The `FOOTPRINT` state is abbreviated as "FP". *Source: Gusmao et al. (2014)* (modified to fit thesis format and/or clarify key points).

BACK	0.0025	-0.0001	0.0001	0.0	**UP (H)**	0.0222	0.0001	0.003	0.0057
	-0.0001	0.0025	0.0	0.0		0.0001	0.0155	0.0006	0.0005
	0.0001	0.0	0.0047	0.0		0.003	0.0006	0.0101	0.0105
	0.0	0.0	0.0	0.0019		0.0057	0.0005	0.0105	0.0341
TOP (H)	0.0216	0.0003	-0.0009	0.0014	**DOWN (H)**	0.0239	0.0001	-0.0033	-0.0002
	0.0003	0.0196	0.0005	0.0003		0.0001	0.009	0.0002	-0.0006
	-0.0009	0.0005	0.0047	-0.001		-0.0033	0.0002	0.0156	-0.0095
	0.0014	0.0003	-0.001	0.0193		-0.0002	-0.0006	-0.0095	0.0313
UP (D)	0.0705	0.0246	-0.0053	0.0025	**TOP (D)**	0.1559	-0.002	-0.0079	0.0052
	0.0246	0.0714	-0.0038	-0.0015		-0.002	0.0384	-0.0008	0.0021
	-0.0053	-0.0038	0.0045	-0.0056		-0.0079	-0.0008	0.007	-0.0096
	0.0025	-0.0015	-0.0056	0.0125		0.0052	0.0021	-0.0096	0.0184
DOWN (D)	0.0687	-0.011	-0.0048	0.004	**FP**	0.0358	-0.0019	-0.0025	0.0007
	-0.011	0.055	0.0039	-0.0		-0.0019	0.0225	0.0001	0.0002
	-0.0048	0.0039	0.0039	-0.0044		-0.0025	0.0001	0.0068	-0.0069
	0.004	-0.0	-0.0044	0.0109		0.0007	0.0002	-0.0069	0.0121

tipede (Pique-Regi et al., 2011), Cuellar (Cuellar-Partida et al., 2012), PIQ (Sherwood et al., 2014), FLR (Yardımcı et al., 2014) and BinDNase (Kähärä and Lähdesmäki, 2015)). The competing computational footprinting methods use either DNase-seq data or a combination of DNase-seq and histone modification ChIP-seq data. Only Centipede uses extra genomic information such as distance to the nearest gene and conservation scores. To allow for a fair comparison (Nature Methods Editorial, 2015), we only used DNase-seq data, as experimental input data, for all methods. All the competing methods were applied to the `Comparative Dataset` cell types (Section 4.1.1): H1-hESC (SH), K562 (SH) and GM12878 (SH). Computational resources necessary for the execution of segmentation and site-centric competing methods are summarized in Table 4.4. The table shows the additional steps needed to execute the footprinting method, the total execution CPU time in hours, the maximum memory used during the execution and the total input storage necessary before the execution of each method.

In addition to the published segmentation and site-centric methods, we also tested a few baseline methods. These methods serve as control experiments, given their simplicity. The site-centric baseline methods (PWM-Rank, TC-Rank and FS-Rank) consist on ranking MPBSs (defined in Section 4.2.1) based on footprint quality scores. Furthermore, given the lack of a segmentation baseline method in the literature, we devised a novel segmentation baseline method which uses signal processing filter techniques. All baseline methods use only DNase-seq data and were applied to the `Analysis Dataset` cell types (Section 4.1.1).

Neph Method

We obtained the footprint predictions for cell type K562 (SH) in Neph et al. (2012). As predictions were not available for cell types H1-hESC (SH) and GM12878 (SH), we obtained the scripts and parameterization in `https://github.com/StamLab/footprinting2012` (Neph et al., 2012).

Table 4.4: Summary of computational resources. The computational resources were evaluated on 88 TFs binding on cell types H1-hESC (SH) and K562 (SH). *Source: Gusmao et al. (2016)* (modified to fit thesis format and/or clarify key points).

Method	Additional Steps	CPU time (hours)	Max. Memory (GB)	Input Storage (GB)
BinDNase	1,2,4	7034	8	95.7
Boyle	NA*	NA*	NA*	NA*
Centipede	1,2,4	7100	8	157.7
Cuellar	1,2,4	575	32	25.4
DNase2TF	3	31	32	29.3
FLR	2,4	870	16	43.1
HINT	3	56	4	17.7
Neph	3	47	4	14.6
PIQ	-	386	32	18.7
Wellington	3	117	16	14.6

[1] Requires extra input file processing.

[2] Requires extra motif matching (Section 4.2.1).

[3] Requires extra DNase-seq peak calling (DHSs).

[4] Requires execution of method for each TF.

* Implementation not available.

Briefly, we used the DNase-seq read overlap signal as input with the parameters from the original publication: flanking component length varied between 3–10 bp and central footprint region length varied between 6–40 bp. Afterwards, the footprints were filtered by a false discovery rate of 1%, which was estimated based on the distribution of footprint scores (FSs) in each cell type (Neph et al., 2012). Finally, we consider only predictions that occurred within DNase-seq hotspots, which were obtained using the method described in Sabo et al. (2004a). We will refer to this framework as "Neph".

Boyle Method

Since no source code or software is provided, we used footprint predictions from Boyle et al. (2011) available at `http://fureylab.web.unc.edu/datasets/footprints/`. We will refer to this method as "Boyle".

Centipede

Centipede software was obtained at `http://centipede.uchicago.edu/` (Pique-Regi et al., 2011) and executed to generate posterior probabilities of regions being bound by TFs. The experimental and genomic data used include DNase-seq, position weight matrix (PWM) bit-score, sequence conservation and distance to the nearest TSS. The experimental data input was generated by obtaining the read overlap DNase-seq signal surrounding a 200 bp window centered on each MPBS. Additionally, we used conservation score, distance to the nearest TSS and the PWM bit-score to create the required prior probabilities. These additional genomic data were obtained from PhastCons conservation score (placental mammals on the 46-way multiple alignment) (Siepel et al., 2005) and Ensembl gene annotation from ENCODE (Hubbard et al., 2002).

All parameters were set to their default values, with exception of the level of shrinkage of multino-

mial parameters (L) and the level of shrinkage of negative binomial parameters (N). We observed that Centipede is very sensitive to these parameters and we performed an extensive computational analysis to estimate these parameters (Gusmao et al., 2014). Our analyses showed that the best parameterization for Centipede is: $L = 0.75$ and $N = 0$ for H1-hESC (SH) and GM12878 (SH) cell types; and $L = 0.75$ and $N = 0.25$ for K562 (SH) cell type (Gusmao et al., 2014).

Cuellar Method

We applied this method as described in Cuellar-Partida et al. (2012). We created a smoothed DNase-seq input signal by evaluating the number of DNase-seq cleavage based on a 150 bp window with 20 bp steps. We obtained their scripts at `http://tlbailey.bitbucket.org/supplementary_data/Cuellar2011/` and created priors using the smoothed version of the DNase-seq signal. As suggested by the authors, the priors were submitted to the "find individual motif occurrences" (FIMO) software (Grant et al., 2011) to obtain the predictions. We will refer to this method as "Cuellar".

We also observed that the predictions are very sensitive to the p-value cutoff threshold from the program FIMO. Therefore, we performed an extensive computational analysis to estimate this parameter. It was found that the best cutoff threshold is at a p-value of 10^{-5} (Gusmao et al., 2014).

Wellington

We have obtained Wellington's source code in `http://jpiper.github.com/pyDNase` (Piper et al., 2013) and executed it with default parameters. Briefly, we used a footprint false discovery rate ($\log_{10}$) cutoff of -30, footprint sizes varying between 6 and 40 with 1 bp steps and shoulder size (flanking regions) of 35 bp.

Protein Interaction Quantification (PIQ)

We obtained PIQ's implementation in `http://piq.csail.mit.edu` (Sherwood et al., 2014) and executed it with default parameters, which are located in the script *c*ommon.r. Briefly, MPBSs were generated with the script *p*wmmatch.exact.r. The DNase-seq signal was created using the script *b*am2rdata.r. And the footprints were detected with the script *p*ertf.r.

Footprint Mixture (FLR)

Method implementation was obtained in `https://ohlerlab.mdc-berlin.de/software/FootprintMixture_109/` (Yardımcı et al., 2014). We executed the method using the 6-mer cleavage bias frequencies for initialization of the background models. The width of the window surrounding the TFBSs (*P*adLen) was set to the default value of 25 bp. Also, we use the expectation maximization to re-estimate background during training (argument *F*ixed set to FALSE). We will refer to this method as "FLR".

DNase2TF

We obtained DNase2TF's code from `http://sourceforge.net/projects/dnase2tfr/` (Sung et al., 2014) and executed DNase2TF with a 4-mer cleavage bias correction. Other parameters were set to their default values: $m$inw $= 6$, $m$axw $= 30$, z_threshold $= -2$ and FDR $= 10^{-3}$.

BinDNase

BinDNase's method implementation was obtained at `http://research.ics.aalto.fi/csb/software/bindnase/` (Kähärä and Lähdesmäki, 2015). As a supervised approach, the method requires positive and negative examples, which are obtained from TF ChIP-seq data (Section 4.2.2).

We have used DNase-seq data around MPBSs on chromosome 1 for training. These MPBSs were subsequently removed from the evaluation procedure. Note that this is the only method evaluated here which requires TF ChIP-seq examples for training. We also point the fact that BinDNase did not successfully execute for 19 TFs of our evaluation dataset (POU5F1, REST, RFX5, SP1, SP2, SRF, TCF12 and ZNF143 binding in H1-hESC; ARID3A, CTCF, IRF1, MEF2A, PU1, REST, RFX5, SP1, SP2, STAT2 and ZNF263 binding in K562) given our maximum running time criteria (one month using 40 computational cluster nodes for each execution, i.e. each TF).

Site-centric Baseline Methods

Site-centric baseline methods consist on ranking the MPBSs for a particular TF based on a quality metric. MPBSs can be seen as a set of genomic regions $R = \{r_1, \cdots, r_m\}$ in which each $r_i = [u, v]$ represents a binding site prediction (genomic region from u to v) based solely on the DNA sequence and the protein's binding affinity to that DNA sequence. MPBSs were obtained through the motif matching algorithm, which is described in Section 4.2.1.

PWM-Rank

The PWM-Rank is a baseline method which consists on ranking MPBSs based on their motif match bit-score. Such metric was obtained directly from the motif matching procedure (Section 4.2.1). The terminology "PWM" stands for "position weight matrix", which is the binding affinity structure that is used to calculate the bit-scores in the motif matching procedure. This method is considered the "absolute control", since it does not use any experimental evidence of chromatin structure to detect active TFBSs. Consequently, the results for the PWM-Rank method are the same for all the cell types in the same organism, since they share the same DNA sequence.

TC-Rank

The TC-Rank method consists on ranking the MPBSs based on the number of aligned reads (referred to as tag count; TC) within their vicinity. The method's rationale is that the more TCs in the vicinity of a MPBS, the more likely it is to be inside an open-chromatin region and therefore be an active TFBS. In this thesis we used the most predominant window size for the TC calculation in the literature, which is 100 bp in total (Cuellar-Partida et al., 2012; Yardımcı et al., 2014; He et al., 2014). Let $R = \{r_1, \cdots, r_l\}$ be a set of l MPBSs for a particular TF and $\mathbf{x}$ the read overlap DNase-seq signal, the TC score for the MPBS $r_i = [u, v]$ was calculated as

$$\mathrm{TC}_{r_i} = \sum_{j=\frac{(u+v)}{2}-50}^{\frac{(u+v)}{2}+50} x_j. \tag{4.1}$$

FS-Rank

The FS-Rank method consists on ranking the MPBSs based on the footprint score (FS) metric, which was used in previous works (Neph et al., 2012; He et al., 2014) as a quality score to rank footprints predictions. The method's rationale is that a MPBS with few DNase I cleavage within the binding site region in comparison to its flanking regions corresponds to the pattern described as the grammar of active TFBSs and therefore is more likely to be an active TFBS. Let $R = \{r_1, \cdots, r_l\}$ be a set of l MPBSs for a particular TF and $\mathbf{x}$ the read overlap DNase-seq signal, the FS for the MPBS $r_i = [u, v]$

was calculated as

$$\mathrm{FS}_{r_i} = \left(\frac{n^C_{r_i}+1}{n^R_{r_i}+1} + \frac{n^C_{r_i}+1}{n^L_{r_i}+1} \right), \tag{4.2}$$

where $n^C_{r_i}$, $n^L_{r_i}$ and $n^R_{r_i}$ are the number of DNase I cleavage hits within the MPBS r_i, in the left (upstream) region of the MPBS r_i and in the right (downstream) region of the MPBS r_i. These values were calculated as

$$n^C_{r_i} = \sum_{j=u}^{v} x_j, \qquad n^R_{r_i} = \sum_{j=v}^{2v-u} x_j, \qquad n^L_{r_i} = \sum_{j=2u-v}^{u} x_j. \tag{4.3}$$

Signal Processing Filters

The rationale of using signal processing filters for computational footprinting is to remove inadequate frequencies in order to make the DNase-seq peaks more pronounced and detectable by simpler window-based approaches. We applied the method as follows. First, a filtering technique is applied to DHSs. We tested four different filtering techniques: Butterworth, Chebyshev, Elliptic and Bessel (Lutovac et al., 2000). Preliminary analyses showed that the Butterworth filtering technique provided higher accuracies. Therefore, here we describe only the signal processing footprinting method using the Butterworth filter.

As we wanted to investigate the accuracy of the filtering technique itself, we did not perform any further signal processing methodology. Consequently, we could not use an HMM or techniques which involve scoring genomic regions with sliding windows (such as FS or TC) to detect footprints in the filtered signal. The reason is that the signal frequency and time-domain transformations affect the absolute signal magnitude significantly. Since the transformations performed by the filtering technique do not significantly affect the signals' standard deviation within small window frames (Shenoi, 2005), we used a standard deviation-based windowing approach to detect the significant depletions in the data, i.e. the footprint pattern.

Signal Filtering

First, we applied the Butterworth signal filtering technique for each DHS. The rationale behind the Butterworth filter is that an ideal signal processing filter should not only reject unwanted frequencies but should also have uniform sensitivity for the wanted frequencies. Such an ideal filter can not be achieved but it can be shown that successively closer approximations are obtained with increasing numbers of filter elements of the right values. It was shown (Shenoi, 2005) that a low-pass filter could be designed whose cutoff frequency was normalized to 1 radian per time unit and whose frequency response (gain) was

$$G_n(\omega) = \sqrt{\frac{1}{1+\omega^{2n}}}, \tag{4.4}$$

where ω is the angular frequency in radians per time unit t (which corresponds to our genomic coordinates) and n is the number of poles in the filter.

Within this framework we were able to perform the signal frequency and time-domain transformations (Lutovac et al., 2000). We applied the Butterworth's implementations of its high-pass, low-pass and band-stop filters: (1) the high-pass filter removes background noise in the data; (2) the low-pass filter attenuates the peaks in the genomic signal and (3) band-stop filter normalizes the signals in order to prepare them for the standard deviation-based footprinting. All filters output real-valued signals which contains negative values. In order to prevent numerical problems on the standard deviation-based footprinting which these negative values might cause, we searched the global minimum value and summed the absolute version of this value for all values of the genomic signal.

Standard Deviation Footprinting

The second part of the method corresponds to the statistical analysis of the filtered signal to obtain the footprint predictions. First, we measured the average standard deviation within the filtered signal for: (1) a 20 bp window centered at the beginning of all MPBSs (Section 4.2.1) in the human chromosome 1, (2) a 20 bp window centered at the ending of all MPBSs in the human chromosome 1. We call these values, respectively $\bar{\alpha}$ and $\bar{\beta}$. We considered all MPBSs obtained by applying motif matching in cell type K562 and we considered the true MPBS the ones that contained ChIP-seq evidence (Section 4.2.2). The human chromosome 1 was removed from all subsequent evaluation experiments.

Then, we were able to perform a window-based search within the genomic signal for 20 bp regions in which the standard deviation estimated at a 20 bp window from the region's start site (and region's end site) did not exceed a certain threshold value $\hat{\alpha}$ (and $\hat{\beta}$) from the experimentally-estimated standard deviations $\bar{\alpha}$ (and $\bar{\beta}$). Such approach consists on a slightly modified version of the algorithm proposed by Neph et al. (2012) using the standard-deviation technique proposed by Shenoi (2005). Such modifications were performed in order to fit the filtered signals. The standard deviations were calculated dynamically as the window slides within the selected regions. We will refer to this method as "Filter".

4.2 Evaluation of Computational Footprinting Methods

In this section we discuss the methodology used to evaluate the footprint predictions from the computational footprinting methods, which is depicted in Figure 4.3. We used two evaluation approaches. The first is based on TF ChIP-seq data (ChIP-seq evaluation) and was generally used to perform comparative analyses in the literature (Pique-Regi et al., 2011; Boyle et al., 2011; Cuellar-Partida et al., 2012). Nevertheless, TF ChIP-seq experiments has a few caveats. First, TF ChIP-seq peaks are also observed in indirect binding events (Yardımcı et al., 2014). Second, the ChIP-seq low spatial resolution makes that false binding sites might be regarded as true binding sites by proximity the actual binding site (Cuellar-Partida et al., 2012; Yardımcı et al., 2014). To avoid the biases which stem from TF ChIP-seq evaluation, we devised a second evaluation approach which does not require TF ChIP-seq data. Instead, it is based on gene expression differences between pairs of cells (gene expression evaluation).

Both evaluation strategies use MPBSs, which are TFBSs predicted using only DNA sequence information and the TF's DNA sequence affinity. In this thesis we used the computational sequence-based method termed motif matching (Section 4.2.1). Then we proceed by defining the evaluation methodologies based on ChIP-seq (Section 4.2.2) and gene expression (Section 4.2.3).

4.2.1. Motif-Predicted Binding Sites

MPBSs are predictions of TFBSs made using only the genomic DNA sequence and the proteins' DNA sequence binding affinity. MPBSs are obtained applying a computational sequence-based method. In this thesis we use an algorithm termed motif matching. The motif matching algorithm takes as input the TF DNA sequence binding affinity, represented as a PFM (Figure 4.4). First, the PFM is normalized, generating another structure called position weight matrix (PWM). Then, the genomic DNA sequence is scanned using the PWM to find substrings which are likely to represent binding sites, given the PWM model. The putative binding sites obtained with the motif matching algorithm are termed MPBSs.

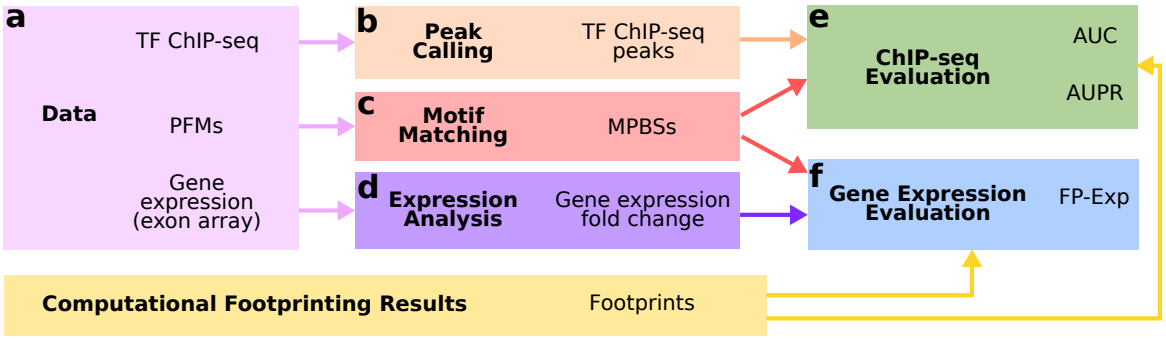

Figure 4.3: Experimental framework of the evaluation of the computational footprinting methods. (**a**) The evaluation methodologies use transcription factor (TF) ChIP-seq, position frequency matrices (PFMs) and gene expression data. (**b**) A peak calling algorithm is applied to the TF ChIP-seq data to find TF ChIP-seq peaks (enriched regions). (**c**) The motif matching algorithm is performed using the PFMs to generate motif-predicted binding sites (MPBSs). (**d**) An expression analysis is performed on the gene expression data to find gene expression fold chance between pairs of cell types. (**e**) The ChIP-seq evaluation methodology uses MPBSs and TF ChIP-seq peaks as the ground truth. When combined with the footprint predictions, we are able to calculate statistics such as the area under the receiver operating characteristic (ROC) curve (termed AUC) and area under the precision-recall (PR) curve (termed AUPR). (**f**) The gene expression evaluation methodology uses MPBSs and gene expression fold change (FC) as ground truth. When combined with the footprint predictions, we are able to calculate the FP-Exp statistics (defined in Section 4.2.3).

Input Data – Position Frequency Matrix

PFMs are created by gathering experimentally verified biological sequences which are known to be bound by the target TF of interest. Then, a multiple alignment algorithm is applied to these experimentally verified DNA sequences. After that, conserved positions within this multiple alignment are determined and a specific window, that varies from 5–25 bp is estimated to be the start and end position of the binding affinity representation (Figure 4.4a and b). At this point, the PFM $\mathbf{X}^{4\times m}$ (Figure 4.4c) is calculated in the following way: each matrix row $i \in D$ with $D = \{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\}$, correspond to each one of the possible four DNA nucleotides; and the each matrix column $j \in \{1,\cdots,m\}$ correspond to a position within the motif in the multiple aligned DNA sequences. Consequently, m represents the total length of the motif. Each entry x_{ij} of the PFM corresponds to the number of nucleotides of type i in position j of the DNA fragments' multiple alignment.

We obtained all PFMs used on our experiments from the repositories Jaspar (Mathelier et al., 2014), Uniprobe (Robasky and Bulyk, 2011) and Transfac (Matys et al., 2006). Each TF has its own PFM representation. These non-organism-specific data were obtained for the subphylum *Vertebrata*. See Supplementary Tables A.3 and A.4 for a full description on the PFMs used in this work.

Position Weight Matrices

From PFMs, we are able to create normalized logarithmic representations termed PWMs $\mathbf{W}^{4\times m}$ (Figure 4.4d). The most common method to create PWMs consists on the calculation of the corrected probability p_{ij} of finding the nucleotide i in the position j, which is given by

$$p_{ij} = \frac{f_{ij} + s(i)}{m + \sum_{i' \in D} s(i')}, \tag{4.5}$$

where f_{ij} is the frequency of base i at position j and $s(i)$ is a pseudocount function. Pseudocounts are small values used to avoid null probabilities. After the evaluation of the corrected probabilities, the entries w_{ij} of the PWM are calculated as

$$w_{ij} = \log_2 \frac{p_{ij}}{b(i)}, \tag{4.6}$$

where $b(i)$ is the genomic frequency of nucleotide i in the genome. The background correction function $b(\cdot)$ is used to correct the PWM for biases regarding the genomic imbalance between the frequencies of the nucleotides.

PWMs are used to score any DNA sequence of length m by a summation of the corresponding nucleotides between the DNA sequence and the PWM (Figure 4.4e). Such a score is called the PWM's bit-score.

Furthermore, we are able to assess the information content $\mathbf{l} = \langle l_1, ..., l_m \rangle$ of each position j of the PWM $\mathbf{W}$ by applying

$$l_j = 2 + \sum_{i \in D} p_{ij} \log_2 p_{ij}, \tag{4.7}$$

where the number 2 is obtained from the total possible information content of the 4-character alphabet D, i.e. $\log_2 4 = 2$. Based on the total information content for every position of the PWM, we are able to create graphical representations of the binding affinity – termed logo graphs (Figure 4.4f) – by multiplying the corrected probability of a certain nucleotide i at a certain position of the PWM j by the total information content at that position (l_j).

Motif Matching

From a PWM it is possible to estimate the sequence-based probability of the particular TF of binding in the genome. We call this procedure motif matching. For each sequence of nucleotides of length m, a bit-score is calculated. There are also many strategies to perform this calculation. The simplest one is the summation of all the entries in w_{ij} matching the nucleotide sequence of length m. More formally, given a sequence of characters $\mathbf{g}$ representing the genome, where $\mathbf{g} = \langle g_1, ..., g_n \rangle \forall g_i \in D$. We are able to define a vector of bit-scores $\mathbf{y} = \langle y_1, ..., y_{n-m} \rangle$ as

$$y_i = \sum_{j=1}^{m} \sum_{k \in D} \mathbf{1}(k = g_i) w_{kj}, \tag{4.8}$$

where $\mathbf{1}(\cdot)$ is an indicator function.

A genome-wide application of a PWM creates bit-scores for every possible contiguous nucleotide sequence of length N within the genome. Then, several statistical techniques can be used to determine a cutoff threshold to accept particular sequences as being bound by the protein, given the PWM. A well-known statistical procedure is to estimate a bit-score cutoff that corresponds to the false positive rate (FPR) of the distribution of the bit-scores from all possible m-mers (Wilczynski et al., 2009). More formally, let $C = \{\mathbf{c}^1, ..., \mathbf{c}^{4^m}\}$ be the set of all m-mers constructed by picking m elements from the set D with order and repetition, where each m-mer $\mathbf{c}^i = \langle c_1^i, \cdots, c_m^i \rangle$. Therefore, we are able to calculate the set $B = \{b_1, \cdots, b_{4^m}\}$ of all the possible bit-scores for the PWM $\mathbf{W}$ as

$$b_i = \sum_{j=1}^{m} \sum_{k \in D} \mathbf{1}(k = c_j^i) w_{kj}. \tag{4.9}$$

Then, it is easy to find the false discovery rate threshold by finding the p-value that corresponds to

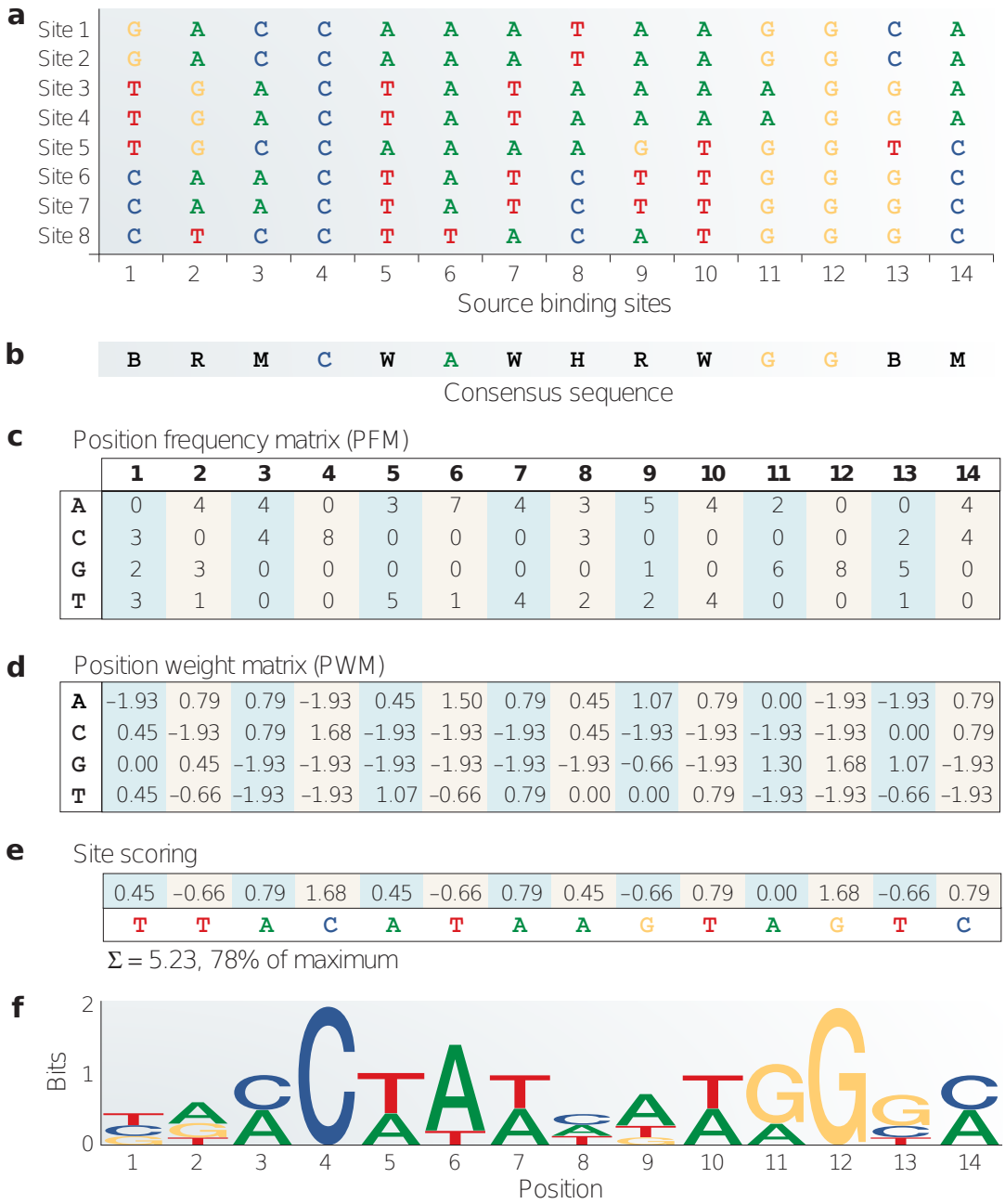

Figure 4.4: PFMs and PWMs used in the motif matching technique. (**a**) A set of experimentally validated binding sites was collected and aligned. The sequence variability of the collection of binding sites strongly affects the downstream models for predicting additional sites. Note the diversity between the sites; for instance, only 50% of the nucleotides are identical between sites one and eight. (**b**) The consensus sequence model using IUPAC symbols. For instance, the symbol "B" means preference for either nucleotides C, G or T. (**c**) PFMs created based on the source binding sites. (**d**) PWMs created using the procedure described by Equations 4.5 and 4.6. (**e**) Using the PWM, a quantitative score for any DNA sequence is generated by summing the values that correspond to the observed nucleotide at each position. (**f**) A sequence logo scales each nucleotide by the total bits of information multiplied by the relative occurrence of the nucleotide at the position (Equation 4.7). Sequence logos enable fast and intuitive visual assessment of pattern characteristics. *Source: Wasserman and Sandelin (2004)* (modified to fit thesis format and/or clarify key points).

B fitted to a certain distribution, say normal

$$B \sim \mathcal{N}(\mu, \sigma^2). \tag{4.10}$$

We observed that the p-value choice significantly affects many aspects of the evaluation procedure. Therefore, we made a careful parameter selection analysis. Such an analysis indicated that a p-value of 10^{-4} resulted in a significant amount of TF ChIP-seq peaks overlapping with MPBSs (Gusmao et al., 2014). Higher p-values generate a very high number of MPBSs which impacts on computational time and increases the imbalance between true and false MPBSs (Gusmao et al., 2014). The set of MPBSs after the application of the false discovery rate cutoff threshold is represented by a genomic region set $R = \{r_1, \cdots, r_l\}$, where each MPBS $r_i = [u, v]$, is an interval from genomic positions u to v.

4.2.2. ChIP-seq Evaluation

The ChIP-seq evaluation approach uses MPBSs in conjunction with TF ChIP-seq peaks as ground truth (Pique-Regi et al., 2011; Boyle et al., 2011; Cuellar-Partida et al., 2012). By evaluating the overlap between MPBSs, TF ChIP-seq peaks and footprint predictions we are able to assess the accuracy of computational footprinting methods (Figure 4.5). The advantage of this approach is that it provides a straightforward scenario for the evaluation of computational footprinting methods. Furthermore, this evaluation approach enables the comparison between different methods for each individual TF.

