Computational models of gene expression regulation

by
Fatemeh BEHJATI ARDAKANI

A dissertation submitted towards the degree

Doctor of Natural Sciences (Dr. rer. nat)

of the Faculty of Mathematics and Computer Science

of Saarland University

Saarbrücken, 2019

Day of Colloquium	21.07.2020
Dean of the Faculty	Univ.-Prof. Dr. Thomas Schuster
Chair of the Committee	Prof. Dr. Kurt Mehlhorn
Reporters	
First Reviewer	Prof. Dr. Marcel Schulz
Second Reviewer	Prof. Dr. Tobias Marschall
Academic Assistant	Dr. Erisa Terolli

"You know that children are growing up when they start asking questions that have answers."

John J. Plomp

SAARLAND UNIVERSITY

Abstract

Mathematics and Informatics
Department of Computer Science

Doctor rerum naturalium

Computational models of gene expression regulation

by Fatemeh BEHJATI ARDAKANI

Throughout the last several decades, many efforts have been put into elucidating the genetic or epigenetic defects that result in various diseases. Gene regulation, i.e., the process of how genes are turned on and off in the right place and at the right time, is a paramount and prevailing question for researchers. Thanks to the discoveries made by researchers in this field, our understanding of interactions between proteins and DNA or proteins with themselves, as well as the dynamics of chromatin structure under different conditions, have substantially advanced. Even though there has been a lot achieved through these discoveries, there are still many unknown aspects about gene regulation. For instance, proteins called transcription factors (TFs) recognize and bind to specific regions of DNA and recruit the transcriptional machinery, which is essential for gene regulation. As there have been more than 2000 TFs identified in the human genome, it is important to study where they bind to or which genes they target. Computational approaches are important, in particular, as the biological experiments are often very expensive and cannot be done for all TFs. In 2016, a competition named *DREAM Challenge* was held encouraging researchers to develop novel computational tools for predicting the binding sites of several TFs. The first chapter of this thesis describes our machine learning approach to address this challenge within the scope of the competition. Using ensembles of random forest classifiers, we formulated our framework such that it is able to benefit from the tissue specificity inherent in the data leading to better generalization. Also, our models were tailored for spotting cofactors involved in the binding of TFs of interest. Comparing the important TFs that our computational models suggested with protein-protein association networks revealed that the models preferentially select motifs of TFs that are potential interaction partners in those networks.

Another important aspect beyond predicting TF binding is to link epigeneomics, such as histone modification (HM) data, with gene expression. We, particularly, concentrated on predicting expression in a subset of genes called bidirectional. Bidirectional genes are referred to as pairs of genes that are located on opposite strands of DNA close to each other. As the sequencing technologies advance, more such bidirectional configurations are being detected. This indicates that in order to understand the gene regulatory mechanisms, it would be beneficial to account for such promoter architectures. In the second and third chapters, we focused on genes having bidirectional promoter architectures utilizing high resolution epigenomic signatures and single cell RNA-seq data to dissect the complex epigenetic architecture at these promoters. Using single-cell RNA-seq data as the estimate of gene expression, we were able to generate a hypothetical model for gene regulation in bidirectional promoters. We showed that bidirectional promoters can be categorized into three architecture types with distinct characteristics. Each of these categories corresponds to a unique gene expression profile at single cell level.

The single cell RNA-seq data proved to be a powerful means for studying gene regulation. Therefore, in the last chapter, we proposed a novel approach for predicting gene expression at the single cell level using cis-regulatory motifs as well as epigenetic features. To achieve this, we designed a tree-guided multi-task learning framework that considers each cell as a task. Through this framework we were able to explain the single cell gene expression values using either TF binding affinities or TF ChIP-seq data measured at specific genomic regions. This allowed us to identify distinct TFs that show cell-type specific regulation in induced pluripotent stem cells. Our approach does not only limit to TFs, rather it can take any type of data that can potentially be used in explaining gene expression at single cell level. We believe that our findings can be used in drug discovery and development that can regulate the presence of TFs or other regulatory factors, which lead the cell fate into abnormal states, to prevent or cure diseases.

SAARLAND UNIVERSITY

Kurzfassung

Mathematics and Informatics
Department of Computer Science

Doctor rerum naturalium

Computational models of gene expression regulation

von Fatemeh BEHJATI ARDAKANI

In den letzten Jahrzehnten wurden große Anstrengungen unternommen, um die genetischen oder epigenetischen Defekte aufzuklären, die zu verschiedenen Krankheiten führen. Die Genregulation, d.h. der Prozess der Ein- und Abschaltung der Gene am richtigen Ort und zur richtigen Zeit reguliert, ist für die Forscher eine Frage von zentraler Bedeutung. Dank der Entdeckungen von Forschern auf diesem Gebiet ist unser Verständnis der Wechselwirkungen zwischen zwischen den Proteinen und der DNA oder der Proteine untereinander sowie der Dynamik der Chromatinstruktur unter verschiedenen Bedingungen wesentlich fortgeschritten. Obwohl durch diese Entdeckungen viel erreicht wurde, gibt es noch viele unbekannte Aspekte der Genregulation. Beispielsweise erkennen Proteine, sogenannte Transkriptionsfaktoren (Transcription Factors, TFs), bestimmte Bereiche der DNA und binden an diese und rekrutieren die Transkriptionsmaschinerie, die für die Genregulation erforderlich ist. Da mehr als 2000 TFs im menschlichen Genom identifiziert wurden, ist es wichtig zu untersuchen, wo sie binden oder auf welche Gene sie abzielen. Rechnerische Ansätze sind insbesondere wichtig, da die biologischen Experimente oft sehr teuer sind und nicht für alle TFs durchgeführt werden können. Im Jahr 2016 fand ein Wettbewerb namens *DREAM Challenge* statt, bei dem Forscher aufgefordert wurden, neuartige Rechenwerkzeuge zur Vorhersage der Bindungsstellen mehrerer TFs zu entwickeln. Das erste Kapitel dieser Arbeit beschreibt unseren Ansatz des maschinellen Lernens, um diese Herausforderung im Rahmen des Wettbewerbs anzugehen. Unter Verwendung von Ensembles von Random Forest Klassifikatoren haben wir unser Framework so formuliert, dass es von der Gewebespezifität der Daten profitiert und damit zu einer besseren Generalisierung führt. Außerdem wurden unsere Modelle auf das Erkennen von Kofaktoren angepasst, die an der Bindung von TFs beteiligt sind, die für uns von Interesse sind. Der Vergleich der wichtigen TFs, die unsere Computermodelle mit Protein-Protein-Assoziationsnetzwerken vorschlugen, ergab, dass die Modelle bevorzugt Motive von TFs auswählen, die potenzielle Interaktionspartner in diesen Netzwerken sind.

Ein weiterer wichtiger Aspekt, der über die Vorhersage der TF-Bindung hinausgeht, besteht darin, epigeneomische Faktoren wie Histonmodifikationsdaten (HM-Daten) mit der Genexpression zu verknüpfen. Wir konzentrierten uns insbesondere auf die Vorhersage der Expression in einer Untergruppe von Genen, die als bidirektional bezeichnet werden. Bidirektionale Gene werden als Paare von Genen bezeichnet, die sich auf gegenüberliegenden DNA-Strängen befinden und nahe beieinander liegen. Mit dem Fortschritt der Sequenzierungstechnologien werden immer mehr solche bidirektionalen Konfigurationen erkannt. Dies weist darauf hin, dass

es zum Verständnis der Genregulationsmechanismen vorteilhaft wäre, solche Promotorarchitekturen zu berücksichtigen. Im zweiten und dritten Kapitel konzentrierten wir uns auf Gene mit bidirektionalen Promotorarchitekturen, um mit Hilfe von epigenomischen Signaturen und Einzelzell-RNA-Sequenzdaten die komplexe epigenetische Architektur an diesen Promotoren zu analysieren. Unter Verwendung von Einzelzell-RNA-Sequenzdaten als Schätzung der Genexpression konnten wir ein hypothetisches Modell für die Genregulation in bidirektionalen Promotoren aufstellen. Wir haben gezeigt, dass bidirektionale Promotoren in drei Architekturtypen mit unterschiedlichen Merkmalen eingeteilt werden können. Jede dieser Kategorien entspricht einem eindeutigen Genexpressionsprofil auf Einzelzellebene.

Die Einzelzell-RNA-Sequenzdaten erwiesen sich als leistungsstarkes Mittel zur Untersuchung der Genregulation. Daher haben wir im letzten Kapitel einen neuen Ansatz zur Vorhersage der Genexpression auf Einzelzellebene unter Verwendung von cis-regulatorischen Motiven sowie epigenetischen Merkmalen vorgeschlagen. Um dies zu erreichen, haben wir ein baumgesteuertes Multitasking-Lernsystem entwickelt, das jede Zelle als eine Aufgabe betrachtet. Durch dieses Gerüst konnten wir die Einzelzellgenexpressionswerte entweder mit TF-Bindungsaffinitäten oder mit TF-ChIP-Sequenzdaten erklären, die in bestimmten Genomregionen gemessen wurden. Dies ermöglichte es uns, verschiedene TFs zu identifizieren, die eine zelltypspezifische Regulation in induzierten pluripotenten Stammzellen zeigen. Unser Ansatz beschränkt sich nicht nur auf TFs, sondern kann jede Art von Daten verwenden, die potentiell zur Erklärung der Genexpression auf Einzelzellebene verwendet werden können. Wir glauben, dass unsere Erkenntnisse für die Entdeckung und Entwicklung von Arzneimitteln verwendet werden können, die das Vorhandensein von TFs oder anderen regulatorischen Faktoren regulieren können, die die Zellen abnormal werden lassen, um Krankheiten zu verhindern oder zu heilen.

Acknowledgements

A very special gratitude goes out to my supervisor Prof. Dr. Marcel H. Schulz who guided me through my PhD journey.

Many thanks to the committee members of my thesis for their willingness to invest their time and effort in the doctoral examination procedure.

I am grateful to my parents, who have provided me through moral and emotional support in my life. I am also grateful to my other family members and friends who have supported me along the way.

With a special mention to my dear colleagues, Dilip Durai, Sivarajan Karunaniti, and Florian Schmidt who provided an enjoyable working atmosphere. It was fantastic to have the opportunity to do my research in the HGSB lab.

I am also grateful to the following collaborators Jörn Walter, Kathrin Kattler, Karl Nordström, Nina Gasparoni, Gills Gasparoni, everyone involved in the DEEP consortium, and the following university staff: Georg Friedrich, Ruth Schneppen-Christmann, and Joachim Büch for their unfailing support and assistance and the Saarland University for funding the PhD research.

And finally, last but by no means least, also to my other friends Narges Jafari, Jonas Fischer, Peter Ebert, Prabhav Kalaghatgi, Valentina Galata, Hosna Sattar, Nina Baumgarten, Janis Kalofolias, Kailash Budhathoki. Thank you for your scientific/moral support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

For centuries naturalists, geologists, and biologists have been inquisitive to fathom out the function behind mechanisms pertaining to vitality. Their efforts to unravel the basis of these mechanisms led to ample discoveries as well as innovative ideas for building suitable apparatus that can measure these mechanisms. For instance, in the case of gene expression, numerous sequencing protocols have been developed throughout the course of time to facilitate the task of measuring gene activity by estimating the abundance of mRNA molecules produced from a particular gene.

The mechanisms that are involved in gene expression are regarded to be so complex and intricate that there are still ongoing research focusing on this area. Until today, there have been ground breaking discoveries in decomposing the transcription machinery of a gene into its building blocks. Transcription factors (TFs), DNA methylation, histone modifications (HMs), small nuclear RNAs, etc., are examples of the components that are involved in transcriptional or post-transcriptional regulation.

TFs are essential components of the transcription machinery. They are proteins that bind to the DNA sequence either directly or indirectly, via forming complexes bound to the DNA, in order to recruit other elements to facilitate the transcriptional regulation. For the past decades, the problem of TF binding site detection has become popular and has attracted many researchers' attention. As a matter of fact, in 2016, a competition was held by the *DREAM challenge* (ENCODE-DREAM, 2017) organizers who provided the data for a competition seeking for an accurate computational model that is capable of predicting the binding sites of TFs. My colleague, Florian Schmidt, together with my supervisor, Marcel Schulz, and I formed a team to participate in this competition. We established a machine learning framework using random forest (RF) classifiers, extending the work of Liu, 2017. Since, Liu, 2017 and Waardenberg, 2016 have shown that tissue-specific cofactor interactions are appropriate for modeling TF binding, we designed our framework such that it was able to benefit from the tissue specificity inherent in the data to achieve a better generalization power. Using numerous Position Weight Matrices (PWMs) together with the DNase-hypersensitive sites (DHSs), which helped us locating the accessible regions of DNA, we were able to compute the TF binding affinity scores. It has been shown that even low binding affinities can deliver biologically relevant insights (Tanay, 2006; Crocker, 2015). Using this data together with the true binding classes provided by the competition, we were able to train the random forest classifiers predicting the binding status of a given TF. We published the results of this work in the *F1000Research* journal in 2019 (Behjati Ardakani, Schmidt, and Schulz, 2019).

The task described above was essential concerning the binding status of the TFs, however it did not explicitly involve gene expression. As a result, no clear association between TF binding and gene regulation was provided by the computational models. In order to investigate the gene regulation mechanism, we embarked on a new task aiming to predict gene expression using the histone modification (HM) data. The reason why we chose histone modifications over the TFs was the study conducted by Budden et al., 2014, which revealed that the TFs and HMs are statistically redundant for predicting gene expression. Given that we had a variety of novel HM data produced and made accessible by the German Epigenome Program consortium (DEEP), we decided to leverage such HM data for predicting gene expression. Karlić et al., 2010 developed a computational method for predicting gene expression from several histone modification data contrasting two distinct groups of promoters, low CpG content and high CpG content. They discovered that specific HMs are highlighted by their predictive models distinguishing the two promoter groups.

Inspired by Karlić et al., 2010, we were intrigued to establish an accurate and interpretable learning setup to predict gene expression for a different group of promoters from the HM data. In this work, we focused on pairs of promoters located on opposite strands of DNA that their transcription start sites are in proximity to each other. This particular set of promoters, which we refer to as bidirectional promoters, became the center of attention in several studies due to their peculiarity in their promoter architecture (Core, Waterfall, and Lis, 2008; Preker et al., 2008; Seila et al., 2008; Core et al., 2014b; Duttke et al., 2015a; Scruggs et al., 2015), raising the following question: Are the regulations of the genes at bidirectional promoters coupled?

To seek an answer for this question, we used the HM read abundance from ChIP-seq data measured across the promoter area of the bidirectional genes, while preserving their spatial distribution in this region. As we were interested in deciphering the spatial HM associations with expression of either of the genes at a bidirectional promoter, we exploited the fused LASSO optimization approach (Tibshirani et al., 2005) that provides interpretable models when features, the genomic locations in this problem, are correlated. The fused LASSO models we built, accompanied by the partial correlation analysis, pointed to a prominent trend of unidirectional association between HM localization at the promoter and gene expression.

Since we used the conventional bulk sequencing data for studying the gene regulation in bidirectional promoters, we could not rule out the possibility of our findings being restricted by the average signal obtained over a population of cells. This lack of satisfaction nudged us towards investigating the bidirectional gene regulation at the single cell resolution. We exploited two of the available single cell RNA-seq data, one produced by DEEP and the other obtained from Pollen et al., 2014, for measuring gene expression at bidirectional promoters. Using single cell RNA-seq data, we were able to derive novel transcriptional states specific to bidirectional promoters. This allowed us to further investigate other characteristics attributed to the identified states, such as DNA methylation and DNase-seq signatures as well as several genomic-related features. We published our findings in the journal of *Epigenetics & Chromatin* in 2018 (Behjati Ardakani et al., 2018).

Given that the single cell gene expression data helped us to delineate distinct promoter architectures in bidirectional promoters, we considered furthering our studies in a slightly different direction by asking a different question. We wondered if we could develop a statistical model that is able to inform us about the regulatory elements that are essential in deriving gene expression levels in a cell specific manner. In other terms, a cell-specific association between gene expression and a certain regulatory feature. There has been already similar work that aimed at inferring such associations for identified cell types (Mohammadi et al., 2018; Aibar et al., 2017; Suo et al., 2018). Even though interesting discoveries have been made, no integrative model was used to incorporate the gene expression data from single cells simultaneously. To reach this goal, we designed a multi-task learning (Caruana, 1997; Kim and Xing, 2010) framework, where each cell corresponds to an individual task, as defined in the optimization formula, and used various feature types to establish the desired associations. Our results led to identifying distinct regulatory elements, from each feature type, that were specific to individual cells. We were able to show that multi-tasking was indeed advantageous over using many single-task models. This indicates that the information sharing attained through multi-tasking could most likely handle the issue of missing values, so called dropouts, which is a common error of the current single cell sequencing protocols.

The outline of this dissertation is as follows. Chapter 2 embraces the definition and description of basic biological and statistical concepts that are helpful in grasping the approaches explained in the subsequent chapters. Chapter 3 provides details related to the *DREAM challenge* competition. Chapters 4 and 5 are devoted to unraveling the dilemma of bidirectional gene regulation. Chapter 6 describes our multi-task learning approach used in discovering cell-specific associations based on multiple regulatory elements. Finally, Chapter 7, concludes the projects presented and described in this dissertation.

Chapter 2

Background

2.1 Biological Basics

2.1.1 DNA structure

DNA (deoxyribonucleic acid) is a double-helical molecule consisting of four nucleotides, adenine (A), cytosine (C), guanine (G), and thymine (T) (Dahm, 2008; Watson and Crick, 1953). The DNA helix is also referred to as DNA strand, to which various names are assigned in order to distinguish them from each other. For instance, some may refer to them as *Watson* and *Crick* strands in the honor of DNA pioneers. Researchers also use *forward* and *reverse*, as well as *plus* and *minus* strands. The length of DNA strand, which is a sequence of nucleotides, can go up to several billion base pairs. For instance, the human DNA contains approximately 3 billion base pairs that stretches to roughly 2 meters long. Therefore, an efficient packing mechanism is essential to arrange the DNA molecule inside the nucleus of the cell. These 3 billion base pairs form 23 pairs of chromosomes that are embedded inside the nucleus of the cell. Each chromosome itself consists of DNA tightly wrapped around proteins called histones that maintain its structure. Complexes of histone proteins around which the DNA is wrapped are called nucleosomes. The nucleosome is often considered as the basic packaging unit in DNA (Figure 2.1). To be precise, there are four types of core histone proteins, H2A, H2B, H3, and H4, with two copies each in a nucleosome. Approximately 147 base pairs of DNA wind around a nucleosome (Zhou, Goren, and Bernstein, 2010). These histone proteins have an N-terminal tail that can be chemically modified via processes such as acetylation, methylation, phosphorylation, etc. Through these modifications the contact between the DNA molecule and the histone cores alters, which facilitates changes in DNA conformation.

2.1.2 Definition of a gene

In 1909, the term *gene* was, initially, coined by Wilhelm Johannsen, a Danish botanist, referring to the essential units of heredity. At the time, *gene* had no physical specifications and was mainly considered in an abstract form to determine heredity. Later, the geneticists were able to identify the location of several genes on the DNA and attribute certain traits to them. Currently, there have been over 65,000 genes identified and annotated in the human genome, of which more than 20,000 are protein-coding. There has been a number of databases that provide the genome annotation, such as Ensembl (Kersey et al., 2018), GENCODE (Harrow et al., 2012), RefSeq (Pruitt, Tatusova, and Maglott, 2005), Gene Ontology (GO) (Ashburner et al., 2000; Consortium, 2018), etc.

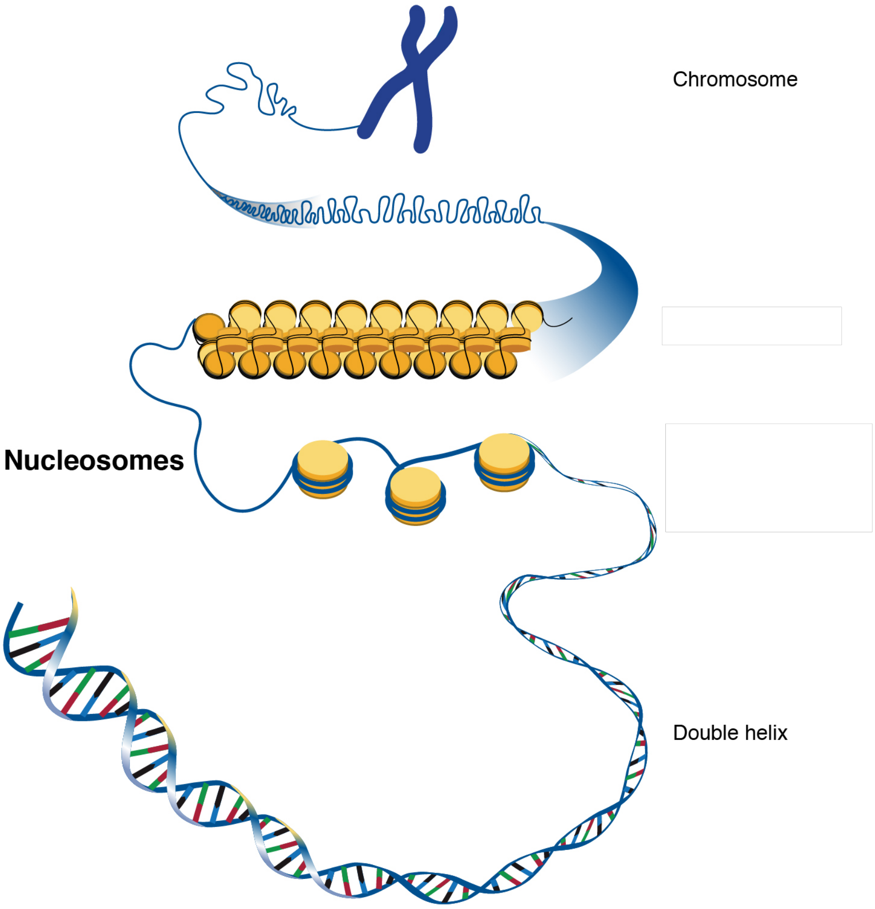


FIGURE 2.1: Schematic illustration of DNA packing achieved through repeating units called nucleosome. Each nucleosome refers to approximately 147 base pairs of DNA that are wrapped around histone proteins shown by yellow cylinders. Courtesy: *National Human Genome Research Institute*

To access the information embedded in a gene, the double helix structure of the DNA needs to be unwinded, such that the gene regulatory elements that are required to read the information off the gene be able to bind to that specific location on the DNA and transcribe the gene. The next section briefly describes the basic steps of transcription and gene regulation.

2.1.3 Gene expression and its regulation

Studying genes is usually coupled with evaluating their activity. But, what does a gene activity mean or how a gene can become active? This is achieved through a mechanism called gene regulation. The first step in gene regulation is transcription. There are three main stages of transcription: initiation, elongation, and termination. During initiation, transcriptional activators, which are gene regulatory proteins, bind to specific sequences in DNA (called enhancers) to facilitate attracting an essential enzyme called RNA polymerase. RNA polymerase binds to the promoter of a gene, which is a sequence of DNA near the gene's start site. Then, RNA polymerase unwinds the double strands of DNA, through which the single-stranded DNA becomes accessible to the transcription complexes to bind. Through the elongation stage, one of the DNA strands acts as a template for RNA polymerase. This strand is, conventionally, referred to as the "template strand". As the RNA polymerase "reads" the bases off this template one at a time, a single stranded molecule, called Ribonucleic acid (RNA), is created from the DNA. The RNA molecule is also composed of four nucleotides, adenine (A), cytosine (C), guanine (G), and uracil (U). There exist sequences on DNA that are called terminators and their role is to signal that the RNA transcript is complete. Once these sequences are transcribed, they result in releasing the RNA molecule from the RNA polymerase. This RNA molecule is considered as pre-mRNA and needs to be processed into a messenger RNA (mRNA). Several mechanisms need to take place to determine the stability and distribution of the generated transcript. One of these mechanisms is capping. Through capping the $5'$ end of the messenger RNA (mRNA) changes to a $3'$ end via a $5' - 5'$ linkage. This protects the RNA molecule from degradation. In addition, a $3'$ poly-A tail is added to the end of the RNA molecule. This tail makes the RNA molecule more stable and prevents its degradation. Eventually, through the translation process, the protein molecule is synthesized from mRNA. Figure 2.2 illustrates a schematic view of the transcription and translation steps pertaining to gene regulation.

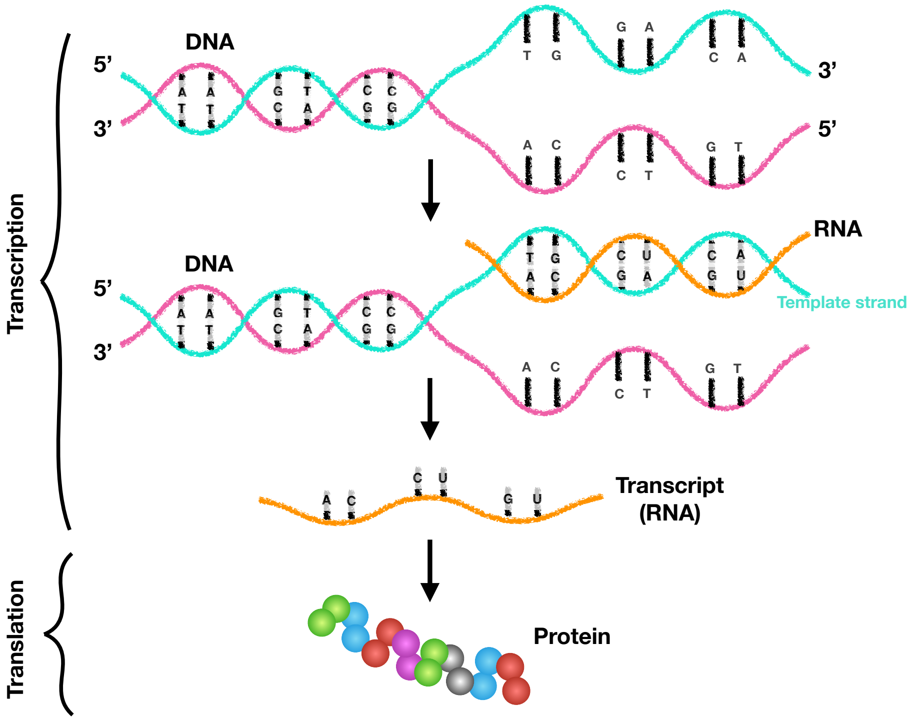


FIGURE 2.2: From DNA to protein. Gene regulation involves transcription and translation as the first two steps. During the transcription, the DNA molecule unwinds and through the transcription mechanism, an RNA molecule is generated. This RNA molecule is essential to proceed with the translation step, which results in production of polypeptide (protein) molecules.

The gene activity is often determined from the mRNA level that is produced through the transcription process. Whether or not it is the right time for a gene to become active, and if so, how much mRNA is needed to be produced from it, is regarded as one of the biggest conundrums in the fields of genetics and molecular biology.

To a great extent, understanding the origin of diseases, in particular cancer that is mostly recognized as aberrant expression of certain genes, is the reason why researchers in the field are curious to solve the mystery of gene regulation. The expression of a gene is a function of several genetic and epigenetic factors. In a very high level epigenetic point of view, nutrients, physical exercises, fraternizing, stress, and plenty other examples, act as external stimuli that can affect the signaling pathways and, as a result, a change in the gene expression pattern. However, in a more low level view, the factors that are involved and conduct the gene regulation are several types of proteins, referred to as transcription factors, that come together and form transcription complexes. More precisely, RNA polymerase II (pol II) is initially required to encode the mRNA through the transcription process starting from a gene's promoter. Before transcription starts, pol II together with several transcription factors such as TFIIB, TFIID, TFIIE, TFIIF, and TFIIH must be assembled on the core promoter (Zawel and Reinberg, 1995; Patikoglou and Burley, 1997). Among which, the only transcription factor that is able to form a direct contact to the DNA sequence (sequence-specific DNA binding) is TFIID (Burley and Roeder, 1996; Patikoglou and Burley, 1997). TFIID identifies and binds to an element called TATA box at the core promoter. The TFIID-TATA complex recruits the rest of the general TFs together with pol II to form a pre-initiation complex (PIC), a large multi-protein assembly that facilitates precise initiation of transcription at the TSSs (Patikoglou and Burley, 1997). The formation of PICs as well as the histone modifications facilitate the chromatin accessibility and pave the way for the transcription mechanism.

There are many other proteins and RNAs that are involved in regulation of mRNA processing. For instance, newly discovered regulators such as miRNAs (Bartel, 2004), generally bind to their target mRNA and, by destabilizing it, repress protein production (Cannell, Kong, and Bushell, 2008). In addition, the DNA methylation can be involved in gene regulation by repressing the transcription, typically, when methyl groups are added to the gene's promoter.

In order to understand the gene expression, it is important to be aware of all these influential factors, not only as individuals, but also the elaborate interplay between them. There are studies that aim to achieve a comprehensive view of gene expression by integrating the aforementioned factors, but the end has not reached yet, as there are still plenty to be investigated and addressed.

2.1.4 Bidirectional genes

Returning back to the topic of genes and their definition, it is worth adding that there are, in fact, various subsets or classes of genes that come to researchers' interest. One of these classes, which is relevant to this dissertation, is bidirectional genes. If the transcription start site of two distinct genes, located on opposite strands of DNA (plus and minus strands), happen to be in a close proximity of each other, then these genes are called *bidirectional genes*. The promoter area shared by these genes is referred to as *bidirectional promoter*. How close the TSSs of the genes should be to each other can be a matter of debate, but this distance normally does not go beyond 1000 base-pair (bp) according to the literature (Adachi and Lieber, 2002; Trinklein et al., 2004). Microarray experiments revealed that more than 10% of human the genome consists of bidirectional promoters. As the sequencing technologies advanced and became more sensitive in detecting nascent transcribed RNA molecules, this number has risen up to $\sim$ 80%. Therefore, exact classification of bidirectional or unidirectional promoters is not straightforward, as the sensitivity of the sequencing assay to recognize unstable, nascent RNAs plays a critical role (Andersson et al., 2015; Duttke et al., 2015b).

2.1.5 Histone modifications

As described in 2.1.1 the nucleosome, which is the basic unit of chromosome packing, is made of eight core histone proteins whose N-terminal amino acids can be modified to enable access to DNA as and when needed. If through these modifications the DNA molecule loses its contact to the nucleosome and becomes more accessible to regulatory elements, the foundation is laid for the gene to be transcribed. There are over 200 different histone modifications identified and the list continues to grow as the number of discovering antibodies used for the chromatin immunoprecipitation experiments (Collas, 2010) keeps on increasing. Zhao and Garcia, 2015 provide a comprehensive list of cataloged histone modifications (also referred to as histone marks, or marks alone) accompanied by their associated function. Not all histone modifications are associated with the gene expression the same way, as there are certain modifications that are associated with other cellular processes such as DNA replication. Histone marks, such as H3K4me3, are associated with initiation of transcription, whereas some others are associated with gene repression by tightening the contact of the DNA's backbone to the nucleosome, e.g. H3K27me3. As I focused working on six different histone marks throughout the course of this dissertation, I am going to briefly refer to the known functions associated to them in the following section.

H3K4me1

Heintzman et al., 2007 showed that the presence of modification caused by addition of one methylation group to the lysine at position four of the amino acid chain of histone H3 is linked to enhancers, the distal regulatory elements that affect gene expression. Moreover, it has been hypothesized that the H3K4me1 modification is in fact involved in a poised enhancer state, where it shares significant similarities with active enhancers. However, it is incapable of driving gene expression in cells that are in the differentiation state (Calo and Wysocka, 2013).

H3K4me3

In contrast to H3K4me1, this mark has three methylation groups bound to the lysine at position four of H3. H3K4me3 is generally associated to transcription activation at the gene's promoter (Lee and Skalnik, 2005; Lee et al., 2007; Xiao et al., 2011) as the ChIP-seq experiments strongly indicate presence of this mark around the transcription start site of the genes (Zhao and Garcia, 2015).

H3K36me3

Methylation of the lysine at position 36 of histone H3 is also associated with transcriptional activity as well as the two aforementioned marks. However, the main difference that sets H3K36me3 apart from the other two marks is that it spreads along the transcribed regions of the genes and often peaks toward the transcription termination site (3′ end of the gene) (Bannister et al., 2005).

H3K27me3

Genome-wide ChIP-seq experiments performed in human and mouse suggest that the modification induced by binding of three methylation groups to the lysine at position 27 of histone H3 is often associated with gene inactivation (Kim and Kim, 2012). In contrast to the H3K4me3, this specific mark shows, generally, a broad accumulation of ChIP-seq tags across genic and nongenic regions.

H3K9me3

Similar to H3K27me3, the three methylation groups added to the lysine at position nine of histone H3 is mainly associated to gene repression (Kim and Kim, 2012). The average signal obtained from the ChIP-seq experiments often shows a depletion at the transcription start site (Pjanic et al., 2011).

H3K27ac

When the acelytated lysine at position 27 of histone H3 is accompanied by the H3K4me1 histone mark, there is an indication that the corresponding genomic region is an active enhancer (Pradeepa, 2017). According to the study done by Creyghton et al., 2010 the regions that have only H3K27ac without any significant enrichment of the H3K4me1 modification should still be regarded as potential enhancer elements.

A comprehensive summary of the colocalization of the histone marks at the genome is depicted in Figure 2.3.

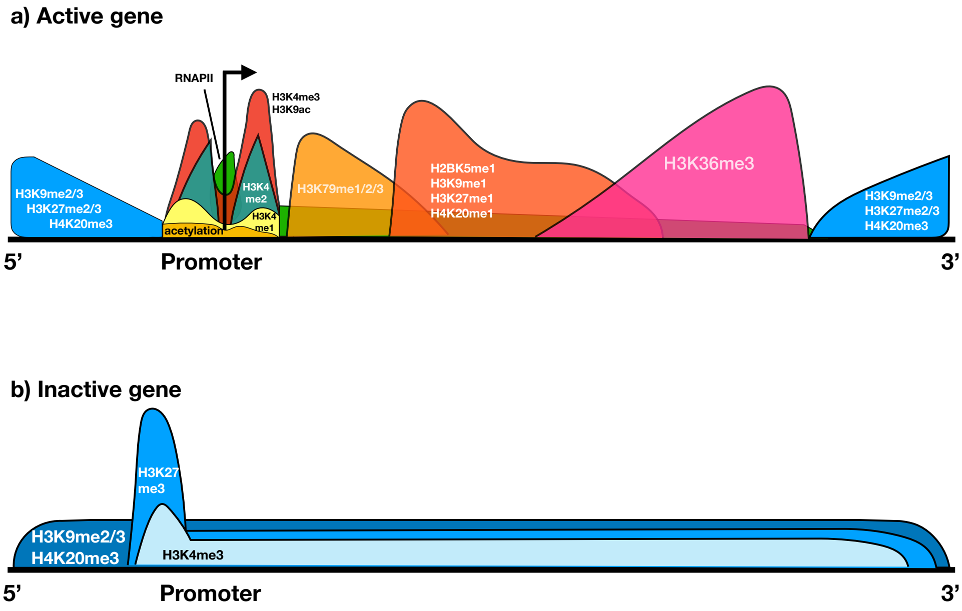


FIGURE 2.3: Distribution of histone modification signals on active (a) and inactive (b) genes. Image remade from Barth and Imhof, 2010.

2.1.6 Sequencing technologies

As described in 2.1.3 the gene activity is often interpreted as mRNA expression level. For quantifying the levels of produced mRNA molecules, several techniques have been developed, which the three most relevant ones are briefly described as following.

RNA-seq

In order to read the content of a mature RNA molecule, i.e., the sequence of A, C, G, and U, which contains the poly(A) tail, it needs to be shredded into smaller fragments such that the sequencing machines are able to process them. This population of RNA fragments, where some contain the poly(A) tail, is converted to a library of cDNA (complementary DNA synthesized from a single stranded RNA) fragments. Next, short, chemically synthesized, single-stranded or double-stranded oligonucleotide that can be ligated to the ends of DNA or RNA molecules, referred to as adaptors, are attached to either one or both ends of each fragment (depending on whether paired- or single-sequencing was performed). Eventually, sequencing reads are obtained from amplifying these fragments that are sequenced in a high-throughput manner (Bainbridge et al., 2006; Mortazavi et al., 2008; Weber, 2015). The length of these reads are typically ranging from 30 bp up to 400 bp, depending on the sequencing technology used (Mortazavi et al., 2008). Figure 2.4 schematically depicts the RNA-seq protocol.

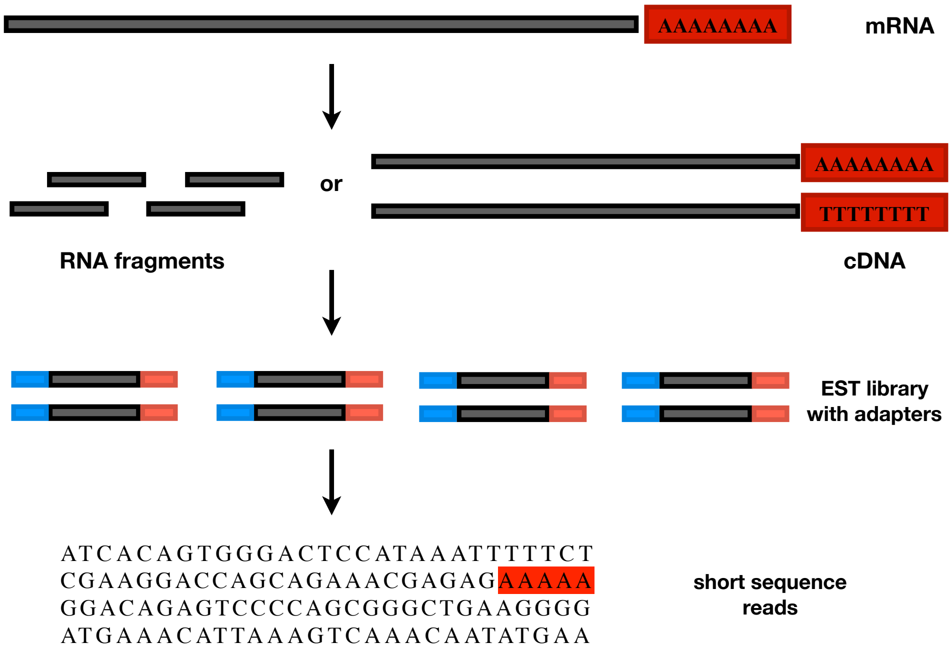

AAAAAAAA
mRNA
or
AAAAAAAA
TTTTTTTT
RNA fragments
cDNA
EST library
with adapters
ATCACAGTGGGACTCCATAAATTTTTCT
CGAAGGACCAGCAGAAACGAGAGAAAAA
GGACAGAGTCCCCAGCGGGCTGAAGGGG
ATGAAACATTAAAGTCAAACAATATGAA
short sequence
reads

FIGURE 2.4: RNA-seq workflow. RNA molecule is first converted into a library of cDNA fragments through either RNA fragmentation or or DNA fragmentation. Sequencing adaptors (blue) are thereafter added to each cDNA fragment and a short sequence is derived from each cDNA using high-throughput sequencing technology. Figure remade from Wang, Gerstein, and Snyder, 2009 and modified.

CAGE

The Cap Analysis of Gene Expression (CAGE) is another sequencing method developed by Shiraki et al., 2003a. Similar to RNA-seq, the cDNA is first synthesized, but then the full-length cDNAs are captured using the biotinylated cap-trapper. Next, a primeable sequence is added at the 5′ ends of the cDNAs. After performing a series of steps involving linker addition, cleavage, and PCR, the CAGE tags that represent the initial 20 nucleotides from the 5′ end of mRNAs are used to generate the CAGE libraries meant to be sequenced (Figure 2.5). What makes CAGE distinct from other sequencing protocols is that it is based on profiling the 5′ end of RNAs with a cap structure, which includes mRNAs and a large fraction of non-coding RNAs (Shiraki et al., 2003a). In addition, CAGE provides accurate estimates of the original mRNA concentration, since it is based on sequencing of concatamers (long continuous molecule containing multiple copies of the same sequence linked in series) of DNA reads deriving from the initial 20 nucleotides from 5′ end mRNAs. This means that not only the overall RNA expression level, but also the expression of each alternative promoter within a gene can be estimated.