Data

We obtained TF ChIP-seq datasets consisting on the enriched regions (peaks). On total, 144 ChIP-seq peaks datasets were obtained to create the ChIP-seq evaluation datasets. All peaks were obtained in ENCODE Project Consortium (2012) with exception of the following TFs: (1) AR – obtained in Yu et al. (2010); (2) ER – obtained in Guertin et al. (2014); and (3) GR – obtained in John et al. (2011). See Supplementary Table A.3 for a full description of TF ChIP-seq data and the TF PFMs matching each ChIP-seq experiment.

Application of the ChIP-seq Evaluation Methodology

MPBSs with ChIP-seq evidence (located within 100 bp from the ChIP-seq peak summit) are considered "true" binding sites; while MPBSs without ChIP-seq evidence are considered "false" binding sites. Every TF prediction (footprint) that overlaps a true binding site is considered a correct prediction (true positive – TP) and every prediction that overlaps a false binding site is considered an incorrect prediction (false positive – FP). Therefore, true negatives (TN) and false negatives (FN) are, respectively, false and true binding site without overlapping predictions (Figure 4.5a). We consider overlaps of at least one bp.

The contingency table (TPs, FPs, TNs and FNs) enables the creation of receiver operating characteristic (ROC) curves, which describe the sensitivity increase as we decrease the specificity of the method (Figure 4.5b). The area under the ROC curve (AUC) metric was calculated at 1%, 10% and 100% FPRs. By evaluating the AUC at different FPRs we avoid misleading interpretations due to the rate in which the specificity decreases with sensitivity increase. The contingency table also enables the creation of precision-recall (PR) curves (Figure 4.5c). The area under the PR curve (AUPR) is a statistic indicated for problems with imbalanced datasets (distinct number of positive and negative examples) (Davis and Goadrich, 2006; Fawcett, 2006). In summary, the ChIP-seq evaluation methodology provides four performance statistics for each TF: (1) AUC at 1% FPR; (2) AUC at 10% FPR; (3) AUC at 100% FPR; and (4) AUPR. For any of these statistics, higher values indicate higher method performance.

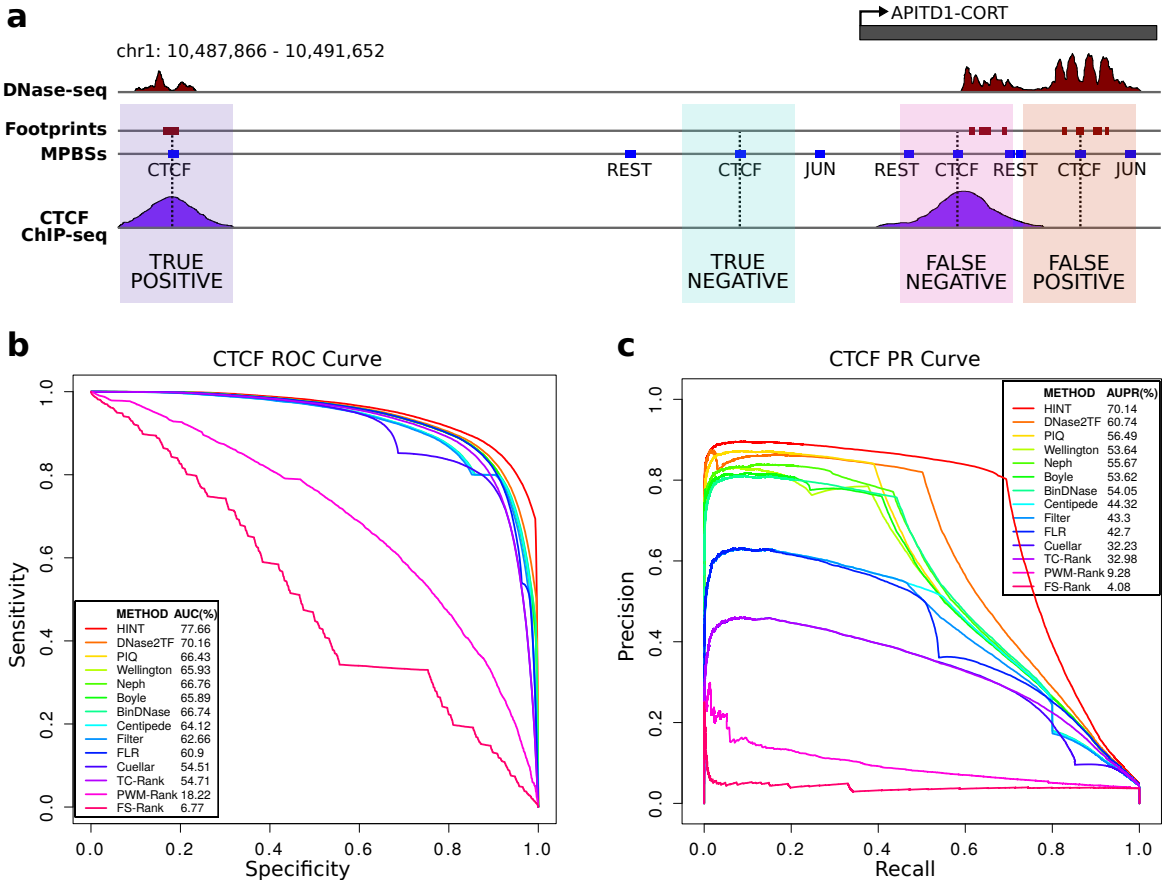

Figure 4.5: ChIP-seq evaluation methodology. (**a**) Example of the ChIP-seq evaluation approach with the real biological data. We are able to calculate a contingency table by checking the overlap between the footprint predictions and the MPBSs with and without cell-specific ChIP-seq evidence. By ranking the MPBSs based on the overlapping footprint's quality, we are able to create: (**b**) receiver operating characteristic (ROC) curves and (**c**) precision-recall (PR) curves. These structures can be used to rank the computational footprinting methods and evaluate their overall performance on identifying each TF individually.

Segmentation-based approaches (Boyle, DNase2TF, HINT, Neph and Wellington) provide footprint predictions that do not necessarily encompass all MPBSs. To create full ROC curves for these methods, we first ranked all predicted sites (MPBS that overlapped a footprint) by their TC (Equation 4.1) followed all non-predicted sites ranked by their TC. In order to present a fair comparison, this approach was also applied to all site-centric methods (Centipede, Cuellar, FLR and PIQ). For that, we considered distinct probability thresholds of the footprint quality scores reported by each method. We performed additional experiments to select the best threshold per method. We observed that a p-value threshold of 0.9 was best for all site-centric methods except BinDNase, which was best at a p-value of 0.8. These analyses were performed in chromosome-1 which was further removed from the comparative evaluation analyses.

ChIP-seq Evaluation Scenarios

Different experiments required a different number of TFs being evaluated. For instance, experiments in which a high number of competing methods were executed, a lower number of TFs were evaluated, given the high demand of computational time. On the other hand, experiments in which a particular computational footprinting feature was being tested using only HINT, we used an evaluation set with

a higher number of TFs to enhance statistical significance. Therefore, we created two ChIP-seq evaluation scenarios, described below.

`Benchmarking Dataset`: For comparative analysis of several competing methods, we selected the two cell types from the `Comparative Dataset` (Section 4.1.1) with the highest number of TF ChIP-seq datasets evaluated in our study: K562 with 59 TFs and H1-hESC with 29 TFs. This was required due to the high computational demands of the execution of some competing methods. All methods described in this study were compared under this evaluation scenario (executed using the SH DNase-seq data).

`Comprehensive dataset`: We have compiled a comprehensive dataset containing 235 combinations of cell types and TFs with matching cellular background. We used the cell types from the `Analysis Dataset` (Section 4.1.1). This dataset was built from a catalog of 144 TF ChIP-seq. This dataset was used in analyses which required a large dataset for statistical significance. In this scenario we only evaluated HINT and the baseline methods.

4.2.3. Gene Expression Evaluation

The ChIP-seq evaluation approach requires TF ChIP-seq experiments which, as indicated by Yardımcı et al. (2014), has some intrinsic biases. First, TF ChIP-seq peaks are also observed in indirect binding events. Second, they have a lower spatial resolution than DNase-seq. Therefore, false MPBSs might be regarded as true MPBSs by proximity to an active TFBS. Recently, Yardımcı et al. (2014) indicated that footprint quality scores, as measured by their method's metric – the footprint likelihood ratio (FLR) – were significantly higher in cells where the TF was expressed. This observation indicates that comparing changes in expression and quality of footprints in a pair of cells provides an alternative footprint evaluation measure. This led us to the development of a novel evaluation methodology based on gene expression by applying this idea systematically for a large set of TFs.

Data

Expression profiling by array (Affymetrix Human Exon 1.0 ST Array) data was obtained in ENCODE Project Consortium (2012). We obtained data for all `Comparative Dataset` cell types: H1-hESC, K562 and GM12878. All samples from each cell type was used to infer the overall gene expression profile. See Supplementary Table A.5 for a full gene expression data description.

Application of the Gene Expression Evaluation Methodology

We used limma (Ritchie et al., 2015) version 3.28.4 to perform between-array normalization on expression of H1-hESC, K562 and GM12878 cells and obtain gene expression fold change (FC) estimates. This generated pairwise FCs between all three cell type pairs: H1-hESC *vs* K562, H1-hESC *vs* GM12878 and K562 *vs* GM12878. We used the R programming language version 3.1.2 implementation of limma. The source code of this software is found at `https://bioconductor.org/packages/release/bioc/html/limma.html`.

Then, we retrieved all non-redundant PFMs from Jaspar in which gene symbol is a perfect match with genes present in the array platform. This leads us to 143 PFMs (Supplementary Table A.4). We applied a genome-wide motif matching (Section 4.2.1) using these PFMs to create MPBSs.

Afterwards, we calculated footprint quality scores for all footprints from all computational footprinting methods, which intersect with MPBSs of a particular motif. In this thesis we used three different metrics as footprint quality scores: (1) the FLR score (Yardımcı et al., 2014); (2) the TC and (3) the FS. We only considered the footprints within DHSs that are in common between the cell type

pair being evaluated, as described in Yardımcı et al. (2014). We expect that TFs with higher expression values in a particular cell type would present higher values regarding footprint quality metrics with DNase-seq from that cell type.

A two-sample Kolmogorov-Smirnov (KS) test was used to assess the difference between each metrics' distribution between the two cell types being evaluated. The KS statistic, which varies within $[0,1]$, is used to indicate the difference between two distributions; higher values indicate higher differences. As the KS score do not indicate the direction of the changes in distribution, we obtained a signed version by multiplying KS statistic by -1, in cases where the median of the quality scores calculated in cell type A $<$ median calculated in cell type B. We calculated the Spearman correlation between the signed KS test statistic and the FC for each TF. Positive values indicate an association between expression of TFs and quality of footprint predictions. We will call this correlation "FP-Exp". The higher the FP-Exp, the better the computational footprinting method. Figure 4.6 exhibits a graphical description of the gene expression evaluation methodology.

4.3 Downstream Analyses

The predicted footprints from a computational footprinting approach represent a map of active TFBSs. In possession of such footprints we are able to perform a number of different downstream analysis. In this section we show two common downstream analyses that we are going to explore in this thesis: the TF enrichment analysis and the *de novo* motif finding. The main goal of the TF enrichment analysis is to identify TFs which are more likely to bind in footprints from a particular cell type when compared to other cell type (Section 4.3.1). On the other hand, the *de novo* motif finding consists on searching for novel TF DNA affinity sequences which do not match any known affinity sequence in the literature (Section 4.3.2).

4.3.1. Transcription Factor Enrichment Analysis

The TF enrichment analysis is divided in two parts: (1) the application of a statistical test, on each cell type or biological condition, to verify if TFs bind more than expected by chance at genomic regions of interest (i.e. footprints); and (2) the comparison between the results of the statistical test for all TFs between the different cell types or biological conditions being investigated.

We start by defining two genomic region sets: the target genomic region set and the background genomic region set. The target genomic region set $X = \{x_1, \cdots, x_n\}$ is composed of the genomic regions associated to the target biological condition being tested (Figure 4.7a). It can be, for instance, footprints identified in a group of differentially expressed genes. The background genomic region set $Y = \{y_1, \cdots, y_m\}$ (Figure 4.7b) is composed of a collection of random genomic regions throughout the genome. The rationale is that the background genomic region set acts as a "control" to which we can compare our target genomic region set against. By comparing the occurrence of putative active TF binding within our target genomic region set X against the background genomic region set Y we can perform a statistical test to assess the enrichment of TFs in X. For statistical power, the number of background genomic regions m should be higher than the number of target genomic regions n.

After the definition of our target and background genomic region sets, we apply the motif matching algorithm to identify MPBSs within these regions (Figure 4.7c). In the analyses presented in this thesis, the motif matching was performed using all the PFMs available from Jaspar (Mathelier et al., 2014) and Uniprobe (Robasky and Bulyk, 2011). Then, by overlapping the MPBSs with the target and background genomic region sets, we create the following statistics for each TF t (Figure 4.7d):

a_t – The number of target genomic regions overlapping at least one MPBS from TF t.

b_t – The number of target genomic regions which do not overlap any MPBS from TF t.

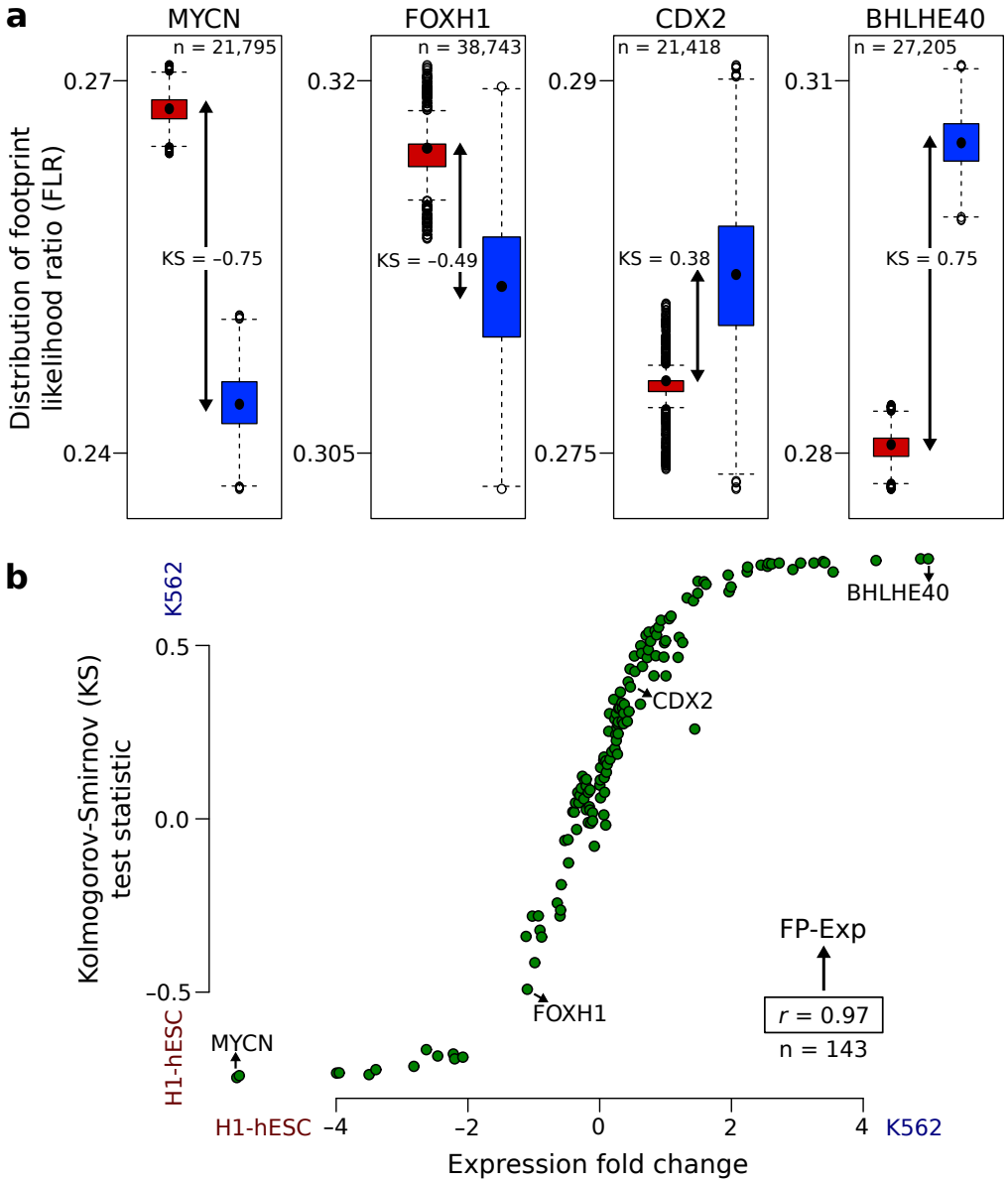

Figure 4.6: Gene expression evaluation methodology. This figure shows an example of the gene expression evaluation methodology using the footprint likelihood ratio (FLR) as a footprint quality score, calculated on DNase-seq data from cell types H1-hESC (SH) and K562 (SH). (**a**) FLR score distribution of footprints predicted with HINT overlapping with MPBSs of selected TFs. These TFs have increasing expression in K562 (red) compared with H1-hESC cell types (blue). The signed Kolmogorov-Smirnov (KS) statistic quantifies the separation of both distributions. The box plot depicts the distribution median value (middle dot) and first and third quartiles (box extremities). Box plots' whiskers represent the 1.5 interquartile region (IQR) and external dots represent outliers (data greater than or smaller than 1.5 IQR). (**b**) Scatter plot with signed KS statistic and expression fold change (FC) for 143 TFs. There is a clear association between TF expression and KS statistic ($r = 0.97$, adjusted p-value $< 10^{-10}$). We call this correlation FP-Exp. The higher the FP-Exp, the better the computational footprinting method. *Source: Gusmao et al. (2016)* (modified to fit thesis format and/or clarify key points).

c_t – The number of background genomic regions overlapping at least one MPBS from TF t.

d_t – The number of background genomic regions which do not overlap any MPBS from TF t.

Then, we apply the Fisher's exact test on the aforementioned statistics a_t, b_t, c_t and d_t. The null hypothesis is defined as: the proportion TF binding at target genomic regions is not greater than the proportion of TF binding at background genomic regions. Nevertheless, since we test a high number of TFs (~600 PFMs from Jaspar and Uniprobe) and each one requires a different and independent statistical test, we perform a multiple testing correction. For that, we use the Benjamini and Hochberg method (Benjamini and Hochberg, 1995) (also known as false discovery rate (FDR) control method). The final result is a list of corrected p-values which describes the likelihood of the tested TFs to be associated to the target genomic regions in comparison to the background genomic regions.

In possession of the corrected p-value list of TF enrichment for all cell types / biological conditions being tested, we search for the TFs that presented significant p-values (< 0.05) in particular cell types / biological conditions. For that, we filter the list of TFs for the ones which: (1) present a significant p-value in at least one of the conditions tested and (2) present a non significant p-value in at least one of the conditions tested. The list of filtered TFs are likely to contain the regulators of specific cell types / biological conditions.

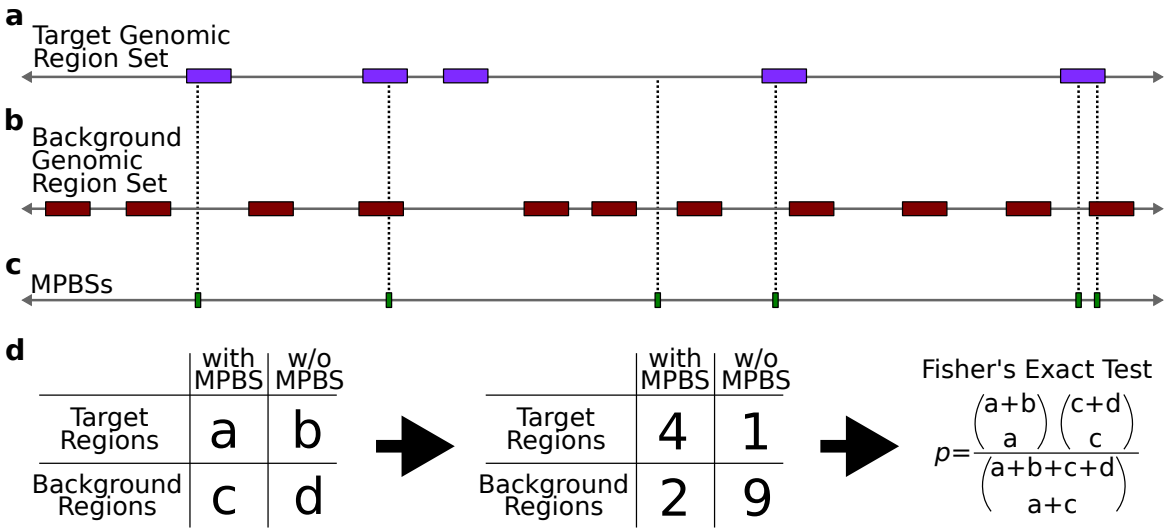

Figure 4.7: TF enrichment analysis. (**a**) The target genomic region set is composed of the genomic regions under study. (**b**) The background genomic region set is composed of "control" genomic regions. It can be, for instance, random genomic regions in the same organism's genome. (**c**) MPBSs are created for a particular TF in which we are interested in evaluating if it is enriched in the target genomic regions in contrast to the background genomic regions. (**d**) Based on the overlap between the target genomic region set, the background genomic region set and the MPBSs for a particular TF, we create a contingency table and perform the Fisher's exact test. The test's p-value gives an indication on the enrichment of the TF at the target genomic regions.

4.3.2. *De Novo* Motif Finding

We show here a very simple protocol to search for novel TF DNA sequence affinity motifs within footprint predictions. Such analysis extends our knowledge on the regulatory elements that binds a particular cell type.

First, we apply the motif matching algorithm on all predicted footprints using all PFMs from the Jaspar (Mathelier et al., 2014) and Uniprobe (Robasky and Bulyk, 2011) repositories. Before the motif matching, we extend all footprints by 10 bp to each side to be able to recognize larger sequence motifs. The goal of this initial motif matching analysis is to eliminate all footprints which correspond to known TF affinity motifs.

Then, we apply the *de novo* motif finding tool "discriminative regular expression motif elicitation" (DREME) (Bailey, 2011) on the footprints that do not present any known motif. Such tool is optimized to perform *de novo* motif analysis in datasets containing many sequences and is able to find multiple different motifs. Briefly, DREME finds substrings that appear in a target genomic region set (in our case, the footprints) more frequent than by chance given a background genomic region set (in our case, random genomic regions with the same length of the footprints but with 100 times more sequences). DREME outputs a number of novel motifs found on the sequence.

Since we performed an extension on the footprints prior to the execution of DREME, we might find a couple of "artifact motifs", i.e. small motifs that do not correspond to a footprint which are in the border of the footprint prediction. To filter for these artifact motifs, we execute the "local motif enrichment analysis" (CENTRIMO) software (Bailey and Machanick, 2012) tool on all sequences associated to the *de novo* motifs found by DREME. This tool makes sure that the motifs found are centrally enriched within the footprints' regions and are not a product of the 10 bp extension.

The *de novo* motifs found by DREME which are significantly centrally enriched according to CENTRIMO correspond to the results of our *de novo* motif analyses. These resulting motifs are represented by PFMs based on the DNA sequence on the binding sites within the footprints.

4.4 Statistical Methods

All method comparison in this work is performed using the non-parametric Friedman-Nemenyi hypothesis test. This hypothesis test is indicated when using multiple gold standard datasets and methods (Demšar, 2006). Such test provides a rank of the methods as well as the statistical significance of whether a particular method was outperformed by other method. The test provides results for the significance levels of 0.95 and 0.99. We obtained an implementation of the Friedman-Nemenyi hypothesis test written in Java and it was run in the Java SE Runtime Environment (build 1.8.0_45-b14).

All correlations calculated in this work are based on the Spearman's rank correlation coefficient (denoted as r) (Duda et al., 2000). The Spearman's rank correlation was chosen since it assesses monotonic relationships (whether linear or not). All Spearman correlation p-values are based on a two sided test with significance of 0.95. We used the R programming language version 3.1.2 implementation of the Spearman correlation test with the function *cor.test*.

All differences between the distribution of two samples (or more samples in a pair-wise manner) are analyzed using the non-parametric Mann–Whitney–Wilcoxon hypothesis test (Duda et al., 2000). All test p-values are based on a two sided test with significance of 0.95. We used the R programming language version 3.1.2 implementation of the Mann–Whitney–Wilcoxon hypothesis test with the function *wilcox.test*. The only exception is regarding the gene expression evaluation approach (Section 4.2.3), in which the Kolmogorov-Smirnoff hypothesis test was used, in accordance to Yardımcı et al. (2014).

All p-values are corrected for multiple comparisons using the Benjamini and Hochberg method (Benjamini and Hochberg, 1995) (also known as false discovery rate (FDR) control method). Multiple test correction is necessary since we perform many hypothesis tests given our evaluation framework. We used the R programming language version 3.1.2 implementation of the Benjamini-Hochberg multiple test correction method with the function *p.adjust*.

4.5 Discussion

In this chapter we defined our computational experimental framework. We described the datasets used as input for our computational footprinting method and for the method evaluation experiments. We defined the signal processing, training and execution of our computational footprinting method HINT. Moreover, we described the execution of nine competing methods, categorized as either segmentation-based or site-centric, and four baseline methods, which are regarded as control experiments. In total, our comparative analysis encompasses 14 different computational footprinting methods. Furthermore, we described the evaluation methodologies, which are based on either TF ChIP-seq or gene expression as ground truth to test footprint predictions. By having two different and independent evaluation approaches we expect to provide a clear picture of the comparison between the tested computational footprinting methods.

We have observed that there are a number of parameters from the computational footprinting methods or from the strategies used to generate the evaluation gold standard datasets that have a significant impact on the performance and further analyses results. Therefore, we have performed multiple empirical tests on parameter selection (Gusmao et al., 2014). The parameters shown in this chapter correspond to the ones which maximize the performance of all computational footprinting methods on empirical analyses performed using data from chromosome 1. To avoid overfitting and interpretation biases, we have excluded the chromosome 1 from all our subsequent comparative analyses results. Important results from parameter selection empirical analyses will be shown in the next chapter.

We close this chapter highlighting the main contributions performed in this thesis with regard to the experimental design of computational footprint methods:

- So far, all studies that perform computational footprinting method comparison used the AUC metric for the ChIP-seq evaluation strategy as a single resource to assess method performance. In this thesis, in addition to using the AUC at various levels of the FPR, we use the AUPR, which is indicated when there is a considerable gold standard data imbalance (very different numbers of positive and negative instances) (Davis and Goadrich, 2006; Fawcett, 2006).

- We devised a novel evaluation strategy which does not rely on ChIP-seq data. Such evaluation strategy uses gene expression fold change from cell type pairs to assess the overall quality of computational footprinting method's performance.

- No study so far have evaluated such a comprehensive number of different computational footprinting methods. Such a comprehensive analysis provides a full picture of the state-of-the-art strategies for computational footprinting.

- Finally, we performed multiple parameter selection empirical analyses (Gusmao et al., 2014). Such analyses resulted in maximally efficient footprint predictions for all computational footprinting methods tested, without adding overfitting biases.

CHAPTER 5

Results

In this chapter we present the results generated by our experimental analysis on computational footprinting methods. First, we performed empirical analyses to determine the best parameters to our computational footprinting methodology HINT (Section 5.1). Then, we investigated two current major challenges in the area: the selection of an optimal footprint ranking strategy and the correction of DNase-seq sequence cleavage bias (Section 5.2). Afterwards, we present our comprehensive comparative analysis, which includes our method, nine competing methods and four baseline methods (Section 5.3). We also provide an insightful discussion on an unexplored and critical challenge of computational footprinting methods – the transcription factor (TF) binding residence time (Section 5.4). Then, we show an example of downstream analysis – the *de novo* motif finding (Section 5.5). After, we present two case studies in which our computational footprinting method was applied successfully to unravel key regulatory TFs on two different biological experiments (Section 5.6). A full discussion on the results presented in this study will be performed thoroughly in the next chapter (Chapter 6).

5.1 HINT Parameter Selection

We performed a number of preliminary analyses to find the best parameters for our computational footprinting framework HINT. First, we studied the hidden Markov model (HMM) topology which optimizes the accuracy on identifying correct footprints (Section 5.1.1). Then, we tested a number of different combination of input histone modification data. Such test not only determined the best data input types for our computational method but also provided interesting insights on the underlying biological problem (Section 5.1.2). Finally, we investigated the level of dependence our method has on the training data (Section 5.1.3).

In the analyses presented in this section, we used the ChIP-seq evaluation approach to assess performance. All scenarios were tested with regard to their accuracy on predicting footprints using the area under the receiver operating characteristic (ROC) curve (AUC) at 100% false positive rate (FPR) on the `Comprehensive Dataset`. Since we are evaluating the impact on performance of different data types (DNase-seq, different combinations of histone modifications ChIP-seq and different combinations of both DNase-seq and histone modifications) we opted to use the bit-score of the motif-predicted binding sites (MPBSs) as the ranking metric to create the receiver operating characteristics (ROC) curves. The rationale for the usage of such ranking strategy is that it is independent of the experimental open chromatin datasets. This is necessary because: (1) this ranking strategy does not bias the analysis towards specific HMM topologies/input data type and (2) it makes the analysis interpretation simpler. In this scenario, we divided the gold standard dataset into two groups: the MPBSs that contain at least one base pair (bp) overlap with HINT's predicted footprints and the ones that do not overlap. Both groups were sorted based on the motif matching bit-score. A single list is then obtained by combining the ranked list of predicted sites before the ranked list of the non-predicted sites. The ROC curve is created based on this combined ranked list of MPBSs. These analyses were performed using data from chromosome 1 only, which was removed from the comparative analyses (in Section 5.3) to allow a fair comparison.

5.1.1. HMM Topology

We investigated the different HMM topologies presented in this thesis (Section 3.2.2). The empirical test consists on evaluating the AUC using the `Comprehensive Dataset` from the ChIP-seq evaluation approach. The distribution of the 233 AUCs at 100% FPR (one for each TF from the `Comprehensive Dataset`) for all the HMM topologies can be seen in Figure 5.1. All the HMM topologies that use histone modification data were tested using a combination of H3K4me1 + H3K4me3.

We are able to observe in Figure 5.1 that the ORIGINAL DNASE + HISTONE provides higher accuracies than all other models. Furthermore, the HISTONE-ONLY MODEL provides the lowest accuracies. This relates to the fact that the HISTONE-ONLY MODEL'S footprints are large and do not capture the spatial specificity provided by the higher-resolution DNase-seq data.

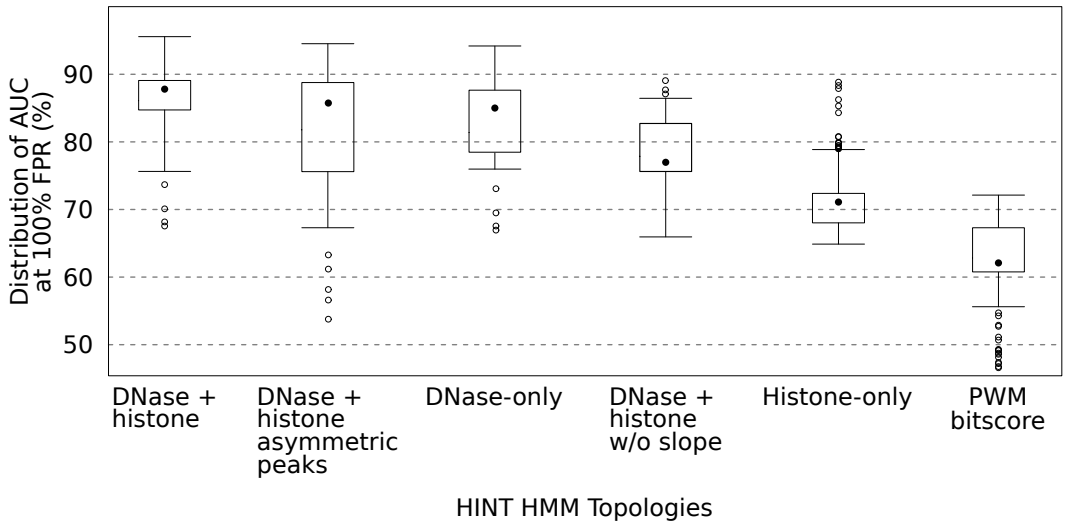

Figure 5.1: Performance of different HINT HMM topologies. Distribution of AUC at 100% FPR of the ROC curves generated using the ChIP-seq evaluation `Comprehensive Dataset` on different HINT HMM topologies. The histone modification combination H3K4me1 + H3K4me3 was used on the HMM topologies which use such data. *Source: Gusmao et al. (2014)* (modified to fit thesis format and/or clarify key points).

We performed a Friedman-Nemenyi test on the distribution of accuracies from different HMM topologies to assess statistical significance. The results can be seen in Table 5.1. We observed that indeed the ORIGINAL DNASE + HISTONE MODEL significantly outperforms all other topologies. This analysis shows that the proper integration of DNase-seq and histone modifications results in significantly higher accuracies than using each of these data separately.

The possible reason for the good results of the ORIGINAL DNASE + HISTONE topology in relation to the DNASE + HISTONE ASYMMETRIC peaks topology is the fact that the normalization methodology emphasizes even little increases in histone modification levels, leveraging the asymmetry issue (Figure 5.2). Furthermore, the poor performance of the DNASE + HISTONE WITHOUT SLOPE topology in comparison to the other DNASE + HISTONE topologies indicates that even with more complex models (4 *vs* 2 variables and 8 *vs* 4 states, respectively) the slope signal and the additional states are crucial in the accurate delineation of the footprints.

It is interesting to observe that the DNASE-ONLY HMM topology provides high accuracies (significantly better than the DNASE + HISTONE WITHOUT SLOPE and HISTONE-ONLY topologies) and is competitive with the ORIGINAL DNASE + HISTONE and ASYMMETRIC PEAKS DNASE + HISTONE topologies. This fact shows the power of the DNase-seq data on identifying TF footprints.

Table 5.1: Friedman-Nemenyi test on different HINT HMM topologies. Friedman-Nemenyi hypothesis test results on AUC at 100% FPR of the ROC curves generated using the ChIP-seq evaluation `Benchmarking Dataset` on different HINT HMM topologies. The asterisk and the cross, respectively, indicate that the method in the column outperformed the method in the row with significance levels of 0.01 and 0.05. *Source: Gusmao et al. (2014)* (modified to fit thesis format and/or clarify key points).

	Original DNase+histone	DNase+histone asymmetric	DNase-only	DNase+histone w/o slope	Histone-only
Original DNase+histone					
DNase+histone asymmetric	*				
DNase-only	*				
DNase+histone w/o slope	*	*	*		
Histone-only	*	*	*	*	

On the other hand, although the HISTONE-ONLY topology presents significantly lower accuracies than all other topologies (median AUC = 71%), it is still a better choice than purely sequence-based approaches (median AUC = 62%).

5.1.2. Combination of Histone Modifications

We have performed an empirical test on the predictive power of combining different histone modifications using HINT's ORIGINAL DNASE + HISTONE HMM topology. In this test we considered the histone modifications H3K4me1, H3K4me3, H3K9ac, H3K27ac and the histone variant H2A.Z (referred to with the term "histone modification", for simplicity of notation). The presence of such histones modifications are associated with active regulatory regions. We compared the performance using each of these histone modifications individually (5 prediction sets, one for each histone mod-

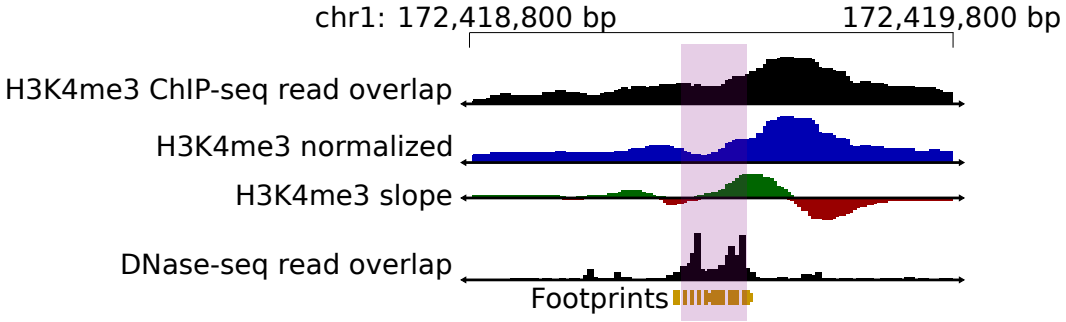

Figure 5.2: Histone modification asymmetry example. H3K4me3 ChIP-seq raw count (top black), normalized (blue) and slope (green for positive and red for negative) signals; DNase-seq count signal (bottom black) and footprints (yellow) predicted in a region with asymmetrical histone modification profile. Although the leftmost peak from the peak-dip-peak pattern of the H3K4me3 contains very small count signals (very close to the background signal, in this region), ORIGINAL DNASE + HISTONE HINT topology is still capable of predicting the footprints within the DNase hypersensitivity site (DHS). Note that after the normalization of the H3K4me3 counts, the resulting signal delineates the DHS more clearly (depicted by the purple box). *Source: Gusmao et al. (2014)* (modified to fit thesis format and/or clarify key points).

ification). We also evaluated the combination of all pairs and triples of histone modification signals by simply merging all predicted sites (as described in Section 4.1.3). Such combinatorial analysis generates 20 additional prediction sets (10 pairs and 10 triples). Note that extending to further combinations would deviate from one of the main goals of this study, which is to create a consistent map of active transcription factor binding sites (TFBSs) with few genome-wide assays. We tested all the 25 combinations using the ChIP-seq evaluation `Comprehensive Dataset` gold standard.

Figure 5.3 presents the distribution of AUCs for all histone modification models tested plus the DNASE-ONLY HMM topology for comparison purpose. We observed that most methods presented the region between the first and third quartiles approximately between AUCs 80%–95%. In order to test the statistical relevance of these differences, we performed a Friedman-Nemenyi test. The Table 5.2 shows the accuracy ranking for all histone modification combinations in decreasing order, providing information on which models significantly outperformed others.

Overall, results indicate that combinations with more histone modifications are better than single-histone models. Several combinations of three marks (H3K4me1+H3K4me3+H3K9ac, H3K4me1+H3K4me3 +H3K27ac, H2A.Z+H3K4me1+H3K4me3, H2A.Z+H3K4me3+H3K9ac and H3K4me3+H3K9ac+H3K27ac) were similarly good, i.e. their AUC are not significantly lower than any other combination. Similarly, if we only consider individual and pairs of histone modification, H3K4me1+H3K4me3, H3K4me3+H3K9ac, H3K4me3+H3K27ac, H2A.Z+H3K4me3 and H3K4me1+H3K9ac have similar AUCs. This indicates that any combination of these histone modifications, whenever available, would perform equally well. Nevertheless, we observed that the DNASE-ONLY topology outperforms a few DNASE + HISTONE modification combinations significantly. This represents further evidence of the importance of such high-resolution signal and might explain previous failed attempts to improve the accuracy of TF predictions by introducing histone modifications individually (Pique-Regi et al., 2011; Cuellar-Partida et al., 2012).

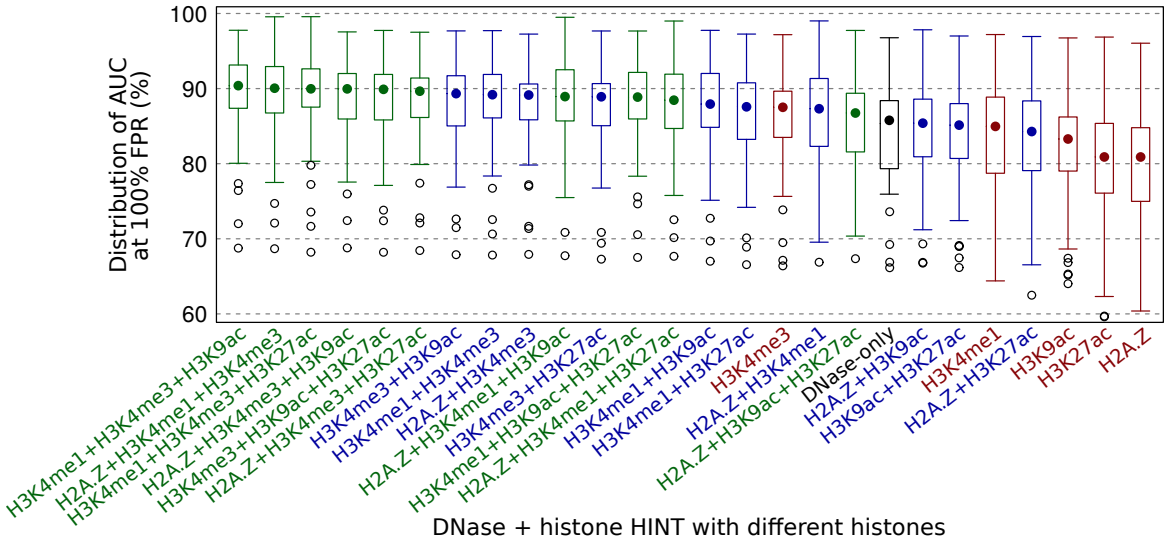

Figure 5.3: Performance of different histone modification combinations. Distribution of AUC at 100% FPR of the ROC curves generated using the ChIP-seq evaluation `Benchmarking Dataset` on DNASE + HISTONE HINT topology with different combinations of histone modifications. We show single, pairs and trios of histone modifications in red, blue and green, respectively. The DNASE-ONLY accuracy is also shown (in black) for comparison purpose. *Source: Gusmao et al. (2014)* (modified to fit thesis format and/or clarify key points).

Table 5.2: Friedman-Nemenyi test on different histone modification combinations. Friedman-Nemenyi hypothesis test results on AUC at 100% FPR of the ROC curves generated using the ChIP-seq evaluation `Benchmarking Dataset` on DNASE + HISTONE HINT topology with different combinations of histone modifications. We show single, pairs and trios of histone modifications in shades of red, blue and green, respectively. In addition, we also show the DNASE-ONLY HMM topology for comparison purpose (gray). The asterisk and the cross, respectively, indicate that the method in the column outperformed the method in the row with significance levels of 0.01 and 0.05. *Source: Gusmao et al. (2014)* (modified to fit thesis format and/or clarify key points).

	H3K4me1+H3K4me3+H3K9ac	H3K4me1+H3K4me3+H3K27ac	H2A.Z+H3K4me1+H3K4me3	H2A.Z+H3K4me3+H3K9ac	H3K4me3+H3K9ac+H3K27ac	H2A.Z+H3K4me3+H3K27ac	H3K4me1+H3K4me3	H3K4me1+H3K9ac+H3K27ac	H2A.Z+H3K4me1+H3K9ac	H3K4me3+H3K9ac	H2A.Z+H3K4me1+H3K27ac	H3K4me3+H3K27ac	H2A.Z+H3K4me3	H3K4me1+H3K9ac	H3K4me1+H3K27ac	H2A.Z+H3K4me1	H2A.Z+H3K9ac+H3K27ac	H3K4me3	DNase-only	H2A.Z+H3K9ac	H3K9ac+H3K27ac	H3K4me1	H2A.Z+H3K27ac	H3K9ac	H3K27ac	H2A.Z
H3K4me1+H3K4me3+H3K9ac																										
H3K4me1+H3K4me3+H3K27ac																										
H2A.Z+H3K4me1+H3K4me3																										
H2A.Z+H3K4me3+H3K9ac																										
H3K4me3+H3K9ac+H3K27ac																										
H2A.Z+H3K4me3+H3K27ac	*																									
H3K4me1+H3K4me3	*																									
H3K4me1+H3K9ac+H3K27ac	*	+																								
H2A.Z+H3K4me1+H3K9ac	*	*	*																							
H3K4me3+H3K9ac	*	*	*																							
H2A.Z+H3K4me1+H3K27ac	*	*	*	*	+																					
H3K4me3+H3K27ac	*	*	*	*	*																					
H2A.Z+H3K4me3	*	*	*	*	*																					
H3K4me1+H3K9ac	*	*	*	*	*	*																				
H3K4me1+H3K27ac	*	*	*	*	*	*	*																			
H2A.Z+H3K4me1	*	*	*	*	*	*	*	*																		
H2A.Z+H3K9ac+H3K27ac	*	*	*	*	*	*	*	*	*																	
H3K4me3	*	*	*	*	*	*	*	*	*	*																
DNase-only	*	*	*	*	*	*	*	*	*	*	*															
H2A.Z+H3K9ac	*	*	*	*	*	*	*	*	*	*	*	*														
H3K9ac+H3K27ac	*	*	*	*	*	*	*	*	*	*	*	*	*													
H3K4me1	*	*	*	*	*	*	*	*	*	*	*	*	*	*												
H2A.Z+H3K27ac	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*											
H3K9ac	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*										
H3K27ac	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*									
H2A.Z	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*								

5.1.3. HMM Training

The annotation of certain genomic regions with the HMM states in order to train HINT is laborious. Therefore, we decided to analyze whether HINT's performance is impacted by training and applying the method to data from different cell types. In this particular empirical test, we used `Comprehensive Dataset`'s evaluation data for four cell types: H1-hESC (29 TFs), HeLa-S3 (20 TFs), HepG2 (21 TFs) and K562 (59 TFs).

We have compared the AUC values of HINT when it was trained in a particular cell type and executed in the same cell type it was trained *vs* the other three cell types. The Friedman-Nemenyi test

was applied to assess statistical significance.

Figure 5.4 shows the results for all models applied to all cell types tested. Each set of four boxplots represent one of the four HINT models (trained with data from one of the four cell types), which was applied to the signal generated from the cell type labeled on the bottom of the set. Statistical significance are directly represented in the graph.

We can observe in Figure 5.4 that only in one out of twelve cases the AUC levels are significantly different (p-value < 0.05). This corresponds to the HepG2 model, when used to generate footprints in the same cell type and in K562 cell type. These results suggest that our signal processing workflow and HMM model are able to robustly mitigate differences between distinct cell type signals. Consequently, we can consider HINT cell-type training-independent. The practical implication of such an important characteristic is that a simple application of a model already stored in our software tool, trained for a particular cell type, is sufficient to generate accurate predictions for any other cell type, without the need to re-train the model. Furthermore, this is evidence that the patterns that make the grammar of active TFBSs are similar between different cells.

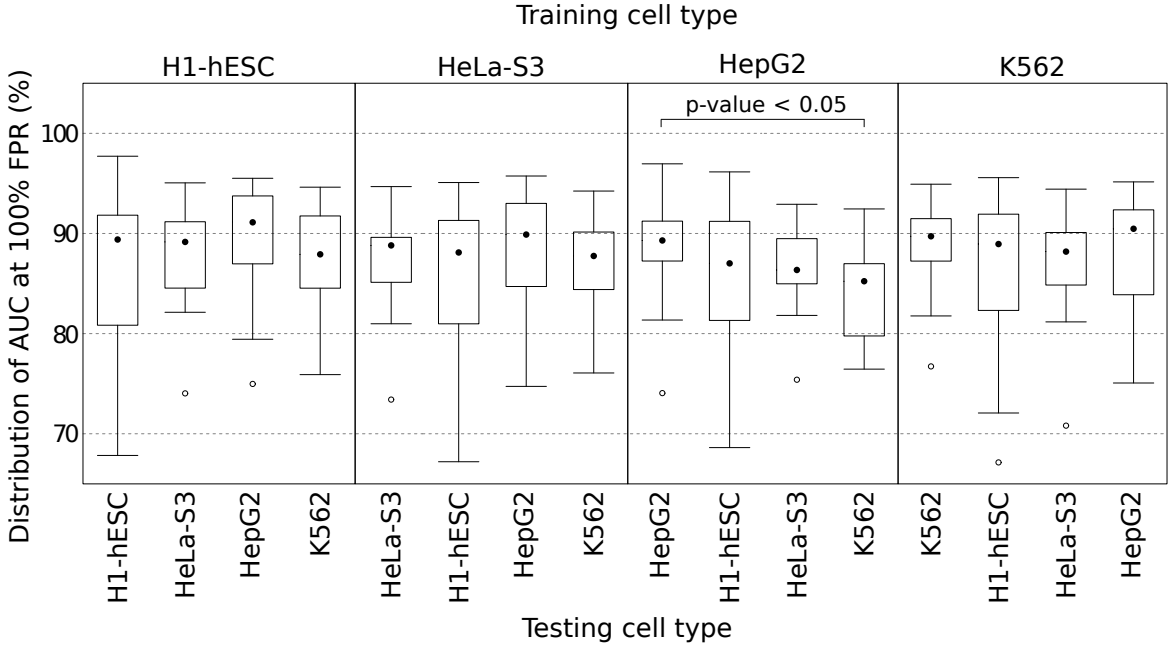

Figure 5.4: Performance of different HMM training/testing scenarios. Distribution of AUC at 100% FPR of the ROC curves generated using the ChIP-seq evaluation `Benchmarking Dataset` on HINT models trained in four different cell types (top x-axis labels) and applied to (tested on) the same four cell types (bottom x-axis labels). The first boxplot within each set represents the model trained in the same cell type as the one it was applied to. The significant Friedman-Nemenyi test p-values are shown on the top of the boxplot. *Source: Gusmao et al. (2014)* (modified to fit thesis format and/or clarify key points).

5.2 Footprint Scoring and Sequence Cleavage Bias Correction

In this section we focused on two important challenges regarding computational footprinting methods. First, we performed a series of empirical tests on HINT and competing methods to identify an optimal footprint scoring metric (Section 5.2.1). We investigated such optimal scoring metric for both

evaluation methodologies proposed (ChIP-seq and gene expression). Second, we investigated the impact of the DNase-seq sequence cleavage bias on computational footprinting and whether such bias could be corrected to improve the performance of HINT (Section 5.2.2). All DNase-seq sequence cleavage bias analyses used only the ChIP-seq evaluation strategy, since we require the TF-wise results provided by this evaluation strategy.

Regarding the ChIP-seq evaluation analyses presented in this section; since such experiments required a few comparisons between HINT and other competing methods which only used DNase-seq data as input, we opted to use the DNASE-ONLY HMM topology to provide a fairer comparison. In analyses that involved HINT and competing methods we used the `Benchmarking Dataset` as gold standard; while in analyses that involved only HINT we used the full `Comprehensive Dataset`. Furthermore, we decided to use the AUC at 10% FPR to capture more subtle differences in reported accuracies. These analyses were performed using data from chromosome 1 only, which was removed from the comparative analyses (in Section 5.3) to allow a fair comparison.

5.2.1. Footprint Ranking Strategy

Some competing footprinting methods also provide statistics to rank footprint predictions. Wellington and DNase2TF use read count statistics to provide p-values for each footprint. Several site-centric approaches provide either probabilities (BinDNase, Centipede and PIQ) or log-odds scores (FLR) of footprints. Other methods use statistics such as the footprints score (Neph) or position weight matrix (PWM) bit-score (Cuellar), to rank predicted footprints.

The main goal of this empirical analysis is to identify the best scoring metric for footprint predictions. For that, we used the TF-wise ChIP-seq evaluation approach to search for such scoring metric by performing an empirical test on the accuracy of HINT and competing methods. Since the gene expression evaluation differs in nature with regard to the scoring metric (footprint quality score), we also evaluated the best scoring metric using such evaluation approach.

ChIP-seq Evaluation

The ChIP-seq evaluation scheme requires a metric to rank the footprint predictions. Since the ChIP-seq evaluation scheme is calculated in a TF-wise manner, we can regard an optimal ranking score to create the ROC curve as the best footprint ranking scheme.

We investigated three footprint scoring metrics: the tag count (TC), the footprint score (FS) and the PWM bit-score (PWM). We assigned a quality score for each footprint predicted using the DNASE-ONLY HINT method. The assignment of the TC and FS to each footprint can be performed straightforwardly using the DNase-seq data. Regarding the PWM metric, each footprint was assigned to the bit-score of its overlapping MPBS. The PWM score assignment was performed as a "control" experiment, since it requires MPBSs, which are only available for known TFs.

The test consists on ranking the footprints by each of these three different metrics and creating the ROC curves based on each different ranking. Figure 5.5 shows the distribution of the AUC at 10% FPR for different footprint ranking strategies using HINT's footprint predictions. The statistical significance assessment in the graph corresponds to the Friedman-Nemenyi test.

We are able to observe that the TC is the best footprint ranking strategy (average AUC $= 90\%$), outperforming both FS and PWM (p-value < 0.01). Furthermore, it is clear that both FS and PWM have significantly lower accuracies than the TC strategy (average AUC $= 50\%$ and 37%, respectively; p-value < 0.01). It is clear that the PWM bit-score is the worst scoring metric, since it does not use the cell-specific open chromatin information provided by the DNase-seq data. However, the success of the TC metric in comparison to the FS is not so straightforward. The FS is defined as a ratio between the DNase-seq at the center of the footprint and its flanking regions. The first problem that the FS encounters is that it does not recover the absolute signal intensity, as the simpler TC metric

does. The second problem of the FS metric regards the window length in which the average signal in the center of the motif and flanking regions are calculated. This issue is similar to the one discussed for window-based segmentation methods (Section 2.4.1). We believe these issues are related to the higher observed accuracies for the TC metric.

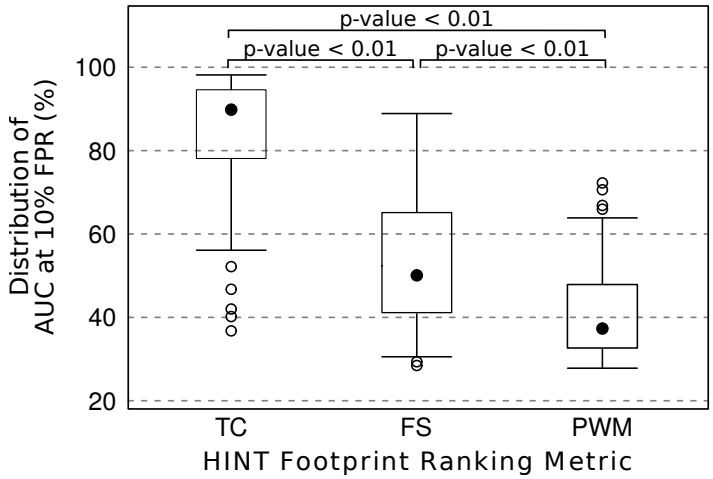

Figure 5.5: Performance of different footprint ranking strategies on HINT Distribution of AUC at 10% FPR of the ROC curves generated using the ChIP-seq evaluation `Comprehensive Dataset` on different footprint ranking strategies on HINT. The significant Friedman-Nemenyi test p-values are shown on the top of the boxplot. *Source: Gusmao et al. (2016)* (modified to fit thesis format and/or clarify key points).

Given its good performance, we evaluated the use of TC as the ranking strategy instead of each method's own ranking for the competing methods that present an intrinsic footprint scoring metric: BinDNase, Centipede, Cuellar, DNase2TF, FLR, PIQ and Wellington. Previous to ranking by TC, site-centric methods required the definition of a minimum probability score to define active footprints. We tested the probability cutoff thresholds of $80\%, 85\%, 90\%, 95\%$ and 99% for the site-centric methods. The results can be seen in Figure 5.6. In all cases, using TC-based strategies/cutoff was significantly better than the methods original ranking (p-value < 0.01; Friedman-Nemenyi test). Concerning site-centric methods, the use of a probability threshold of 90% was best for all methods except BinDNase, where 80% was best.

Given the results obtained in these empirical analyses, we selected the TC as the best footprint ranking metric. Furthermore, the TC is used for HINT and all competing methods with regard to the ChIP-seq evaluation approach on our comparative study.

Gene Expression Evaluation

The gene expression evaluation consists on correlating differences in gene expression with a footprint quality score between two different cell types. Such correlation is termed FP-Exp. In this analysis, we evaluated three footprint quality scores: the TC, the FS and the footprint likelihood ratio (FLR) metric as suggested by Yardımcı et al. (2014). Figure 5.7 shows a selection of graphs that exhibit the correlation between gene expression fold change (FC) and the Kolmogorov-Smirnov (KS) statistic applied to the difference on the distribution between footprint quality scores for 143 evaluated TFs. In this figure, we are able to observe that the FLR score presents high correlations ($r > 0.9$); while the TC metric presents low correlations ($r < 0.4$) and several cases in which the signal of KS and fold change disagree (off diagonal points).

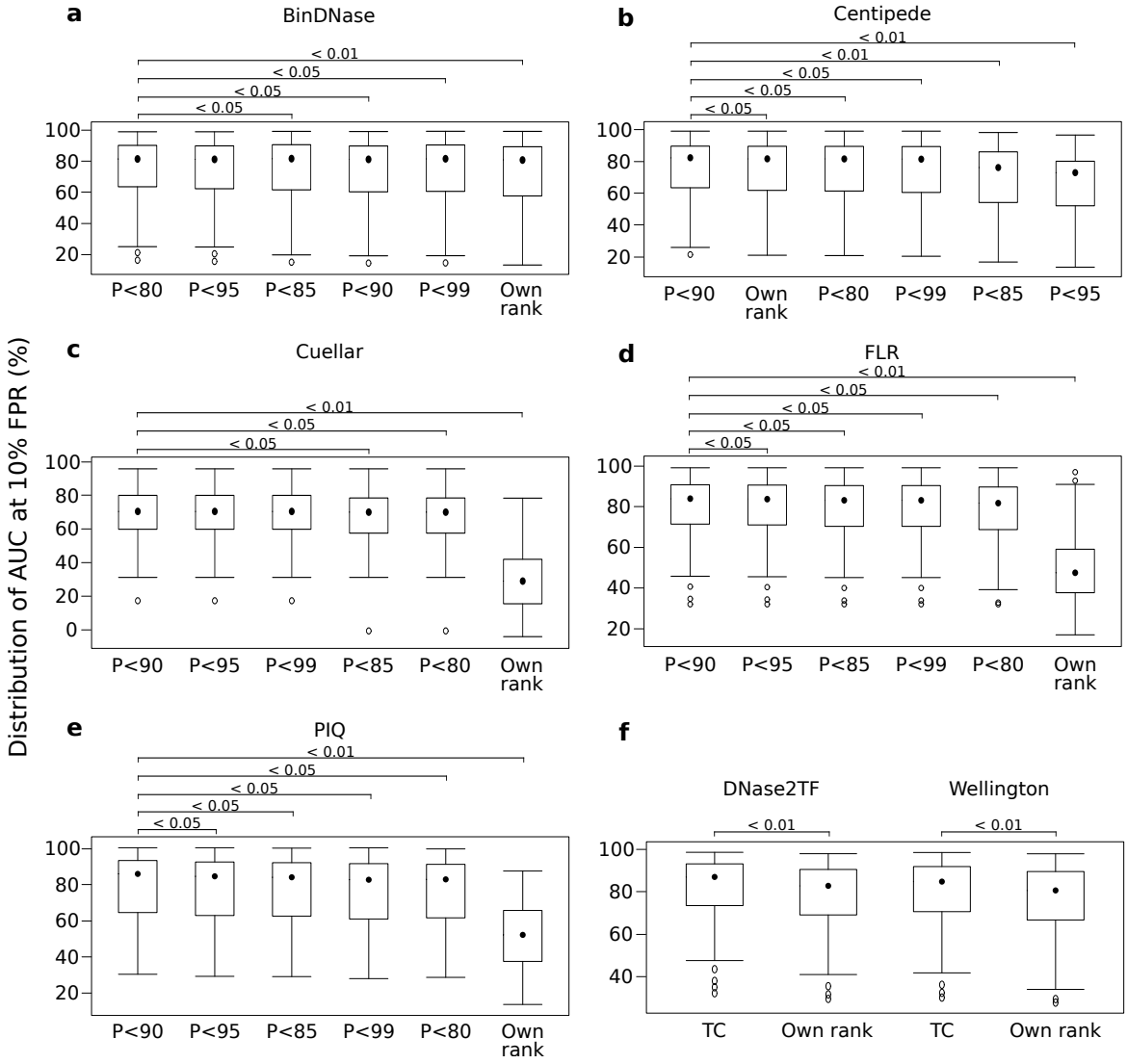

Figure 5.6: TC *vs* competing method's own ranking strategy. Distribution of AUC values (10% FPR) by using distinct ranking strategies for site-centric methods (**a**) BinDNase, (**b**) Centipede, (**c**) Cuellar, (**d**) FLR, (**e**) PIQ and (**f**) segmentation methods DNase2TF and Wellington. Probability cutoff thresholds of $80\%, 85\%, 90\%, 95\%$ and 99% were used for the site-centric methods ranked with the TC metric. Ranking strategies (x-axis) are ranked by decreasing median AUC. The significant Friedman-Nemenyi test p-values are shown on the top of each boxplot. *Source: Gusmao et al. (2016)* (modified to fit thesis format and/or clarify key points).

To investigate the footprint quality scores on the gene expression evaluation more thoroughly we generated the distribution of the FP-Exp using each of the tested footprint quality score metrics on all computational footprinting methods and all cell type pair combinations possible within the cell types GM12878, H1-hESC and K562 (Figure 5.8). Furthermore, to assess statistical significance we performed a Friedman-Nemenyi hypothesis test.

We observed that the FLR metric results in higher FP-Exp scores (average FP-Exp $= 0.79$) and significantly outperforms the results generated with the other footprint quality scores (p-value < 0.05 for FS and p-value < 0.01 for TC). The FS presented a lower average FP-Exp ($= 0.73$) than the FLR metric; however significantly outperformed the TC metric (p-value < 0.01). We also observed that the ranking of methods by FP-Exp using FLR metric and FS are very similar ($r = 0.89$). Moreover,

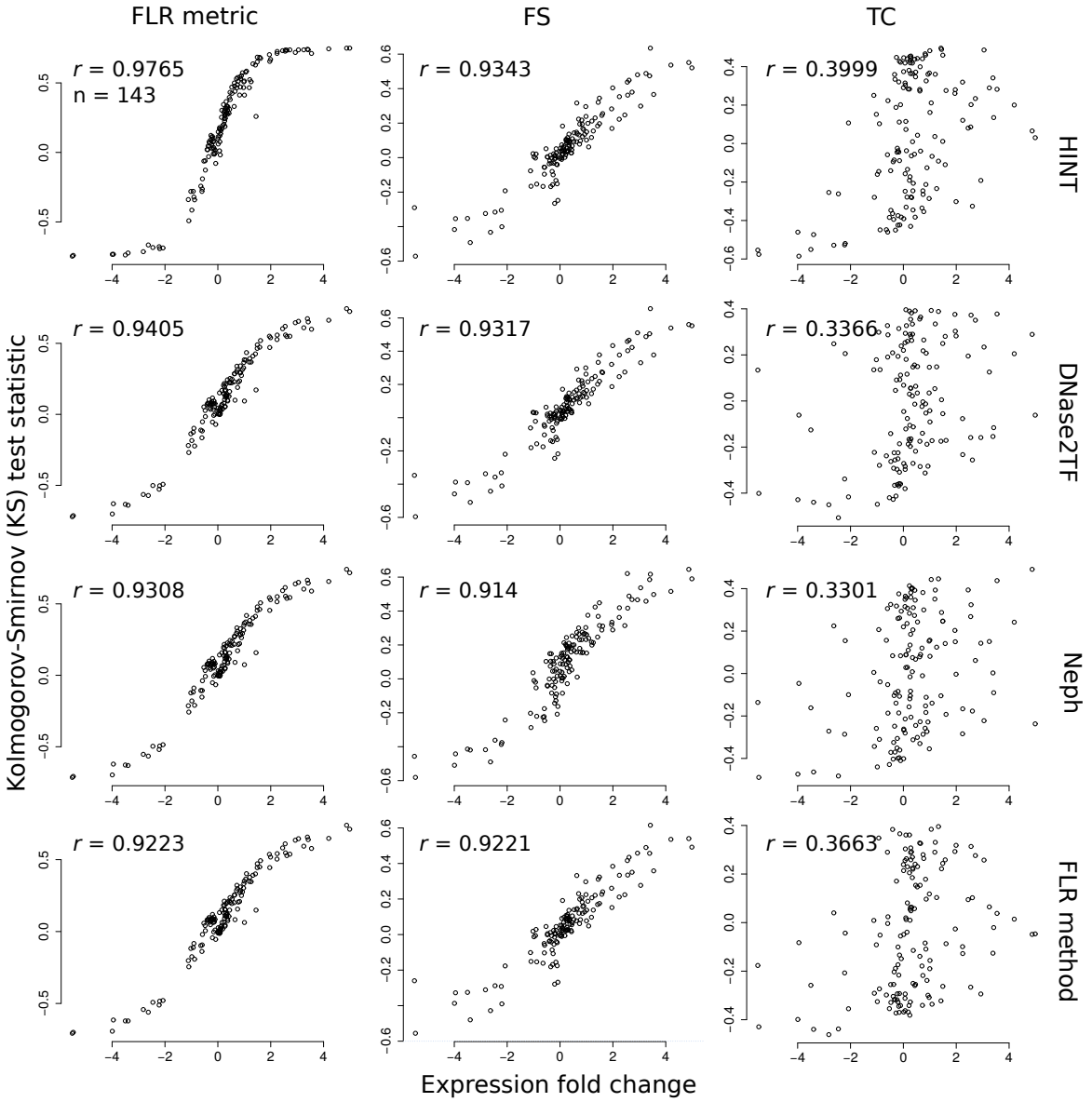

Figure 5.7: Correlation between KS statistic and FC expression for different scoring metrics. Correlation between KS statistic *vs* fold change expression for cell type pair H1-hESC *vs* K562 by evaluating either the FLR metric (left), FS (middle) and TC (right) as quality metric for the footprints. Footprints were predicted with HINT, DNase2TF, Neph and FLR (from top to bottom, respectively). *Source: Gusmao et al. (2016)* (modified to fit thesis format and/or clarify key points).

differently from what was observed for the ChIP-seq evaluation approach, the TC presents the lowest FP-Exp scores (average FP-Exp $= 0.35$).