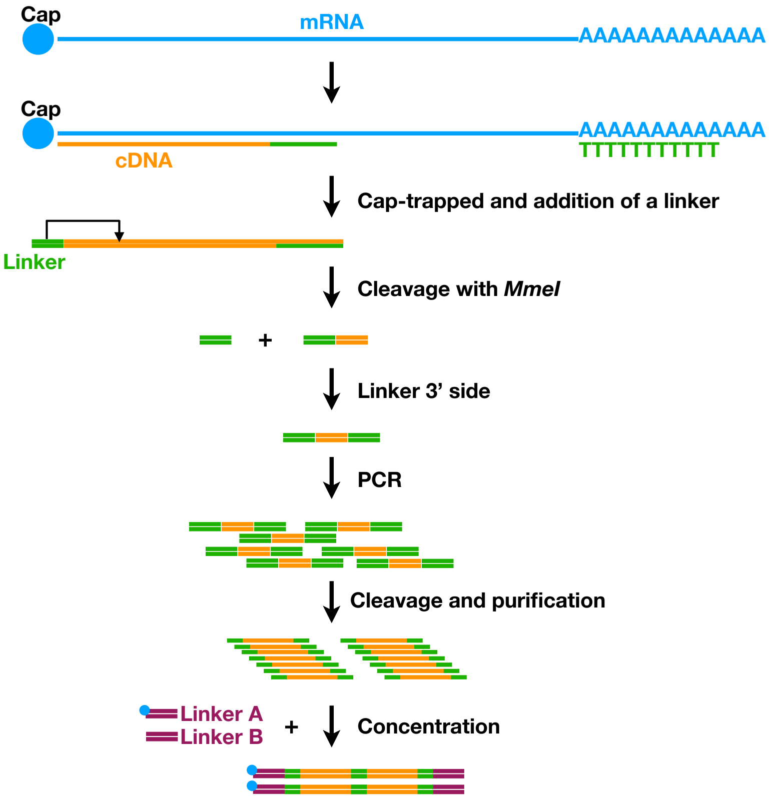


FIGURE 2.5: CAGE workflow. Image remade from the FANTOM website (`http://fantom.gsc.riken.jp/protocols/basic.html`).

GRO-cap

Global Run On (Gro)-cap protocol is mainly known for its ability to detect nascent molecules of RNA that are being generated by the RNA polymerase. In contrast to poly(A)-enriched mRNA-sequencing methods, GRO-cap also captures incomplete transcripts, which lack the poly(A) tail (unstable transcripts that are susceptible for degradation). The ability to detect transcripts from intronic regions makes GRO-cap an attractive sequencing candidate for detecting enhancer-related transcripts (Core et al., 2014a).

ATAC-seq

Assay for Transposase-Accessible Chromatin (ATAC)-seq is another sequencing protocol that is dedicated to assess and map the chromatin accessibility throughout the genome (Buenrostro et al., 2013). The key step is the *in vitro* insertion of Tn5 (Reznikoff, 1993) to a mixture consisting of DNA molecules and adapters. This enzyme can fragment the genome and, simultaneously, attach the adapters to these fragments. When the chromatin is accessible the chances that this transposition takes place is higher compared to when the chromatin is less accessible. This means that the amplified fragments obtained after the PCR step would more likely be enriched in the regions where the chromatin was accessible (Figure 2.6). This sequencing protocol together with other chromatin accessibility assays (Song and Crawford, 2010; Simon et al., 2012) allow a genome-wide profiling of the epigenetic landscape.

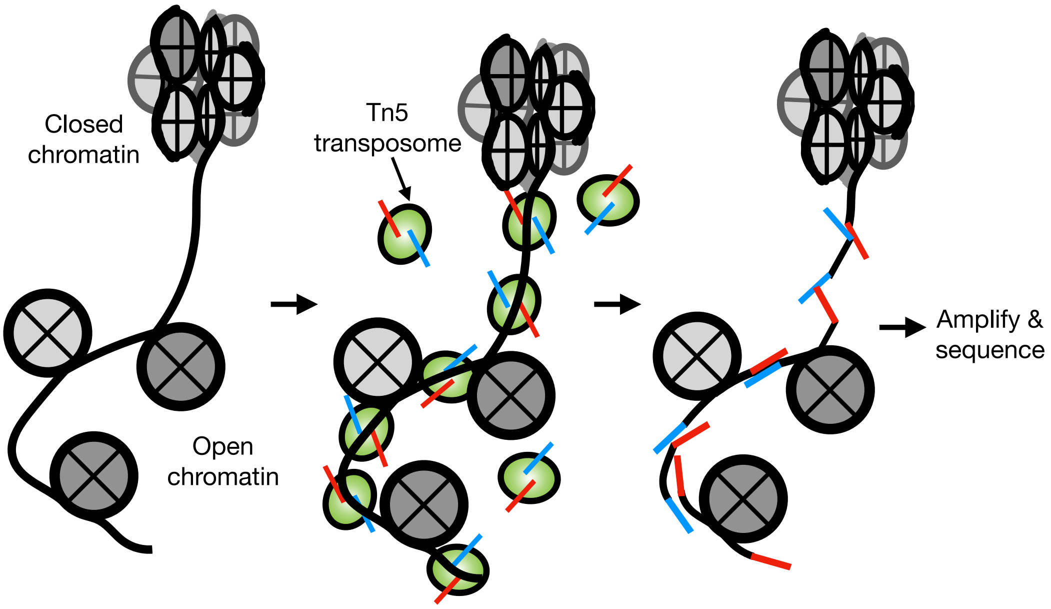


FIGURE 2.6: ATAC-seq workflow. The transposon Tn5 (green) loaded with sequencing adapters (blue and red) transposes into the open (accessible) chromatin regions. Image remade from Buenrostro et al., 2013

ChIP-seq

The Chromatin immunoprecipitation (ChIP)-seq protocol was developed by Barski et al., 2007 to determine the genome-wide binding sites of proteins of interest. This involves using antibodies to fish out DNA fragments that are bound by the protein that the antibody targets were used. Figure 2.7 illustrates the workflow for the ChIP-seq experiment.

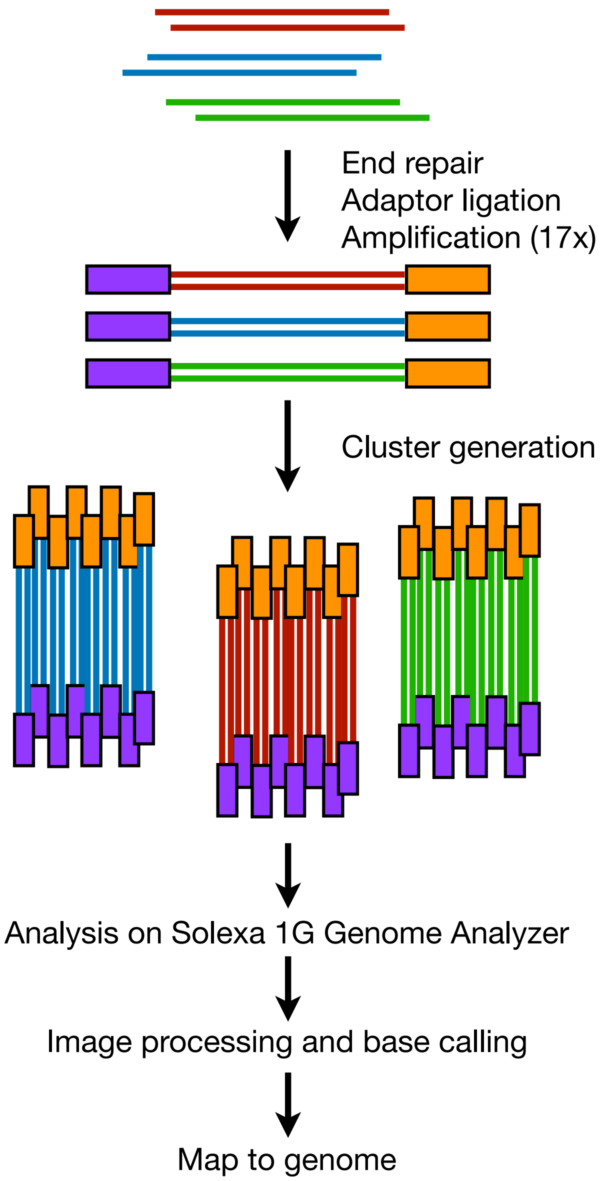


FIGURE 2.7: ChIP-seq workflow. The sequencing procedure requires an adaptor ligation and together with 17 cycles of PCR amplification of ChIP DNA molecules. Next, the ligated and amplified molecules are used to determine clusters to be given to the Solexa Genome Analyzer. Finally, after the image processing and base calling step, the reads are mapped to the genome. Image remade from Barski et al., 2007.

2.1.7 Bulk and single-cell sequencing

In all the sequencing protocols described in 2.1.6, a bulk of DNA or RNA materials were used to perform the experiment. This is due to providing sufficient DNA or RNA molecules for the sample to be sequenced such that the technical noise and artifacts would be averaged out, and therefore more accurate and reliable results be obtained as output. Although there is a good intention behind performing bulk sequencing, this results in a loss of details about different cell populations, which may be of interest in the study. And, sometimes, the original cell population is rare compared to other cell types in the mixture, for instance, the embryonic cells obtained from the very early hours after fertilization. In such scenarios, the sequencing protocols need to be adapted to address the small number of available cell.

Tang et al., 2009 performed a whole-transcriptome analysis of RNA-seq in a single mouse blastomere. Their procedure starts off by lysing the cell, where in a mixture containing primer sequences, the cDNA can be synthesized from the released mRNA molecules. Next is performing a reverse transcription step to obtain full-length cDNAs of the first strand tagged with the primers. After adding poly(A) tails to the tagged cDNAs, the cDNA from the second strand is synthesized are amplified through a PCR amplification step. Finally, the cDNA libraries are generated after cDNA shearing and adapter ligation steps (Figure 2.8).

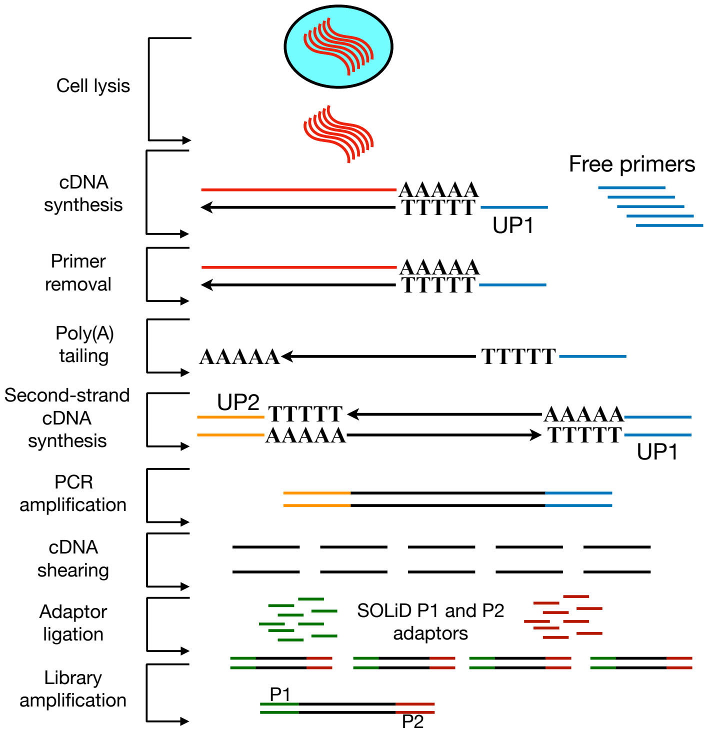


FIGURE 2.8: scRNA-seq workflow. Image remade from Tang et al., 2009.

Ever since Tang et al., 2009 was published, numerous studies have been carried out using single-cell sequencing technologies, in particular single-cell RNA-seq (scRNA-seq), investigating diverse aspects of gene regulation at the single cell resolution. Ramsköld et al., 2012 proposed an RNA-seq protocol called Smart-seq that can be applied at the single cell level. Using their method they could identify distinct gene expression patterns in tumor cells, which could then lead to discovering novel biomarkers for those tumor cells. Another sequencing method was designed by Hashimshony et al., 2012 to sequence a large number of cells in parallel. They applied their method of sequencing, CEL-seq, on mammalian cells and nematode embryonic blastomeres. This allowed them to analyze transcriptomics at single cell resolution and train a classifier to predict the identities of blastomeres sister pairs.
The challenges faced in single-cell sequencing do not limit only to designing new protocols or machines, but also the computational methods that can handle the noisy data as a result of the low-coverage reads. The problem arises from having insufficient materials as the first step. This, then, leads to getting low-coverage of the sequenced reads. In other words, when the estimated expression of a gene based

on the mapped reads to the corresponding genomic location is equal to zero, it can either mean that the gene was truly silent, or due to the technical noise the mRNA molecules could not be detected. The latter scenario is referred to as dropouts problem, a very common and intrinsic difficulty faced in single-cell sequencing. There have been various studies that focused on addressing the effects of dropouts on the gene expression profile. This has been most commonly achieved by modeling the dropout distribution into the density function reflecting the gene expression. So, these methods usually tackle a dual challenge, one is detecting the dropouts, and the other is imputing their values (Mongia, Sengupta, and Majumdar, 2019; Li and Li, 2018a; Gong et al., 2018; Tracy, Yuan, and Dries, 2019).

2.2 Statistical Background

This section primarily focuses on various basic and advanced statistical learning methods that come of relevance to this dissertation. There are two major types of algorithms in the field of machine learning, 1) unsupervised, and 2) supervised. The following will address these two types in more detail and extent.

2.2.1 Unsupervised learning

Imagine we have collected N observations describing weight (kilograms) and height (centimeters) measurements from a mixed population of wrestlers and basketball players. However, there is no data available for the type of sport they are engaged in. Figure 2.9 shows a distribution of these points in the space defined on the weight and height variables. By inspecting the distribution of the points in this plot, one can see that there are two clusters of points, one reflecting shorter and heavier subjects, and another representing taller and slightly lighter subjects, compared to each other. Since, the observations at hand are only describing the features (weight and height) and no data is available about the category the subjects are from, the algorithms that try to fit a model to cluster the data are considered as unsupervised learning algorithms. In more mathematical terms, let $X \in \mathbb{R}^{N \times p}$ represent a data matrix of N samples with p features, where $x_i \in \mathbb{R}^p$ for $i = \{1, \cdots, N\}$ denotes an individual observation corresponding to i^{th} row of X. Then the aim of any unsupervised learning algorithm is to divide these N observations into k clusters such that similar data points are assigned to the same cluster. The choice of k and definition of similarity metric are two interesting and yet challenging aspects of a clustering method. In the following section two of the most common and well-established clustering methods are described.

k-means clustering

For a given k and observation matrix X, k-means clustering tries to partition the data points into k distinct sets, using the Euclidean distance as the measure of similarity in an iterative manner. The algorithm starts by randomly selecting k data points as the cluster centers, or cluster representatives. Then it computes the distance of each point to these k centers. The assignment of each data point to a cluster is then based on finding the cluster in where it had the smallest distance to. For each cluster a new center is defined by taking the mean of the data points in that cluster. Through this iterative process, the cluster centers gradually move towards a point where they can best partition the points in the space, i.e., the assignments do not get changed anymore.

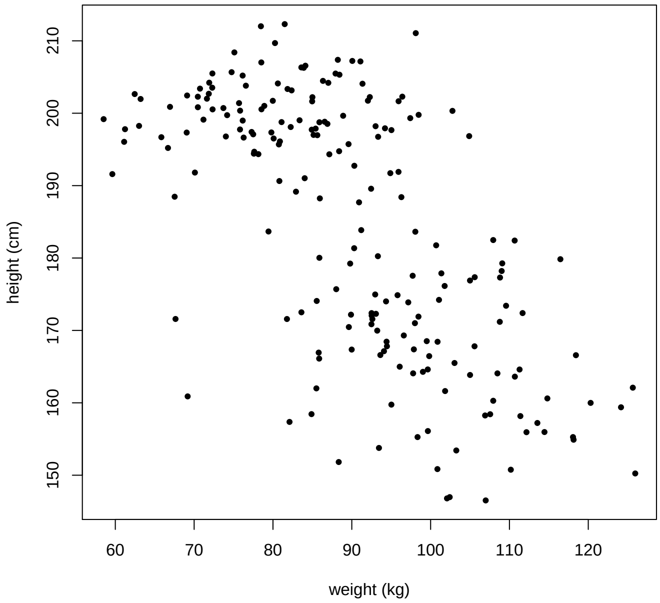


FIGURE 2.9: Example of unsupervised learning. Each point represents the weight and height of an individual sampled from wrestlers and basketball players. The distribution of the points in the space suggests two clusters of populations.

Algorithm 1 provides the pseudo-code for k-means clustering algorithm, where the Euclidean distance between two vectors of size p can be defined as $||x - y||^2 = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$. The objective function can be written as follows:

$$\min_C \sum_{j=1}^{k} N_j \sum_{C(i)=j} ||x_i - \bar{x}_j||^2, \tag{2.1}$$

where N_j denotes the number of data points assigned to cluster j, and $C(i) = j$ returns the data points, i, that are assigned to cluster j.

One of the drawbacks of k-means algorithm is its dependence on the selection of initial centers, which in turn may lead to different partitions. To tackle this problem, several starting points are chosen to run the algorithm with. The clustering that minimizes the value of the objective function shown in formula 2.1 is eventually used as the final clustering result.

Hierarchical clustering

In contrast to k-means clustering that required a predefined k to initiate the clustering procedure, hierarchical clustering methods are not restricted to such specifications. There are two strategies to conduct a hierarchical clustering, 1) bottom-up, or agglomerative, 2) top-down, or divisive. In the bottom-up approach, each data point is, first, considered as a singleton cluster. Pairs of clusters are then merged together to form larger clusters, according to the provided distance metric between disjoint groups of observations. The selected pair for merging would be the two clusters with the highest inter-group similarity. This procedure is continued until all data

```
Input: $X \in \mathbb{R}^{N \times p}$ denoting $N$ observations with $p$ features, and $k < N$, the
  desired number of clusters
Output: Assignment of observations into $k$ clusters
Procedure:
Initialization: Randomly choose $k$ observations from $X$ representing cluster
  centers;
while New assignment is possible do
    1. Assign each data point, $x_i \in \mathbb{R}^p$, to a cluster, $C(i)$, such that
       $C(i) = arg \min_{1 \leq j \leq k} \|x_i - \bar{x}_j\|^2$, where $\bar{x}_j \in \mathbb{R}^p$ denotes the mean of the $j^{th}$
       cluster;
    2. Recompute and update the cluster means according to the new
       assignment of points to each cluster;
end
```

Algorithm 1: *k*-means clustering pseudocode.

points are clustered into one. The top-down strategies, on the contrary, regard the entire observations as one whole cluster and recursively break it into smaller clusters down to singletons. One of the main reasons for the reputation of hierarchical clustering methods is its ability to provide a highly interpretable complete description of the hierarchical clustering in a graphical format. This graphical format is depicted as dendrogram, which is a tree structure describing the clustering steps, in either agglomorative or divisive way, with preserving the dissimilarity distance between the clusters reflected in the length of the edges of this tree. In order to obtain a certain number of clusters, the dendrogram must be cut at that certain level. For instance, if the clustering results of the *k*-means method are to be compared with a hierarchical clustering approach, the dendrogram obtained from the hierarchical clustering should be cut at level *k*. The partitioning of the leaf nodes derived from this cut, forms the desired clustering that will be compared with the *k*-means clustering.

2.2.2 Supervised learning

Unlike unsupervised learning, the observations are coupled with response values, in the supervised learning scenario. This response variable can either be continuous real valued measurements (regression task) or categorical values (classification task), such as the *wrestlers* or *basketball players* in the example above. In mathematical terms, the observations are tuples of the form (x_i, y_i), where $x_i \in \mathbb{R}^p$ and $y_i \in \{1, \cdots, k\}$ in a classification task with k classes, or $y_i \in \mathbb{R}$ in a regression task.

The classification algorithms take both feature and response values as their input in order to train a model that given the features can predict (explain) the response. The training procedure involves minimizing an appropriate loss function, which is tailored to either the classification or the regression task, through which certain parameters defining the model will be adjusted. The following sections describe some of the regression and classification methods in more depth.

Regression models

As mentioned above, a regression task in the supervised learning is referred to predicting the response variable that consists of continuous real values. One of the most elementary regression models is called Ordinary Least Squares (OLS) that seeks a linear association between the feature and response variables. This linear association is described by a hyperplane fitted to the data, such that the least squared errors between the fitted values and actual measurements of response are achieved. This hyperplane is defined by a slope vector, $\beta \in \mathbb{R}^p$, and an intercept, $\beta_0 \in \mathbb{R}$. The fitted (predicted) values of the response variable can then be obtained via the following:

$$\hat{y}_i = \beta_0 + < \beta, x_i >, \tag{2.2}$$

where $< .,. >$ denotes the inner product between two vectors. The OLS describes the association between x_i and y_i using the following equation:

$$y = \hat{y}_i + \epsilon_i, \tag{2.3}$$

where ϵ_i defines the prediction error corresponding to observation i, also known as residuals. Figure 2.10 shows the result of an OLS model applied on univariate observations of number of hours of study per week, as feature, and the grade obtained, as response. The deviation of the data point i from its projection on the fitted line is equal to ϵ_i in equation 2.3.

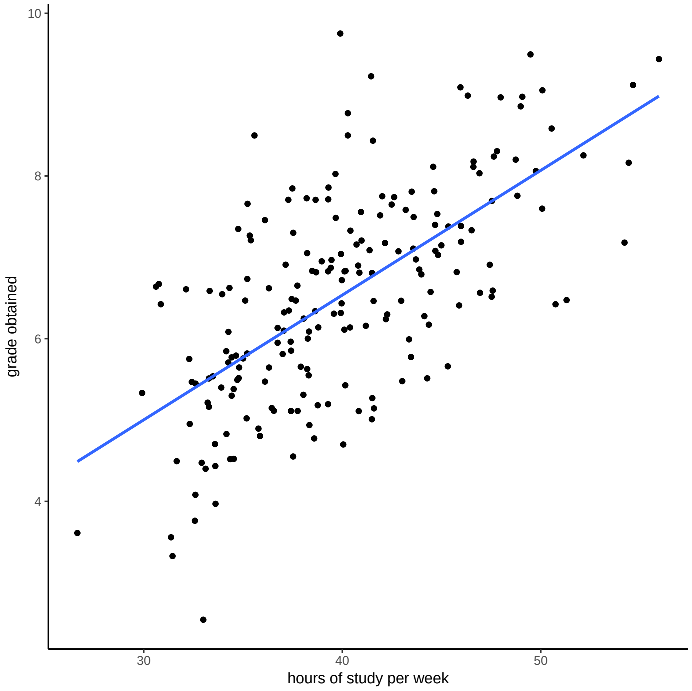


FIGURE 2.10: OLS applied on an example data.

To estimate the optimal values for the unknown parameters β_0 and β the following objective function needs to be optimized:

$$L(\beta_0, \beta) = \sum_{i=1}^{N} \epsilon_i^2 + \beta_0, \tag{2.4}$$

where the term ϵ_i^2 denotes the squared errors, $(y_i - \hat{y}_i)^2$. The formula 2.4 aims to minimize the sum of squared errors, a common and basic penalty term in regression loss functions. The optimal values of the regression coefficients, β_0^* and β^*, can be computed analytically by setting the first order derivatives of the objective function with respect to β_0 and β to zero. In other words:

$$\frac{\partial L(\beta_0, \beta)}{\partial \beta_0} = 0, \tag{2.5}$$

and

$$\frac{\partial L(\beta_0, \beta)}{\partial \beta} = 0. \tag{2.6}$$

This is a linear system of $p + 1$ equations with $p + 1$ unknown variables and can be solved easily for β_0:

$$\frac{\partial L(\beta_0, \beta)}{\partial \beta_0} = 0 \Rightarrow \sum_{i=1}^{N} (y_i - \beta_0 - \beta \times x_i) = 0 \Rightarrow \beta_0^* = \sum_{i=1}^{N} (y_i - \beta \times x_i) \tag{2.7}$$

and for β:

$$\begin{gathered} \frac{\partial L(\beta_0, \beta)}{\partial \beta} = 0 \Rightarrow \sum_{i=1}^{N} (y_i - \beta_0 - \beta \times x_i) x_i = 0 \Rightarrow \\ \beta^* = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} = \frac{Cov(x, y)}{Var(x)} = X^T X^{-1} X^T y, \end{gathered} \tag{2.8}$$

where X is the matrix containing x_i's ($X \in \mathbb{R}^{N \times p}$). One of the major problems with OLS models is the lack of regularization in its objective function. Regularization helps building a statistical model with better generalizability power.

Next, another linear regression model called Least Absolute Shrinkage and Selection Operator (LASSO) will be described. This method is built upon the OLS with an additional penalty term, L_1 norm of the coefficients, that induces sparsity in the β coefficients of the model. The objective function for LASSO is as follows:

$$L(\beta_0, \beta) = \sum_{i=1}^{N} \epsilon_i^2 + \lambda \sum_{i=1}^{N} |\beta_i| + \beta_0. \tag{2.9}$$

Penalizing the sum of absolute values of the coefficients encourages the model to set the coefficients corresponding to irrelevant features to zero, thus sparser models. This phenomenon is of course a function of the regularizer's parameter, λ as well. The larger the λ, the sparser the models. Setting the right value for λ can be achieved via a technique called cross validation, a, concept that will be discussed later in this chapter. Unlike OLS, LASSO doesn't have an explicit closed form for estimating the optimal coefficients for a given λ, even though the objective function is convex. But, fortunately, the gradient descent algorithms come to rescue, and can be used to approximate the optimal β coefficients.

LASSO can be considered as a feature selector, due to the fact that the L_1 norm penalty in its objective function sets the coefficients of irrelevant features to zero. This L_1 norm regularization becomes effective in cases where correlated features exist in the data. For instance, a data consisting of feature variables such as the number of hours of study, h, and the grade a student obtained in the exam, g. Assume that in the data collected from several students the h and g variables are correlated. Therefore, in order to predict whether the student failed or passed in the exam, one of the variables, either g or h, would be enough to get the job done. LASSO seems to be the right tool to achieve this, but the question is which variable it will choose? The answer is that it heavily depends on the data. Imagine, in our data set explained above, g slightly better associates with the response which is pass or fail. But, if we perturb the data, for instance collect the data from another school, then h might be better explaining the response. These perturbations in the data hurt the interpretability one expects from LASSO. Meaning that the features it selects aren't consistent across perturbed data sets with correlated features. Therefore, another variation of LASSO approach, called group LASSO, was proposed by Yuan and Lin, 2006 that takes predefined groups of variables, most likely to be correlated with each other, and assigns the coefficients such that the members of each group are either all included or all excluded. Assuming that there are G predefined groups, and $X_i \in \mathbb{R}^{N \times |G_i|}$ denotes the subset of $X \in \mathbb{R}^{N \times p}$ corresponding to the i^{th} group, the objective function for group LASSO is as follows:

$$L(B) = \sum_{i=1}^{N} \epsilon_i^2 + \lambda \sum_{i=1}^{G} \|B_i\|, \tag{2.10}$$

where $B \in \mathbb{R}^p$ is the coefficient vector and $B_i \in \mathbb{R}^{|\mathbb{G}|}$ denotes the vector of coefficients corresponding to the features of i^{th} group. The $\|.\|$ denotes the L_2 norm of a vector that is equal to the square root of sum of squared elements of the vector.

Although group LASSO is able to incorporate the information about the groups of variables that are related to each other into its optimization criteria, it does not capture the existing structure in the data. In other words, if the data at hand consists of temporal or spatial features, such that the order of variables in the feature space matters, group LASSO will not be able to reflect those properties onto its model coefficient assignments. However, there is another extension to standard LASSO, called fused LASSO that addresses this issue. The fused LASSO's objective function is as follows:

$$L(\beta_0, \beta) = \sum_{i=1}^{N} \epsilon_i^2 + \lambda \sum_{(i,j) \in E} |\beta_i - \beta_j| + \gamma \dot{\lambda} \sum_{i=1}^{p} |\beta_i| + \beta_0, \tag{2.11}$$

where E denotes the edges of a given graph $G = (V, E)$ defined over the feature variables as its node set (V). The features connected via an edge between them are subject to the fusion penalty $(\lambda \sum_{(i,j) \in E} |\beta_i - \beta_j|)$. The regularizer for fusion is denoted by λ, whereas γ controls the amount of sparsity and fusion and sparsity in the solution space. In other words, $\gamma \dot{\lambda}$ dictates the amount of sparsity induced in the model. Pure fusion (no sparsity) can be achieved by setting γ to zero, while $\gamma = 1$ results an equal level of sparsity and fusion.

Elastic net (Zou and Hastie, 2005) is another extension to the OLS models that imposes both L_1 and L_2 norm regularizations on the coefficients. Its objective function is as follows:

$$L(\beta_0, \beta) = \sum_{i=1}^{N} \epsilon_i^2 + \lambda_1 \|\beta\|^2 + \lambda_1 |\beta|, \tag{2.12}$$

where λ_1 and λ_2 are the regularization parameters for the L_1 and L_2 norms, respectively.

So far, the assumption was that the response variable is a vector comprising of the target values for each sample in the data. However, this does not always have to hold. There can be cases where multiple measurements of the response are available for a data point. For instance, in the wrestler/basketball player scenario, imagine that we are also interested in predicting the gender (female or male) of each individual as well as their level (under-19 or professional). In other words, the response variable corresponding to each data point holds three values, e.g., *basketball player*, *female*, and *professional*.

The statistical models that are learnt based on multivariate response variables are called multi-task learning (MTL) models. The obvious difference of MTL objective function compared to the single-task (univariate response) models is the nature of the β coefficients, which is going to be a matrix of dimensions $p \times k$, where p and k are the number of features and number of tasks, respectively. An interesting case of MTL was proposed by Kim and Xing, 2010 with the assumption that the response variables could be grouped hierarchically defined by a tree structure. They showed that this method is capable of finding a sparse estimate of β coefficients while maintaining the relation between the response variables according to the tree structure.

All the models explained above are unanimous on one assumption, which is the association between the feature and the response variables are linear ($y = \beta^T X + \epsilon$). But, this linearity might not always hold. To address this, there has been several other methods that do not require such assumption (non-linear models). Artificial neural networks (ANNs) are an example of such non-linear models. The idea was initially inspired from the neural network of a brain and how the neurons communicated with each other via synapses. Within each ANN, each neuron is considered as a computational unit, where it receives an input and after applying an activation function it produces the output. The architecture of ANNs consists of three types of layers: input layer, hidden layer, and output layer. The input layer forms the first layer of a feed forward network and it has as many neurons (nodes) as the number of features. The output layer comes as the last layer in network's architecture and contains only one node for the single-task regression scenario or k nodes for a multi-task learning setup with k tasks. The hidden layers are those appear between the input and output layer and can vary in terms of the number of layers and the number of node in each layer. The more the number of hidden layers, the deeper the network gets (not to be confused with deep neural networks that often have convolutional, pooling, etc. layers), which leads to more complex (non-linear) models. The activation function used in each layer (except the input layer) mimics the biological firing that takes place in biological neurons. When the chemical level that a neuron receives reaches a certain threshold, it causes the neuron to fire and transmit a signal to the neighboring neurons. This phenomenon motivated the mathematicians to define a non-linear function, which resembles the neuron's firing event, called sigmoid function as following:

$$f(x) = \frac{1}{1 + e^x}. \tag{2.13}$$

The non-linearity that was promised for ANNs is delivered through this non-linear activation function. Each layer is connected to the next layer through a weight matrix that defines the parameters of the model. These weights are typically assigned randomly and through an iterative algorithm, called back propagation, get updated until they converge to a local optimal. The loss function that is being optimized is the squared error similar to the aforementioned models.

Classification models

Up until now, the models described above were trying to solve an optimization problem where the response were continuous real values. But, as already mentioned, there is another type of supervised learning, called classification, where the response is in fact (finite) discrete values referred to as class labels. Here, an interpretable and promising classification method, called random forests, will be described. But before that, it is worth briefly mentioning about the decision trees that the random forests are built up on.

Decision trees are amongst the most favorable classifiers, mainly due to their appealing visualization and interpretation power. They try to solve the problem by building a tree over the feature variables until it reaches the leaf node where the class label is assigned. For instance, in the wrestler and basketball player classification problem, where the height and weight are the features and wrestler and basketball player the labels, a decision tree would pose the questions on the variables as shown in 2.11.

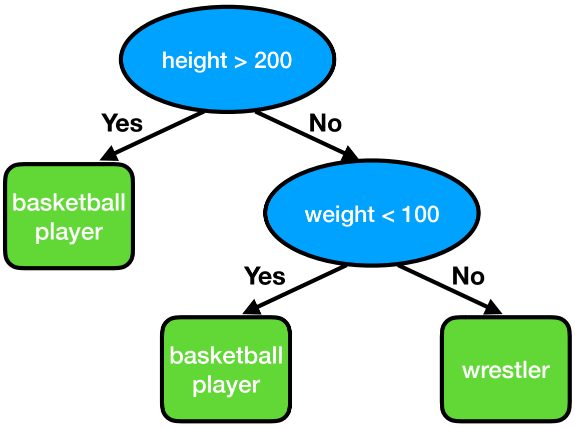


FIGURE 2.11: Example of a decision tree performed on the basketball player and wrestler classification problem.

Obviously, in more complex problems where the classes are not perfectly separable, the decision made at the leaf nodes would not be pure depending on how heterogeneous the samples satisfying the implied rule are (in a certain path from root to leaf). In order to construct a tree the entropy and information gain measurement are used to find the best variables for splitting. The entropy is used to measure the homogeneity of a sample. If a sample is entirely homogeneous, then the entropy is 0, otherwise it takes a value larger yet bounded by 1. The information gain is used to determine what variable to choose for splitting such that the highest information is gained eventually.

Ensemble learning

In 1875, Marquis de Condorcet, a French mathematician and philosopher, proposed a theorem called *Condorcet's jury theorem*. The theorem refers to a group of independent voters who need to reach a decision by majority voting, for instance a jury deciding whether a defendant should be found guilty or not. Assuming that p is the probability of each voter being correct and L is the probability of the majority vote obtained from the voters being correct, the theorem states the following:

- having each voter deciding better than random ($p > 0.5$) implies that the final vote would be better than each individual vote ($L > p$), and
- L approaches 1, for all $p > 0.5$ as the number of voters increases (i.e., this number approaches infinity).

Even though the *Condorcet's jury theorem* was conceived, originally, to provide a theoretical basis for democracy, supervised learning methods adopted the same

classifier	$class_1$	$class_2$	$class_3$
$classifier_1$	0	0	1
$classifier_2$	0.2	0.2	0.6
$classifier_3$	0.3	0.4	0.3
$classifier_4$	0.4	0.3	0.3
$classifier_5$	0.5	0.1	0.4

TABLE 2.1: Decision profile of five classifiers deciding on a 3-class problem for a given sample x.

Operator	$class_1$	$class_2$	$class_3$
Min	0	0	0.3
Max	0.5	0.4	1
$Average$	0.28	0.2	0.52

TABLE 2.2: Final votes obtained from applying the Min, Max, and $Average$ operators on the decision profile shown in Table 2.1.

principle to improve their prediction power and robustness (Rokach, 2010). The methods that employ and combine multiple learners in order to derive more accurate and robust prediction results are referred to as ensemble learning methods.

There are various operators for combining the decision of classifiers (voters), such as Min, Max, and $Average$. As their names suggest, these operators, take the min, max, and average of the votes across the classifiers for each class label, respectively. For instance, with $c = 3$ being the number of classes and $v = 5$ the number of classifiers, Table 2.1 provides an exemplary profile of the classifiers' decision for each of the three classes on a given sample x.

Applying the Min, Max, and $Average$ operators on the decision profile presented in Table 2.1 yields the final combined results shown in Table 2.2.

Another interesting method of ensemble learning was introduced by Wolpert, 1992, called *stacked generalization*. In stacked generalization, the output pattern of an ensemble of trained classifiers serves as an input to the second level classifier. Having seen the mistakes the individual classifiers in the first level have made, the ensemble classifier is able to adjust for those mistakes during the training phase.

In general, the more diverse the classifiers are, the more effective ensemble approaches get. This diversity can be obtained either by choosing distinct classifiers that are diverse in terms of their classification algorithms or through the training phase by letting them become expert on subsamples of the data. Random forest (RF) classifiers (Breiman, 2001) are examples of the latter case. Essentially, RFs are ensembles of many decision trees that each are trained on subsets of data obtained from bootstrapping (Efron, 1979). Bootstrapping, generally, is referred to random sampling of the data with replacement. After the bootstraps of the data are created, in order to increase diverse (less correlated) trees, a subset of features are then chosen to perform the node splitting upon. Algorithm 2 provides the pseudocode for constructing an RF.

One of the advantages of using random forests is that it can provide the out-of-bag (OOB) samples as well as out-of-bag error. An OOB sample is defined as the following: For each observation $z_i = (x_i, y_i)$, a random forest model is built by averaging only those trees corresponding to bootstrap samples in which z_i did not appear. Analogously, the OOB error is the average prediction error using only the trees that did not have z_i in their bootstrap sample.

```
Input: $X \in \mathbb{R}^{N \times p}$ denoting $N$ observations with $p$ features, $B$ the number of
  bootstraps, $m < p$ the number of features used for node splitting, and $n_{min}$
  the minimum node size
Output: Ensemble of trees $\{T_b\}_{b=1}^{B}$
Procedure:
Initialization: $b = 1$;
while $b \leq B$ do
    1. Draw a bootstrap sample $Z^*$ of size $N$ from the training data;
    2. Using $Z^*$ grow a tree $T_b$, by recursively repeating the following steps for
      each terminal node of the tree, until the minimum node size $n_{min}$ is
      reached;
        • Select $m$ features at random from the $p$ features.
        • Pick the best feature among the $m$ as the split point.
        • Split the node into two daughter nodes.
end
To make a prediction for an individual sample $x$:
  • Classification:
    $\hat{y}(x) =$ majority vote$\{\hat{C}_b(x)\}_{b=1}^{B}$, where $\hat{C}_b(x)$ is the class prediction of $T_b$.
  • Regression:
    $\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$.
```

Algorithm 2: Pseudocode for building a random forest

Feature importance is another advantage that random forests bring. At each split in each tree, the improvement in the split-criterion is the importance measure attributed to the splitting variable, and is accumulated over all the trees in the forest separately for each variable. Using the feature importance, it is possible to interpret the model by exploring the importance of each feature and assess their contribution on predicting the response.

2.2.3 Model assessment and selection

One of the most popular methods for estimating prediction error is cross validation. To execute a cross validation procedure, the data needs to be, initially, partitioned into training and test sets. The test set must not be used in the model tuning and selection at all. The purpose of keeping this portion of data aside is to have a fair comparison between different models that their prediction were being made on this test set.

The training set, however, as the name suggests, should be used throughout the training phase. When cross validation is employed, the training set itself will further be partitioned into subsets called validation. One of the well-known cross validation methods is called *k-fold* cross validation. For a given k, the training set is partitioned into k (almost) equal size subsets referred to as folds. The $k - 1$ folds are dedicated to train and fit a model and the k^{th} fold is used to validate the model on. Through an iterative process, each time a different fold will be picked and set aside and the final prediction accuracy is computed by taking the average over the predictions obtained on the validation fold.

2.3 Related Work

2.3.1 Modeling gene regulation using epigenetic data

Ever since the genes were identified in human DNA or other organisms and the gene transcription concept was noted, the researchers have been relentlessly focusing on understanding this mechanism. Ample studies investigated the genetic and epigenetic associations with gene expression and numerous statistical models have been proposed to describe such associations. Ouyang, Zhou, and Wong, 2009 built a regression model based on the principal components of the transcription factor (TF) ChIP-seq signal to predict the gene expression measured as RNA-seq. They used the mouse embryonic stem cell data to identify two groups of TFs that either act as activators in general or those with dual (activator or repressor) functionalities. Ferdous et al., 2018 applied several statistical learning algorithms, naming feed forward neural networks, decision trees, random forests, on an integrated data set consisting of ChIP-seq time-series data for six protein markers (including histone modifications). They used the corresponding gene expression data evaluated under several biological conditions as the response for their statistical models. Their results support the key role that histone modifications play in transcriptional regulation. Also, the better accuracy that their models obtained on later time point profiles hints at the temporal aspect of the regulatory mechanism. Using their comprehensive study, they were also able to pinpoint at several protein profiles, such as CDK9 and Brd4, that were not strongly involved in transcriptional regulations.

2.3.2 Bidirectional gene regulation

Elison et al., 2018 particularly focus on understanding the bidirectional gene expression by studying GAL1 and GAL10; two bidirectional genes discovered in yeast *Saccharomyces cerevisiae*. Using a two-step CRISPR method (Elison, Song, and Acar, 2017), they experimentally investigated the effects of editing the regulatory sites within the promoter region shared by GAL1 and GAL10. Several other studies concentrated on unraveling the mechanisms of bidirectional gene regulation in different organisms and different promoter architectures (Wei et al., 2011; Xu et al., 2009; Neil et al., 2009; Fux and Fussenegger, 2003; Park et al., 2014; Yan et al., 2015; Amendola et al., 2005). The findings, however, were incomprehensive and inconsistent to some extent, which motivated us to systematically study this phenomenon as part of this thesis.