Given these results, we opt to use the FLR metric as the footprint quality score for our comparative study with regard to the gene expression evaluation approach. However, we point that the FS metric can be used as an alternative footprint quality score for the gene expression evaluation procedure given its simplicity and similar accuracies to FLR metric.

5.2.2. Impact of DNase-seq Sequence Cleavage Bias

In this section we investigate the impact of the DNase-seq sequence cleavage bias on the performance of HINT and whether the correction of such bias improves the footprint prediction accuracy.

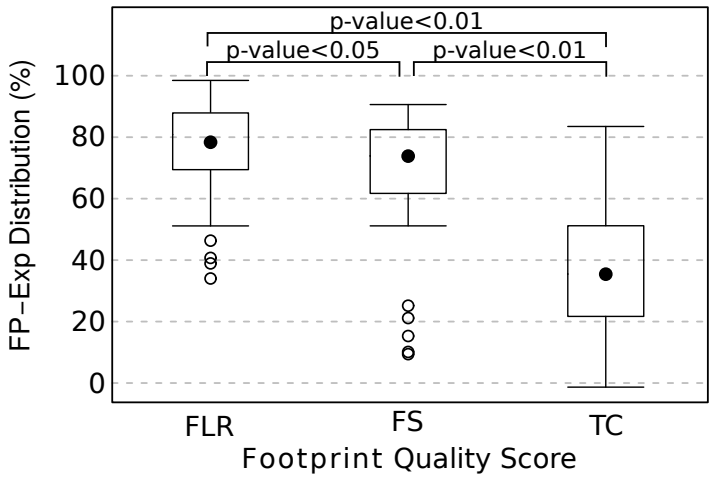

Figure 5.8: Performance of different FP-Exp footprint quality scores Distribution of the FP-Exp scores generated using the gene expression evaluation with different footprint quality score metrics. The distributions are based on the values from the three cell type combinations and all the competing methods used. The significant Friedman-Nemenyi test *p*-values are shown on the top of the boxplot. *Source: Gusmao et al. (2016)* (modified to fit thesis format and/or clarify key points).

We tested the two approaches described in Section 4.1.2: the "DHS sequence bias" and the "naked deoxyribonucleic acid (DNA) sequence bias". The DHS sequence bias considers the sequence bias estimates within DNase hypersensitivity sites (DHSs) of each DNase-seq experiment. This approach captures DNase I cleavage bias, read fragmentation and sequence complexity bias of DHSs of each DNase-seq experiment. The naked DNA sequence bias considers the sequence bias estimates within naked DNA DNase-seq experiments. In this case, all DNA regions are open, therefore the sequence bias estimates will mainly capture the DNase I cleavage bias.

Throughout this section we used the following HINT variations: (1) HINT was applied without any DNase-seq sequence bias correction (HINT w/o BC), (2) HINT was applied with the DHS sequence bias correction approach (HINT bias-corrected; HINT-BC) and (3) HINT was applied with the naked DNA sequence bias correction approach (HINT bias-corrected on naked DNase-seq; HINT-BCN). This nomenclature will be used within this section.

DNase-seq Sequence Cleavage Bias is Protocol-Specific

First, to understand the nature of artifacts on DNase-seq experiments, we analyzed the DNase-seq sequence cleavage bias estimates on the `Full Dataset`, i.e. all 61 Tier 1 and Tier 2 DNase-seq datasets from ENCODE Project Consortium (2012) (Supplementary Table A.1). The sequence cleavage bias corresponds to the 6-mer estimations as shown in Equation 3.7. These experiments include two existing DNase-seq protocols: the single-hit and double-hit techniques. For every DNase-seq dataset we calculated the 6-mer bias estimates for the cell-specific DHS sequence bias. Furthermore, we also included the naked DNA sequence bias estimates in this analysis from naked DNA DNase-seq experiments for three cell types. A clustering analysis of the correlation between the pairwise 6-mer sequence bias estimates forms two clear groups, which splits experiments from single-hit and double-hit protocols (Figure 5.9). This indicates that sequence biases are protocol-specific. Naked DNA sequence bias estimates forms a sub-cluster within estimates from the double-hit experiments.

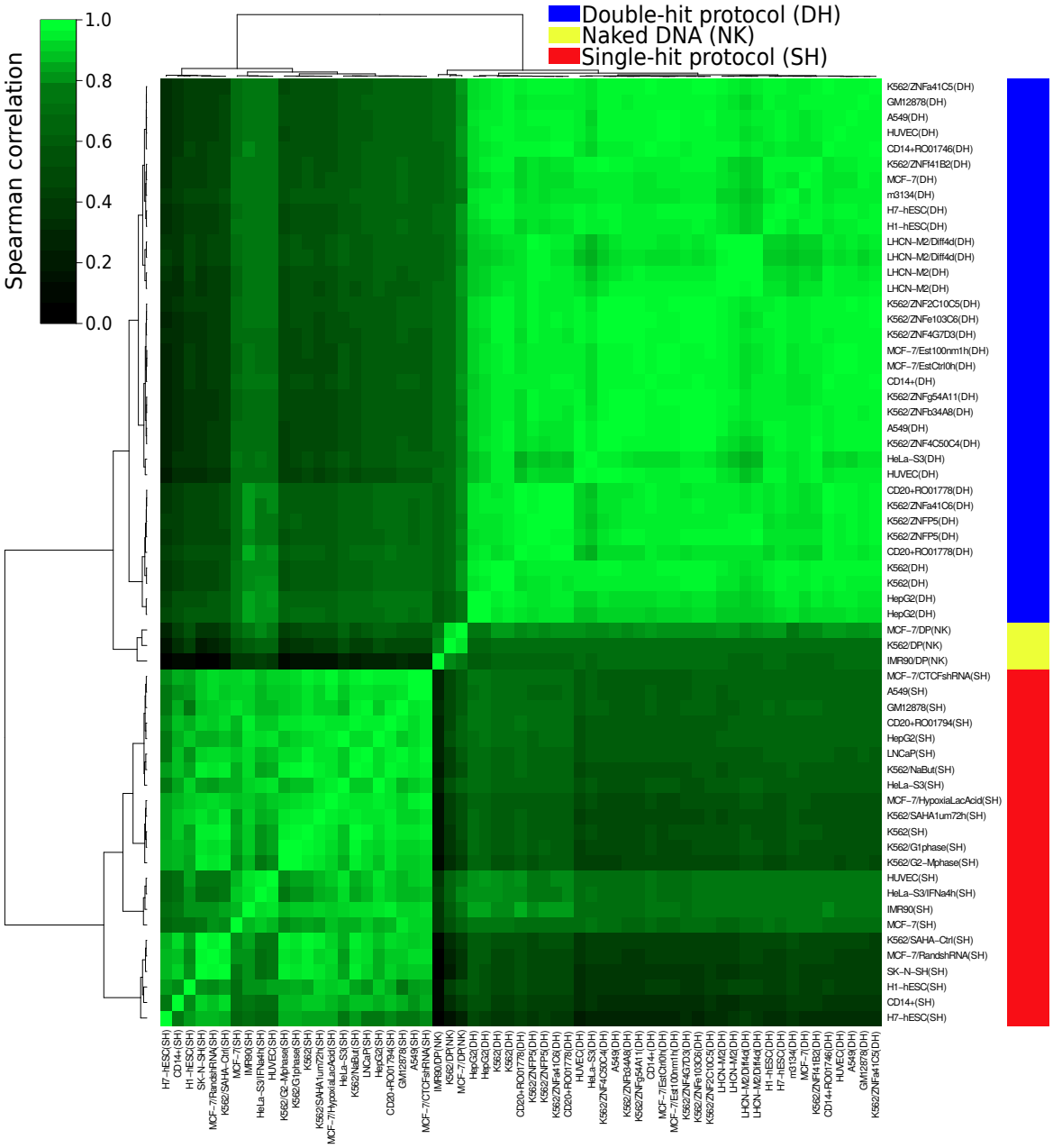

Figure 5.9: Clustering of bias estimates. Ward's minimum variance clustering based on pairwise Spearman correlation coefficient (r) from bias estimates (all possible 6-mers within the DNA alphabet $\{A,C,G,T\}$) of all encyclopedia of DNA elements (ENCODE) DNase-seq data and three naked DNA DNase-seq data obtained from different sources. DNase-seq experiments were based on single-hit (red), double-hit (blue) protocols or naked DNA (yellow). *Source: Gusmao et al. (2016)* (modified to fit thesis format and/or clarify key points).

Impact of the DNase-seq Sequence Cleavage bias on the Accuracy of Computational Footprinting Methods

Next, we evaluated the influence of sequence bias on all footprinting methods. In this analysis we plot, for all computational footprinting methods, the TF-wise amount of bias *vs* the TF-wise footprint prediction accuracy (Figure 5.10a). The amount of bias is calculated as the correlation between the uncorrected DNase-seq signal and the bias signal (Equation 3.7). Such correlation between observed

DNase-seq and predicted bias signal is called "OBS" (observed *vs* bias signal). The footprint prediction accuracy is measured through the ChIP-seq evaluation approach using the `Benchmarking Dataset`.

Our analysis shows that only six out of 14 evaluated methods (Wellington, Neph, Boyle, DNase2TF, Centipede and FS-Rank) present a significant negative Spearman correlation ($r = -0.35$, -0.32, -0.28, -0.28, -0.24 and -0.22, respectively) between their accuracy performance and amount of sequence bias (Figure 5.10a; p-value < 0.05). Methods explicitly using 6-mer sequence bias statistics (HINT-BC, HINT-BCN and FLR) or performing smoothing (Cuellar, BinDNase and PIQ) are not significantly influenced by sequence bias. Moreover, the performance of HINT-BC is the least affected by sequence bias ($r = -0.06$).

Nevertheless, we noticed an increase in accuracy for bias-corrected versions of HINT. The reason for such accuracy increase became clear when we examined the DNase-seq average signals, bias-corrected and without correction, surrounding active TFBSs. As an example, we show sequence bias estimates, corrected and uncorrected DNase-seq average profiles around TFBSs with the highest AUC gain between HINT-BC and HINT w/o BC (Figure 5.10b–c). The NRF1 and EGR1 DNase-seq profiles indicate that the bias-corrected signal fits better their sequence affinity than the uncorrected signal. This means that the higher-affinity parts of NRF1's and EGR1's motif are located in the regions with lowest DNase-seq cleavage. As a consequence, the distinctive pattern of active TF binding (i.e. grammar of active TFBSs) are more clearly recognizable in the bias-corrected DNase-seq signals than in the non-corrected DNase-seq signals.

Statistical Evaluation of DNase-seq Sequence Cleavage Bias Correction Strategies

Using the same experimental settings as explained in Figure 5.10, we have investigated more thoroughly the best DNase-seq sequence cleavage bias correction strategy. For that, we calculated the distribution of the AUC at 10% FPR for HINT w/o BC, HINT-BC and HINT-BCN (Figure 5.11). Furthermore, we performed a Friedman-Nemenyi hypothesis test on these three HINT scenarios.

By analyzing these results, we are able to observe that, although HINT-BC presents slightly higher accuracies than HINT-BCN, these differences are not statistically significant. However, we are able to observe that the accuracies of the HINT-BC strategy significantly outperforms the HINT w/o BC. This is a strong indication that the DNase-seq sequence cleavage bias correction improves the performance of computational footprinting methods. This is an important result since no previous computational footprinting method that treated DNase-seq sequence cleavage bias observed significant gain in accuracy (Yardımcı et al., 2014; Sung et al., 2014; Kähärä and Lähdesmäki, 2015). Given these results, the DHS sequence cleavage bias correction is the strategy of choice for the application of HINT in all other sections of this chapter.

Evaluation of whether the DNase-seq Sequence Cleavage Bias Correction is an Artifact of the Genomic Nucleotide Frequency Distribution

Since the TFs have a sequence binding affinity preference, which in many cases is composed of C and G nucleotides, the DNase-seq sequence cleavage bias correction could be simply creating "artifact" peak-dip-peak patterns on the DNase-seq data, since the DNase I enzyme also has a preference to bind CG-rich motifs.

To investigate such claim we calculated the distribution of the pairwise AUC differences between the three HINT versions (HINT w/o BC, HINT-BC and HINT-BCN) for all TFs of the `Comprehensive Dataset` gold standard (Figure5.12a). Furthermore, we also calculated the CG content of these TF motifs (Figure 5.12b).

By analyzing Figure 5.12a–b we observe no correlation between CG content of the motifs and the individual AUC of each method: HINT (w/o BC) $r = -0.0144$, HINT-BC $r = 0.0254$ and HINT-

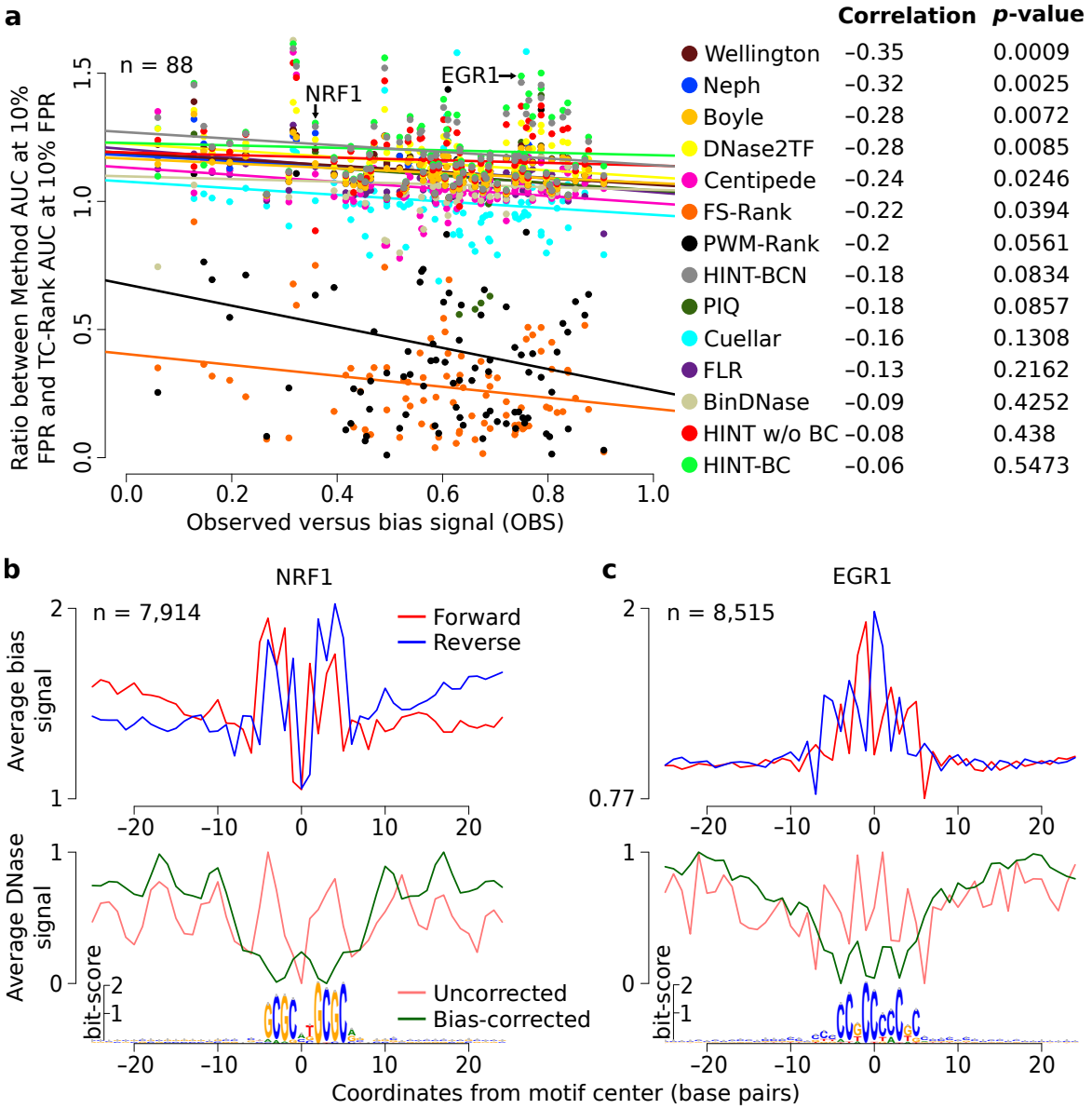

Figure 5.10: Effects of DNase I sequence cleavage biases on computational footprinting methods. (**a**) Association between the performance of footprinting methods (relative to TC-Rank performance) and their sequence bias estimated for the TF if the `Benchmarking Dataset`. The *x*-axis represents the correlation between the uncorrected and bias signal (observed *vs* bias signal; OBS). The OBS is calculated for each TF by measuring the uncorrected DNase-seq signal and the bias signal for every MPBS that overlaps a footprint from the evaluated method. Then, the Spearman correlation is calculated between the average uncorrected and bias signals. Higher OBS values indicate higher bias. The *y*-axis represents the ratio between the AUC at 10% FPR for each evaluated method and the TC-Rank method; higher values indicate higher accuracy. (**b**–**c**) Average bias signal (top) and uncorrected/bias-corrected DNase-seq signal (bottom) for the TFs: (**b**) NRF1 and (**c**) EGR1. Signals in the top graph are DNA strand-specific (forward strand in red and reverse strand in blue). Signals in the bottom graph were standardized to be in the interval $[0, 1]$. The motif logo represents all underlying DNA sequences centered on the TFBSs. *Source: Gusmao et al. (2016)* (modified to fit thesis format and/or clarify key points).

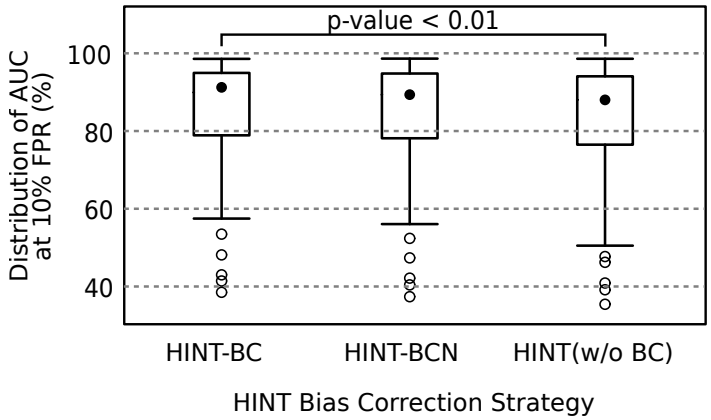

Figure 5.11: Performance of different bias correction strategies. Distribution of AUC at 10% FPR of the ROC curves generated using the ChIP-seq evaluation `Comprehensive Dataset` on HINT using: the DHS sequence bias correction (HINT-BC), the naked DNA sequence bias correction (HINT-BCN) and no DNase-seq sequence cleavage bias correction (HINT w/o BC). The significant Friedman-Nemenyi test p-values are shown on the top of the boxplot. *Source: Gusmao et al. (2016)* (modified to fit thesis format and/or clarify key points).

BCN $r = 0.0108$ (p-value > 0.05; Spearman correlation test). Furthermore, we observe no correlation between CG content of motifs and differences in AUC: HINT-BC $-$ HINT-BCN $r = 0.0188$, HINT-BC $-$ HINT (w/o BC) $r = 0.0724$ and HINT-BCN $-$ HINT (w/o BC) $r = 0.0644$ (p-value > 0.05; Spearman correlation test). This is evidence that the significantly higher performance of HINT-BC over HINT is not due to artifacts generated by the bias correction strategy.

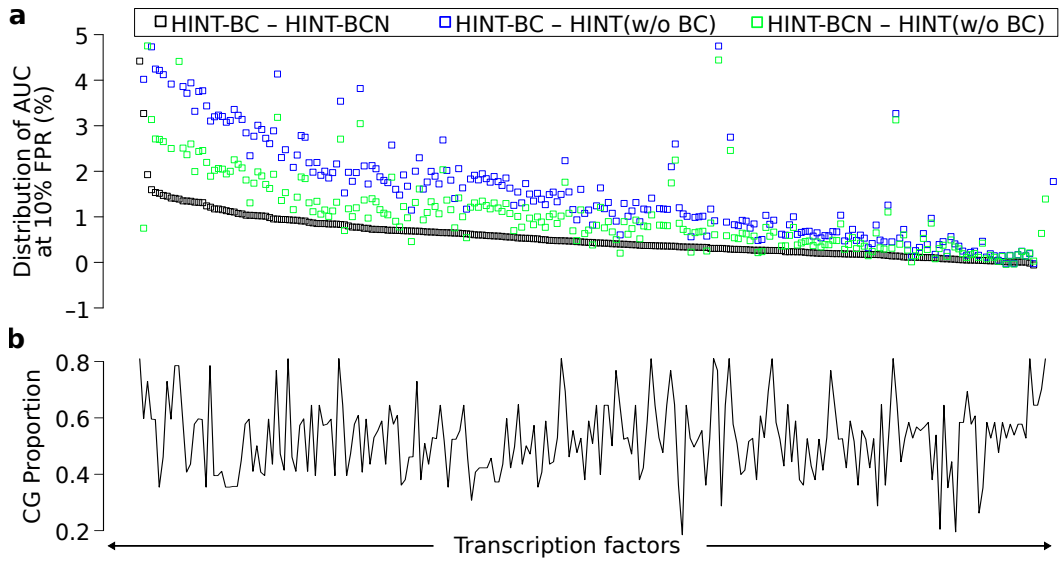

Figure 5.12: Evaluation of bias correction strategies and CG content contribution. (**a**) Distribution of AUC (10% FPR) differences bettween HINT-BC and HINT (w/o BC), HINT-BCN and HINT (w/o BC); HINT-BC and HINT-BCN for the `Comprehensive Dataset`. TFs are ranked by the difference between HINT-BC and HINT-BCN. (**b**) CG content of TFs. The CG content is calculated as $\frac{n_C+n_G}{n_A+n_C+n_G+n_T}$, where n_X is the frequency of the nucleotide X in all TFBSs. *Source: Gusmao et al. (2016)* (modified to fit thesis format and/or clarify key points).

DNase-seq Sequence Cleavage Bias Correction Decreases the Number of False Negatives

To better understand in which experimental cases the DNase-seq sequence cleavage bias correction improves the accuracy of HINT, we generated DNase-seq profiles with uncorrected and bias-corrected signals in TFBSs that matched true positive (TP), false negative (FN), false positive (FP) and true negative (TN) predictions (i.e. the contingency table from the ChIP-seq evaluation scheme). The results of such analysis is presented in Figure 5.13 for three selected TFs: GABP, NRF1 and EGR1. This figure shows, for all TFs and signal types, the clear peak-dip-peak DNase-seq profile (i.e. the grammar of active TFBSs) for the true positive predictions and a complete lack of such average pattern for the true negatives.

The profile around the false positive predictions are virtually the same between the uncorrected and bias-corrected DNase-seq signals. However, we are able to observe that the profiles around false negative predictions are significantly lower in intensity and different in shape (p-value < 0.01; Mann-Whitney-Wilcoxon on the DNase-seq signal distribution at flanking regions and motif center) regarding the bias-corrected signal, in comparison to the uncorrected case. This shows that the DNase-seq sequence cleavage bias correction strategy enhances the accuracy by correctly predicting TFs that would not be otherwise predicted without such correction. Such observation is in line with the fact that we observe a clearer dip-peak-dip pattern in the average DNase-seq signal for multiple TFs (Figure 5.10b–c).

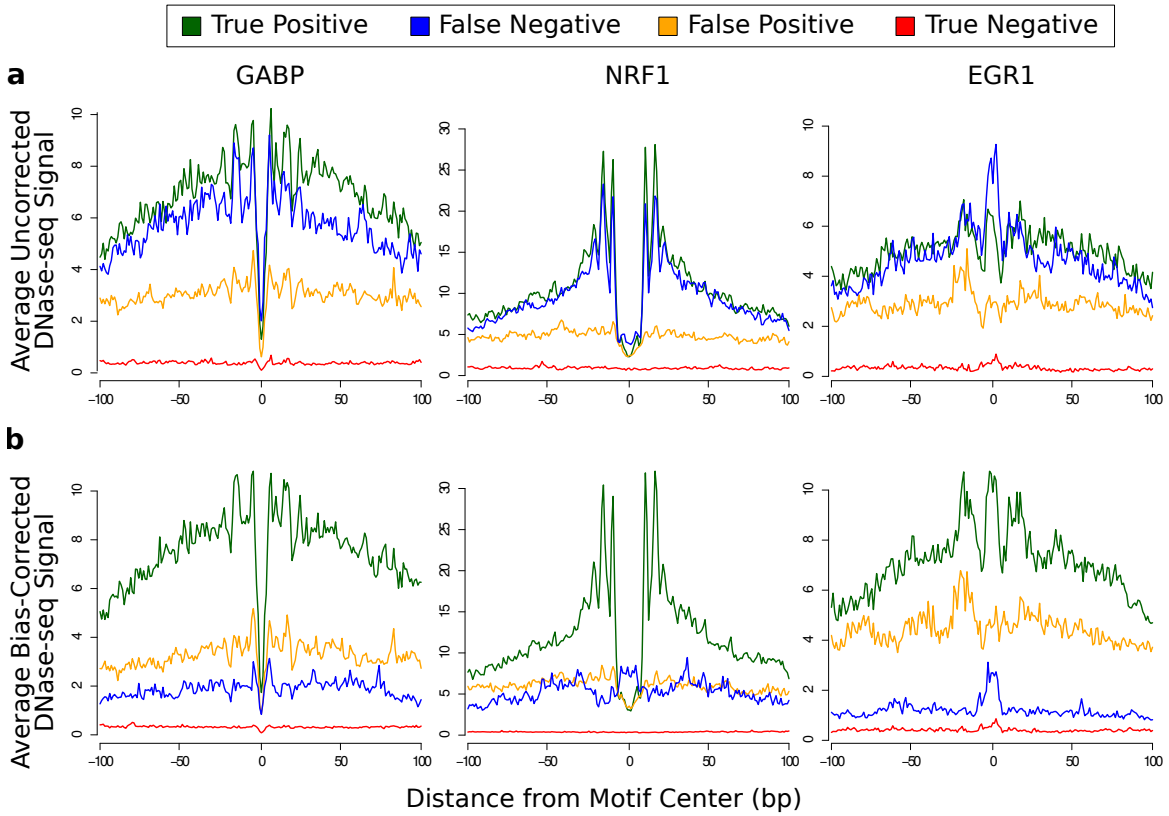

Figure 5.13: Uncorrected and bias-corrected DNase-seq profile between different contigency table statistics. This figure shows the: (**a**) uncorrected and (**b**) bias-corrected DNase-seq average signal centered at HINT's true positive (TP), false negative (FN), false positive (FP) and true negative (TN) predictions of the binding site of transcription factors GABP, NRF1 and EGR1.

Uncorrected DNase-seq Signal may Lead to Inaccurate TF Binding Predictions

An example that ignoring experimental artifacts might lead to false predictions can be seen in Figure 5.14. In this figure we show the DNase-seq profile for two motifs (termed 0458 and 0500) found using Neph's footprint predictions which did not match any existing known motif (i.e. *de novo* TF motifs). These *de novo* motifs were reported in Neph et al. (2012). Bias corrected DNase-seq profiles reveal very weak footprint shape. Furthermore, we compared the overlap between footprints generated by HINT-BC and Neph in the same cell type in which these *de novo* motifs were found (H7-hESC). We observed that 24.99% (motif 0458) and 28.58% (motif 0500) of MPBSs associated with a Neph footprint. In contrast, only 0.73% (motif 0458) and 1.71% (motif 0500) of MPBSs overlapped with a HINT-BC footprint. Altogether, this indicates that these motifs are indeed potential artifacts of sequence cleavage bias (as Neph's method do not use any bias correction strategy) and reinforces the importance of bias correction prior to any DNase-seq analysis.