2.3.3 Revolutionizing discoveries using scRNA-seq

Inspecting the gene regulation in bidirectional promoters requires a very fine grained dissection of transcription mechanism preferably at the single cell resolution. There have been a myriad of studies that generated the scRNA-seq data as well as tools and methods aiming to analyze such data. As previously mentioned in 2.1.7, new sequencing protocols have emerged with the ability to capture a single cell, and sequence its mRNA, resulting in obtaining scRNA-seq data (Tang et al., 2009; Ramsköld et al., 2012; Hashimshony et al., 2012). Liu et al., 2019 produced and analyzed scRNA-seq data for bone marrow stromal cells that led to identification of three subpopulations based on the single cell gene expression profiles of known markers. Another interesting study concentrates on understanding the cell heterogeneity of midbrain dopamine neurons that are associated with neurological diseases such as Parkinson's disease. They prepared data sets of scRNA-seq for human and mouse in order to discover new subtypes among midbrain dopamine cells and characterize them at the genome-wide level (Tiklová et al., 2019). They utilized monocle (Trapnell et al., 2014) to infer the pseudotime trajectories in order to identify distinct temporal profiles across cells. Monocle is an all-round tool developed to analyze many different aspects related to the single cell data. Analyses such as clustering the cells based on their gene expression profile, identifying differentially expressed genes across the discovered cell subpopulations, building branching trajectories across cells reflecting the cell fate decisions.

As highlighted earlier, dropouts are common errors of single cell sequencing protocols and there have been differently many approaches to detect and impute the missing value of gene expression in a cell. Mongia, Sengupta, and Majumdar, 2019 impute dropouts in scRNA-seq data through a low-rank matrix completion based technique. They test their approach, mcImpute, on a number of real data sets and showed that mcImpute is capable of discerning true zeros from dropouts as well as imputing their missing values. Another study addressing the dropout issue in single cell sequencing is called scImpute (Li and Li, 2018a). The authors proposed a statistical method that automatically detects potential dropouts without introducing new biases to the rest of the data. scImpute also identifies outlier cells and discards them ensuring that the imputation would not be applied on them. Gong et al., 2018 proposed a different approach to attenuate the effect of dropouts on the scRNA-seq data. They considered different clustering configurations obtained from combinations of similarity metrics (Pearson and Spearman correlation) together with a varying range of number of clusters to perform cell-wise clustering on the data. For each combination, they estimated the zero values in the input matrix. The resulting clusters were all averaged to obtain the final imputation for the putative dropout events. Later Tracy, Yuan, and Dries, 2019 proposed another imputation method called RESCUE. In their workflow, they first, through a greedy approach, find the most variable genes across cells and subsample these genes using a bootrstapping procedure. The samples are then clustered in order to generate the imputed gene expression by averaging the within-cluster expression values. Finally. through an ensemble approach, all these imputed samples are averaged in order to obtain the final imputed data. The authors show that RESCUE outperforms the two aforementioned imputation methods, scImpute and DrImpute, owing to retaining the nature of the single cell data without imposing any strict model assumptions.

The potential of single cell data does not limit only to identifying new subtypes or cell subpopulations, even though it is a very interesting and worthwhile area to explore, rather it can be further exploited to better understand the underlying gene regulatory mechanism for a single cell. One of the studies that concentrated on delivering a tool that can provide biological insights into the mechanisms driving cellular heterogeneity is called SCENIC (Aibar et al., 2017). SCENIC was developed to map gene regulatory networks using scRNA-seq and discover stable cell states through evaluating the activity of those networks in single cells. They first identify subsets of genes, which are co-expressed with transcription factors. Then, in order to remove false positives due to indirect binding events, they retain the putative direct binding targets by contrasting the subsets with *cis*-regulatory motifs and identifying the significant cases. These subsets are then given a score based on their proposed scoring algorithm, which results in a binary activity matrix of genes versus TFs.

Chapter 3

Transcription factor binding - DREAM CHALLENGE

The work that is presented in this chapter has been published in F1000Research in 2019 (Behjati Ardakani, Schmidt, and Schulz, 2019).

3.1 Introduction

Transcription factors (TFs) are essential elements of transcriptional regulation. They play crucial roles in establishing and maintaining cellular identity and the aberrant changes in their activity can result in several diseases (Vaquerizas, 2009). Some TFs form a direct bind to the DNA molecule at distinct positions, mostly in open chromatin regions, where the chromatin is accessible (Natarajan, 2012), and regulate transcription by recruiting additional proteins. General TFs are involved in altering chromatin organization as well as recruiting RNA polymerase to initiate transcription (Vaquerizas, 2009). Therefore, to understand the function of TFs, it is necessary to identify the TF binding sites (TFBS) on the genome. Depending on the tissue, TFs bind and regulate distinct genes, i.e., these binding sites are tissue-specific (Natarajan, 2012).

Nowadays, ChIP-seq experiments are widely used to experimentally (*in vivo*) determine tissue-specific TFBS genome-wide. But the downside is that the ChIP-seq experiments can become very challenging as performing the experiments are expensive and require an antibody for the target TF. To tackle these limitations, several computational methods have been proposed to identify TFBS. The majority of these methods are established based on Position Weight Matrices (PWMs) describing the sequence preference of TFs (Mathelier, 2016). PWMs specify the occurrence of each nucleotide at each position of a TF binding motif. Unfortunately, screening the entire genome using PWMs, in order to identify the TFBS, introduces too many false positive predictions. Therefore, several methods have been developed to mitigate the prediction error by combining PWMs with epigenetics data, such as DNase1-seq, ATAC-seq, or Histone Modifications, representing chromatin accessibility. It has been shown that including additional features, such as nucleotide composition, sequence conservation, and DNA shape, can remarkably enhance the task of TFBS prediction (Pique-Regi, 2011; Luo, 2013; Gusmao, 2014; Kahara, 2015; Yardımcı, 2014; Cuellar-Partida, 2012; O'Connor and Bailey, 2014; Liu, 2017). Jayaram, 2016, indeed, provides a non-exhaustive overview of this topic.

Even though PWM based models are amongst the most common tools to evaluate the likelihood of a TF binding to genomic sequences, more involved approaches such as Slim models, which capture nucleotide dependencies, have been successfully used as well (Keilwagen and Grau, 2015). Alipanahi, 2015 proposed a *de novo* approach to learn TF binding specificities from large scale datasets using deep learning.

Given the importance of identifying TFBS, there has been a challenge held by the DREAM challenge organizers to encourage researchers focusing on tackling this issue. The challenge's title was *ENCODE-DREAM in vivo Transcription Factor binding site prediction challenge* (ENCODE-DREAM, 2017) through which participants had the opportunity to develop and evaluate their method on the provided data, which consisted of TF-ChIP seq data for 31 TFs, accompanied with RNA-seq and DNase1-seq data in 12 different tissues. The class labels were deduced from the TF ChIP-seq data. The *challenge* organizers had a systematic comparison between the different approaches on TFBS prediction through evaluating the prediction power of the proposed models on completely different tissue/cell types that were not used for training.

My colleague, Florian Schmidt, and I with the supervision of Prof. Dr. Marcel Schulz formed a team to participate in this challenge. The methodology we used and experiments we designed to address the TFBS prediction as well as the results are described in this chapter. Briefly, we proposed an ensemble learning approach using random forest (RF) classifiers, extending the work of Liu, 2017. Since, Liu, 2017 and Waardenberg, 2016 have shown that tissue-specific cofactor interactions are appropriate for modeling TF binding, we designed our ensemble model such that it was able to exploit the tissue specificity inherent in the data and gain a better generalizability power.

The summary of the procedure is as follows. First TF affinities are computed using TRAP (Roider, 2007) for 557 PWMs in DNase-hypersensitive sites (DHSs) identified with JAMM (Ibrahim, 2015). The computed TF affinity scores can capture low affinity binding sites, which were shown to be biologically relevant (Tanay, 2006; Crocker, 2015). This data, which was generated by Florian Schmidt, forms our feature space to be used in training random forest classifiers in order to predict the binding site of the TF of interest. Generating the models and their evaluation was performed by me.

3.2 Methods

3.2.1 Data

For 31 TFs, the ChIP-seq data was provided, as well as DNase1-seq and RNA-seq data for 13 different tissues. The TFs used for training the classifiers are listed in Table 3.1. It provides the number of bins labeled as bound for each tissue or cell line for which the TF ChIP-seq was available to infer the class labels. Except for the held-out chromosomes 1, 8, and 21, all chromosomes are used for the training phase.

The evaluation on the unseen test data was performed on eight TFs listed in Table 3.2. The predictions were made in bins of size 200 bp, each shifted by 50 bp, spanning the entire genome, as specified by the competition. In addition to bound and unbound labels, the bins were annotated with another class called *ambiguous*, which we excluded them from our study. Further details on the data can be accessed from the *challenge* website ENCODE-DREAM, 2017.

3.2.2 Data preprocessing and feature generation

Due to the memory limitation as well as drastic imbalance in the class distribution, we shrank the data into a balanced and reasonably smaller subset. For each TF, we

TF	Number of bins labelled as bound per tissue
ATF7	272,2234 (GM12878), 218,239 (HepG2), 345,775 (K562)
CREB1	164,968 (GM12878), 103,752 (H1-hESC), 178,080 (HepG2), 98,554 (K562)
CTCF	179,672 (A549), 271,097 (H1-hESC), 206,336 (HeLa-S3), 208,868 (HepG2), 215,238 (K562), 305,547 (MCF-7)
E2F1	93,117 (GM12878), 55,391 (HeLa-S3)
EGR1	72,595 (GM12878), 52,733 (H1-hESC), 175,994 (HCT116), 58,793 (MCF-7)
EP300	126,409 (GM12878), 69,247 (H1-hESC), 157,629 (HeLa-S3), 168,173 (HepG2), 137,369 (K562)
GABPA	26,467 (GM12878), 51,666(H1-hESC), 31,202 (HeLa-S3), 60,552 (HepG2), 109,423 (MCF-7), 78,403 (SK-N-SH)
JUND	203,665 (HCT116), 179,999 (HeLa-S3), 183,558 (HepG2), 193,814 (K562), 92,905 (MCF-7), 222,013 (SK-N-SH)
MAFK	34,054 (GM12878), 97,659 (H1-hESC), 62,124 (HeLA-S3), 291,337 (HepG2), 201,157 (IMR90)
MAX	301,615 (A549), 98,327 (GM12878), 224,379 (H1-hESC), 321,501 (HCT116), 211,590 (HeLa-S3), 317,579 (HepG2), 318,318 (K562), 250,775 (SK-N-SH)
MYC	57,512 (A549), 91,325 (HeLa-S3), 183,627 (K562), 151,748 (MCF-7)
REST	71,251 (H1-hESC), 47,654 (HeLa-S3), 67,453 (HepG2), 59,640 (MCF-7), 48,946 (Panc1), 94,082 (SK-N-SH)
RFX5	161,689 (GM12878), 22,948 (HeLa-S3), 54,961 (MCF-7)
SRF	21,495 (GM12878), 40,201 (H1-hESC), 176,158 (HCT116), 22,593 (HepG2), 18,895 (K562)
TAF1	87,109 (GM12878), 185,027 (H1-hESC), 93,824 (HeLa-S3), 110,385 (K562), 83,276 (SK-N-SH)
TCF12	51,798 (GM12878), 104,834 (H1-hESC), 82,102 (MCF-7)
TCF7L2	100,926 (HCT116), 165,264 (HeLa-S3), 143,025 (Panc1)
TEAD4	66,198 (A549), 103,483 (H1-hESC), 174,716 (HCT116), 125,917 (HepG2), 186,759 (K562)
YY1	136,621(GM12878), 195,489 (H1-hESC), 63,293 (HCT116), 133,943 (HepG2)
ZNF143	197,385 (GM12878), 178,088 (H1-hESC), 48,154 (HeLA-S3), 103,755 (HepG2)

TABLE 3.1: Number of bins labeled as bound per transcription factor (TF) and tissue.

TF	Tissue(s)
CTCF	PC-3, Induced pluripotent stem cell
E2F1	K562
EGR1	liver
GABPA	liver
JUND	liver
MAX	liver
REST	liver
TAF1	liver

TABLE 3.2: Test data shown per transcription factor (TF) and tissue.

randomly sampled as many negative sites as there were positive binding sites for training the classifiers.

We explored two different approaches for designing the feature data: (1) with and (2) without considering DNase-hypersensitive sites. In none of the approaches, we have used the provided RNA-seq data nor did we compute DNA shape features. The TF binding affinities are computed using *TRAP* (Roider, 2007) for 557 distinct TFs executed with the default parameter settings. The position specific energy matrices (PSEMs) used in our computation are converted from position weight matrices (PWMs) obtained from JASPAR (Mathelier, 2016), UniPROBE (Hume, 2015), and Hocomoco (Kulakovskiy, 2016). The code to perform the conversion and running TRAP was included by Florian Schmidt in our github repository [1]. Figure 3.1a shows the workflow for the first feature setup, where tissue-specific DHSs using the peak caller JAMM (Ibrahim, 2015) (version 1.0.7.2) are computed and then the called peaks are merged using the *bedtools merge* (Quinlan and Hall, 2010) command (bedtools version 2.25.0). Next, TF affinities are calculated in the specified DNase-hypersensitive sites using TRAP and then the median DNase1-seq signal per peak is computed from the *bigwig* files provided by the competition. Using a *left outer join* command from bedtools, the resulting data is intersected with the binned genome structure required for training and testing provided by the competition.

The second setup, as shown in Figure 3.1b, does not rely on DHSs, instead TF binding affinities and the DNase1-seq signal are computed in the given bins described in *Data*. To address the variability between biological and technical replicates, the median coverage of the DNase1-seq signal across replicates is computed using the *bedtools coverage* command. To sum up, in setup 2, the features computed for each bin are TF affinities attributed to that bin and the DNase1-seq signal that is measured in that bin as well as the bins located to the left and right (DNase1L and DNase1R).

3.2.3 Ensemble Random Forest classifier

The Random Forest models, implemented using the *randomForest* R-package (Liaw and Wiener, 2002) (version 4.6-12), are trained on either of the feature setups explained in the previous section. Training the RF models consists of two main steps, irrespective of feature setup. In order to prevent the classifiers inclining to the major class, i.e., over-fitting to the class with larger size, we first balanced the two *bound* and *unbound* classes. To fit the RF classifiers, we used $4,500$ trees, and at most $30,000$ positive (bound) and negative (unbound) samples. This restriction was imposed due to the limitations of the *randomForest* R-package. As illustrated in Figure 3.2a, for a given target TF, we first learn the tissue-specific RF classifiers using all available features, $T_i \in R^{n\times 557}$ $;i \in \{1,\cdots,m\}$, where n is the number of bins forming the training set, and m denotes the number of training tissues for the target TF:

$$RF_i = RandomForest(T_i, Binding(T_i)),$$

where $Binding(T_i)$ is a vector of length n, holding the binding labels for the target TF in tissue i, and $RandomForest(.,.)$ generates the RF model trained on the features and labels provided by the first and second argument, respectively. An example of the input matrix T_i, for three tissues, i.e., $m = 3$, and the response vector $Binding(T_i)$ is shown in Figure 3.2b. The resulting models are then used to perform an intermediate feature selection step. In the second step, by shrinking the feature space to

[1] `https://github.com/SchulzLab/TFAnalysis`

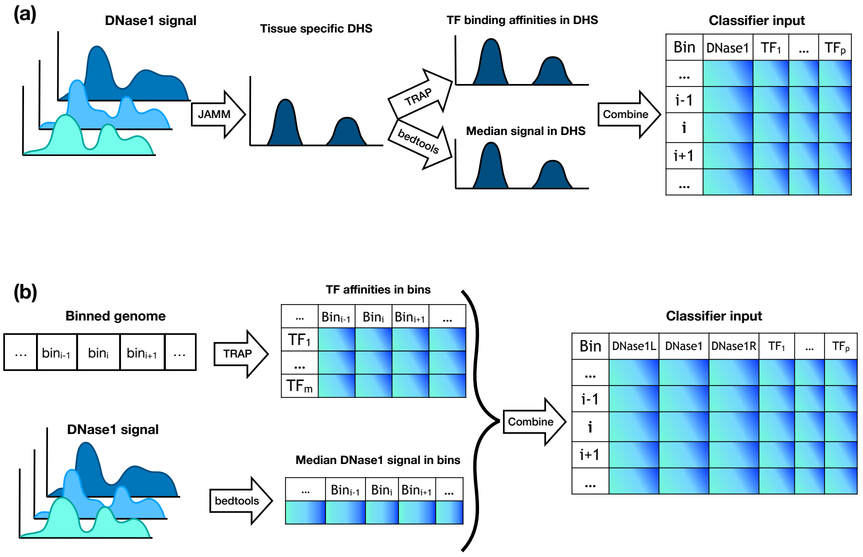


FIGURE 3.1: Two data preprocessing workflows (a) Using the JAMM peak caller, DHSs are called for all replicates of a distinct tissue. TF affinities in the identified DHSs are computed using TRAP for $m = 557$ TFs, the median signal of DHSs is computed using bedtools. The concatenation of both TF affinities and median DHS signal forms the input feature. (b) Similar to (a) but instead of DHSs, TF affinities and median DNase1-seq signal are computed per bin.

the union of top 20 regulators (c.f. Figure 3.3a) among T_i' classifiers, we put focus on the essential regulators to be further used in our subsequent models. The top 20 regulators are obtained through ranking the feature variables according to their *Gini index* (Figure 3.2c):

$$T_i' = Subset(T_i, \bigcup_{j=1}^{m} TopFeatures(RF_j)),$$

where $TopFeatures(RF_j)$ denotes the top 20 features of RF_j and $Subset(.,.)$ generates the reduced feature matrix based on the union of the top TFs. In the following, we refer to training datasets consisting of only one tissue and training datasets consisting of multiple tissues as *single tissue* and *multi tissue*, respectively. Considering the single tissue case, we train an RF model, RF_i', on the reduced feature space and use this as the final model for the respective target TF (no ensembling step):

$$RF_i' = RandomForest(T_i', Binding(T_i)).$$

In the multi-tissue scenario, we retrain tissue-specific RF models on the reduced feature space and apply them across all available training tissues:

$$T_E' = \{Prediction(RF_i', T_i'); i = \{1, \cdots, m\} \in [0,1]^{n \times m}\},$$

where $Prediction(RF_i', T_i')$ returns the predictions made by RF_i' when applied on the T_i'. Their predictions are combined in a new feature matrix that is used as input to train an ensemble RF, RF_E. Note that the input matrix contains predictions of all tissue-specific RF models on all training tissues (Figure 3.2d):

$$RF_E = RandomForest(T_E', Binding(T_E')).$$

By design, the ensemble model combines the tissue-specific RF classifiers in a non-linear fashion for a better generalizability across all training tissues.

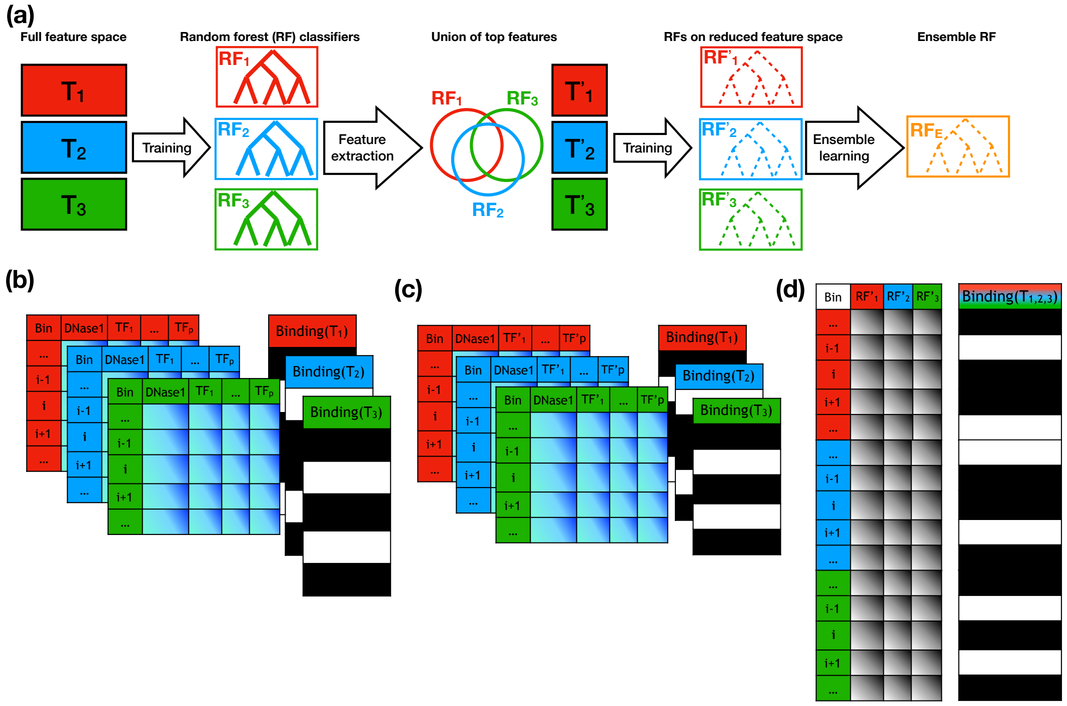


FIGURE 3.2: a) An overview of model training for a given target TF with multiple training tissues. b) Full feature matrices, T_1, T_2, T_3 from (a), are used to train tissue-specific Random Forest (RF) classifiers. From those RF classifiers (RF_1, RF_2, RF_3 from a), the union of the top 20 features is determined from each RF. c) Based on the union ($m' \leq m$), reduced tissue-specific feature matrices, T_1', T_2', T_3' as ina(a), are produced. Subsequently, tissue-specific RF classifiers (RF_1', RF_2', RF_3' from a) are trained on these reduced feature sets. d) The predictions of tissue-specific RF classifiers applied on all training tissues are aggregated to form the feature matrix T_E', which is then used to train the ensemble model (RF_E from a). Note that the feature matrices represent feature setup (1), where the DHS sites were used.