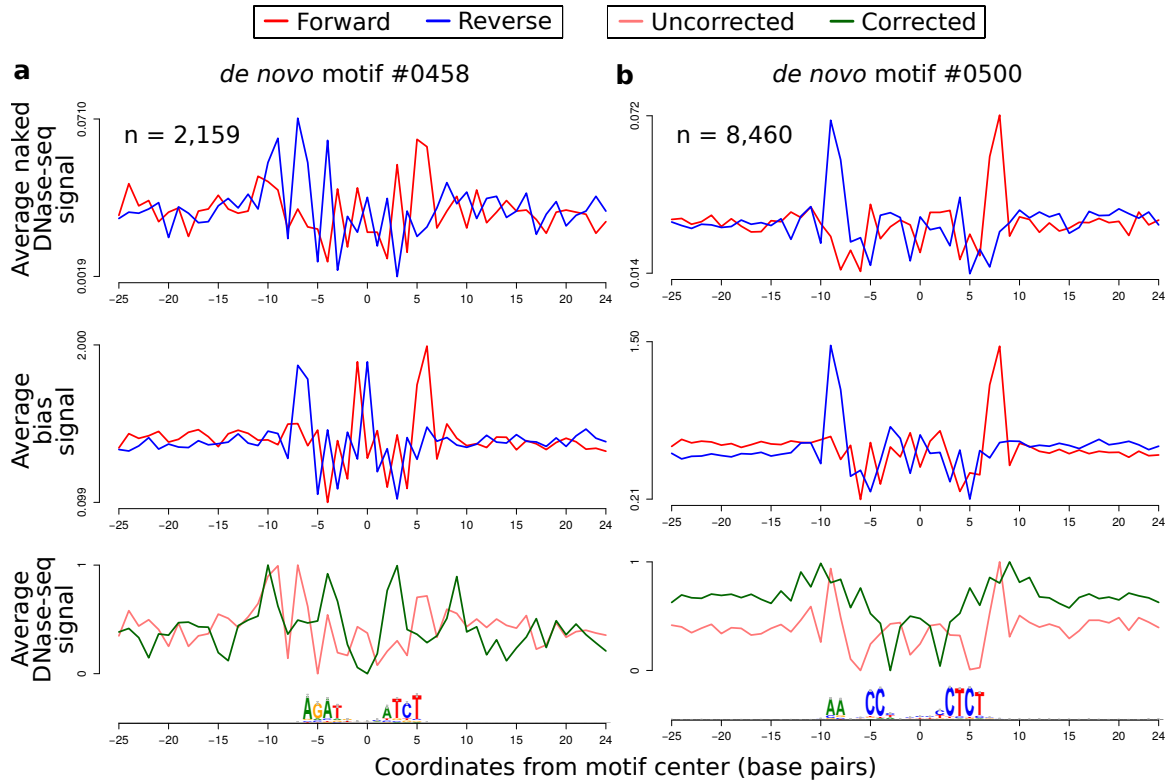

Figure 5.14: Average bias and DNase-seq signals around binding sites of Neph's *de novo* motifs. Average bias and DNase-seq signals around binding sites of *de novo* motifs 0458 and 0500 on cell type H7-hESC (Neph et al., 2012). In the top panel, we show the strand-specific average DNase-seq signal on naked DNA DNase-seq experiments (MCF-7 cell type); the middle panel shows the strand-specific estimated sequence cleavage bias signal; and the bottom panels shows the (1) uncorrected – observed DNase-seq I cleavage signal and (2) corrected – DNase-seq signal after the bias correction by using Equation 3.7. Signals in the top and middle graphs are DNA strand-specific (forward strand in red and reverse strand in blue). Bottom panel signals were standardized to be in $[0,1]$. Below the graphs, it is shown the motif logo estimated on the DNA sequences of these regions. These motifs were discovered in the footprint analysis of Neph et al. (2012) and indicated in He et al. (2014) to be possible artifacts of sequence cleavage bias. *Source: Gusmao et al. (2016)* (modified to fit thesis format and/or clarify key points).

5.3 Computational Footprinting Methods Comparison

In this section we present a comprehensive comparative analysis of HINT and all competing computational footprinting methods. Since most competing methods use only DNase-seq data, we used the DNASE-ONLY HINT topology for a fairer comparison Nature Methods Editorial (2015). Furthermore, the accuracies presented in this section were calculated using all chromosomes but the chromosome 1, since data from such chromosome was used to perform method parameter selection. HINT's DNase-seq sequence cleavage bias correction strategy followed the DHS cleavage bias scheme. Both HINT and competing methods were executed exactly as described in Chapter 4.

The computational footprinting method comparison was performed using: (1) the ChIP-seq evaluation method with the `Benchmarking Dataset` (Section 5.3.1) and (2) the gene expression evaluation method (Section 5.3.2). We close this section with a discussion on both evaluation methodologies in a general comparison showing the full experimental result's picture (Section 5.3.3).

5.3.1. ChIP-seq Evaluation

In the ChIP-seq evaluation approach, we create ROC and precision-recall (PR) curves for each method on the prediction of each TF in our `Benchmark Dataset`. Figure 5.15 shows examples of ROC and PR curves for the TFs EGR1, GABP and C-JUN. At first glance we are able to observe that the site-centric baseline methods FP-Rank and PWM-Rank present the lowest accuracies. Furthermore, the site-centric baseline method TC-Rank and the segmentation baseline method Filter exhibit a good performance. The Filter method often outperformed more complex computational footprinting methods (Cuellar, FLR and Centipede). The higher-ranked computational footprinting methods (HINT, DNase2TF and PIQ) have very close AUC at 100% FPR. However, the PR curves in Figure 5.15 show us that these methods compete with regard to the delay in sensitivity decrease as specificity increases. This is the main reason in which we also calculated the AUC at lower FPR levels and the area under the PR curve (AUPR).

To have a better perspective of the results for all methods and TFs tested we calculated the distribution of the AUC at 100%, 10% and 1% FPR as well as the AUPR (Figure 5.16). AUC at lower FPRs favors methods with higher sensitivity in expense of specificity. Also, AUC at lower FPRs tend to get closer results to the AUPR, which is ideal for very imbalanced classification problems. We observe the importance of using cell-specific open chromatin data (in this case, DNase-seq) by analyzing the dramatic increase in accuracy from the PWM method to all other methods that use such open chromatin data. The only exception to this remark is the FS-Rank method, which uses DNase-seq data but does present lower average accuracies than the PWM method. However, note that the FS metric, when combined with footprint predictions, generally present higher accuracies than the PWM metric (see "Gene Expression Evaluation" in Section 5.2.1). The reason for such lower FS-Rank accuracies stems from its inability to model the length of the DNase-seq depletion and peaks given that the FS-Rank relies on a fixed-window length strategy.

Figure 5.16 shows that the HINT method is the best method with regard to all metrics tested. HINT is closely followed by DNase2TF, PIQ and Wellington. It is interesting to observe that all segmentation methods are in the top-six positions of the rank for all metrics. This suggests that segmentation methods outperform site-centric approaches on the detection of active binding sites. Moreover, the results from the different evaluation statistics (AUC at different FDR thresholds and AUPR) result in very similar rankings ($r > 0.98$).

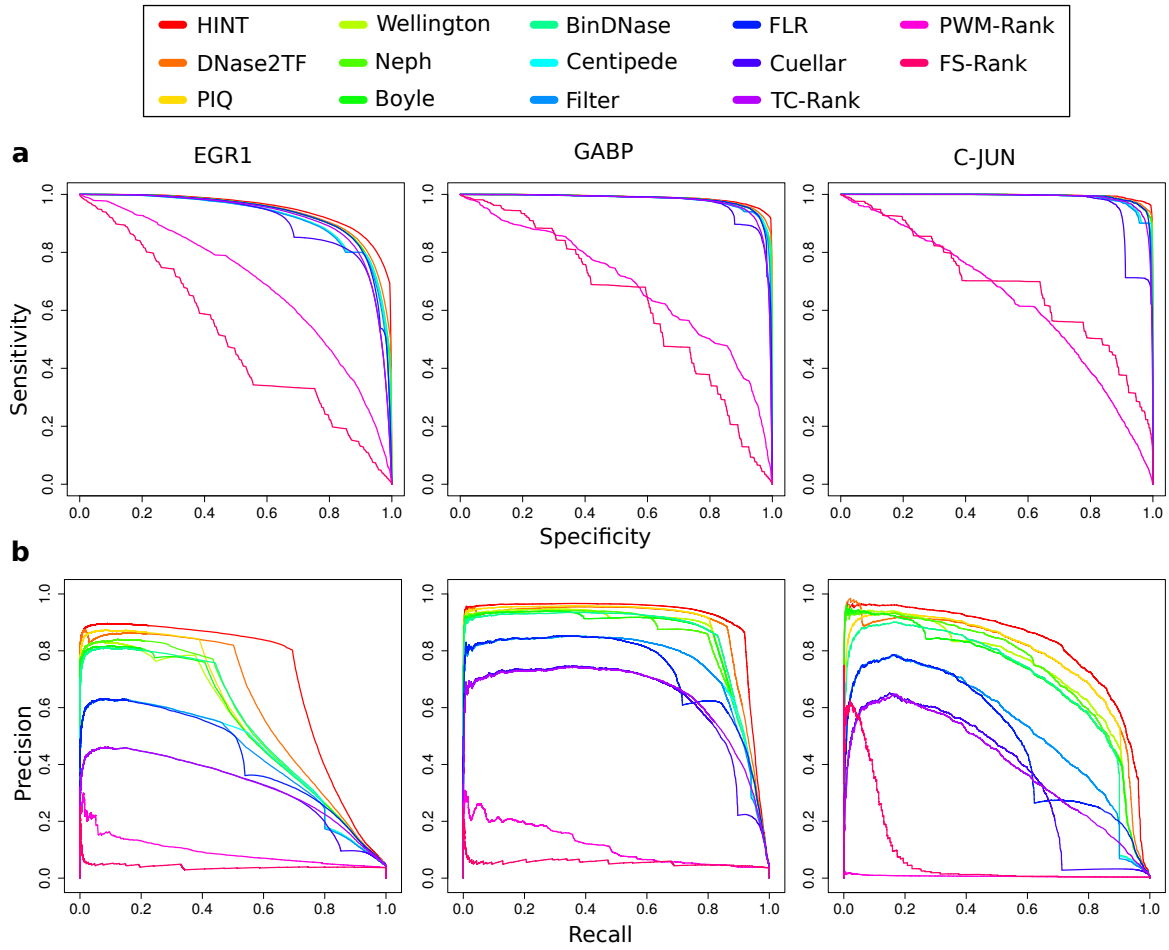

Figure 5.15: Example of ROC and PR curves. (**a**) Example of ROC curves for the TFs EGR1, GABP, and C-JUN. Each graph depicts a curve of a different color for each of the 14 computational footprinting methods evaluated. (**b**) Example of PR curves for the TFs EGR1, GABP, and C-JUN.

5.3.2. Gene Expression Evaluation

In the gene expression evaluation approach, we calculate the correlation between changes in gene expression for a number of TFs with differences in the quality of footprint predictions for these factors. To measure the change in gene expression we use the gene expression fold change (FC). The quality metric of footprints used is the footprint likelihood ratio (FLR) metric (Yardımcı et al., 2014). Differences in such footprint quality metric is measured with the Kolmogorov-Smirnov (KS) test statistic. The Spearman correlation between FLR score difference and expression FC, which we refer to as FP-Exp, will be used to rank footprinting methods. Higher FP-Exp values indicate better performance. The gene expression evaluation methodology only requires expression data and is therefore more generally applicable than the ChIP-seq evaluation. However, differently from the ChIP-seq evaluation, the gene expression approach cannot evaluate footprint predictions of individual TFs.

In Figure 5.17 we show the top four methods with regard to the gene expression evaluation. Since a single FP-Exp is evaluated for a collection of TFs, it is not possible to graphically display the distribution of accuracies as in the ChIP-seq evaluation scheme. Nevertheless, we are able to observe that HINT is again the top-ranked method, followed by DNase2TF, Neph and FLR.

We observed high average FP-Exp values for the majority of evaluated methods (FP-Exp $= 0.79$) and very high FP-Exp values (FP-Exp > 0.9) for the four top performing methods on comparisons

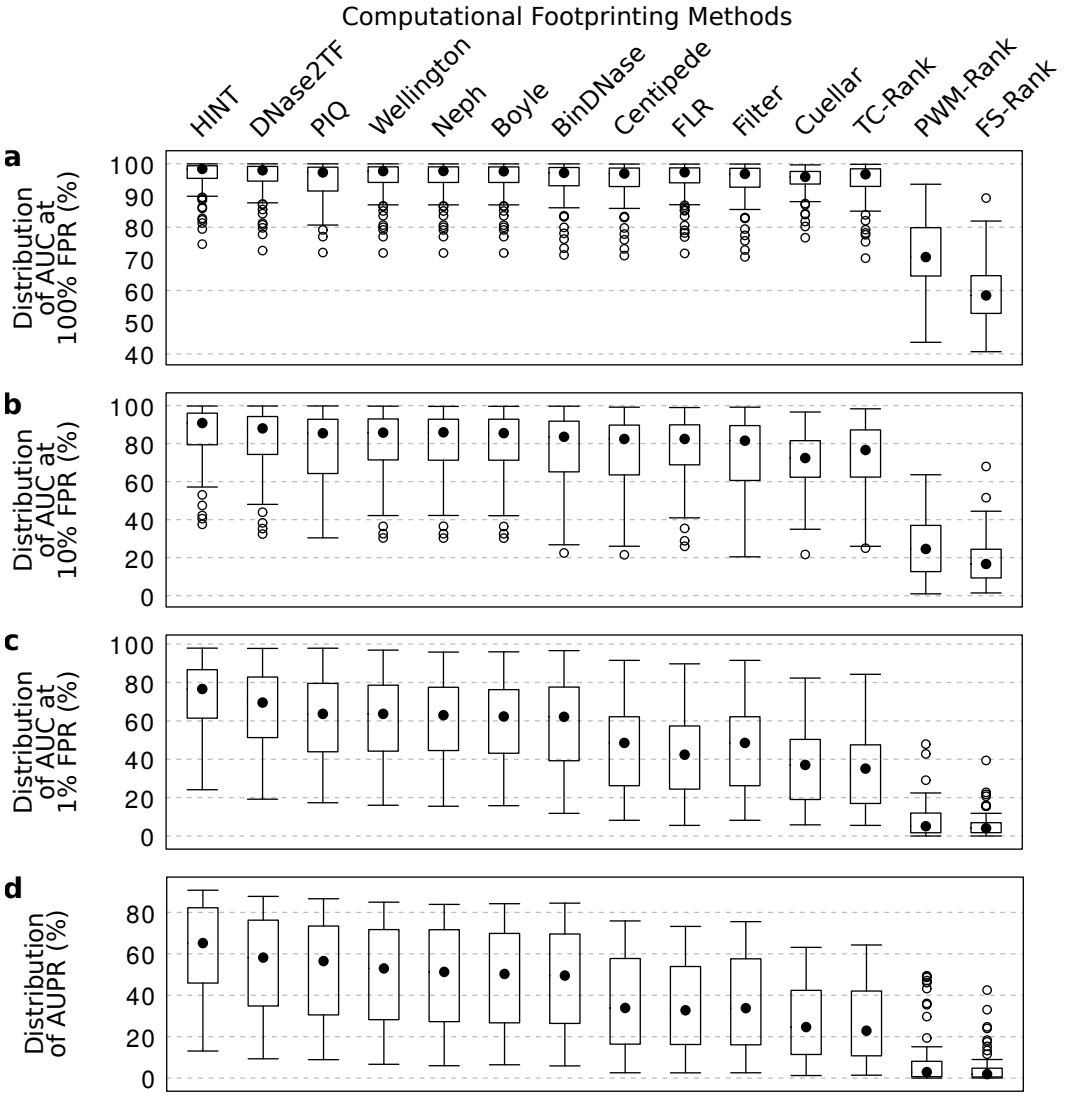

Figure 5.16: ChIP-seq evaluation accuracy distributions. Accuracy distribution for 14 footprinting methods regarding all validation sets (ordered by Friedman Ranking). The accuracy is given by the statistics: (**a**) AUC at 100% FPR (**b**) AUC at 10% FPR (**c**) AUC at 1% FPR and (**d**) AUPR. *Source: Gusmao et al. (2016)* (modified to fit thesis format and/or clarify key points).

between pairs of cell types H1-hESC, K562 and GM12878 (Figure 5.17). Moreover, similar rankings of methods are obtained for each cell pair: H1-hESC/K562 *vs* H1-hESC/GM12878 $r = 0.99$, H1-hESC/K562 *vs* GM12878/K562 $r = 0.96$, and H1-hESC/GM12878 *vs* GM12878/K562 $r = 0.97$. We also observed a high agreement between the ranking of computational footprinting methods using the gene expression evaluation methodology and the ranking of methods using the ChIP-seq evaluation approach ($r > 0.88$).

5.3.3. General Comparison

We integrated all the computational footprinting method's results to perform a global comparison. Figure 5.18 shows the ranking of all computational footprinting methods with regard to all evaluation metrics: FP-Exp, AUCs (at 100%, 10% and 1% FPR) and AUPR. Furthermore, we combined all these results and performed a Friedman-Nemenyi test for statistical significance (Table 5.3).

HINT has the highest FP-Exp, AUC and AUPR values and significantly outperforms all methods

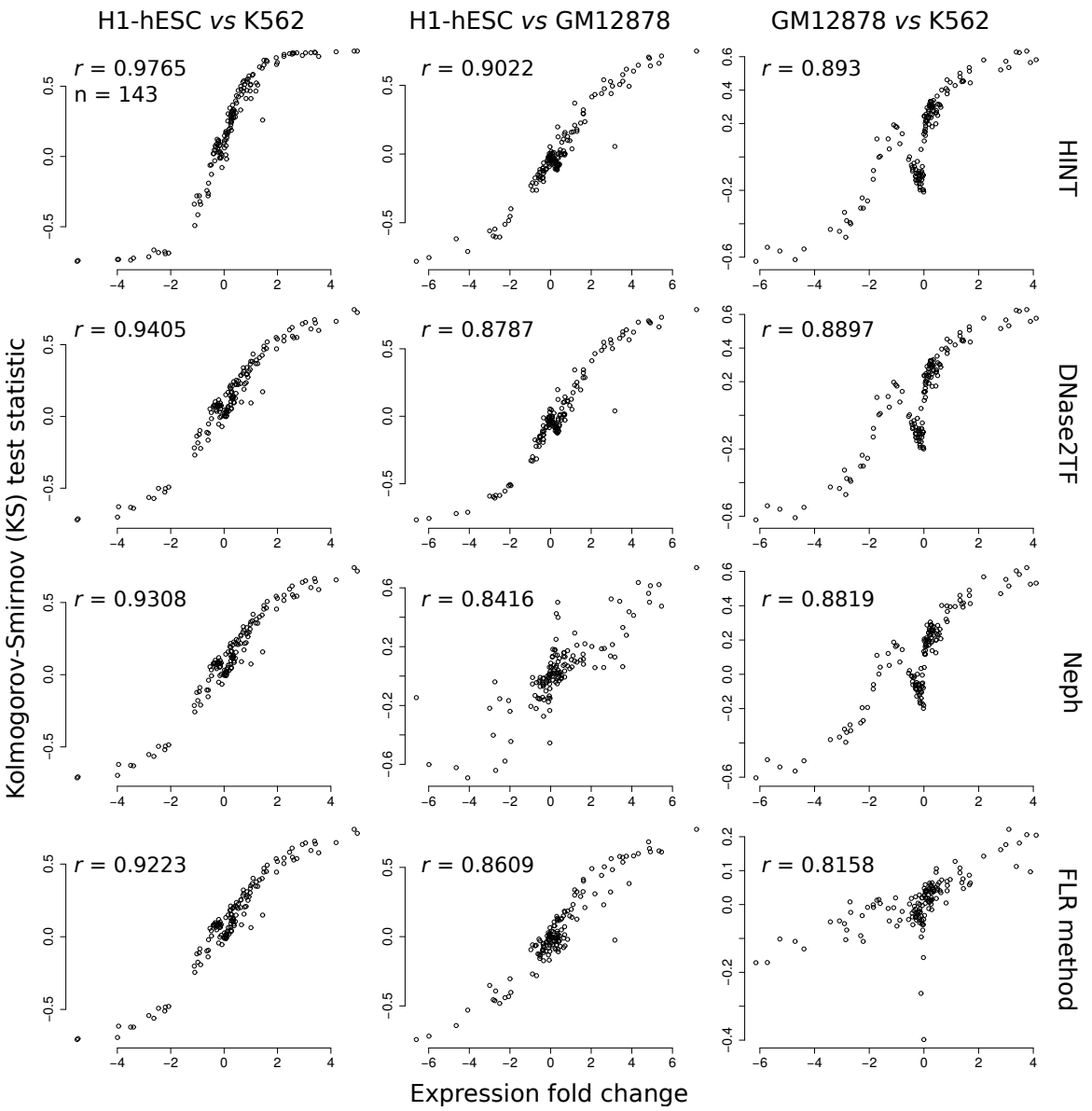

Figure 5.17: Gene expression evaluation accuracy correlations (FP-Exp). Correlation between KS statistics from FLR scores *vs* fold change expression for cell type pairs H1-hESC *vs* K562 (left), H1-hESC *vs* GM12878 (middle) and GM12878 *vs* K562 (right) for footprints predicted by: HINT, DNase2TF, Neph and FLR (from top to bottom, respectively). *Source: Gusmao et al. (2016)* (modified to fit thesis format and/or clarify key points).

(p-value < 0.01). The next top performing method is DNase2TF, which significantly outperforms all other methods except PIQ (p-value < 0.05 for Wellington; p-value < 0.01 for all others). PIQ outperforms all of its lower ranked competitors but Wellington (p-value < 0.05 for Neph; p-value < 0.01 for all others). The segmentation methods HINT, DNase2TF, Wellington and Neph are ranked within the top five methods regarding all evaluation metrics, individually. Boyle is the only segmentation method not included within the top five; however it displayed good accuracies (placed 6th in the global ranking). The site-centric method PIQ obtained the best accuracies among the site-centric methods (placed 3rd in the global ranking). All site-centric baseline methods (FS-Rank, PWM-Rank and TC-Rank) are in the bottom four positions of the ranks. These results lead us to claim that the segmentation approach is preferable over the site-centric approach.

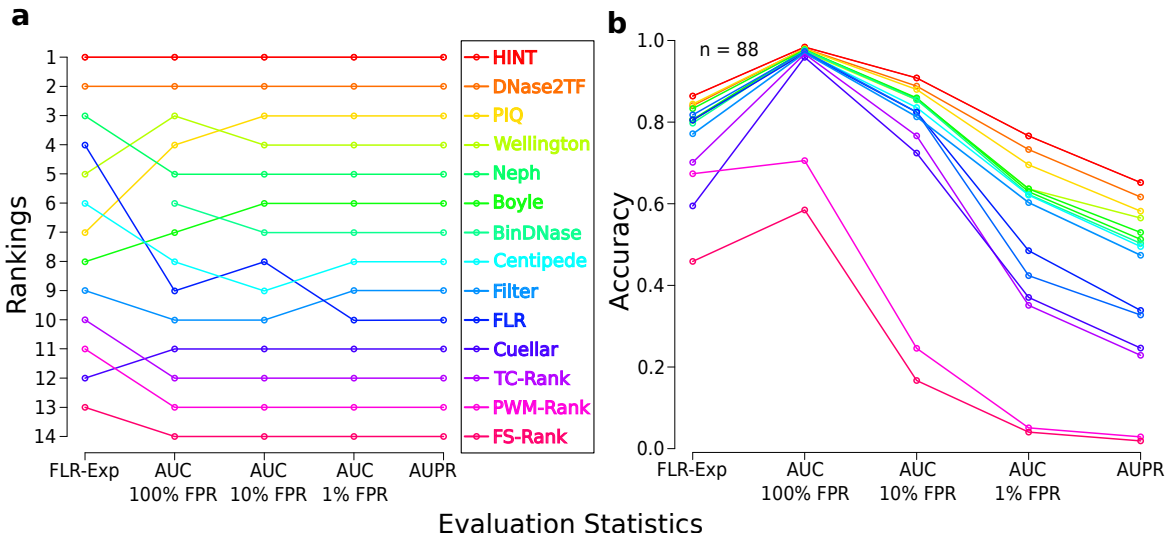

Figure 5.18: Evaluation of computational footprinting methods. (**a**) Average rankings for the evaluated computational footprinting methods. The rankings are given for all evaluation criteria: FP-Exp, ChIP-seq evaluation AUC (at 100%, 10% and 1% FPR) and AUPR. (**b**) For all evaluated methods we show the FP-Exp values (as a combination of all pairwise comparison within cell types H1-hESC, K562 and GM12878), median ChIP-seq evaluation AUC (at 100%, 10% and 1% FPR) values and median AUPR values. Note that BinDNase could not be evaluated with the gene expression approach, as it requires TF ChIP-seq data for training. *Source: Gusmao et al. (2016)* (modified to fit thesis format and/or clarify key points).

Table 5.3: Friedman-Nemenyi hypothesis test on different computational footprinting methods. Friedman-Nemenyi hypothesis test results for all computational footprinting methods evaluated on the distribution of all tested metrics: FP-Exp, AUCs and AUPR. The asterisk and the cross, respectively, indicate that the method in the column outperformed the method in the row with significance levels of 0.01 and 0.05. *Source: Gusmao et al. (2016)* (modified to fit thesis format and/or clarify key points).

	HINT	DNase2TF	PIQ	Wellington	Neph	Boyle	BinDNase	FLR	Centipede	Filter	Cuellar	TC	PWM	FS
HINT														
DNase2TF	*													
PIQ	*													
Wellington	*	+												
Neph	*	*	+											
Boyle	*	*	*											
BinDNase	*	*	*											
FLR	*	*	*	*	*	*	*							
Centipede	*	*	*	*	*	+	+							
Filter	*	*	*	*	*	*	*	*						
Cuellar	*	*	*	*	*	*	*	*						
TC	*	*	*	*	*	*	*	*	+					
PWM	*	*	*	*	*	*	*	*	*	*				
FS	*	*	*	*	*	*	*	*	*	*	+			

5.4 Impact of Transcription Factor Residence Time

Despite the high average prediction values of top performing footprint methods, they consistently perform worst in a similar set of TFs, i.e. HINT, DNase2TF and PIQ have 89% of TFs in common in the lower quartile of AUC at 10% FPR. This list includes nuclear receptors, which has low residence binding time (Sung et al., 2014) and display a lower DNase I cleavage protection pattern (Figure 5.19). A careful analysis of Figure 5.19 shows that, while corrected DNase-seq profiles from ER have a better match with the underlying motif, this is not the case for AR and GR. However, we observed a small gain in the AUC score comparing HINT (with DHS sequence cleavage bias correction) and HINT (without bias correction). This difference is in the upper quartile range for all 233 TFs analyzed from the `Comprehensive Dataset`. These results indicate that sequence cleavage bias correction also brings improvements to footprint prediction of nuclear receptors. However, all these nuclear receptor TFs have low AUC scores in all footprinting methods, i.e. lower quartiles for HINT or TC AUC score. This indicates that short binding time indeed poses a challenge in footprint prediction.

To further investigate this, we propose a statistic – termed protection score – inspired by the concepts presented by Sung et al. (2014) to detect TFs with potential short residence time. The protection score measures the difference between the amounts of DNase I digestion in the flanking regions and within the TFBS on bias-corrected DNase-seq signals. More formally, the protection score for a genomic region $r_i = [u, v]$ is defined as

$$\text{PROTECTION}_{r_i} = \frac{(n^R_{r_i} - n^C_{r_i}) + (n^L_{r_i} - n^C_{r_i})}{2N}, \tag{5.1}$$

where $n_{C,i}$, $n_{L,i}$ $n_{L,i}$ are the number of DNase-seq reads within, upstream and downstream of the genomic region r_i, respectively (Equation 2.2).

In short, the protection score indicates the average difference of DNase-seq counts in the flanking region and the within TFBSs. Positive values will indicate protection in the flanking regions, while values close to zero or negative indicate no protection. The protection score is similar to the FS. The main difference is that the FS score measures the ratio between reads in flanking *vs* binding sites, while the protection score measures the difference.

We used the protection score to analyze the predictive performance of methods on TFs with distinct residence time. For this, we used the TFs from the `Comprehensive Dataset`. We observed that TFs with known short residence time on DNA, such as nuclear receptors AR (Tewari et al., 2012), ER (Sharp et al., 2006) and GR (McNally et al., 2000), present a negative protection score (Figure 5.20a). TFs with intermediate and long residence time on DNA (C-JUN (Malnou et al., 2010) and CTCF (Nakahashi et al., 2013), respectively) present a positive protection score. The amount of protection is clearly reflected in the bias-corrected DNase-seq profiles (Figure 5.20b–d). In addition, Figure 5.20a also reveals an association of the protection score and the performance of HINT. Overall, the protection score positively correlates with the AUC values of evaluated methods, such as TC-Rank ($r = 0.19$) and HINT ($r = 0.26$), and negatively correlates ($r = -0.49$) with the sequence bias (adjusted p-value < 0.05). These results reinforce the concept that TFs with potential short residence time are poorly detected via DNase-seq footprints in comparison to TFs with higher residence time. Nevertheless, in the absence of biological experimental data on the residence time of TFs, the protection score can be used to identify TFs with potential short residence time and can be an important tool on experiments involving computational footprinting methods.

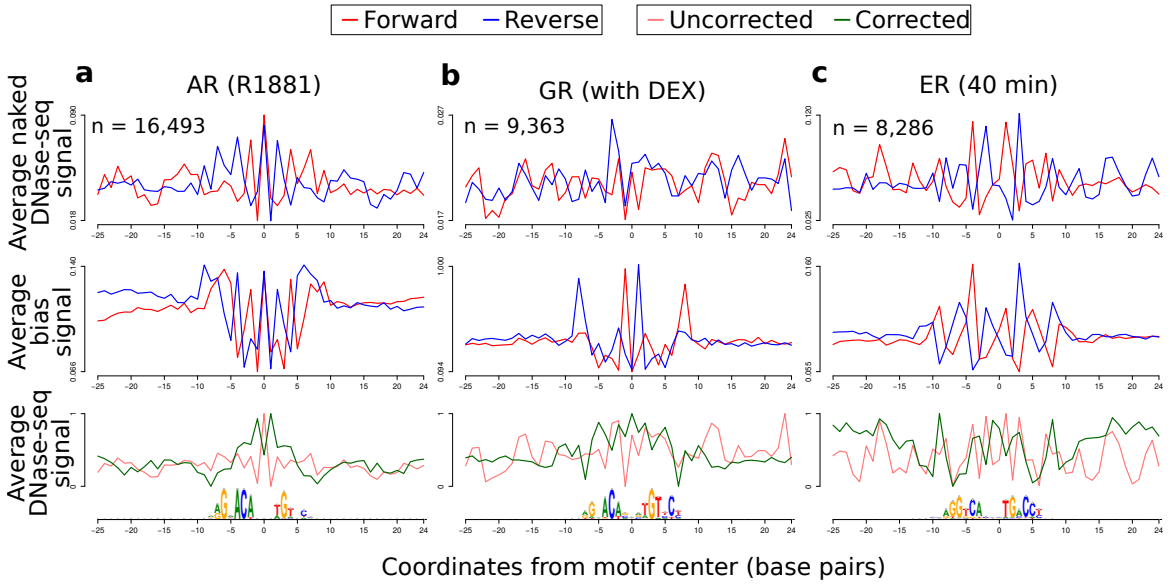

Figure 5.19: Average DNase-seq signals around nuclear receptors. Average DNase-seq signals around nuclear receptor TFs with ChIP-seq evidence in LNCaP, m3134 and MCF-7 cell types. In the top panel, we show the strand-specific average DNase-seq signal on naked DNA DNase-seq experiments (MCF-7 for datasets from single-hit and IMR90 for datasets with double-hit protocol); the middle panel shows the strand-specific estimated DNase-seq sequence cleavage bias signal; and the bottom panels shows the (1) uncorrected – observed DNase-seq I cleavage signal and (2) corrected – DNase-seq signal after sequence cleavage bias correction. Signals in the top and middle graphs are DNA strand-specific (forward strand in red and reverse strand in blue). Bottom panel signals were standardized to be in $[0, 1]$. Below the graphs, it is shown the motif logo estimated on the DNA sequences of these regions. *Source: Gusmao et al. (2016)* (modified to fit thesis format and/or clarify key points).

5.5 *De Novo* Motif Finding on Predicted Footprints

The predicted footprints from a segmentation computational footprinting approach represent a map of active TFBSs. In possession of such regulatory landscape one might be able to perform many downstream analysis. Here we present an example of such analysis – the *de novo* motif finding. This analysis consists on searching for novel TF DNA affinity sequences which do not match any known affinity sequence in the literature.

We performed a *de novo* motif finding analysis on all 738,707 footprints predicted by HINT on cell type H1-hESC as described in Section 4.3.2. After the initial motif matching to filter out footprints that match with known TFs, we are left with ~5.37% (39,703) of H1-hESC's footprints. The tools "discriminative regular expression motif elicitation" (DREME) and "local motif enrichment analysis" (CENTRIMO) were applied to these filtered footprints.