3.2.4 Performance assessment

The model performances are assessed in two different scenarios. The first scenario addresses the evaluation on the out-of-bag (OOB) data. The OOB error is defined as the mean prediction error for each training sample using trees that were not trained on that sample. Using the PRROC (Grau, Grosse, and Keilwagen, 2015) package, the performance on OOB data is computed in terms of the area under the precision recall curve (PR-AUC) and the area under the receiver operator characteristic curve (ROC-AUC). The former measures the accuracy by contrasting precision against recall, whereas the latter contrasts false positive rate against true positive rate. A ROC-AUC value around 0.5 suggests a random classifier, however there is no such baseline defined for PR-AUC. The higher values of PR-AUC suggest better classification power and the lower values indicate weaker prediction power. In addition to the curve based measurements, we evaluate the models based on the misclassification rate for the Bound and Unbound classes, corresponding to the false negative and false positive rate, respectively:

$$Bound = (\frac{FN}{TP + FN}), \quad Unbound = (\frac{FP}{TN + FP}), \tag{3.1}$$

where TP is the number of bins correctly predicted as bound, TN is the number of bins correctly predicted as unbound, FP and FN denote the number of bins incorrectly predicted as bound and unbound, respectively.

Secondly, we compute the aforementioned performance measurements for the test data shown in Table 3.2. As previously mentioned, the test data consists of three held-out chromosomes, 1, 8, and 21, which have not been used for training. Beside the distinct chromosomes allocated for the testing round, the TF binding prediction is performed on unseen tissues, i.e. tissues that were not used for training. Unlike the training data, test data is not balanced, i.e. the unbound class is larger than the bound class. Given this class imbalance, the PR-AUC is considered as a more appropriate metric for measuring performance compared to other metrics, such as ROC-AUC, false positive and false negative rates. Due to memory limitation of the PRROC package the test data had to be downsampled to 100,000 samples, while preserving the original Bound to Unbound ratio.

The two feature setups illustrated in Figure 3.1 are evaluated on the same gold standard (the same test data sets), therefore a fair comparison can be made to contrast their performance against each other.

3.2.5 Protein-protein-interaction score

The purpose of the feature reduction step was to select TFs that are likely to interact with the target TF. To systematically test whether this goal has been achieved, we utilized a protein-protein-interaction score.

A customized protein-protein-interaction (PPI) probability matrix R was introduced by Köhler, 2008, where a random walk analysis was conducted on the protein-protein-association network obtained from STRING (Szklarczyk, 2017) (database version 9.05). An entry $R_{i,j}$ in this matrix represents the probability for which protein i interacts with protein j. This probability is not symmetric, meaning $R_{i,j} \neq R_{j,i}$. To generate a score describing how likely it is that a subset of proteins P contained in R interact with a distinct TF t, guided by the feature importance scores the RF models provide, we define the PPI score $S_{t,P}$ as follows

$$S_{t,P} = -log(\frac{\sum_{p \in P}((R_{p,t} + R_{t,p}) \times GI(p))}{2|P|}), \tag{3.2}$$

where $GI(p)$ represents the Gini index values of p obtained from the RF model corresponding to t. This means that for smaller values of $S_{t,P}$ the chance that the regulators in P interact with TF t is higher.

3.3 Results

In this section, we first show that narrowing the feature space down to those TFs essential for training does not significantly hurt model accuracy. Then, we exhibit the benefits of the ensemble learning to tackle the TFBS prediction problem and how its accuracy varies as a function of number of training tissues. We further examine the top selected TFs by the RF classifiers and discover known interaction partners based on their high PPI scores. Finally, we compare the two feature setups, described in the *Methods* section, and investigate their influences on model performance. If not stated otherwise, all figures presented in the following are based on annotation setup (1), obtained in DHSs.

3.3.1 Reducing the feature space to a small subset does not affect classification performance

Since having a sparse feature space eases model interpretation, we reduce the feature space to a hand full of important ones. As mentioned previously, we determined top features based on the Gini index obtained from the RF models, resulting in TF and tissue-specific sets containing either the top 10 or top 20 features. As depicted in Figure 3.3a (also Appendix A.3a) the difference in OOB error between the feature set with top 10, top 20, and all features is negligible. Intriguingly, on test data a slight increase in model performance can be observed for the reduced feature space models in comparison with the model where all features were used. This observation hints to gaining generalizability power through the feature extraction procedure (Appendix Figure A.1). Because of this performance gain on test data, as well as a substantial assistance to the interpretability of the models and runtime, we decided to choose the case where the union of top 20 TFs is taken across models to proceed with the learning procedure. Our results reveal that, for feature setup (1), the DNase1-seq signal within the DHSs is the most important feature across all TFs. Similarly, for setup (2), the features corresponding to DNase1-seq signals measured in left, center, and right bins are amongst the most important ones.

3.3.2 Ensemble learning improves model accuracy

As shown in Figure 3.3b (Appendix Figure A.3b), the RF ensemble classifiers outperform the tissue-specific ones, which suggests that the ensemble models are capable of generalizing across tissues. Figure 3.3c provides the performance of the models on the test tissues that were linked to multiple training tissues. These results show that ensemble models have higher PR-AUC compared to tissue-specific classifiers. Since the test data was imbalanced, the ROC-AUC measurements tend to be in favor of the tissue-specific classifiers. However, this is an example of a case where ROC-AUC is not a suitable performance metric, since it is biased due to the high number of negative (i.e. unbound) cases in the test data. The superiority of the ensemble models are also reflected by false positive and negative rates (Appendix A.3c). Taking these results into account, it can be concluded that ensemble learning is a promising approach to address the tissue specificity of TF binding.

3.3.3 Prediction accuracy and its relation to the number of training tissues

Although the results in Figure 3.3b and 3.3c suggest that the ensemble methods perform well, it would be interesting to understand how the number of training tissues influence the performance. To address this, permutation experiments were performed to train classifiers with all possible combinations of training tissues. As this is a computationally expensive task, we executed the procedure on only three, arbitrarily selected, TFs: MAX, TEAD4, and E2F6. The results in figure 3.4a (Appendix A.4a) show that the OOB AUC value increases as the number of training tissues increases. This observation suggests that the ability of an ensemble RF to generalize across tissues improves with more training tissues.

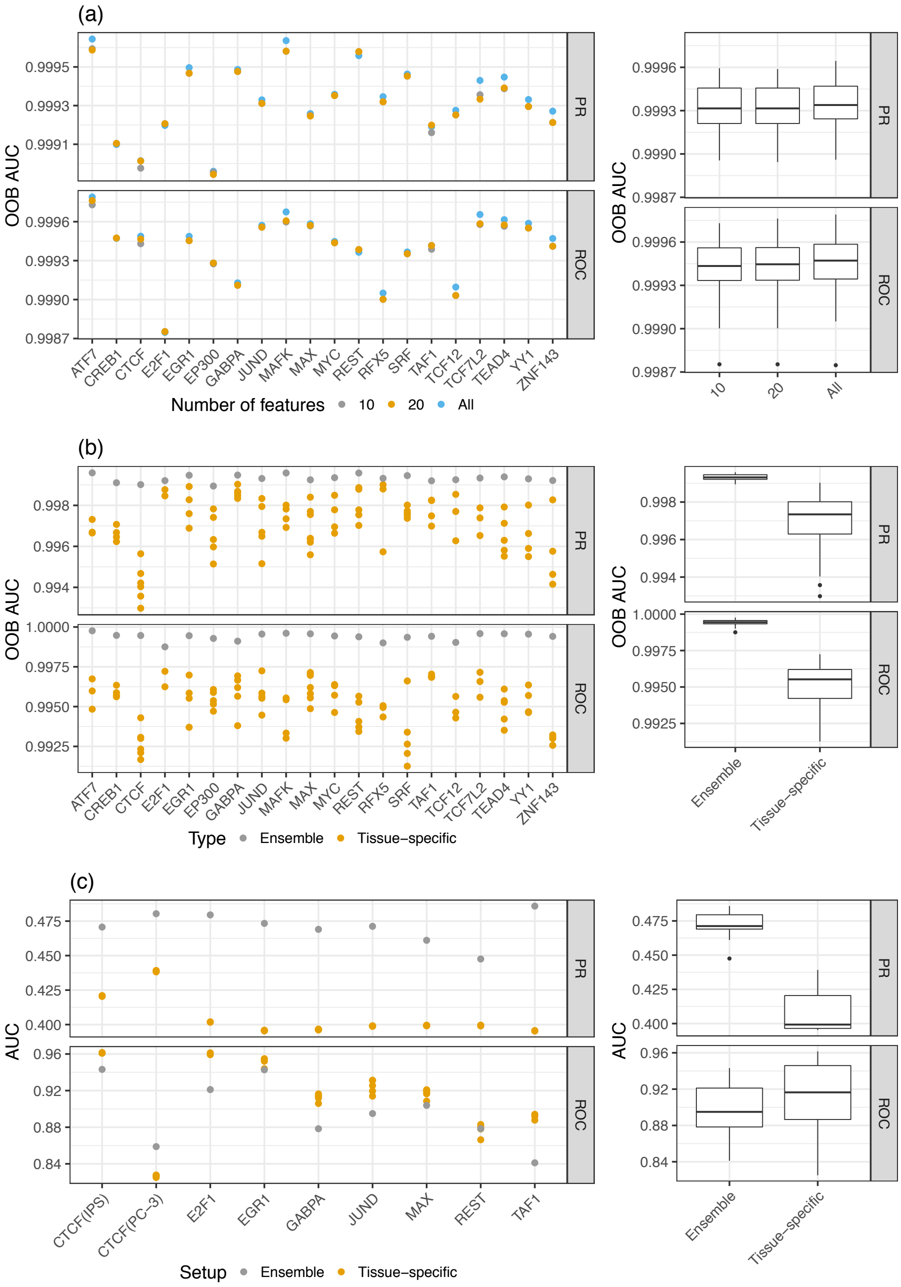


FIGURE 3.3: a) PR-AUC and ROC-AUC for different sets of features (all features, top 10, and top 20). Differences in performance between the top 20 and all features models are minor. b) Comparison of the out of bag (OOB) error between ensemble models and tissue-specific random forest (RF) classifiers. Ensemble models outperform tissue-specific RF classifiers. c) PR-AUC and ROC-AUC computed on unseen test data for ensemble and tissue-specific RF classifiers. Note that the scale of the y-axes is different for the subplots.

In order to test whether the improved accuracy obtained from the ensemble RF classifiers was in fact because of the ensemble learning two additional learning setups were designed. Firstly, we aggregated all tissue-specific data sets into one. In other words, we pooled the training data for one TF across all available tissues into one data set. Then, we used this pooled data set to train a new RF model. Secondly,

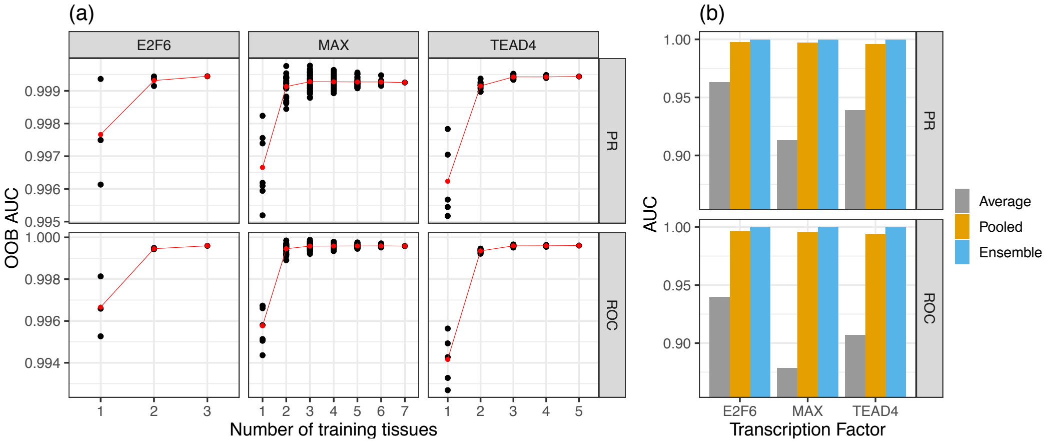


FIGURE 3.4: Comparison of number of tissues and classifier setups for the three TFs E2F6, MAX, and TEAD4. a) Model performance as a function of number of tissues used for training. The OOB AUC values increase when more tissues are included in the ensemble learning. Red dots represent the mean classification error across all tissue-specific classifiers. The black points represent individual models. b) Comparison between two ensemble models: averaging (takes the average of all individual RF predictions) and the RF ensemble model. In addition, one RF classifier was trained on pooled data sets comprised of training data for all available tissues for one target TF. The ensemble models perform better than the models based on aggregated data.

we examined another ensemble approach, which we consider to be a baseline for our proposed ensemble model. In detail, we computed the average of predictions over tissue-specific models in order to obtain the final prediction. As depicted in Figure 3.4b the proposed ensemble models perform better than both tested alternatives. This shows that the RF ensemble technique is better suited to capture tissue-specific information than any other tested methods.

3.3.4 Predictors selected by the RF classifiers are associated to the target TF

As previously stated, we speculated that the top predictors selected by the RF classifiers should represent regulators that exist either in protein complexes with the target TF through direct or indirect binding, or bind directly to DNA in close proximity to the target TF. To address this speculation, we computed a PPI score $S_{t,P}$ (see 3.2.5) for the selected predictors P of target TF t and compared it against PPI scores obtained for randomly sampled sets of TFs (based on 100 TF subsets drawn randomly). The PPI score $S_{t,P}$ for target TF t is small, if t is likely to interact with factors in P. Conversely, the score is high if t is less likely to interact with factors in P. As shown in Figure 3.5a, except for three TFs (MAX, TAF1, ZNF143), the PPI scores of the TFs suggested by our RF classifiers are better (i.e., smaller) than the scores for the randomly selected set. This supports that the features selected by RF classifiers represent regulators that are likely to interact with the target TF, either directly or with indirect contacts.

Figure 3.5b shows an example of a PPI network focused on the TF MAFK. The network was obtained from the STRING database (Szklarczyk, 2017), with the settings *highest confidence* and *no more than 10 interactors*. The top features selected by the RF classifiers contain all known regulatory proteins in this network, except for

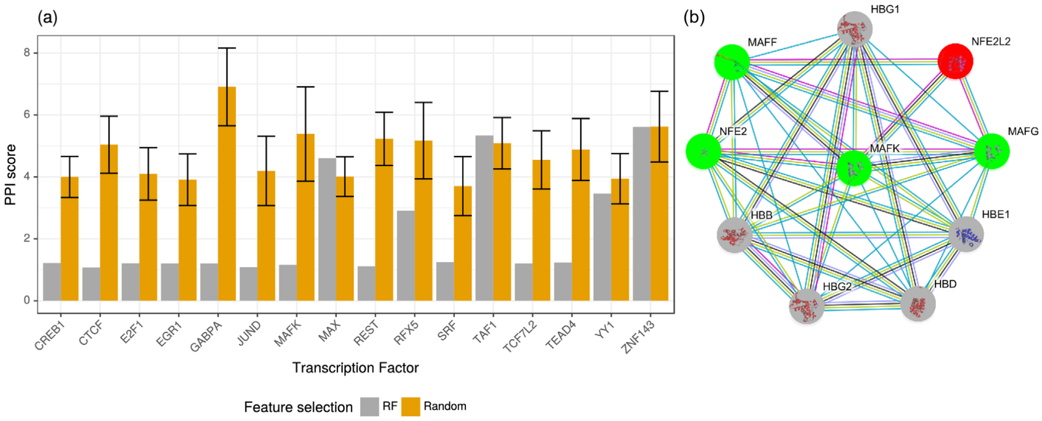


FIGURE 3.5: a) PPI scores calculated for a set of TFs. In the *Random* case, mean and standard deviation of PPI scores are shown across 100 random draws. Except for three TFs (*MAX*, *TAF1*, *ZNF143*), the PPI scores obtained from the TFs suggested by the RF classifiers are better than *Random*. b) protein-protein-association network obtained from STRING database centered around the TF *MAFK*, highlighting proteins that interact with *MAFK* with high confidence. Proteins colored in green were reported as important features by the RF classifiers, proteins shown in gray could not be retrieved by our model, either due to the fact that they are not DNA-binding proteins, or we do not have their PWM in our set. Regulators shown in red could have been detected by the RF, but did not appear in the top set of regulators.

NFE2L2, shown in red. Among these TFs are MAFK itself, MAFF, MAFG and NFE2 (highlighted in green). The strong interactions among the small MAF proteins (Kannan, 2012) as well as the dimerization of those with NFE2 (Igarashi, 1994) have been previously reported in the literature.

Interaction partners marked in gray can not be detected by our approach as either these are proteins without known regulatory functions or we do not have their PWMs to include in our features.

3.3.5 Feature setups influence the FP and FN

Figure 3.6, delivers the PR-AUC and ROC-AUC values computed for the bin based and peak based feature setups evaluated on the test data. Since the test data was largely unbalanced, the ROC-AUC assessments appear to be rather inconclusive. According to the PR-AUC values shown in this figure, the bin based models outperform the peak based models in the *Bound* case, whereas the peak based models exhibit superiority in the *Unbound* case. In contrast, bin based models perform poorly in the *Unbound* case, which is most likely due to the strong dependence of the RF classifiers on the DNase1-seq signal. On the other hand, the peak based models perform well in the *Unbound* case, as the search space for TFBSs is restricted only to DHSs. This elevates the precision of the predictions, but simultaneously lowers the recall, which is reflected by the large misclassification rate in the *Bound* samples (Appendix Figure A.5).

The bin based models (setup 2) outperform the peak based models when predicting the *Bound* class, whereas the peak based models show superior performance for predicting the *Unbound* labels. Correspondingly, bin based models perform poorly in the *Unbound* case, which is probably driven by the strong dependence of the RF

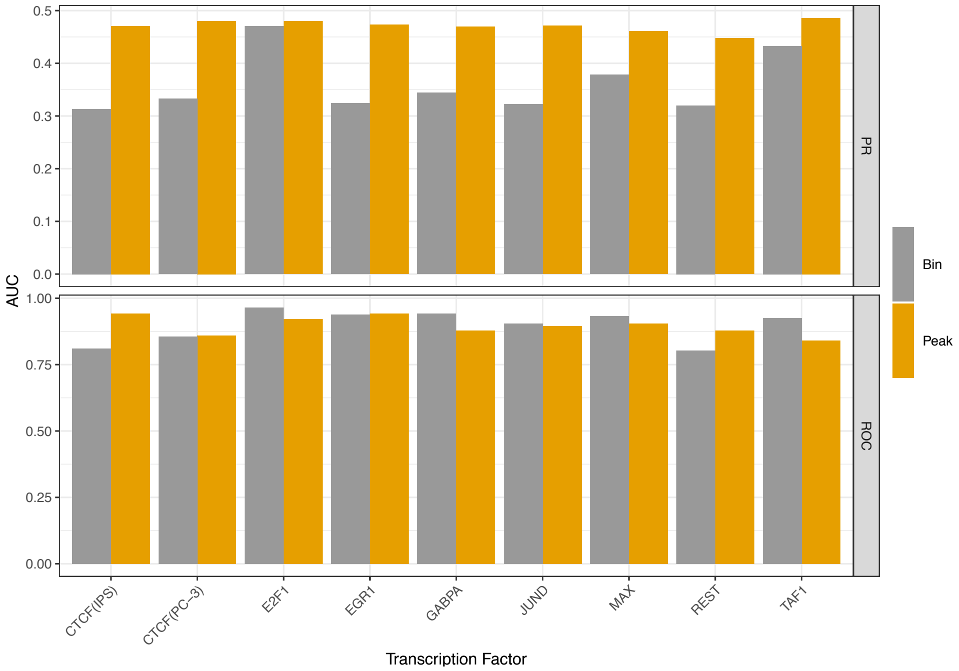


FIGURE 3.6: Comparison of PR-AUC and ROC-AUC values for the feature setups, Bin and Peak, measured on test data. Peak based models clearly outperform the bin based models when PR-AUC measurements are considered. This trend becomes somewhat vague when ROC-AUC values are examined. However, in this case, the ROC-AUC assessment is less reliable than PR-AUC due to the largely unbalanced test data.

classifiers on the DNase1-seq signal. On the other hand, models based on DHSs perform well in the *Unbound* case, because the search space for TFBSs is limited to only DHSs. This increases the precision of the predictions, but simultaneously lowers the recall, reflected by the high misclassification rate in the *Bound* case.