Given the quality scores given by DREME and CENTRIMO, we were able to find six frequent motifs in H1-hESC which did not match with any existing motif from the repositories used. Figure 5.21 shows these six motifs and their DREME p-value. Each motif is named after its IUPAC consensus sequence standards. We also show in Figure 5.21 the average DNase-seq signal on a 200 bp window around each of these *de novo* motifs (DNase-seq profile graph). We are able to see that, with exception of the TACCCR motif, all other motifs presented a very clear peak-dip-peak DNase-seq pattern, consistent with the grammar of active TFBSs. Furthermore, the motifs CKCSGAG and CCGGAGHC present very clear signs of co-binding. This can be seen as a pattern of multiple dips with a 10–15 bp

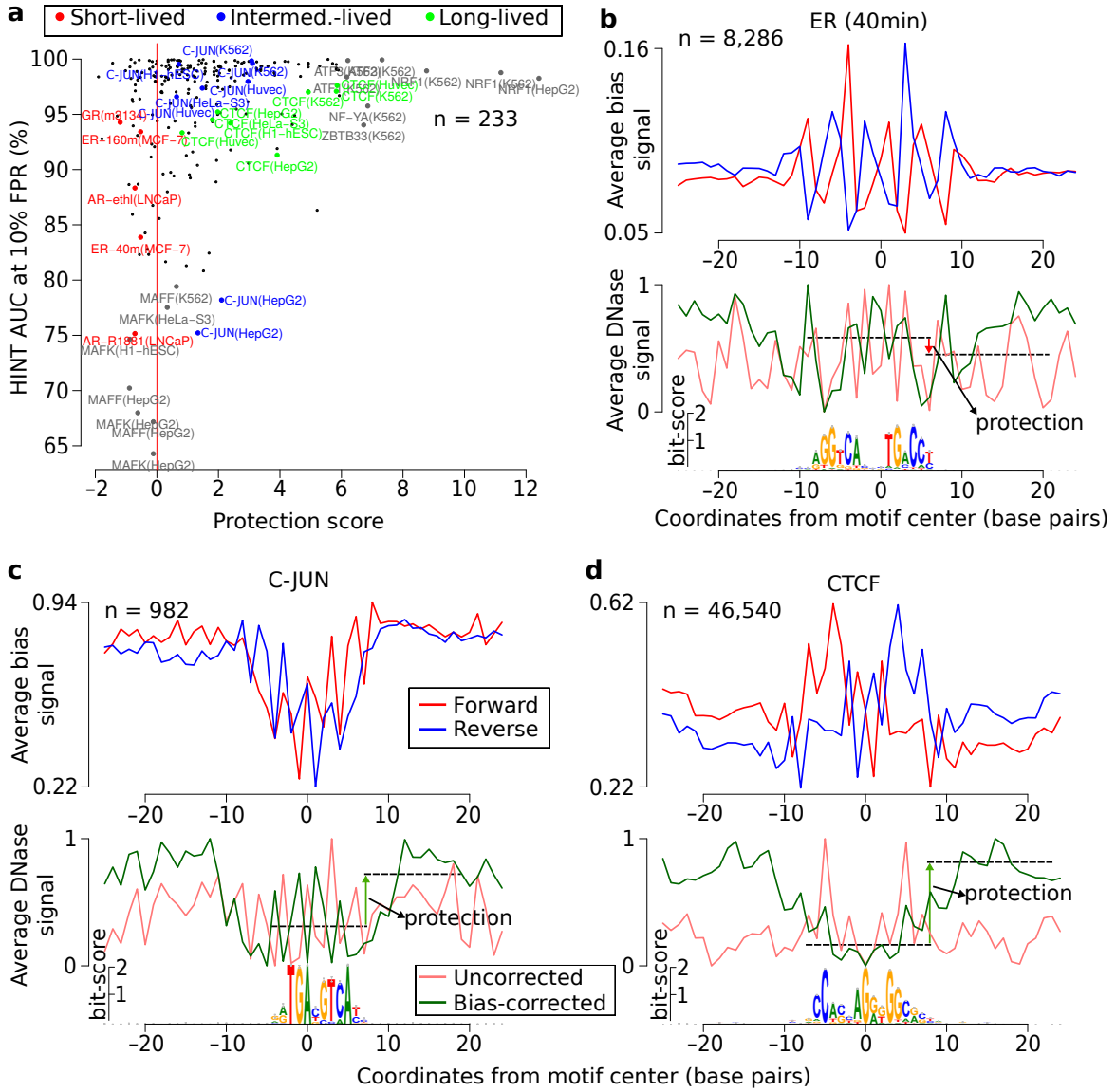

Figure 5.20: Impact of transcription factor residence binding time on computational footprinting. (**a**) Scatter plot with the protection score (*x*-axis) *vs* the AUC (at 10% FPR) of HINT (*y*-axis) for the TFs from the `Comprehensive Dataset`. We highlight nuclear receptors AR, ER and GR (short residence time, red); C-JUN (intermediate residence time, blue); CTCF (long residence time, green) and other TFs with either high (> 6) protection score or low (< 0.8) AUC values (grey). (**b–d**) Average bias signal (top) and uncorrected/bias-corrected DNase-seq signal (bottom) for the TFs (**b**) ER, (**c**) C-JUN and (**d**) CTCF. Signals in the top graph are DNA strand-specific (forward strand in red and reverse strand in blue). Signals in the bottom graph were standardized to be in the interval $[0, 1]$. The motif logo represents all underlying DNA sequences centered on the TFBSs. *Source: Gusmao et al. (2016)* (modified to fit thesis format and/or clarify key points).

spacing between them, instead of a single dip in the middle of the DNase-seq profile graphs.

Needless to say, such *de novo* motifs need to be experimentally validated using biological methods. However, the fact that we are able to see a peak-dip-peak pattern on the DNase-seq profile is a clear indication of active TF binding. The intent of this analysis is to show an example of downstream analysis using footprints predicted with HINT. In Section 5.6 we will show the application of footprints on real biological scenarios in which positive inferences could be performed.

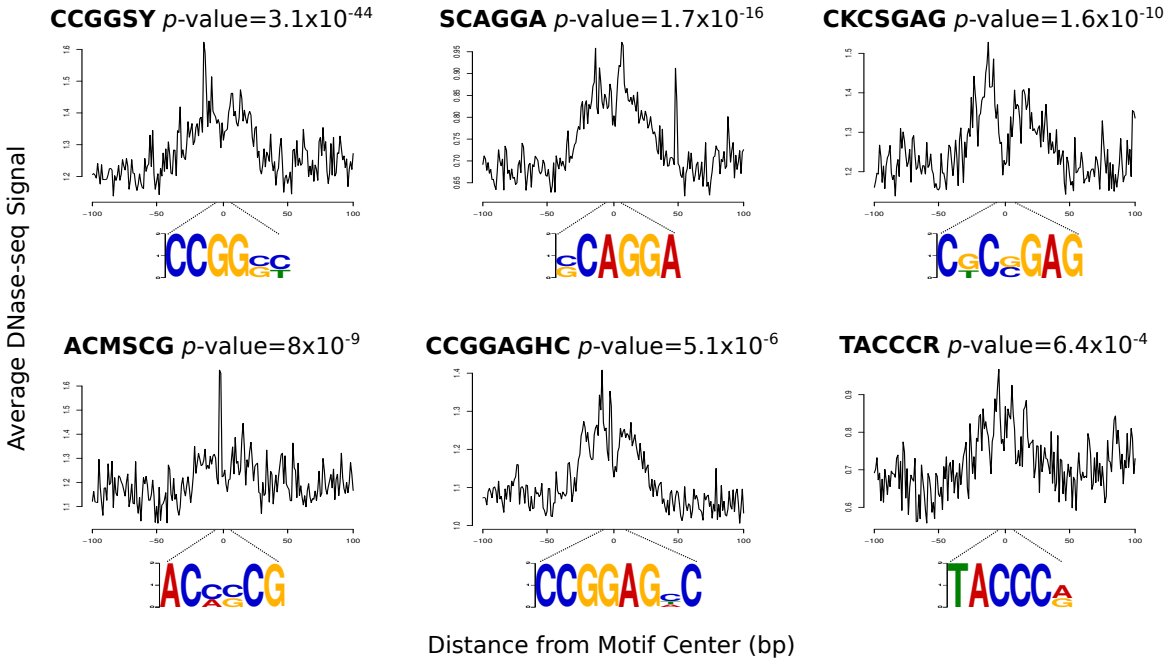

Figure 5.21: *De novo* TF motifs predicted on H1-hESC with HINT's footprints. We show six *de novo* motifs which satisfied all quality checks after the application of DREME and CENTRIMO softwares on H1-hESC's HINT footprint predictions which did not match any known DNA sequence affinity motif. Each motif is named after its IUPAC consensus sequence (bold, top of each graph) and the *p*-value of the DREME analysis is also shown. The graphs represent the average DNase-seq signal around a 200 bp window centered on MPBSs found in the whole genome after applying each *de novo* position frequency matrix. The motif sequence logo is shown below each *de novo* motif's graph.

5.6 HINT Case Studies – Identification of Regulatory TFs involved in Different Biological Conditions

In this section we show two case studies in which our computational footprinting method was successfully used to identify key regulatory players on two different biological analyses. The first case study regards the identification of key regulatory TFs on the differentiation of dendritic cells in mouse (Section 5.6.1). The second case study concerns the identification of TFs associated to (i.e. binds together with) the NF-κB TF, which is a key regulator on the mammalian inflammatory response (Section 5.6.2). Both case studies exhibit a similar experimental workflow. Briefly, we first apply HINT to detect footprint predictions in different cellular conditions. Then, we compare these different cellular conditions to find the TFs more likely to be associated with each condition. For that, we use the TF enrichment analysis as described in Section 4.3.1.

All the results shown in this section represent a subset of the analyses which were published in papers co-authored by this thesis' author. More specifically, the results shown in Section 5.6.1 were published and authored by Lin et al. (2015) and the results shown in Section 5.6.2 were published and authored by Kolovos et al..

5.6.1. Case Study: Regulatory Network during Differentiation of Dendritic Cells

This case study focuses on dendritic cells (DCs). DCs are professional antigen-presenting cells that develop from hematopoietic stem cells through successive steps of lineage commitment and differen-

tiation (Merad et al., 2013; Belz and Nutt, 2012). Here, we focus on four cell types which represent such lineage commitment steps: multipotent progenitors (MPPs), common DC progenitors (CDPs), classical DCs (cDCs) and plasmacytoid DCs (pDCs). Multipotent progenitors (MPPs) are committed to common DC progenitors (CDPs), which further differentiate into specific DC subsets: the classical DCs (cDCs) and plasmacytoid DCs (pDCs) (Belz and Nutt, 2012; Lin et al., 2015). The understanding of DC differentiation is important since it has impact on the further understanding of adaptive immune responses (Merad et al., 2013).

The goal of this study is to understand the regulatory circuitry that determines the differentiation of MPPs to CDPs and CDPs to either cDCs or pDCs. It is known that the TF PU.1 is an important factor within this cell differentiation framework (Belz and Nutt, 2012). The TF PU.1 is a master regulator, since it initiates differentiation events and are associated to many other TFs. In this section we show the results regarding our investigation of TFs which are significantly associated to PU.1.

In this study, ChIP-seq experiments were performed for several histone modifications, including H3K4me1 (Lin et al., 2015). With the goal to capture the regulatory landscape of dendritic cell differentiation associated to the PU.1 master regulator, we performed a TF enrichment analysis using:

- Only the PU.1 ChIP-seq peaks (i.e. PU.1 TFBSs). This analysis do not involve any footprinting. The rationale of this analysis is to find TFs associated to the PU.1 master regulator.
- Footprints predicted with HINT on data from the H3K4me1 ChIP-seq that are also associated with (i.e. close to) PU.1 TFBSs. The rationale of this analysis is to verify if footprints can enhance the specificity of the TF enrichment analysis.
- Footprints predicted with HINT on data from the H3K4me1 ChIP-seq that are not associated with PU.1 TFBSs. The rationale of this analysis is to find TFs which are associated to DC differentiation but do not necessarily associate with the PU.1 master regulator.

This analysis, which is a part of the study performed by Lin et al. (2015), is presented as a case study for our computational footprinting framework in the next subsections.

Computational Footprinting

We applied the HISTONE-ONLY HINT model (see "HISTONE-ONLY MODEL" in Section 3.2.2) to the H3K4me1 ChIP-seq data. We followed the experimental settings as described in Chapter 4. Briefly, we extended the H3K4me1 enriched regions by $5,000$ bp to each side and applied our HISTONE-ONLY HMM model trained with H3K4me1 data from random genomic regions. Given the lower resolution of ChIP-seq data and the nature of the probabilistic model, footprints from H3K4me1 tend to span larger regions. Therefore, we further reduced the footprint predictions by considering only 250 bp to the left (downstream) and right (upstream) of its center.

As aforementioned, we performed three TF enrichment analyses: (1) on PU.1 peaks, (2) on H3K4me1 footprints that overlap PU.1 peaks and (3) on H3K4me1 footprints that do not overlap PU.1 peaks. These three definitions of target genomic region sets for the TF enrichment analysis were used for data on the four different cells being analyzed: MPPs, CDPs, cDCs and pDCs. In all tests, the size of the background genomic region sets were 100 times higher than the size of the target genomic region sets.

Results

Figure 5.22a shows the overlap between H3K4me1 footprints and PU.1 ChIP-seq enriched regions (peaks). We can observe different levels of overlap. A higher overlap was found on more specialized cells (cDC; ~68% of H3K4me1 footprints overlap with PU.1 peaks) in contrast to less specialized

cells (MPP; ~23% of H3K4me1 footprints overlap with PU.1 peaks). These results are consistent with gene expression information obtained with DNA microarray analyses, which shows higher expression of PU.1 in cDC than MPP (Lin et al., 2015).

We present here the three TF enrichment analyses results, using as the target genomic region set: (1) only PU.1 peaks (Figure 5.22b); (2) H3K4me1 footprints that overlap PU.1 peaks (Figure 5.22c) and (3) in H3K4me1 footprints that did not overlap PU.1 peaks (Figure 5.22d). The p-values from the TF enrichment analyses are presented as a heatmap. The enrichment is represented in a gray to blue scale. A gray heatmap entry represent no enrichment (p-value > 0.05) for the TF represented in the row at the cell type represented in the column. A blue heatmap entry represents evidence of enrichment (p-value < 0.05). Enriched TFs were separated in different clusters (numbered I to VI) given their different enrichment levels in different cells.

We were able to detect many TFs involved in DC differentiation. For instance, in the footprint+PU.1 TF enrichment analysis shown in Figure 5.22c we observed the binding of the pioneer PU.1 alongside evidence of KLF4 and RUNX1 in MPP. The AP1-like TFs (FOS and JUN) and some IRF factors (IRF2, IRF4 and IRF5) appear to be cDC-specific. This means that these factors might have some role on the differentiation from CDP to cDC cell type. On the other hand, TCF factors (TCF3 and TCF4), EGR1 and KLF4 appear to play a role in the differentiation from CDP to pDC cell type.

Interestingly, the TF enrichment analysis performed in H3K4me1 footprints captured most of the PU.1-only enrichment analysis. Furthermore, the H3K4me1 footprint analyses recovered two TFs (CEBPB and BHLHE40; marked in red in Figure 5.22c–d) which were not found by the PU.1-only enrichment analysis.

The results presented here demonstrate the power of computational footprinting coupled with the TF enrichment analysis to increase the specificity of biological analyses. The H3K4me1 footprints represent regulatory regions and were shown to have a high overlap (~80%) with open chromatin regions (Lin et al., 2015). The H3K4me1 footprint predictions were used to search for TFs which act in conjunction with PU.1 master regulator or independently from the PU.1 master regulator. Such results, combined with other experimental data and knowledge from previous studies, were used to devise a regulatory network on the differentiation of dendritic cells. The complete results of these experiments are in Lin et al. (2015).

5.6.2. Case Study: Multimodal Role of NF-κB during Intermmediate-Early Inflammatory Response

This case study focuses on the inflammatory mechanism of human umbilical vein endothelial cells (HUVECs). In this study, we focused on the TF NF-κB. This TF is a key regulator of inflammatory mechanisms (Hayden and Ghosh, 2012). However, NF-κB has still many unknown features with regard to its interaction with other TFs and chromatin dynamic processes (Hayden and Ghosh, 2012). In HUVECs, the tumor necrosis factor alpha (TNFα) acutely remodels the cell's transcriptional program, but our understanding of how the activation of proinflammatory genes is achieved at the expense of the ongoing transcriptional program is far from complete (Kempe et al., 2005). It is known that NF-κB predominantly "hijacks" the regulatory machinery of the cell by binding already-active enhancers, more than half of which do not carry NF-κB recognition motifs.

The goal of this study is to understand the different regulatory players involved in enhancers (i.e. distal regulatory regions) in which:

- NF-κB is found (ChIP-seq peak evidence) and the NF-κB recognition motif is present (in the DNA).
- NF-κB is found and the NF-κB recognition motif is not present.

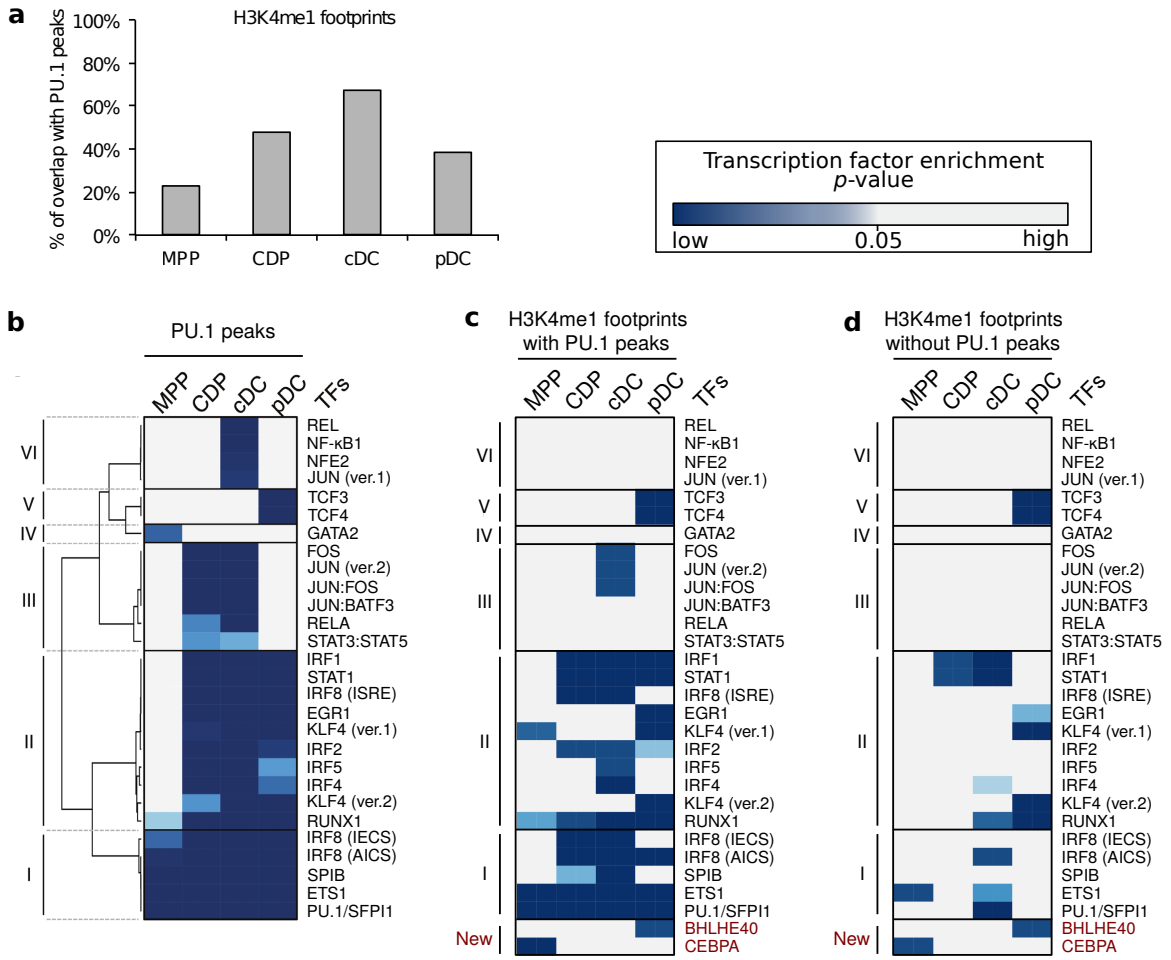

Figure 5.22: Dendritic cells footprint enrichment analysis results. (**a**) The overlap between dendritic cell's master regulator TF PU.1 ChIP-seq enriched regions with H3K4me1 footprints in the context of dendritic cell differentiation. (**b**–**d**) Heatmap depicts the enrichment of TF motifs in MPP, CDP, cDC and pDC based on: (**b**) PU.1 peaks, (**c**) H3K4me1 footprints overlapping PU.1 peaks and (**d**) H3K4me1 footprints not overlapping PU.1 peaks. The *p*-values are plotted and color-coded using a continuous spectrum from gray (p-value > 0.05) to blue (p-value < 0.05). We mark in red the TFs which were found to be enriched in the H3K4me1 footprint analyses and not in the PU.1-only analysis. *Source: Lin et al. (2015)* (modified to fit thesis format and/or clarify key points).

For that, we obtained DNase-seq data from HUVEC cells and applied HINT to identify footprints. Then, we were able to perform a TF enrichment analysis on the genomic regions that carry the NF-κB recognition motifs and that do not carry such motif, to understand the regulatory players involved in assisting the binding of NF-κB in the absence of its canonical DNA motif. This analysis, which is a part of the study performed by Kolovos et al., is presented as a case study for our computational footprinting framework in the next subsections.

Computational Footprinting

DNase-seq raw DNA sequences (reads) from HUVECs were obtained from ENCODE Project Consortium (2012). The data is available in the gene expression omnibus (GEO) repository with accession number GSM816646. The short reads from the DNase-seq experiment were aligned to the human reference genome (ENCODE Project Consortium, 2012) using Bowtie (Langmead and Salzberg, 2012). To identify the DHSs (DNase-seq enriched regions), the software F-seq (Boyle et al., 2008) was

applied to the aligned reads (using the procedure described in Section 4.1.2).

Then, we applied the DNASE-ONLY HINT model (see "DNASE-ONLY MODEL" in Section 3.2.2) to the DNase-seq data. We followed the experimental settings as described in Chapter 4. Briefly, we extended the DHSs by 5,000 bp to each side and applied our DNASE-ONLY HMM model trained with DNase-seq data from the HUVEC experiments presented in this thesis.

The resulting footprint predictions from the DNASE-ONLY HINT were separated in two categories. The footprint predictions that overlaps NF-κB ChIP-seq enriched regions (peaks) (Papantonis et al., 2012) that: (1) carries the canonical NF-κB motif and (2) do not carry such motif. Then we performed a TF enrichment analysis in these two different conditions, considering the footprint predictions (overlapping NF-κB peaks) as our target genomic region set. In both TF enrichment analyses the background genomic regions correspond to random regions in the human genome. The size of the background genomic region sets were 100 times higher than the size of the target genomic region sets.

Results

NF-κB binding predominantly occurs at already-active (upon inflammation stimuli) distal regulatory regions called enhancers. These are mostly intragenic, display little overlap with CTCF-bound sites, and half carry the canonical motif or remain bound by NF-κB at 60 min after inflammatory stimulation. To obtain a more precise view of NF-κB binding choices, we performed a TF enrichment analysis on DNase-seq footprints. The analysis consists on comparing two different genomic region sets: (1) NF-κB ChIP-seq enriched regions (peaks) at enhancer regions with the canonical NF-κB motif and (2) NF-κB ChIP-seq enriched regions (peaks) at enhancer regions without the canonical NF-κB motif.

The result of the TF enrichment analysis can be seen in Figure 5.23. We exhibit a heatmap that combines the two conditions tested. The color code is a gradient from blue (TFs enriched in NF-κB peaks with motif) to white (no enrichment) to red (TFs enriched in NF-κB peaks without motif).

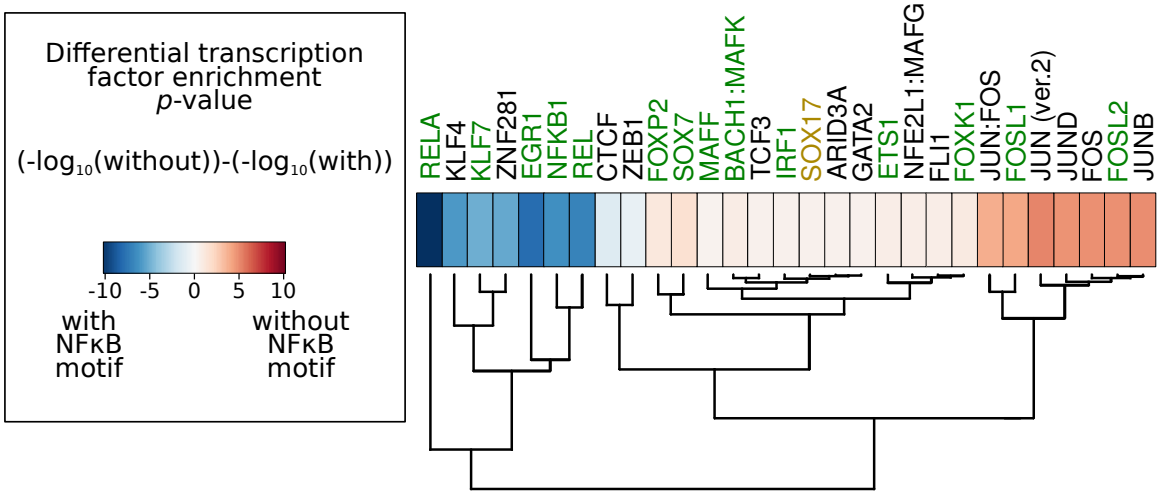

Figure 5.23: HUVEC cells footprint enrichment analysis results. Heatmap showing the TF enrichment analysis results at HUVEC enhancer regions that overlap the footprints predicted with the DNASE-ONLY HINT model. The heatmap represents enrichment between ChIP-seq enriched regions with (blue) and without the canonical NF-κB motif (red). TFs induced or repressed by the inflammatory response factor TNFα are demarcated green and yellow, respectively. *Source: Kolovos et al.* (modified to fit thesis format and/or clarify key points).

As expected, REL-like TFs (REL, RELA and NF-κB1) are enriched in the NF-κB peaks with canonical motif, since their DNA recognition motif are very similar to the canonical NF-κB motif. Furthermore, the TFs KLF4, KLF7, ZNF281 and EGR1 appear to be significantly enriched in these regions. Among these factors, there are the TNFα-induced REL, RELA, NF-κB1, KLF7 and EGR1. On the other hand, we observe that the JUN-related TFs JUN, JUND, JUNB, FOS, FOS, FOSL1 and FOSL2 are significantly associated to NF-κB-binding enhancer regions without the canonical NF-κB motif. The only TNFα-repressed TF that appeared in our enrichment analysis – SOX17 – appears to be slightly associated to the regions without NF-κB motif. The implications of such discovery are still under investigation. The mechanisms behind these NF-κB-associated TFs and human inflammatory response are further explored in Kolovos et al..

In summary, the analysis have shown that REL-like TFs (RELA and NF-κB1) are markedly enriched at the NF-κB peaks with canonical motif; whereas JUN-like TFs (JUN and FOS) appear to be enriched at the NF-κB peaks without canonical motif. Further analyses show that the footprint enrichment analysis prediction is backed by ENCODE Project Consortium (2012) ChIP-seq data from HUVECs, where co-binding of NF-κB and JUN/FOS (and to a lesser extent GATA2) was most prominent at enhancers without NF-κB recognition sites.

Differently from the previous case study, in which different cell types were being analyzed for regulatory elements within their open chromatin regions; the analysis presented in this case study was based on two conditions which differed only in the presence/absence of the NF-κB motif. In this case, the usage of DNase-seq alleviates much of the noise from intervening unbound sequences. The results presented in this section, combined with other experimental data, were used to understand the mechanisms behind hijacked enhancers and the regulatory role of NF-κB in the human inflammatory mechanism. The complete results of these experiments are found in Kolovos et al..

CHAPTER 6

Conclusion

This work aimed at analyzing relevant features of computational footprinting methods, which use mathematical models to predict active transcription factor binding sites (TFBSs) with open chromatin data. We devised a novel computational footprinting framework – HINT – which uses DNase-seq and histone modification ChIP-seq data to predict active TFBSs. HINT is the first method to integrate the full resolution data of both DNase-seq and histone modification ChIP-seq. We performed a comprehensive evaluation of 14 different computational footprinting methods, which showed that our method HINT significantly outperformed its competitors. Furthermore, we addressed a number of relevant characteristics on computational footprinting methods. Finally, we presented real case scenarios in which footprint predictions obtained with HINT aided in the understanding of regulatory mechanisms.

HINT Method

We devised HINT – a novel HMM-based computational footprinting method that segments the genome based on the full resolution signals of DNase-seq and histone modification ChIP-seq data. We investigated the performance of five different HMM topologies. The ORIGINAL DNASE + HISTONE topology was created to recognize the grammar of active TFBSs. Furthermore, we devised two topologies which also combine DNase and histone data: the DNASE + HISTONE ASYMMETRIC PEAKS topology considers the intrinsic asymmetry observed for histone modification peaks and the DNASE + HISTONE WITHOUT SLOPE topology consists of a simpler HMM model which considers only signal intensity. Moreover, we designed two additional topologies which: use only DNase-seq data and use only histone modification ChIP-seq data. We observed that the ORIGINAL DNASE + HISTONE HMM topology outperforms all other HMM topologies (Section 5.1.1). However, it is noticeable that the DNASE-ONLY topology's accuracies are very close to the DNASE + HISTONE topologies' accuracies. On the other hand, the HISTONE-ONLY topology presented the lowest accuracies. Our results showed that: (1) The proper integration of DNase-seq and histone modifications increases the accuracy of the prediction of active TFBSs and (2) The DNase-seq data has a great predictive power given its high spatial specificity.

Furthermore, we tested, for the ORIGINAL DNASE + HISTONE topology, a number of different histone modification combinations. We tested models using individual, pairs and triples of the activating histone modifications H3K4me1, H3K4me3, H3K9ac, H3K27ac and H2A.Z. We showed in Section 5.1.2 that many combinations perform equally well. However, the histone modifications H3K4me1, H3K4me3 and H3K27ac seem to be particularly good predictors of open chromatin regions. As expected, models containing more histone modifications generally outperformed models with less histone modifications. However, the increase in performance is smaller when considering higher number of histone modifications. This result, together with the fact that one of the goals of our model is to generate accurate predictions with as few assays as possible, does justify an optimal number of assays between a combination of two to three histone modifications.

With regard to HINT's training, we observed that it is cell-type train-independent (Section 5.1.3). This means that an HMM model trained with data from one cell type does not present a significant change in accuracy when applied to another cell type. In practice, one could use a trained HMM

(for a particular topology) and apply it to data from any other cell type. Although we did not test this claim on different organisms, this seems to be the case, since the training robustness stems from the efficacy of the normalization strategy. This result is very important from a practical perspective because it allows the use of the method without the need to train new HMM models, given new data.

Guidelines for Computational Footprinting

Until now, it was not clear the extent in which experimental issues such as the DNase-seq sequence cleavage bias and the transcription factor (TF) residence time had in the performance of computational footprinting methods. We performed an in-depth investigation of a number of features relevant for the identification of active TFBSs using open chromatin data. We highlighted three insightful experiments: (1) the selection of an optimal scoring strategy for computational footprinting methods and whether such scoring strategy alone could outperform more complex approaches; (2) the impact on performance of the DNase-seq sequence cleavage bias and (3) issues regarding the TF residence time.

The TC-Rank is a computational footprinting method which consists on scoring and ranking motif-predicted binding sites (MPBSs) based on the tag count (TC) metric. In contrast to positive evaluations of the TC-Rank by previous works (Cuellar-Partida et al., 2012; He et al., 2014) we show that it has poor sensitivity performance as indicated by the area under the receiver operating characteristic (ROC) curve (AUC) at low false positive rate (FPR) levels. Such poor sensitivity was further evidenced by observing the very low FP-Exp values of the gene expression evaluation methodology. The ability of a footprint-specific metric, such as the FLR, to distinguish a change in binding events appears to be a distinct advantage of computational footprinting methods over a more general statistic, such as the TC, that only captures overall DNase hypersensitivity in a large window around MPBSs. On the other hand, as pointed in Section 5.2.1 the TC metric outperformed the footprint score (FS), position weight matrix (PWM) bit-score and method-specific scoring metrics on ranking footprints. This shows that, while using TC to rank MPBSs and applying a cutoff strategy does not provide good results, using TC to rank already-predicted footprints is the best strategy observed.

The refined DNase-seq protocol and experimental artifacts presented in He et al. (2014) underscore that robust *in silico* techniques are required to correct for experimental artifacts and to derive valid biological predictions. In Section 5.2.2 we showed that the correction of DNase-seq signal using the "DNase hypersensitivity site (DHS) sequence cleavage bias" approach estimates virtually removes the effects of sequence bias artifacts on computational footprinting. We demonstrated that such correction can be performed prior to the execution of the computational footprinting method. On the other hand, ignoring experimental artifacts might lead to false predictions, as observed previously for Neph et al.'s predicted *de novo* motifs (Neph et al., 2012; He et al., 2014).

It was shown in Sung et al. (2014) that TFs with low residence time do not present a recognizable footprint pattern. Therefore, these factors would not be accurately predicted by computational footprinting methods. This issue was discussed in details in Section 5.4. Although the TF residence time is not an issue that can be solved computationally, we showed that we can use the protection score to indicate footprints of TFs with potential short binding time. Such footprint predictions of TFs with low protection score should be interpreted with caution.

Comparative Analyses on Computational Footprinting Methods

Our comparative evaluation analysis presented in Section 5.3 indicates the superior performance (in decreasing order) of HINT, DNase2TF and PIQ in the prediction of active TFBS in all evaluated scenarios. Moreover, tools implementing these methods were user-friendly and had lower computational demands than other evaluated methods. Clearly, the choice of computational footprinting approaches should also be based on experimental design aspects. For example, PIQ is the only method supporting

analysis of replicates and time-series. On the other hand, studies requiring footprint predictions for latter *de novo* motif analysis should use segmentation approaches as HINT or DNase2TF.

The availability, usability and scalability of software tools implementing the methods are also important features. Neph, HINT, PIQ and Wellington provide tutorials and software to run experiments with few command line calls. Of those, only HINT, PIQ and Wellington natively support standard genomic formats as input. Site-centric methods Cuellar, BinDNase, Centipede and FLR require a single execution and input data per TF and cell type, while segmentation methods require an execution per cell type only. These site-centric methods have computational demands 5 times (FLR and Cuellar) to 50 times (BinDNase and Centipede) higher than the slowest segmentation method (Wellington) in our comparative analysis using the `Benchmarking Dataset` (Table 4.4).

Examples of the infeasibility of site-centric methods on the basis of processing time can also be taken from the case studies presented here (Section 5.6). The segmentation approach HINT was executed four times in the dendritic cell case study (one time for each cell type) and one time in the HUVEC inflammation case study (only the cell type HUVEC was analyzed). The total running time of these five computational footprinting methods was ~140 hours (or ~1.5 hour in a 100-core computational cluster). On the other hand, a site-centric approach would have to be executed for each TF in which we are interested in performing the TF enrichment analysis, for each cell type. This makes a total of ~3000 executions (given a restricted set of 600 tested TFs), with an estimated execution time (based on the fastest site-centric method PIQ) of 579,000 hours (or 241 days in a 100-core computational cluster).

In conclusion, the assessment of computational footprinting methods is a demanding task, both computationally and technically. We have created a fair and reproducible benchmarking dataset for evaluation of TF binding using two validation approaches: using ChIP-seq and using gene expression. Although the rationales of the ChIP-seq and gene expression evaluation procedures are, in principle, very different, we observed a high agreement between their respective ranking of methods. This is evidence that this study provides a robust map of the accuracy of state-of-the-art computational footprinting methods. We provide all statistics, basic data and computational scripts to evaluate future computational footprinting methods. These resources are available at:

`http://costalab.org/hint-bc`

This is an important resource for increasing transparency and reproducibility of research on computational footprinting methods.

Downstream Analyses and Case Studies

We present two common downstream analyses based on footprint predictions: the *de novo* motif finding and the TF enrichment analysis.

We performed a *de novo* motif finding procedure on footprints predicted with HINT combining the tools "discriminative regular expression motif elicitation" (DREME) and "local motif enrichment analysis" (CENTRIMO; Section 5.5). We identified six novel motifs associated to human embryonic cell type H1-hESC. Five of these motifs presented a particularly noticeable peak-dip-peak DNase-seq pattern, indicative of active TF binding. Although this analysis used a particularly simple experiment design, it exemplifies downstream analyses that can only be performed on footprint predictions from segmentation-based computational footprinting methods.

In Section 5.6 we presented two case studies in which our computational footprinting method HINT was successfully applied to identify TFs associated to different biological conditions. Both studies use the same downstream analysis on the predicted footprints: the TF enrichment analysis. We have shown that it is possible to explore different HINT's HMM topologies to address specific biological questions. The inclusion of such case studies had the main goal of showing the flexibility of our

computational footprinting framework towards very different experimental scenarios. There were differences in the organism under study (mouse *vs* human), in the availability of input data (histone modification ChIP-seq *vs* DNase-seq) and in the biological questions asked. Nevertheless, in these two distinct scenarios, HINT's predictions aided in the identification of the respective key regulatory players.

6.1 Future Work

Although we covered a number of different challenges on the detection of active TFBSs with computational genomic footprinting methods, this research area still has some unexplored aspects. In this section we categorize these research opportunities as: computational footprinting method extension and further downstream analyses that can be performed with footprint predictions.

Computational Footprinting Method Extension

We have systematically investigated the DNase-seq sequence cleavage bias. However, as extensively explored in Meyer and Liu (2014), open chromatin genomic data are affected by other artifacts stemming from either the biological protocol or the computational pre-processing steps, such as: (1) chromatin fragmentation and size selection, (2) tissue-specific signal variability generated by the phenol chloroform extraction step commonly used to separate deoxyribonucleic acid (DNA) from protein, (3) DNA amplification biases and duplications, (4) particularities of read mapping algorithms and (5) TF binding characteristics. HINT can still be further expanded to encompass the correction of other experimental artifacts.

Moreover, in this thesis we focused on using the open chromatin data from DNase-seq and histone modification ChIP-seq. However, there are novel experimental biological assays, such as ATAC-seq (assay for transposase-accessible chromatin) (Buenrostro et al., 2013), which are able to generate a nucleotide-resolution genome-wide map of open chromatin regions. ATAC-seq also exhibits active TF's footprint-like patterns similar to DNase-seq; and has two major advantages over DNase-seq: (1) ATAC-seq is less technical and (2) ATAC-seq requires a much lower number of cells to start the protocol. Furthermore, current efforts are being made in order to obtain the genome-wide signal for these experimental assays (DNase-seq, ChIP-seq and ATAC-seq) in a single-cell manner (Buenrostro et al., 2015). In this new paradigm, we are going to be able to study tissue heterogeneity by analyzing open chromatin profiles of individual cells.

Further Downstream Analyses

Here we have shown two common downstream analysis: the *de novo* motif finding (Section 5.5) and the TF enrichment analysis (Section 5.6). Nevertheless, there are a number of different downstream analyses that can be performed on computationally-predicted footprints, such as: (1) integration with TF ChIP-seq data – to determine the exact position where the TF is binding without relying on purely sequence-based metrics (Pique-Regi et al., 2011); (2) differential footprinting – which searches for footprints that occur at particular cell conditions and finds, within these footprints, regulatory elements associated to such conditions (He et al., 2012); and (3) integrative analyses – in which the footprints are integrated with further chromatin dynamics information, such as the spatial configuration of the chromatin, to infer indirect binding events and protein tethering (Thurman et al., 2012).

Furthermore, no effort was made to improve current available downstream analysis, such as the *de novo* motif finding, to handle the massive data generated by computational footprinting methods. The research of novel downstream methods which are devised particularly for footprints is needed to explore the full potential of computational footprint predictions.

APPENDIX A

Appendix – Supplementary Tables

Table A.1: Summary of DNase-seq data. DNase-seq datasets used as input for computational footprinting methods (main comparative experiments and further empirical analyses) are highlighted in bold. The other DNase-seq datasets were used in the DNase-seq bias estimates clustering analysis. We represent both DNase-seq protocols as: single-hit (SH; generated in Crawford lab (ENCODE Project Consortium, 2012)) and double-hit (DH; generated in Stamatoyannopoulous lab (ENCODE Project Consortium, 2012)). Naked deoxyribonucleic acid (DNA) DNase-seq experiments are represented as NK. *Source: Gusmao et al. (2016)* (modified to fit thesis format and/or clarify key points).

Cell Type	Protocol	UCSC ID	GEO/NCBI ID	# Mapped Reads
H1-hESC	**SH**	**wgEncodeEH000556**	**GSM816632**	**110303078**
HeLa-S3	**SH**	**wgEncodeEH000540**	**GSM816643**	**54267867**
HepG2	**SH**	**wgEncodeEH000537**	**GSM816662**	**50838536**
HUVEC	**SH**	**wgEncodeEH000548**	**GSM816646**	**31848532**
K562	**SH**	**wgEncodeEH000530**	**GSM816655**	**365820647**
LNCaP	**SH**	**wgEncodeEH001097**	**GSM816637**	**163625945**
MCF-7	**SH**	**wgEncodeEH000579**	**GSM816627**	**89113893**
K562	**NK**	**–**	**GSM1496625**	**202001412**
MCF-7	**NK**	**–**	**GSM1496626**	**210715393**
H7-hESC	**DH**	**wgEncodeEH000511**	**GSM736638**	**302050785**
			GSM736610	
HepG2	**DH**	**wgEncodeEH000482**	**GSM736637**	**168883956**
			GSM736639	
HUVEC	**DH**	**wgEncodeEH000488**	**GSM736575**	**429088276**
			GSM736533	
K562	**DH**	**wgEncodeEH000484**	**GSM736629**	**179970820**
			GSM736566	
m3134	**DH**	**wgEncodeEM001721**	**GSM1014196**	**127594903**
IMR90	**NK**	**–**	**SRA068503**	**138604440**
H7-hESC	SH	wgEncodeEH002554	GSM1008596	433296955
CD14+	SH	wgEncodeEH003466	GSM1008582	287039145
SK-N-SH	SH	wgEncodeEH003483	GSM1008585	287186739
MCF-7/RandshRNA	SH	wgEncodeEH003468	GSM1008603	288004844
K562/SAHA-Ctrl	SH	wgEncodeEH003489	GSM1008580	503410467
MCF-7	SH	wgEncodeEH003470	GSM1008565	89113893
IMR90	SH	wgEncodeEH003482	GSM1008586	303769598
HeLa-S3/IFNa4h	SH	wgEncodeEH000577	GSM816633	110348694
K562/G2-Mphase	SH	wgEncodeEH003472	GSM1008567	431722812
K562/G1phase	SH	wgEncodeEH003469	GSM1008602	426934260
K562/SAHA1um72h	SH	wgEncodeEH003490	GSM1008558	503301111
MCF-7/HypLacAc	SH	wgEncodeEH001745	GSM816670	244207602

Cell Type	Protocol	UCSC ID	GEO/NCBI ID	# Mapped Reads
K562/NaBut	SH	wgEncodeEH002559	GSM1008601	267722720
CD20+RO01794	SH	wgEncodeEH003465	GSM1008588	256442597
GM12878	SH	wgEncodeEH000534	GSM816665	245090730
A549	SH	wgEncodeEH001095	GSM816649	133567925
MCF-7/CTCFshRNA	SH	wgEncodeEH003467	GSM1008581	295954052
K562/ZNFP5	DH	wgEncodeEH003016	–	70400755
CD20+RO01778	DH	wgEncodeEH001884	GSM1024765 GSM1024766	71398619
HeLa-S3	DH	wgEncodeEH000495	GSM736510 GSM736564	70669968
K562/ZNF4C50C4	DH	wgEncodeEH003009	–	82579252
A549	DH	wgEncodeEH001180	GSM736506 GSM736580	75764710
K562/ZNFb34A8	DH	wgEncodeEH003012	–	95113482
K562/ZNFg54A11	DH	wgEncodeEH003015	–	76873236
CD14+	DH	wgEncodeEH001196	–	33322702
MCF-7/EstCtrl0h	DH	wgEncodeEH003018	GSM1024764 GSM1024767	151170759
MCF-7/Est100nm1h	DH	wgEncodeEH003017	GSM1024783 GSM1024784	164440980
K562/ZNF4G7D3	DH	wgEncodeEH003010	–	83034668
K562/ZNFe103C6	DH	wgEncodeEH003013	–	78100065
K562/ZNF2C10C5	DH	wgEncodeEH003008	–	173334712
LHCN-M2	DH	wgEncodeEH003005	GSM1024786 GSM1024787	89558026
LHCN-M2/Diff4d	DH	wgEncodeEH003006	GSM1024771 GSM1024772	120358720
H1-hESC	DH	wgEncodeEH000496	GSM736582	24431583
MCF-7	DH	wgEncodeEH000502	GSM736581 GSM736588	89482135
K562/ZNFf41B2	DH	wgEncodeEH003014	–	109124535
CD14+/RO01746	DH	wgEncodeEH001196	GSM1024791	67698560
GM12878	DH	wgEncodeEH000492	GSM736496 GSM736620	47899421
K562/ZNFa41C6	DH	wgEncodeEH003011	–	99106989
HepG2	DH	wgEncodeEH000476	GSM646559	69810990
K562	DH	wgEncodeEH000480	GSM646567	71250291
CD20+RO01778	DH	wgEncodeEH002442	GSM1014525	240594387
K562/ZNFP5	DH	wgEncodeEH003153	–	346226678
K562/ZNFa41C6	DH	wgEncodeEH003152	–	372806338
LHCN-M2	DH	wgEncodeEH003149	GSM1014524	255134452
LHCN-M2/Diff4d	DH	wgEncodeEH003154	GSM1014539	357827356
H7-hESC	DH	wgEncodeEH000834	GSM646563	302050785
HUVEC	DH	wgEncodeEH002460	GSM1014528	429088276
A549	DH	wgEncodeEH003146	GSM1014517	350629033

Table A.2: Summary of the histone modification ChIP-seq data. All datasets were generated in Bernstein lab, associated to ENCODE Project Consortium (2012). *Source: Gusmao et al. (2014)* (modified to fit thesis format and/or clarify key points).

Cell Type	Data Type	UCSC Access.	GEO ID	# Mapped Reads
H1-hESC	H3K4me1	wgEncodeEH000106	GSM733782	27286943
H1-hESC	H3K4me3	wgEncodeEH000086	GSM733657	19203931
H1-hESC	H3K9ac	wgEncodeEH000109	GSM733773	30288927
H1-hESC	H3K27ac	wgEncodeEH000997	GSM733718	31993560
H1-hESC	H2A.Z	wgEncodeEH002082	GSM1003579	76761942
K562	H3K4me1	wgEncodeEH000046	GSM733692	29197613
K562	H3K4me3	wgEncodeEH000048	GSM733680	25153055
K562	H3K9ac	wgEncodeEH000049	GSM733778	32634427
K562	H3K27ac	wgEncodeEH000043	GSM733656	24470196
K562	H2A.Z	wgEncodeEH001038	GSM733786	38763180
GM12878	H3K4me1	wgEncodeEH000033	GSM733772	48444878
GM12878	H3K4me3	wgEncodeEH000028	GSM733708	64016296
GM12878	H3K9ac	wgEncodeEH000035	GSM733677	19513948
GM12878	H3K27ac	wgEncodeEH000030	GSM733771	19582373
GM12878	H2A.Z	wgEncodeEH001033	GSM733767	32327975
HeLa-S3	H3K4me1	wgEncodeEH001750	GSM798322	38435440
HeLa-S3	H3K4me3	wgEncodeEH001017	GSM733682	35897578
HepG2	H3K4me1	wgEncodeEH001749	GSM798321	52320612
HepG2	H3K4me3	wgEncodeEH000095	GSM733737	18620773

Table A.3: Position frequency matrices (PFMs) and transcription factors (TFs) ChIP-seq used in the ChIP-seq evaluation methodology. ChIP-seq was obtained from multiple labs within the ENCODE Project Consortium (2012). PFMs were obtained from Jaspar (Mathelier et al., 2014), Uniprobe (Robasky and Bulyk, 2011) and Transfac (Matys et al., 2006). *Source: Gusmao et al. (2016)* (modified to fit thesis format and/or clarify key points).

Cell	Factor	PFM Repository	PFM ID	ChIP-seq Lab	ChIP-seq ID	Number of Motifs	Number of Peaks	Number of Peaks with Motifs	Percentage of Peaks with Motifs (%)
H1-hESC	ATF3	Jaspar	MA0093.2	Myers	wgEncodeEH001566	691899	4804	1777	36.99
H1-hESC	BACH1	Transfac	M00495	Snyder	wgEncodeEH002842	614421	11457	2941	25.66
H1-hESC	BRCA1	Jaspar	MA0133.1	Snyder	wgEncodeEH002801	333055	2025	15	0.74
H1-hESC	CEBPB	Jaspar	MA0466.1	Snyder	wgEncodeEH002825	1342548	15557	9720	62.47
H1-hESC	CTCF	Jaspar	MA0139.1	Myers	wgEncodeEH001649	565933	54070	41994	77.66
H1-hESC	EGR1	Jaspar	MA0162.2	Myers	wgEncodeEH001538	1060314	8743	5225	59.76
H1-hESC	FOSL1	Jaspar	MA0477.1	Myers	wgEncodeEH001660	699220	1111	61	5.49
H1-hESC	GABP	Jaspar	MA0062.2	Myers	wgEncodeEH001534	181503	5652	2165	38.30
H1-hESC	JUN	Jaspar	MA0488.1	Snyder	wgEncodeEH001854	832374	2148	646	30.07
H1-hESC	JUND	Jaspar	MA0491.1	Snyder	wgEncodeEH002023	717223	9550	3784	39.62
H1-hESC	MAFK	Jaspar	MA0496.1	Snyder	wgEncodeEH002828	1221488	11425	7849	68.70
H1-hESC	MAX	Jaspar	MA0058.2	Farnham	wgEncodeEH001757	855374	11124	3126	28.10
H1-hESC	MYC	Jaspar	MA0147.2	Snyder	wgEncodeEH002795	614797	4551	1161	25.51
H1-hESC	NRF1	Jaspar	MA0506.1	Snyder	wgEncodeEH001847	137117	4513	3636	80.56
H1-hESC	POU5F1	Jaspar	MA0142.1	Myers	wgEncodeEH001636	2201678	3994	2757	69.02
H1-hESC	RAD21	Jaspar	MA0139.1	Snyder	wgEncodeEH001836	565933	55674	42657	76.61
H1-hESC	REST	Jaspar	MA0138.2	Myers	wgEncodeEH001498	629168	13269	6440	48.53
H1-hESC	RFX5	Jaspar	MA0510.1	Snyder	wgEncodeEH001835	629248	1695	697	41.12
H1-hESC	RXRA	Jaspar	MA0512.1	Myers	wgEncodeEH001560	1110004	1306	276	21.13
H1-hESC	SIX5	Jaspar	MA0088.1	Myers	wgEncodeEH001528	1032447	3422	1680	49.09
H1-hESC	SP1	Jaspar	MA0079.3	Myers	wgEncodeEH001529	1797400	15103	5303	35.11
H1-hESC	SP2	Jaspar	MA0516.1	Myers	wgEncodeEH002302	1587339	2469	1247	50.50
H1-hESC	SP4	Uniprobe	UP00002	Myers	wgEncodeEH002317	503235	5752	1802	31.32
H1-hESC	SRF	Jaspar	MA0083.2	Myers	wgEncodeEH001533	1024023	5102	2969	58.19
H1-hESC	TCF12	Jaspar	MA0521.1	Myers	wgEncodeEH001531	893836	7829	1904	24.31
H1-hESC	USF1	Jaspar	MA0093.2	Myers	wgEncodeEH001532	691899	26028	18288	70.26
H1-hESC	USF2	Jaspar	MA0526.1	Snyder	wgEncodeEH001837	759040	6952	4488	64.55
H1-hESC	YY1	Jaspar	MA0095.2	Myers	wgEncodeEH001567	1325447	18310	6506	35.53
H1-hESC	ZNF143	Jaspar	MA0088.1	Snyder	wgEncodeEH002802	1032447	30687	3809	12.41
HeLa-S3	BRCA1	Jaspar	MA0133.1	Snyder	wgEncodeEH001814	333055	8114	88	1.08
HeLa-S3	CEBPB	Jaspar	MA0466.1	Snyder	wgEncodeEH001815	1342548	61004	26770	43.88
HeLa-S3	CTCF	Jaspar	MA0139.1	Bernstein	wgEncodeEH001012	565933	52783	38397	72.74
HeLa-S3	E2F4	Jaspar	MA0470.1	Snyder	wgEncodeEH000689	173646	2831	1397	49.34
HeLa-S3	E2F6	Jaspar	MA0471.1	Snyder	wgEncodeEH000692	1051116	4775	1457	30.51
HeLa-S3	ELK1	Jaspar	MA0028.1	Snyder	wgEncodeEH002864	100691	4809	1892	39.34
HeLa-S3	FOS	Jaspar	MA0476.1	Snyder	wgEncodeEH000647	762222	9325	6900	73.99
HeLa-S3	GABP	Jaspar	MA0062.2	Myers	wgEncodeEH001504	181503	6761	3571	52.81
HeLa-S3	JUN	Jaspar	MA0488.1	Snyder	wgEncodeEH000746	832374	21903	3302	15.07
HeLa-S3	JUND	Jaspar	MA0491.1	Snyder	wgEncodeEH000745	717223	31633	21182	66.96
HeLa-S3	MAFK	Jaspar	MA0496.1	Snyder	wgEncodeEH002856	1221488	14185	8658	61.03
HeLa-S3	MAX	Jaspar	MA0058.2	Snyder	wgEncodeEH002830	855374	29647	3204	10.80
HeLa-S3	MYC	Jaspar	MA0147.2	Snyder	wgEncodeEH000648	614797	10226	1647	16.10
HeLa-S3	NFYA	Jaspar	MA0060.2	Snyder	wgEncodeEH002066	428913	5978	2537	42.43
HeLa-S3	NFYB	Jaspar	MA0502.1	Snyder	wgEncodeEH002067	470725	7156	4139	57.83
HeLa-S3	NRF1	Jaspar	MA0506.1	Snyder	wgEncodeEH000723	137117	2915	2369	81.26
HeLa-S3	RAD21	Jaspar	MA0139.1	Snyder	wgEncodeEH001789	565933	43420	30385	69.97
HeLa-S3	REST	Jaspar	MA0138.2	Myers	wgEncodeEH001629	629168	10247	4524	44.14
HeLa-S3	STAT1	Jaspar	MA0137.3	Snyder	wgEncodeEH000614	1272026	16158	5655	34.99
HeLa-S3	USF2	Jaspar	MA0526.1	Snyder	wgEncodeEH001819	759040	12306	6099	49.56
HeLa-S3	ZNF143	Jaspar	MA0088.1	Snyder	wgEncodeEH002028	1032447	7048	1865	26.46
HepG2	ARID3A	Jaspar	MA0151.1	Snyder	wgEncodeEH002858	2112327	17614	1041	5.91
HepG2	ATF3	Jaspar	MA0018.2	Myers	wgEncodeEH001568	496476	3290	270	8.20
HepG2	BHLHE40	Jaspar	MA0464.1	Myers	wgEncodeEH001515	572185	2859	1186	41.48
HepG2	BRCA1	Jaspar	MA0133.1	Snyder	wgEncodeEH001859	333055	1497	15	1.00
HepG2	CEBPB	Jaspar	MA0466.1	Myers	wgEncodeEH002304	1342548	18114	10146	56.01
HepG2	CTCF	Jaspar	MA0139.1	Myers	wgEncodeEH001516	565933	55733	44323	79.52
HepG2	ELF1	Jaspar	MA0473.1	Myers	wgEncodeEH001641	1026618	17998	8728	48.49
HepG2	GABP	Jaspar	MA0062.2	Myers	wgEncodeEH001548	181503	10105	4722	46.72

Cell	Factor	PFM Repository	PFM ID	ChIP-seq Lab	ChIP-seq ID	Number of Motifs	Number of Peaks	Number of Peaks with Motifs	Percentage of Peaks with Motifs (%)
HepG2	JUN	Jaspar	MA0488.1	Snyder	wgEncodeEH001794	832374	12669	7136	56.32
HepG2	JUND	Jaspar	MA0491.1	Myers	wgEncodeEH001470	717223	21606	8490	39.29
HepG2	MAFF	Jaspar	MA0495.1	Snyder	wgEncodeEH001841	1215808	37587	29284	77.90
HepG2	MAFK	Jaspar	MA0496.1	Snyder	wgEncodeEH001842	1221488	61847	44299	71.62
HepG2	MAX	Jaspar	MA0058.2	Snyder	wgEncodeEH002796	855374	11852	2101	17.72
HepG2	MYC	Jaspar	MA0147.2	Iyer	wgEncodeEH000545	614797	4411	1160	26.29
HepG2	NRF1	Jaspar	MA0506.1	Snyder	wgEncodeEH001802	137117	1902	1699	89.32
HepG2	RAD21	Jaspar	MA0139.1	Myers	wgEncodeEH001608	565933	54261	40827	75.24
HepG2	REST	Jaspar	MA0138.2	Myers	wgEncodeEH001549	629168	6021	2848	47.30
HepG2	RXRA	Jaspar	MA0512.1	Myers	wgEncodeEH001506	1110004	17059	4628	27.12
HepG2	SP1	Jaspar	MA0079.3	Myers	wgEncodeEH001561	1797400	25465	5277	20.72
HepG2	SP2	Jaspar	MA0516.1	Myers	wgEncodeEH002264	1587339	2626	567	21.59
HepG2	SRF	Jaspar	MA0083.2	Myers	wgEncodeEH001611	1024023	5311	2693	50.70
HepG2	USF1	Jaspar	MA0093.2	Myers	wgEncodeEH001472	691899	21885	14209	64.92
HepG2	USF2	Jaspar	MA0526.1	Snyder	wgEncodeEH001804	759040	6290	4621	73.46
HepG2	YY1	Jaspar	MA0095.2	Myers	wgEncodeEH001661	1325447	17871	4035	22.57
HUVEC	CTCF	Jaspar	MA0139.1	Iyer	wgEncodeEH000551	565933	43982	36279	82.48
HUVEC	FOS	Jaspar	MA0476.1	Farnham	wgEncodeEH001774	762222	46726	29319	62.74
HUVEC	GATA2	Jaspar	MA0036.2	Farnham	wgEncodeEH001758	1028569	27454	6162	22.44
HUVEC	JUN	Jaspar	MA0488.1	Snyder	wgEncodeEH000719	832374	29502	4220	14.30
HUVEC	MAX	Jaspar	MA0058.2	Snyder	wgEncodeEH000768	855374	9120	2650	29.05
HUVEC	MYC	Jaspar	MA0147.2	Iyer	wgEncodeEH000561	614797	5143	1213	23.58
K562	ARID3A	Jaspar	MA0151.1	Snyder	wgEncodeEH002861	2112327	9026	606	6.71
K562	ATF1	Uniprobe	UP00020	Struhl	wgEncodeEH002865	246442	14864	2609	17.55
K562	ATF3	Jaspar	MA0018.2	Struhl	wgEncodeEH000700	496476	1233	165	13.38
K562	BACH1	Transfac	M00495	Snyder	wgEncodeEH002846	614421	3806	1980	52.02
K562	BHLHE40	Jaspar	MA0464.1	Snyder	wgEncodeEH001857	572185	22497	5958	26.48
K562	CCNT2	Jaspar	MA0140.2	Struhl	wgEncodeEH001864	708983	20057	2284	11.38
K562	CEBPB	Jaspar	MA0466.1	Snyder	wgEncodeEH001821	1342548	38715	24789	64.02
K562	CTCF	Jaspar	MA0139.1	Snyder	wgEncodeEH002797	565933	54387	41122	75.60
K562	CTCFL	Jaspar	MA0139.1	Myers	wgEncodeEH001652	565933	11533	8878	76.97
K562	E2F4	Jaspar	MA0470.1	Farnham	wgEncodeEH000671	173646	8181	2809	34.33
K562	E2F6	Jaspar	MA0471.1	Farnham	wgEncodeEH000676	1051116	16312	4251	26.06
K562	EFOS	Jaspar	MA0476.1	White	wgEncodeEH001207	762222	10256	8796	85.76
K562	EGATA	Jaspar	MA0036.2	White	wgEncodeEH001208	1028569	11478	3846	33.50
K562	EGR1	Jaspar	MA0162.2	Myers	wgEncodeEH001646	1060314	36997	25164	68.01
K562	EJUNB	Jaspar	MA0490.1	White	wgEncodeEH001210	717235	12287	7788	63.38
K562	EJUND	Jaspar	MA0491.1	White	wgEncodeEH001211	717223	26674	11027	41.33
K562	ELF1	Jaspar	MA0473.1	Myers	wgEncodeEH001619	1026618	27780	14324	51.56
K562	ELK1	Jaspar	MA0028.1	Snyder	wgEncodeEH003356	100691	2961	1315	44.41
K562	ETS1	Jaspar	MA0098.2	Myers	wgEncodeEH001580	1319961	10726	1734	16.16
K562	FOS	Jaspar	MA0476.1	Snyder	wgEncodeEH000619	762222	7646	3423	44.76
K562	FOSL1	Jaspar	MA0477.1	Myers	wgEncodeEH001637	699220	11174	8865	79.33
K562	GABP	Jaspar	MA0062.2	Myers	wgEncodeEH001604	181503	14393	5406	37.55
K562	GATA1	Jaspar	MA0035.3	Farnham	wgEncodeEH000638	1040470	4074	1923	47.20
K562	GATA2	Jaspar	MA0036.2	Farnham	wgEncodeEH000683	1028569	10648	4267	40.07
K562	IRF1	Jaspar	MA0050.2	Snyder	wgEncodeEH002798	2330047	8352	3274	39.20
K562	JUN	Jaspar	MA0488.1	Snyder	wgEncodeEH000620	832374	9848	2150	21.83
K562	JUND	Jaspar	MA0491.1	Snyder	wgEncodeEH002164	717223	40052	15395	38.43
K562	MAFF	Jaspar	MA0495.1	Snyder	wgEncodeEH002804	1215808	25074	17425	69.49
K562	MAFK	Jaspar	MA0496.1	Snyder	wgEncodeEH001844	1221488	19317	12423	64.31
K562	MAX	Jaspar	MA0058.2	Snyder	wgEncodeEH002869	855374	31436	4766	15.16
K562	MEF2A	Jaspar	MA0052.2	Myers	wgEncodeEH001663	3210613	5631	2664	47.30
K562	MYC	Jaspar	MA0147.2	Snyder	wgEncodeEH000621	614797	5023	1312	26.11
K562	NFE2	Jaspar	MA0501.1	Snyder	wgEncodeEH000624	796063	2637	2177	82.55
K562	NFYA	Jaspar	MA0060.2	Snyder	wgEncodeEH002021	428913	4286	2770	64.62
K562	NFYB	Jaspar	MA0502.1	Snyder	wgEncodeEH002024	470725	10096	7786	77.11
K562	NR2F2	Uniprobe	UP00009	Myers	wgEncodeEH002382	626663	16678	2971	17.81
K562	NRF1	Jaspar	MA0506.1	Snyder	wgEncodeEH001796	137117	4211	3114	73.94
K562	PU1	Jaspar	MA0080.3	Myers	wgEncodeEH001482	2040890	28677	24657	85.98
K562	RAD21	Jaspar	MA0139.1	Snyder	wgEncodeEH000649	565933	17627	16218	92.00