Conclusion and discussion

Here, we proposed an RF based ensemble learning approach to predict transcription factor binding sites. Our proposed approach pinpoints the advantages of ensemble learning in a multi tissue setting, which is able to pick up the associated cofactors to the target TF.

Evaluated on the OOB and test data, our proposed ensemble approach is able to better generalize across tissues, in contrast to classifiers trained only on a single tissue (Figure 3.3). Moreover, the accuracy of the ensemble classifiers elevates as the number of training tissues increases (Figure 3.4a). We further show that merely pooling all training data to learn one RF does not provide as accurate results as our ensemble model (Figure 3.4b). In this study, we chose random forests, as they are considered powerful and accurate classifiers that generate non-linear predictions in a reasonable time.

RF classifiers have also been proposed recently (Liu, 2017) as a suitable method to predict TF binding. Similar to our work, the authors of Liu, 2017 perform cross cell-type predictions, but not in an ensemble fashion.

As shown in Figure 3.3a, reducing the feature space to the most relevant ones provides similar classification performance as the full feature space. In Figure 3.5, we showed that most of these selected TFs are known interaction partners for the target TF. This is further supported by a recent study showing that most TFs bind in dense clusters around genes, indicating an extensive interaction among them (Yan, 2013).

We further examined the interactions between target TF and feature TFs by comparing their corresponding PPI scores. For most TFs, the results show a better PPI score for the predicted TFs than random. However, the PPI score for TAF1 and MAX, for which the ensemble classifier could improve only marginally over the tissue-specific classifiers, was worse than random. This implies that the models trained for TAF1 and MAX cannot account for their true interaction partners. As a matter of fact, an inspection of the STRING database revealed that only TAF1 and TBP are among the top 20 regulators for TAF1 that are included in our PWM collection. For the remaining interaction partners, which are mostly from the TAF family, no binding motif exists in the public repositories. As a result, these TFs are absent in our PWM collection that were used to generate the features for our RF classifiers. Similarly, for MAX, only 5 out of 20 high confidence interaction partners are included in our PWM collection. Specifically, no PWM is available for 6 TFs interacting with MAX, while the remaining interacting proteins are not categorized as TFs.

Replacing the feature setup for the RF classifiers from (1) DHS-based to (2) bin-based showed that DHSs are necessary to reduce the false positive rate (Figure 3.6) for TFBS predictions. Using only bins, without DHS information, recall could be improved, but only at the cost of poor precision. One obvious explanation for this is the difference in size of the genomic search space between the two setups. The bin based models have a low misclassification rate in the *Bound* case, because they do consider the whole genome without neglecting any sites beforehand, thus improving recall. However, our observations suggest that considering only the raw signal does not adequately correct for false positive sites, as opposed to using DHSs, which yields an improved misclassifcation rate in the *Unbound* case compared to the raw signal. To potentially overcome the strong biases introduced by DHS- and bin-based models, we could train yet another ensemble classifier using the predictions of the DHS- and the bin-based models as input. Depending on the application, the model could be optimized based on different accuracy metrics, such as precision, recall, or a joint metric like PR-AUC.

In general, training and evaluating methods for predicting TFBS is challenging mostly because of the imbalanced nature of the problem. In other words, there are many more *Unbound* (negative) than *Bound*(positive) binding sites in the genome. This requires both (a) proposing approaches that avoid over-fitting by learning only the major class and (b) developing evaluation strategies that can account for this issue.

Aside from the aforementioned technical difficulties, we show that modeling cofactors is beneficial to predict TFBS and that ensemble learning is a promising approach to gain generalizability across tissues.

Chapter 4

Bidirectional genes and their histone map

4.1 Introduction

Promoters are key structures for a coordinated regulation of gene expression. The increasing number of large-scale high resolution epigenomic and RNA-sequencing technologies led to a deeper understanding of genome-wide promoter configurations. Recent studies have shown that the number of bidirectional promoters (BPs) in the human genome is much larger than previously anticipated (Core, Waterfall, and Lis, 2008; Preker et al., 2008; Seila et al., 2008). Sensitive assays, such as sequencing of 5′-ends of capped nascent RNAs (GRO-cap and Start-seq), allow the detection of unstable nascent RNAs produced at promoters, and have revealed more widespread bidirectional transcriptional initiation than previously recognized (Core et al., 2014b; Duttke et al., 2015a; Scruggs et al., 2015). However, exact classification of bidirectional or unidirectional promoters in a sample of interest is challenging, as it depends heavily on the sensitivity of the sequencing assay to recognize unstable, nascent RNAs (Andersson et al., 2015; Duttke et al., 2015b).

Recent studies discuss two types of bidirectional promoters. The first type concerns transcription of two RNAs in opposite direction from one core promoter, *i.e.,* one promoter leads to bidirectional transcription (Bagchi and Iyer, 2016; Duttke et al., 2015a; Lacadie et al., 2016). In the second type, transcriptional initiation of both RNAs occurs at two distinct core promoters that are close to each other, but are oriented in reverse direction, thus sometimes termed divergent bidirectional promoters.

In this work we focus on bidirectional promoters that have two distinct core promoter elements that drive divergent transcription of two nearby genes. We are interested in the question if the regulation of one of these core promoters depends on the regulation of the other by studying the histone modifications (HMs) that are associated with transcription of the two genes at a BP.

In contrast to previous studies that have relied on the comparison of unidirectional against bidirectional promoters, we use a different approach by focusing solely on the bidirectional genes and aiming to associate spatial genomic features to the direction of transcription.

We analyze multiple histone modification data together with their corresponding gene expression measured in bulk RNA-seq to form an interpretable statistical model that is able to delineate how histone modifications might direct the transcription in bidirectional genes.

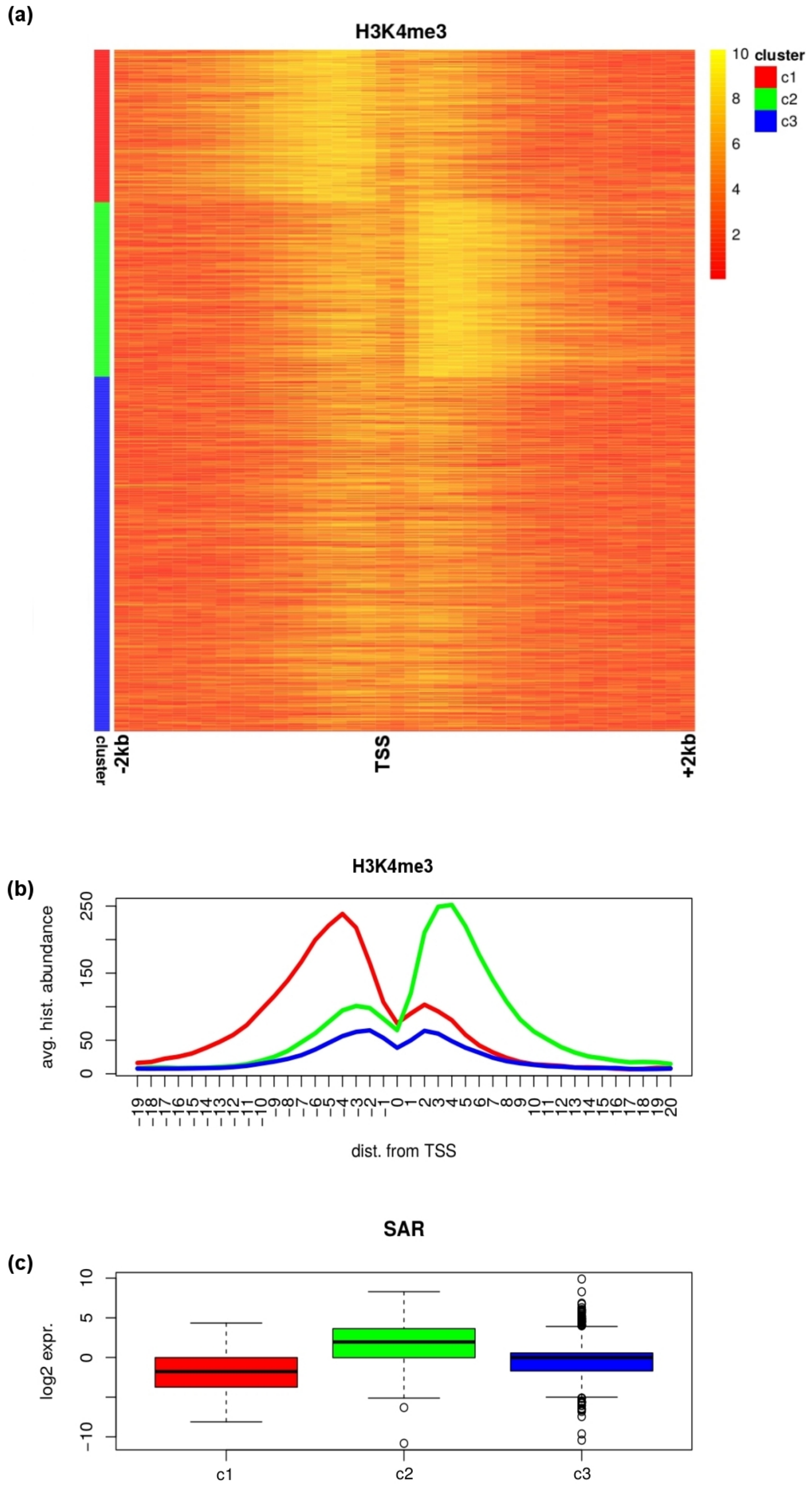


FIGURE 4.1: Bimodal distribution of H3K4me3 mark at bidirectional promoters. a) Heat map showing H3K4me3 signal in a 4kb window anchored on the TSS of the plus gene. The rows represent genes clustered by kmeans. b) H3K4me3 average signals in three clusters. c) The boxplots show the logarithm of plus to minus gene expression ratio for the three clusters.

For the sake of clarity, we refer to the two strands of the DNA as plus (also known as forward or Watson) and minus (also known as reverse or Crick) strands. The plus strand refers to the nucleotide sequence that is stretched from the $5'$ end to the $3'$ end. Conversely, the minus strand is the complementary DNA of the plus strand that stretches from the $3'$ end to the $5'$ end. Figure 4.1a shows the distribution of the H3K4me3 histone modification ChIP-seq signal measured in bins of 100 bp spanning a 4 kb window anchored at the TSS of the plus gene of 1242 BPs. As not all genes are regulated the same, even though they come from a particular subset of genes sharing a common property of being bidirectional, we were keen to find further subsets with the help of clustering. We performed k-means clustering on the matrix shown in the heat map of Figure 4.1a and inspected the cluster representatives (average signal in each cluster) as illustrated in Figure 4.1b. The signal corresponding to the cluster colored in blue has the smallest magnitude with two equal elevations at $\sim$300 bp up and downstream of the TSS. The other two clusters, on the other hand, exhibit uneven peaks with one being remarkably more pronounced than the other at $\sim$400 bp away from the TSS. This observation inspired us to inspect the bidirectional gene expression at these clusters. Figure 4.1c demonstrates the distribution of the log-transformed ratio between the expression values of genes located on plus and minus strands for the clustered BPs. Not surprisingly, the green cluster that showed a pronounced peak of H3K4me3 mark at the downstream of the TSS in Figure 4.1b, has higher expression values of the plus gene compared to minus. Similarly the red cluster, where the HM signal peaks upstream of the TSS (where the minus gene is located), has higher minus gene expressions. The blue cluster, as expected, is associated to bidirectional genes with relatively equal expression of plus and minus genes. These observations raised an interesting question: Are histone modifications directional? In other words, is there an association between the abundance of histone modifications and the direction where the gene is transcribed. This chapter is devoted to address this question, using diverse genetic and epigenetic data, from RNA-seq to ChIP-seq.

4.2 Methods

4.2.1 Primary cell types and cell lines

Epigenomic data for the following primary cell types and cell lines have been produced by the DEEP consortium and are available on the DEEP data portal[1] as well as the IHEC data grid[2], including metadata information with further details on sample sorting and preprocessing: hepatocytes (41_Hf03_LiHe_Ct), monocytes (43_Hm03_BlMo_Ct), macrophages (43_Hm05_BlMa_Ct), central memory T-cells (51_Hf01_BlCM_Ct), HepG2 cell line (01_HepG2_LiHG_Ct1).

The rest of the data, K562 (histone modification, transcription factor, CAGE expression), GM12878 (histone modification, CAGE expression), and HepG2 (transcription factor, CAGE expression) were obtained from the ENCODE portal.

[1] http://deep.dkfz.de/#/home

[2] http://epigenomesportal.ca/ihec/grid.html

4.2.2 Mapping of ChIP-seq data

Reads were mapped to the 1000 genomes phase 2 assembly of the human reference genome (NCBI build 37.1,[3]) with a hardware-accelerated implementation of Burrows-Wheeler Aligner BWA aln version 0.6.2 (Liu et al., 2012) with -q 20, and BWA 0.6.2 sampe with -a 1000. Merging and duplicate marking was performed with Picard version 1.125 ([4]). Laura Arrigoni and sarah Kinkley generated the ChIP-seq data. Bäarbel Felder, Gideon Giacomelli, Karl Nordström, Peter Ebert, Andreas S. Richter, Barbara Hutter, Benedikt Brors, Jürgen Eils performed mapping and management of sequencing data.

4.2.3 Mapping of RNA data

BAM files of RNA-seq reads were produced with TopHat 2.0.11 (Kim et al., 2013), with Bowtie 2.2.1 (Langmead and Salzberg, 2012) and NCBI build 37.1 in –library-type fr-firststrand and –b2-very-sensitive setting. Cufflinks was used for gene expression computation (Trapnell et al., 2012) using GENCODE release 19 (GRCh37.p13).

4.2.4 GRO-cap and CAGE expression estimation

We downloaded GRO-cap data for K562 and GM12878 cell lines from the GEO under accession GSE60456 provided by Core et al., 2014b. The count of the reads overlapping with a window in the region [0,+100] base pairs downstream of a gene TSS is used to define GRO-cap or CAGE derived expression.

4.2.5 Unsupervised clustering of histone data

To group the HM signals heuristically, we incorporated the kmeans clustering implemented in R. The best number of clusters was chosen based on the best average silhouette value for different number of clusters ranging from 2 to 10.

4.2.6 Construction of histone features

For each BP, we remove the region between the TSSs of BP gene pairs, from the estimation of the HM signals. This allows us to place the anchor for our binning method on a single site while capturing the spatial information stored on both sides of the bidirectional TSSs. To perform the binning, we define a window of size 4000 bp centered at the aforementioned locus. This window is partitioned into 40 bins each of size 100 bp. In the end, all the binned regions corresponding to six HMs are joined together forming a long vector of size 240 (6×40) elements.

4.2.7 Regression model learning

We incorporated the *fusedlasso* function from the genlasso R package (Arnold and Tibshirani, 2014) and added an intercept to the model to account for non-standardized data. We use fused Lasso to model dependencies between adjacent genomic bins in our regression. Analysis with the normal Lasso are done with the glmnet R package.

[3] `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/`

[4] `http://broadinstitute.github.io/picard`

As described in 2.2, the objective function for fused LASSO is as follows:

$$L(\beta_0, \beta) = \min_{\beta} \sum_{i=1}^{N} \epsilon_i^2 + \lambda \sum_{(i,j) \in E} |\beta_i - \beta_j| + \gamma \lambda \sum_{i=1}^{p} |\beta_i| + \beta_0, \tag{4.1}$$

where E denotes the edges of a given graph $G = (V, E)$ defined over the feature variables as its node set (V). Precisely, V describes the genomic bins,

$$V = \{B_i^j\} \quad i \in \{1, 2, \cdots, 40\};\ j \in \{1, 2, \cdots, 6\}. \tag{4.2}$$

On the other hand, E defines the connection between the bins.

$$E = \{e^j\} \quad j \in \{1, 2, \cdots, 6\}, \tag{4.3}$$

where

$$\{e^j\} = \{B_i^j, B_{i+1}^j\} \quad i \in \{1, 2, \cdots, 19, 21, \cdots, 39\}. \tag{4.4}$$

In other words, E represents the edges that connect the adjacent bins except for the border where there is the gap between the TSS (no edge between B_{20}^j and B_{21}^j). In addition, the histone modifications should not be connected together, meaning, the last bin of HM_1 should be disjoint from the first bin of HM_2 and so on. Therefore the transition between the concatenated HMs must be interrupted as well.

Next, we partitioned the data (1242 genes and 240 bins in total) into training (80%) and test (20%) sets and trained the fused LASSO models using 5-fold cross validation on the log2-transformed training data with a multi-layer search grid initially defined between 0 and 10^5 with exponential step size to seek an optimal trade off between sparsity and fusion. We performed our method on the biological samples processed through the same data preparation pipeline and investigated the final model. The R scripts for implementing the learning setup are accessible via the github link `http://github.com/fba67/fusedlasso`.

4.2.8 Generating model coefficient heatmap

We performed F-test statistics to assess the significance of the upstream or downstream coefficients obtained from the fused LASSO models. For this purpose, for each data set, we removed the up/down-stream coefficients of each HM and measured the changes in average RSS compared with respect to the full model (keeping all the coefficients in). If removing the up/down-stream coefficients was not significant with a significance cutoff of 0.05 we discarded those coefficients (setting their values to zero). The compilation of all the significant coefficients for each HM and each data set yields the heat map illustrated in Figure 4.8.

4.3 Results

The aim of this study is to integrate histone modification signals and the genomic structural properties of BPs. For this purpose we collected a data set of 1,242 BPs, which are divergent promoters with two core promoter elements, obtained from annotated ENSEMBL genes (GRCh37.75), such that the distance between TSSs of each BP does not exceed 500 bp. We removed cases where their TSS had an overlap with any other annotated gene region (±2kb from TSS). Figure 4.2 schematically illustrates two distinct genomic regions in where they both satisfy the TSS distance but one violates the overlapping constraint.

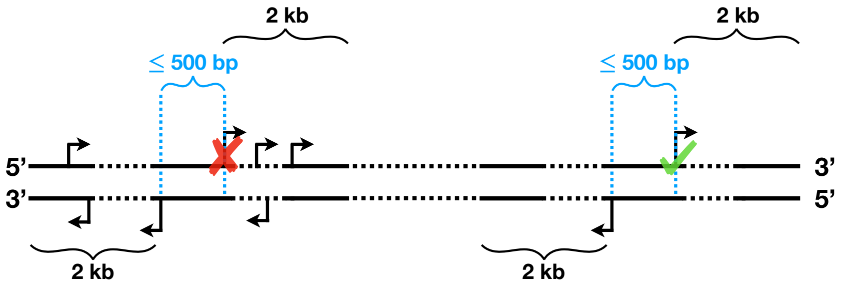


FIGURE 4.2: Constraints necessary to be satisfied for definition of bidirectional genes in this study. The distance between the two TSSs located on opposing DNA strands should not exceed 500 base pairs. In addition, there should not be any other annotated gene in a 2 kb window starting from each TSS of a potential BP (pair of genes that have met the first requirement).

The histone modification reads are measured in bins of 100 bp spanning a 4 kb window anchored on the TSS of bidirectional genes. Figure 4.3a depicts this binning approach applied on N different HMs for a particular TSS. The resulting vectors representing the abundance of HM reads in bins are concatenated to form a larger vector of size $N \times M$. This vector holds the histone modification profile of a single gene. When the histone profiles for all genes, G, are stacked on top of each other, a matrix, $H \in \mathbb{R}^{G\times(N\times M)}$ is constructed, which is used as the feature matrix for our statistical analysis (Figure 4.3b).