Cell	Factor	PFM Repository	PFM ID	ChIP-seq Lab	ChIP-seq ID	Number of Motifs	Number of Peaks	Number of Peaks with Motifs	Percentage of Peaks with Motifs (%)
K562	REST	Jaspar	MA0138.2	Myers	wgEncodeEH001638	629168	15849	4191	26.44
K562	RFX5	Jaspar	MA0510.1	Snyder	wgEncodeEH002033	629248	2201	475	21.58
K562	SIX5	Jaspar	MA0088.1	Myers	wgEncodeEH001483	1032447	4194	1554	37.05
K562	SMC3	Jaspar	MA0139.1	Snyder	wgEncodeEH001845	565933	23598	20753	87.94
K562	SP1	Jaspar	MA0079.3	Myers	wgEncodeEH001578	1797400	7206	3269	45.36
K562	SP2	Jaspar	MA0516.1	Myers	wgEncodeEH001653	1587339	3124	1735	55.53
K562	SRF	Jaspar	MA0083.2	Myers	wgEncodeEH001600	1024023	4717	1473	31.22
K562	STAT1	Jaspar	MA0137.3	Snyder	wgEncodeEH000664	1272026	1476	204	13.82
K562	STAT2	Jaspar	MA0517.1	Snyder	wgEncodeEH000666	3077582	1923	1132	58.86
K562	STAT5A	Jaspar	MA0519.1	Myers	wgEncodeEH002347	1292097	9811	2033	20.72
K562	TAL1	Jaspar	MA0140.2	Snyder	wgEncodeEH001824	708983	26260	11345	43.20
K562	THAP1	Jaspar	MA0597.1	Myers	wgEncodeEH001655	561707	3506	338	9.64
K562	TR4	Jaspar	MA0504.1	Farnham	wgEncodeEH000682	825980	587	170	28.96
K562	USF1	Jaspar	MA0093.2	Myers	wgEncodeEH001583	691899	18521	11966	64.60
K562	USF2	Jaspar	MA0526.1	Snyder	wgEncodeEH001797	759040	3083	2271	73.66
K562	YY1	Jaspar	MA0095.2	Farnham	wgEncodeEH000684	1325447	4948	3035	61.33
K562	ZBTB33	Jaspar	MA0527.1	Myers	wgEncodeEH001569	82928	3285	1454	44.26
K562	ZBTB7A	Uniprobe	UP00047	Myers	wgEncodeEH001620	412506	21711	801	3.68
K562	ZNF143	Jaspar	MA0088.1	Snyder	wgEncodeEH002030	1032447	29069	3628	12.48
K562	ZNF263	Jaspar	MA0528.1	Farnham	wgEncodeEH000630	2577084	3081	1110	36.02
Mcf7	ER(160m)	Jaspar	MA0112.2	Hager	GSM1325251	801832	1450	801	55.24
Mcf7	ER(40m)	Jaspar	MA0112.2	Hager	GSM1325250	801832	10397	4696	45.16
HepG2	ARID3A	Jaspar	MA0151.1	Snyder	wgEncodeEH002858	2112327	17614	1041	5.91
HepG2	ATF3	Jaspar	MA0018.2	Myers	wgEncodeEH001568	496476	3290	270	8.20
HepG2	BHLHE40	Jaspar	MA0464.1	Myers	wgEncodeEH001515	572185	2859	1186	41.48
HepG2	BRCA1	Jaspar	MA0133.1	Snyder	wgEncodeEH001859	333055	1497	15	1.00
HepG2	CEBPB	Jaspar	MA0466.1	Myers	wgEncodeEH002304	1342548	18114	10146	56.01
HepG2	CTCF	Jaspar	MA0139.1	Myers	wgEncodeEH001516	565933	55733	44323	79.52
HepG2	ELF1	Jaspar	MA0473.1	Myers	wgEncodeEH001641	1026618	17998	8728	48.49
HepG2	GABP	Jaspar	MA0062.2	Myers	wgEncodeEH001548	181503	10105	4722	46.72
HepG2	JUN	Jaspar	MA0488.1	Snyder	wgEncodeEH001794	832374	12669	7136	56.32
HepG2	JUND	Jaspar	MA0491.1	Myers	wgEncodeEH001470	717223	21606	8490	39.29
HepG2	MAFF	Jaspar	MA0495.1	Snyder	wgEncodeEH001841	1215808	37587	29284	77.90
HepG2	MAFK	Jaspar	MA0496.1	Snyder	wgEncodeEH001842	1221488	61847	44299	71.62
HepG2	MAX	Jaspar	MA0058.2	Snyder	wgEncodeEH002796	855374	11852	2101	17.72
HepG2	MYC	Jaspar	MA0147.2	Iyer	wgEncodeEH000545	614797	4411	1160	26.29
HepG2	NRF1	Jaspar	MA0506.1	Snyder	wgEncodeEH001802	137117	1902	1699	89.32
HepG2	RAD21	Jaspar	MA0139.1	Myers	wgEncodeEH001608	565933	54261	40827	75.24
HepG2	REST	Jaspar	MA0138.2	Myers	wgEncodeEH001549	629168	6021	2848	47.30
HepG2	RXRA	Jaspar	MA0512.1	Myers	wgEncodeEH001506	1110004	17059	4628	27.12
HepG2	SP1	Jaspar	MA0079.3	Myers	wgEncodeEH001561	1797400	25465	5277	20.72
HepG2	SP2	Jaspar	MA0516.1	Myers	wgEncodeEH002264	1587339	2626	567	21.59
HepG2	SRF	Jaspar	MA0083.2	Myers	wgEncodeEH001611	1024023	5311	2693	50.70
HepG2	USF1	Jaspar	MA0093.2	Myers	wgEncodeEH001472	691899	21885	14209	64.92
HepG2	USF2	Jaspar	MA0526.1	Snyder	wgEncodeEH001804	759040	6290	4621	73.46
HepG2	YY1	Jaspar	MA0095.2	Myers	wgEncodeEH001661	1325447	17871	4035	22.57
HUVEC	CTCF	Jaspar	MA0139.1	Iyer	wgEncodeEH000551	565933	43982	36279	82.48
HUVEC	FOS	Jaspar	MA0476.1	Farnham	wgEncodeEH001774	762222	46726	29319	62.74
HUVEC	GATA2	Jaspar	MA0036.2	Farnham	wgEncodeEH001758	1028569	27454	6162	22.44
HUVEC	JUN	Jaspar	MA0488.1	Snyder	wgEncodeEH000719	832374	29502	4220	14.30
HUVEC	MAX	Jaspar	MA0058.2	Snyder	wgEncodeEH000768	855374	9120	2650	29.05
HUVEC	MYC	Jaspar	MA0147.2	Iyer	wgEncodeEH000561	614797	5143	1213	23.58
K562	ARID3A	Jaspar	MA0151.1	Snyder	wgEncodeEH002861	2112327	9026	606	6.71
K562	ATF1	Uniprobe	UP00020	Struhl	wgEncodeEH002865	246442	14864	2609	17.55
K562	ATF3	Jaspar	MA0018.2	Struhl	wgEncodeEH000700	496476	1233	165	13.38
K562	BACH1	Transfac	M00495	Snyder	wgEncodeEH002846	614421	3806	1980	52.02
K562	BHLHE40	Jaspar	MA0464.1	Snyder	wgEncodeEH001857	572185	22497	5958	26.48
K562	CCNT2	Jaspar	MA0140.2	Struhl	wgEncodeEH001864	708983	20057	2284	11.38
K562	CEBPB	Jaspar	MA0466.1	Snyder	wgEncodeEH001821	1342548	38715	24789	64.02
K562	CTCF	Jaspar	MA0139.1	Snyder	wgEncodeEH002797	565933	54387	41122	75.60
K562	CTCFL	Jaspar	MA0139.1	Myers	wgEncodeEH001652	565933	11533	8878	76.97

Cell	Factor	PFM Repository	PFM ID	ChIP-seq Lab	ChIP-seq ID	Number of Motifs	Number of Peaks	Number of Peaks with Motifs	Percentage of Peaks with Motifs (%)
K562	E2F4	Jaspar	MA0470.1	Farnham	wgEncodeEH000671	173646	8181	2809	34.33
K562	E2F6	Jaspar	MA0471.1	Farnham	wgEncodeEH000676	1051116	16312	4251	26.06
K562	EFOS	Jaspar	MA0476.1	White	wgEncodeEH001207	762222	10256	8796	85.76
K562	EGATA	Jaspar	MA0036.2	White	wgEncodeEH001208	1028569	11478	3846	33.50
K562	EGR1	Jaspar	MA0162.2	Myers	wgEncodeEH001646	1060314	36997	25164	68.01
K562	EJUNB	Jaspar	MA0490.1	White	wgEncodeEH001210	717235	12287	7788	63.38
K562	EJUND	Jaspar	MA0491.1	White	wgEncodeEH001211	717223	26674	11027	41.33
K562	ELF1	Jaspar	MA0473.1	Myers	wgEncodeEH001619	1026618	27780	14324	51.56
K562	ELK1	Jaspar	MA0028.1	Snyder	wgEncodeEH003356	100691	2961	1315	44.41
K562	ETS1	Jaspar	MA0098.2	Myers	wgEncodeEH001580	1319961	10726	1734	16.16
K562	FOS	Jaspar	MA0476.1	Snyder	wgEncodeEH000619	762222	7646	3423	44.76
K562	FOSL1	Jaspar	MA0477.1	Myers	wgEncodeEH001637	699220	11174	8865	79.33
K562	GABP	Jaspar	MA0062.2	Myers	wgEncodeEH001604	181503	14393	5406	37.55
K562	GATA1	Jaspar	MA0035.3	Farnham	wgEncodeEH000638	1040470	4074	1923	47.20
K562	GATA2	Jaspar	MA0036.2	Farnham	wgEncodeEH000683	1028569	10648	4267	40.07
K562	IRF1	Jaspar	MA0050.2	Snyder	wgEncodeEH002798	2330047	8352	3274	39.20
K562	JUN	Jaspar	MA0488.1	Snyder	wgEncodeEH000620	832374	9848	2150	21.83
K562	JUND	Jaspar	MA0491.1	Snyder	wgEncodeEH002164	717223	40052	15395	38.43
K562	MAFF	Jaspar	MA0495.1	Snyder	wgEncodeEH002804	1215808	25074	17425	69.49
K562	MAFK	Jaspar	MA0496.1	Snyder	wgEncodeEH001844	1221488	19317	12423	64.31
K562	MAX	Jaspar	MA0058.2	Snyder	wgEncodeEH002869	855374	31436	4766	15.16
K562	MEF2A	Jaspar	MA0052.2	Myers	wgEncodeEH001663	3210613	5631	2664	47.30
K562	MYC	Jaspar	MA0147.2	Snyder	wgEncodeEH000621	614797	5023	1312	26.11
K562	NFE2	Jaspar	MA0501.1	Snyder	wgEncodeEH000624	796063	2637	2177	82.55
K562	NFYA	Jaspar	MA0060.2	Snyder	wgEncodeEH002021	428913	4286	2770	64.62
K562	NFYB	Jaspar	MA0502.1	Snyder	wgEncodeEH002024	470725	10096	7786	77.11
K562	NR2F2	Uniprobe	UP00009	Myers	wgEncodeEH002382	626663	16678	2971	17.81
K562	NRF1	Jaspar	MA0506.1	Snyder	wgEncodeEH001796	137117	4211	3114	73.94
K562	PU1	Jaspar	MA0080.3	Myers	wgEncodeEH001482	2040890	28677	24657	85.98
K562	RAD21	Jaspar	MA0139.1	Snyder	wgEncodeEH000649	565933	17627	16218	92.00
K562	REST	Jaspar	MA0138.2	Myers	wgEncodeEH001638	629168	15849	4191	26.44
K562	RFX5	Jaspar	MA0510.1	Snyder	wgEncodeEH002033	629248	2201	475	21.58
K562	SIX5	Jaspar	MA0088.1	Myers	wgEncodeEH001483	1032447	4194	1554	37.05
K562	SMC3	Jaspar	MA0139.1	Snyder	wgEncodeEH001845	565933	23598	20753	87.94
K562	SP1	Jaspar	MA0079.3	Myers	wgEncodeEH001578	1797400	7206	3269	45.36
K562	SP2	Jaspar	MA0516.1	Myers	wgEncodeEH001653	1587339	3124	1735	55.53
K562	SRF	Jaspar	MA0083.2	Myers	wgEncodeEH001600	1024023	4717	1473	31.22
K562	STAT1	Jaspar	MA0137.3	Snyder	wgEncodeEH000664	1272026	1476	204	13.82
K562	STAT2	Jaspar	MA0517.1	Snyder	wgEncodeEH000666	3077582	1923	1132	58.86
K562	STAT5A	Jaspar	MA0519.1	Myers	wgEncodeEH002347	1292097	9811	2033	20.72
K562	TAL1	Jaspar	MA0140.2	Snyder	wgEncodeEH001824	708983	26260	11345	43.20
K562	THAP1	Jaspar	MA0597.1	Myers	wgEncodeEH001655	561707	3506	338	9.64
K562	TR4	Jaspar	MA0504.1	Farnham	wgEncodeEH000682	825980	587	170	28.96
K562	USF1	Jaspar	MA0093.2	Myers	wgEncodeEH001583	691899	18521	11966	64.60
K562	USF2	Jaspar	MA0526.1	Snyder	wgEncodeEH001797	759040	3083	2271	73.66
K562	YY1	Jaspar	MA0095.2	Farnham	wgEncodeEH000684	1325447	4948	3035	61.33
K562	ZBTB33	Jaspar	MA0527.1	Myers	wgEncodeEH001569	82928	3285	1454	44.26
K562	ZBTB7A	Uniprobe	UP00047	Myers	wgEncodeEH001620	412506	21711	801	3.68
K562	ZNF143	Jaspar	MA0088.1	Snyder	wgEncodeEH002030	1032447	29069	3628	12.48
K562	ZNF263	Jaspar	MA0528.1	Farnham	wgEncodeEH000630	2577084	3081	1110	36.02
LnCaP	AR(R1881)	Jaspar	MA0007.2	Yu	GSM353644	913583	51799	12978	25.05
LnCaP	AR(ethl)	Jaspar	MA0007.2	Yu	GSM353643	913583	6103	685	11.22
m3134	GR(DEX)	Jaspar	MA0113.2	Stam.	SRP004871	1051822	28078	7270	25.89

Table A.4: PFMs used in the gene expression evaluation methodology. PFMs were obtained from Jaspar (Mathelier et al., 2014). *Source: Gusmao et al. (2016)* (modified to fit thesis format and/or clarify key points).

Factor	PFM ID
MYCN	MA0104.3
PBX1	MA0070.1
TCF7L1	MA0522.1
TP53	MA0106.2
ETS1	MA0098.2
TCF7L2	MA0523.1
AR	MA0007.2
SRY	MA0084.1
TFAP2C	MA0524.1
ZNF354C	MA0130.1
FOXO1	MA0480.1
NR5A2	MA0505.1
FLII	MA0475.1
FOXH1	MA0479.1
CREB1	MA0018.2
EGR2	MA0472.1
REST	MA0138.2
RFX5	MA0510.1
SOX3	MA0514.1
FOXD3	MA0041.1
HNF4G	MA0484.1
SOX9	MA0077.1
NKX3-1	MA0124.1
INSM1	MA0155.1
ERG	MA0474.1
STAT1	MA0137.3
USF1	MA0093.2
EGR1	MA0162.2
CTCF	MA0139.1
MAFB	MA0117.1
E2F1	MA0024.2
STAT4	MA0518.1
MAFK	MA0496.1
NFYA	MA0060.2
GABPA	MA0062.2
YY1	MA0095.2
KLF4	MA0039.2
SRF	MA0083.2
STAT3	MA0144.2
HOXA5	MA0158.1
SREBF2	MA0596.1
HOXA9	MA0594.1
FOSL2	MA0478.1
TCF12	MA0521.1
SOX10	MA0442.1
FOXP2	MA0593.1
ATOH1	MA0461.1
PPARG	MA0066.1
GATA3	MA0037.2
NR2F1	MA0017.1
SOX17	MA0078.1
NKX2-5	MA0503.1
HLF	MA0043.1
HNF1A	MA0046.1
NR2E3	MA0164.1
PAX2	MA0067.1
PAX5	MA0014.2
RXRA	MA0512.1
HINFP	MA0131.1
MYOG	MA0500.1
NKX3-2	MA0122.1
EBF1	MA0154.2
HNF1B	MA0153.1
ESR1	MA0112.2
NR2C2	MA0504.1
FOXC1	MA0032.1
NRF1	MA0506.1
HNF4A	MA0114.2
LHX3	MA0135.1
FOXL1	MA0033.1
RUNX2	MA0511.1
FOXI1	MA0042.1
FOXA2	MA0047.2
HSF1	MA0486.1
E2F4	MA0470.1
ZNF143	MA0088.1
FOXP1	MA0481.1
FEV	MA0156.1
TFAP2A	MA0003.2
FOXQ1	MA0040.1
ELF5	MA0136.1
ZNF263	MA0528.1
E2F3	MA0469.1
PAX4	MA0068.1
ESRRA	MA0592.1
T	MA0009.1
EN1	MA0027.1
FOXD1	MA0031.1
HLTF	MA0109.1
MAX	MA0058.2
CDX2	MA0465.1
FOXO3	MA0157.1
THAP1	MA0597.1
SREBF1	MA0595.1
GFI1	MA0038.1
GATA4	MA0482.1
ZBTB33	MA0527.1
FOSL1	MA0477.1
FOXA1	MA0148.3
FOXF2	MA0030.1
ELK1	MA0028.1
RFX2	MA0600.1
MAFF	MA0495.1
SP2	MA0516.1
NHLH1	MA0048.1
ZFX	MA0146.2
ELK4	MA0076.2
CEBPB	MA0466.1
NFE2L2	MA0150.2
BCL6	MA0463.1
NFIL3	MA0025.1
PRDM1	MA0508.1
NFKB1	MA0105.3
TBP	MA0108.2
BRCA1	MA0133.1
ESR2	MA0258.2
RREB1	MA0073.1
RELA	MA0107.1
JUN	MA0489.1
IRF1	MA0050.2
REL	MA0101.1
SOX5	MA0087.1
E2F6	MA0471.1
TP63	MA0525.1
NR4A2	MA0160.1
PAX6	MA0069.1
KLF1	MA0493.1
NR3C1	MA0113.2
ELF1	MA0473.1
MYC	MA0147.2
NFATC2	MA0152.1
SPI1	MA0080.3
ZEB1	MA0103.2
KLF5	MA0599.1
RUNX1	MA0002.2
MEIS1	MA0498.1
GATA2	MA0036.2
GFI1B	MA0483.1
MYB	MA0100.2
MECOM	MA0029.1
GATA1	MA0035.3
MEF2C	MA0497.1
BHLHE40	MA0464.1

Table A.5: Summary of the gene expression data. Expression profiling by array (Affymetrix Human Exon 1.0 ST Array) data was obtained in ENCODE Project Consortium (2012). *Source: Gusmao et al. (2016)* (modified to fit thesis format and/or clarify key points).

Cell Type	GEO Accession	# Samples
GM12878	GSE12760	20
H1-hESC	GSE14863	4
K562	GSE12760	21

Bibliography

Alberts B., Johnson A., Lewis J., et al. *Molecular Biology of the Cell*. Garland Science, 5th edition, 2007.

Arvey A., Agius P., Noble W. S., and Leslie C. Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Research*, 22(9):1723–1734, 2012.

Ashoor H., Hérault A., Kamoun A., et al. HMCan: a method for detecting chromatin modifications in cancer samples using ChIP-seq data. *Bioinformatics*, 29(23):2979–2986, 2013.

Bailey T. L. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, 27(12): 1653–1659, 2011.

Bailey T. L. and Machanick P. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Research*, 40(17):e128, 2012.

Bánfai B., Jia H., Khatun J., et al. Long noncoding RNAs are rarely translated in two human cell lines. *Genome Research*, 22(9):1646–1657, 2012.

Belz G. T. and Nutt S. L. Transcriptional programming of the dendritic cell network. *Nature Review Immunology*, 12(2):101–113, Feb. 2012.

Benjamini Y. and Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

Bishop C. M. *Pattern recognition and machine learning*. Springer, 1st edition, 2006.

Boyle A. P., Guinney J., Crawford G. E., and Furey T. S. F-seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*, 24(21):2537–2538, 2008.

Boyle A. P., Song L., Lee B.-K., et al. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Research*, 21(3):456–464, 2011.

Buenrostro J. D., Giresi P. G., Zaba L. C., Chang H. Y., and Greenleaf W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12):1213–1218, 2013.

Buenrostro J. D., Wu B., Litzenburger U. M., et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490, 2015.

Charos A. E., Reed B. D., Raha D., et al. A highly integrated and complex PPARGC1A transcription factor binding network in HepG2 cells. *Genome Research*, 22(9):1668–1679, Sep 2012.

Crawford G. E., Holt I. E., Mullikin J. C., et al. Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *Proceedings of the National Academy of Sciences of the United States of America*, 101(4):992–997, 2004.

Cuellar-Partida G., Buske F. A., McLeay R. C., et al. Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics*, 28(1):56–62, 2012.

Davis J. and Goadrich M. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 233–240, New York, NY, USA, 2006.

Demšar J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Derrien T., Estellé J., Sola S. M., et al. Fast computation and applications of genome mappability. *PLoS ONE*, 7(1):e30377+, 2012.

Diaz A., Park K., Lim D. A., and Song J. S. Normalization, bias correction, and peak calling for ChIP-seq. *Statistical Applications in Genetics and Molecular Biology*, 11(3), 2012.

Duda R. O., Stork D. G., and Hart P. E. *Pattern classification*. Wiley, 2nd edition, 2000.

Durbin R., Eddy S. R., Krogh A., and Mitchison G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1st edition, 1998.

ENCODE Project Consortium . An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

Grant C. E., Bailey T. L., and Noble W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.

Guertin M., Zhang X., Coonrod S., and Hager G. Transient estrogen receptor binding and p300 redistribution support a squelching mechanism for estradiol-repressed genes. *Molecular Endocrinology*, 28(9):1522–1533, 2014.

Gusmao E. G., Dieterich C., and Costa I. G. Prediction of transcription factor binding sites by integrating dnase digestion and histone modification. In *Proceedings of the 7th Brazilian Symposium on Bioinformatics*, Campo Grande, Mato Grosso do Sul, Brazil, 2012.

Gusmao E. G., Dieterich C., Zenke M., and Costa I. G. Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. *Bioinformatics*, 30(22):3143–3151, 2014.

Gusmao E. G., Allhoff M., Zenke M., and Costa I. G. Analysis of computational footprinting methods for DNase sequencing experiments. *Nature Methods*, advance online publication, Feb. 2016.

Hayden E. C. Is the $1,000 genome for real?, 2014.

Hayden M. S. and Ghosh S. NF-κB, the first quarter-century: remarkable progress and outstanding questions. *Genes & development*, 26(3):203–234, Feb. 2012.

He H. H., Meyer C. A., Chen M. W., et al. Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics. *Genome Research*, 22(6):1015–1025, 2012.

He H. H., Meyer C. A., Hu S. S., et al. Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nature Methods*, 11(1):73–78, 2014.

Hesselberth J. R., Chen X., Zhang Z., et al. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature Methods*, 6(4):283–289, 2009.

Hon G., Wang W., and Ren B. Discovery and Annotation of Functional Chromatin Signatures in the Human Genome. *PLoS Computational Biology*, 5(11):e1000566+, 2009.

Hubbard T., Barker D., Birney E., et al. The ensembl genome database project. *Nucleic Acids Research*, 30(1):38–41, 2002.

John S., Sabo P. J., Thurman R. E., et al. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature Genetics*, 43(3):264–268, 2011.

Johnson D. S., Mortazavi A., Myers R. M., and Wold B. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–1502, 2007.

Kähärä J. and Lähdesmäki H. BinDNase: A discriminatory approach for transcription factor binding prediction using DNase i hypersensitivity data. *Bioinformatics*, 2015.

Kempe S., Kestler H., Lasar A., and Wirth T. NF-κB controls the global pro-inflammatory response in endothelial cells: evidence for the regulation of a pro-atherogenic program. *Nucleic Acids Research*, 33(16):5308–5319, 2005.

Kim J., Chu J., Shen X., Wang J., and Orkin S. H. An extended transcriptional network for pluripotency of embryonic stem cells. *Cell*, 132(6):1049–1061, 2008.

Kolovos P., Georgomanolis T., Nikolic M., et al. Enhancer hijacking reveals a multimodal role of NF-κB during the immediate-early inflammatory response. *In Review*.

Kundaje A., Kyriazopoulou-Panagiotopoulou S., Libbrecht M., et al. Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome Research*, 22(9):1735–1747, 2012.

Lall S. Primers on chromatin. *Nature Structural & Molecular Biology*, 14(11):1110–1115, Nov 2007.

Langmead B. and Salzberg S. L. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

Li H. and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

Lin Q., Chauvistré H., Costa I. G., et al. Epigenetic program and transcription factor circuitry of dendritic cell development. *Nucleic Acids Research*, 43(20):9680–9693, 2015.

Lodish H., Berk A., Kaiser C. A., et al. *Molecular Cell Biology*. W. H. Freeman, 6th edition, 2007.

Luo J., Ying K., He P., and Bai J. Properties of Savitzky-Golay digital differentiators. *Digital Signal Processing*, 15(2):122–136, 2005.

Lutovac M. D., Tosic D. V., and Evans B. L. *Filter Design for Signal Process Using MATLAB and Mathematica*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition, 2000.

Madden H. H. Comments on the Savitzky-Golay convolution method for least-squares fit smoothing and differentiation of digital data. *Analytical Chemistry*, 50:1383–1386, 1978.

Mahony S. and Benos P. V. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Research*, 35(Web Server issue):gkm272–258, 2007.

Malnou C. E., Brockly F., Favard C., et al. Heterodimerization with different Jun proteins controls c-Fos intranuclear dynamics and distribution. *Journal of Biological Chemistry*, 285(9):6552–6562, 2010.

Maston G. A., Evans S. K., and Green M. R. Transcriptional Regulatory Elements in the Human Genome. *Annual Review of Genomics and Human Genetics*, 7(1):29–59, 2006.

Mathelier A. and Wasserman W. W. The next generation of transcription factor binding site prediction. *PLoS Computational Biology*, 9(9):e1003214+, 2013.

Mathelier A., Zhao X., Zhang A. W., et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 42(D1): D142–D147, 2014.

Matys V., Kel-Margoulis O. V., Fricke E., et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, 34(Database issue):D108–D110, 2006.

McNally J. G., Müller W. G., Walker D., Wolford R., and Hager G. L. The glucocorticoid receptor: rapid exchange with regulatory sites in living cells. *Science*, 287(5456):1262–1265, 2000.

Merad M., Sathe P., Helft J., Miller J., and Mortha A. The dendritic cell lineage: Ontogeny and function of dendritic cells and their subsets in the steady state and the inflamed setting. *Annual Review of Immunology*, 31(1):563–604, 2013.

Meyer C. and Liu X. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews. Genetics*, 2014.

Mitchell T. M. *Machine Learning*. McGraw-Hill Science/Engineering/Math, 1st edition, 1997.

Nakahashi H., Kwon K.-R., Resch W., et al. A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell reports*, 3(5):1678–1689, 2013.

Natarajan A., Yardımcı G. G., Sheffield N. C., Crawford G. E., and Ohler U. Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Research*, 22(9):1711–1722, 2012.

Nature Methods Editorial . The difficulty of a fair comparison. *Nature Methods*, 12:273, 2015.

Neph S., Vierstra J., Stergachis A. B., et al. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489(7414):83–90, 2012.

Ouyang Z., Zhou Q., and Wong W. H. ChIP-seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proceedings of the National Academy of Sciences*, 106(51):21521–21526, 2009.

Papantonis A., Kohro T., Baboo S., et al. TNFα signals through specialized factories where responsive coding and miRNA genes are transcribed. *The EMBO journal*, 31(23):4404–4414, 2012.

Park P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, 2009.

Piper J., Elze M. C., Cauchy P., et al. Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Research*, 41(21):e201, 2013.

Pique-Regi R., Degner J. F., Pai A. A., et al. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research*, 21(3):447–455, 2011.

Rabiner L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Ritchie M. E., Phipson B., Wu D., et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):gkv007–e47, 2015.

Robasky K. and Bulyk M. L. UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Research*, 39(Database issue), 2011.

Rusk N. Torrents of sequence. *Nature Methods*, 8(1):44, 2010.

Sabo P. J., Hawrylycz M., Wallace J. C., et al. Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proceedings of the National Academy of Sciences of the United States of America*, 101(48):16837–16842, 2004a.

Sabo P. J., Humbert R., Hawrylycz M., et al. Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. *Proceedings of the National Academy of Sciences of the United States of America*, 101(13):4537–4542, 2004b.

Schaub M. A., Boyle A. P., Kundaje A., Batzoglou S., and Snyder M. Linking disease associations with regulatory information in the human genome. *Genome Research*, 22(9):1748–1759, 2012.

Sharp Z. D., Mancini M. G., Hinojos C. A., et al. Estrogen-receptor-α exchange and chromatin dynamics are ligand- and domain-dependent. *Journal of Cell Science*, 119(19):4101–4116, 2006.

Shendure J. and Ji H. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145, 2008.

Shenoi B. A. *Introduction to Digital Signal Processing and Filter Design*. Wiley-Interscience, 1st edition, 2005.

Sherwood R. I., Hashimoto T., O'Donnell C. W., et al. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nature Biotechnology*, 32(2):171–8, 2014.

Siepel A., Bejerano G., Pedersen J. S., et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050, 2005.

Stormo G. D. DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, 2000.

Sung M.-H. H., Guertin M. J., Baek S., and Hager G. L. DNase footprint signatures are dictated by factor dynamics and DNA sequence. *Molecular Cell*, 56(2):275–285, 2014.

Tewari A., Yardimci G., Shibata Y., et al. Chromatin accessibility reveals insights into androgen receptor activation and transcriptional specificity. *Genome Biology*, 13(10):R88+, 2012.

Thurman R. E., Rynes E., Humbert R., et al. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, 2012.

Tilgner H., Knowles D. G., Johnson R., et al. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Research*, 22(9):1616–1625, 2012.

Tsankov A. M., Gu H., Akopian V., et al. Transcription factor binding dynamics during human ES cell differentiation. *Nature*, 518(7539):344–349, 2015.

Tucker T., Marra M., and Friedman J. M. Massively parallel sequencing: The next big thing in genetic medicine. *The American Journal of Human Genetics*, 85(2):142–154, 2009.

Vernot B., Stergachis A. B., Maurano M. T., et al. Personal and population genomics of human regulatory variation. *Genome Research*, 22(9):1689–1697, 2012.

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

Whitfield T., Wang J., Collins P., et al. Functional analysis of transcription factor binding sites in human promoters. *Genome Biology*, 13(9):R50+, 2012.

Wilczynski B., Dojer N., Patelak M., and Tiuryn J. Finding evolutionarily conserved cis-regulatory modules with a universal set of motifs. *BMC Bioinformatics*, 10(1):82+, 2009.

Yardımcı G. G., Frank C. L., Crawford G. E., and Ohler U. Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. *Nucleic Acids Research*, 42(19): 11865–78, 2014.

Yip K., Cheng C., Bhardwaj N., et al. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biology*, 13 (9):R48+, 2012.

Yu J., Yu J., Mani R.-S., et al. An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. *Cancer Cell*, 17(5):443–454, 2010.

Zhang Y., Liu T., Meyer C. A., et al. Model-based analysis of ChIP-Seq (MACS). *Genome biology*, 9(9):R137+, 2008.