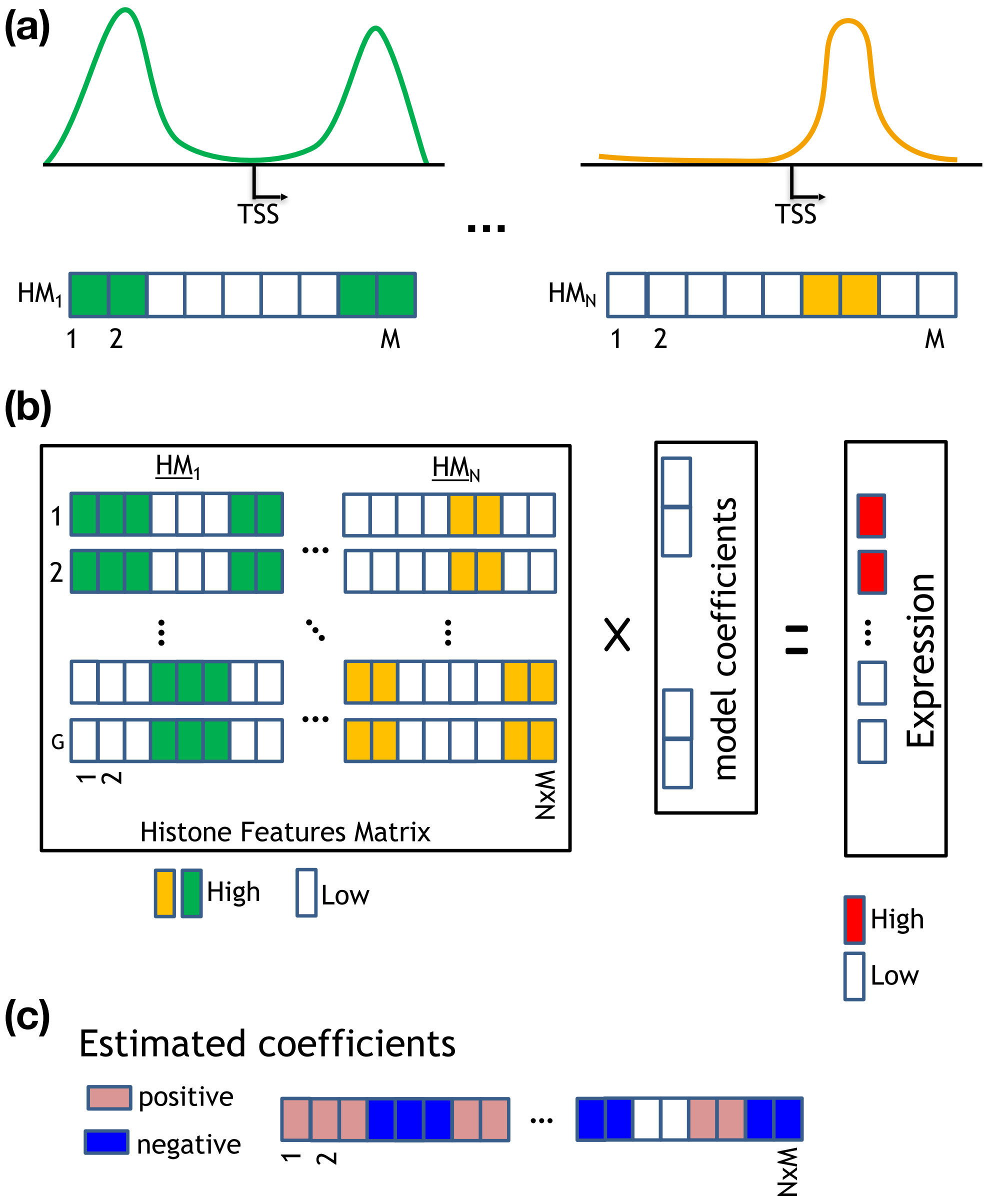


FIGURE 4.3: Statistical learning frame work. a) N distinct histone modifications signals measured around the TSS of a gene. The genomic region where the HM reads are estimated is partitioned into M bins. b) Feature matrix consisting of measurements of the N histones in M bins from (a) at the columns for G genes placed at the rows. This matrix can be used to setup a linear regression framework designed to regress the gene expression on histone modification data through estimating the model coefficients. c) Estimated coefficients can be used for interpreting how the features contribute to explaining the response.

4.3.1 Preliminary correlation analysis suggests a spatial association between HMs and gene expression in BPs

To investigate whether the positioning of histone modifications is associated to direction of gene expression, we first tested if the genomic bins holding HM abundance are correlated with the expression of either plus or minus genes. Figure 4.4

shows the Pearson correlation coefficient values computed between the bins of six histone modifications introduced in 4.2.1 and RNA-seq data for the HepG2 cell line. Each entry, $Cor_{i,j}$, in this heat map, Cor, represents the correlation coefficient between the expression of the plus gene and j^{th} bin of i^{th} histone modification. The results already hint on a directional association between the abundance of histone modifications and direction of transcription. In other words, since the correlation was measured for the plus gene, which is transcribed from 5′ to 3′ (i.e., left to right of the Cor matrix), the fact that the higher correlation values appear at the downstream of the TSS (location $Cor_{.,0}$) indicates such directional association. In mathematical terms, given the matrix H described above and gene expression vector, $Y \in \mathbb{R}^G$, the Pearson correlation matrix is derived as follows:

$$Cor_{i,j} = cor(H[,(i-1)*N+j], Y); \quad i \in \{1, \cdots, N\} \text{ and } j \in \{1, \cdots, M\}, \tag{4.5}$$

where $H[,i] \in \mathbb{R}^G$ is a columns vector denoting the i^{th} column of H.

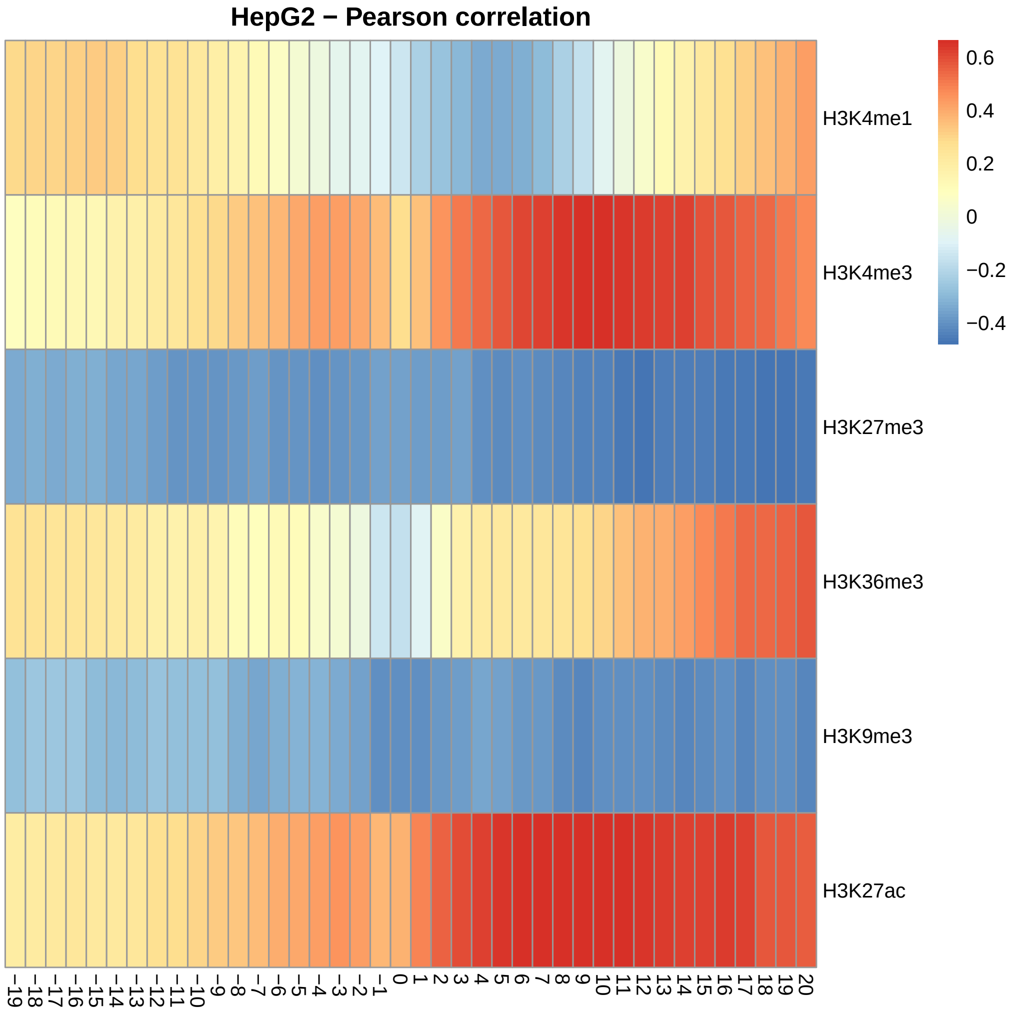


FIGURE 4.4: Heat map showing the Pearson correlation coefficients between bins of six histone modification data and bulk RNA-seq for cell line HepG2. Both HM and RNA-seq data are log-transformed. Each cell of the heat map corresponds to the bins described in 4.3.

Observing the non-zero correlation values upstream of the plus TSS, motivated us to extend correlation coefficients to partial correlation, where the relationship between two variables is conditioned on the third one. Let $PCor(X,Y,Z)$ be the partial correlation between variables X and Y, given Z. Here, we consider X being the logarithm of sum of downstream HM bins (including TSS), Y the vector of logarithm of gene expression measurements, and Z the logarithm of sum of upstream bins (excluding TSS). Thus, $PCor(X,Y,Z)$ reflects the correlation between the sum of downstream HM bins and gene expression, while factoring out the influence of upstream HMs.

We calculated the partial correlation in such described way on the HepG2 cell line, separately for each histone modification. We also were interested to see how the values compare when $PCor(Z,Y,X)$ is computed. The partial correlation values for the six histone modifications are provided in Table 4.1. First of all, it can be seen that the partial correlation values for the direction of interest, $PCor(X,Y,Z)$, is higher than $PCor(Z,Y,X)$. This further supports the fact that downstream bins are predictive of expression for the plus gene.

Second, it can be noted that for the two promoter initiation associated marks, H3K4me3 and H3K27ac, the partial correlation results in a higher value than the correlation coefficient obtained from aggregating the entire considered genomic region, by summing the upstream and downstream HM bins, shown in the third column of Table 4.1. This means that removing the upstream HM data can, as a matter of fact, improve the relation between HMs and gene expression. The repression associated marks, on the other hand, seem to benefit from having the upstream HM bins in the calculations, as the correlation coefficient for the H3K27me3 and H3K9me3 are higher than their partial correlation coefficients conditioned on the upstream HMs. In other words, the abundance of these histone marks upstream of the TSS is positively associated to gene expression. In addition to the repressive marks, the elongation mark, H3K36me3, also has a higher correlation coefficient, but the difference is much slighter than the two aforementioned repressive marks.

As shown in Figure 4.4, the H3K4me1 mark exhibits a peculiar behavior. The first ~10 bins downstream of the TSS show a negative correlation coefficient with gene expression, while most of the upstream bins are positively correlated. This observation is, to some extent, reflected in our partial correlation analysis as well. It can be seen that the partial correlation conditioned on the upstream HM bins holds a very small negative value, whereas when conditioned on the downstream HM bins the value rises up to ~ 0.3.

HM	$PCor(X,Y,Z)$	$PCor(Z,Y,X)$	$Cor(X+Z,Y)$
H3K4me1	-0.058	0.289	0.240
H3K4me3	0.565	-0.167	0.544
H3K27me3	-0.330	-0.014	-0.519
H3K36me3	0.450	0.059	0.481
H3K9me3	-0.380	0.028	-0.504
H3K27ac	0.594	-0.187	0.561

TABLE 4.1: Partial correlation coefficients obtained for the HepG2 cell line, individually for the six histone modifications, conditioned on the upstream HM. X and Z denote the logarithm of sum of downstream and upstream HM bins, respectively. $expr.$ denotes gene expression data for the HepG2 cell line. $Cor(X+Z,Y)$ denotes the unconditional correlation coefficient between the sum of upstream and downstream HM bins $(X+Z)$ and gene expression.

4.3.2 Fused LASSO: a promising tool to investigate spatial dependencies in BPs

The correlation analysis conducted in section 4.3.1 was suggestive of spatial associations between the gene expression and positioning of the histone modifications around the TSS in bidirectional promoters. However, this is not sufficient for building an integrative framework that incorporates the combinatorial role of HMs in driving expression, as the analyses were conducted on each HM separately. In order to cope with such limitation, as well as, having a method that is able to predict the gene expression from histone modification data, we designed a predictive statistical framework. As described in 4.2.7, fused LASSO is a linear regression model that aims to explore the feature space while preserving the spatial dependencies between the adjacent features as given in the underlying graph.

To better understand how the feature selection in fused LASSO works, we designed several simulation studies that can bring a deeper insight into its optimization procedure.

Simulation #1

The first experiment is as follows. We simulated data for two imaginary histone modifications, called A and B, each having 20 bins for 600 genes. This forms the feature data, which is a matrix of 600 rows and 40 columns, $X \in \mathbb{R}^{600\times40}$. Similarly, the response vector for this study has 600 elements. These data are illustrated in figures 4.5a and 4.5b. The underlying graph required for fused LASSO optimization is depicted in Figure 4.5c, where the edges are connecting adjacent bins of A or B. We then computed the optimal coefficients, using 80% of the data for training, with optimal $\gamma = 0.08$ as provided in Figure 4.5d. Note that the coefficient vector is split into two rows, each corresponding to the features for A and B. The results on the test data (remaining 20%) achieves RMSE of zero and consequently, correlation of 1 between the measured and predicted measurements.

To understand the coefficient selection for this particular given data, one needs to examine the patterns in X and Y simultaneously. For instance, it can be seen that in the coefficient heat map illustrated in Figure 4.5d, the first 20 features are all set zero. When comparing this with the data shown in figures 4.5a,b, it becomes clear that those features are not informative in predicting the response, i.e., histone modification A is not relevant in predicting expression when B is given. Therefore, setting values other than zero would only increase the model complexity with no improvement in the performance. However the remaining 20 features (corresponding to B) are the most interesting ones. In the first 300 samples (Figure 4.5a,b), the features 21 to 30, hold small values and their response value is very high, in the last 300 samples, on the other hand, an opposite trend is apparent, the same features hold very high values where their response is extremely low. Moving on to the features 31 to 40, it can be noted that the model assigned positive values to those coefficients in order to contribute to predicting the high values of expression when data from the first 300 rows are given, and assigned negative values to the features 21 to 30 to account for the small response values for the data obtained from the second 300 genes.

Taken all together, this experiment suggests that using fused LASSO can be advantageous for obtaining interpretable results for associating spatial information embedded in the features to the response values.

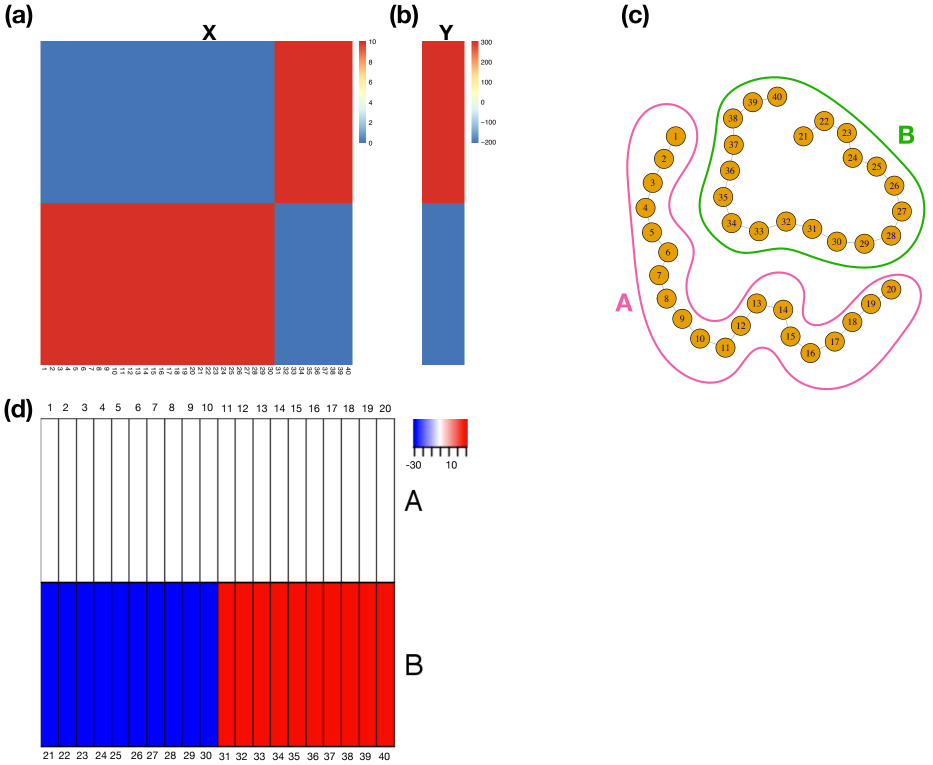


FIGURE 4.5: Simulation study #1. a) Heat map illustrating the simulated features of 600 rows and 40 columns. The heat map shows two equally sized subpopulations, with mutual exclusive patterns. b) Response variable for the simulated data given in (a). c) Underlying graph required for fused LASSO's optimization function, connecting the adjacent bins of each HM. Note that there is no edge between A and B. d) Heat map separating the coefficient of each HM into two rows, thus a 2×20 matrix. Blue and red colors correspond to negative and positive values, respectively.

Simulation #2

In our second simulation study, we generated data slightly different than in the first simulation. Figure 4.6 provides heat maps of the new X and Y variables, as well as the coefficient heat map obtained from a fused LASSO model trained with $\gamma = 0$. In order to understand the model, we need to precisely know how the response variable is generated. Let $q_1, \cdots, q_4$ represent the four consecutive blocks of ten features. In other words, $q_1 = \{1, \cdots, 10\}$, $q_2 = \{11, \cdots, 20\}$, $q_3 = \{21, \cdots, 30\}$, and $q_4 = \{31, \cdots, 40\}$. We define Y as the following:

$$Y_i = 3 \sum_{j \in q_1 \bigcup q_4} X[i,j] + 4 \quad \text{for } i \in \{1, \cdots, 300\},$$
$$Y_i = -3 \sum_{j \in q_2 \bigcup q_3} X[i,j] + 4 \quad \text{for } i \in \{301, \cdots, 600\}.$$

So, the first 300 response values are a linear function of the first and fourth feature blocks in X and the last 300 values are negatively associated to the second and third feature blocks with the similar intensity as the first half (according to the multiplicative term $\times 3 + 4$). This already elucidates why the model picked the coefficients as shown in Figure 4.6c. All features seem to be relevant in explaining the response; the first and last blocks (first 20 features of A and last 20 features of B) are positively associated and the remaining features are negatively associated.

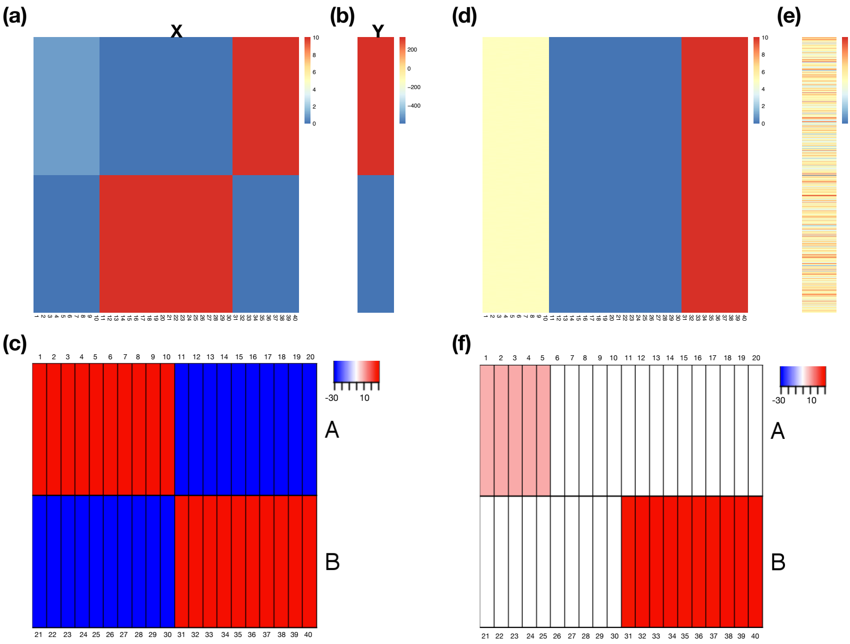


FIGURE 4.6: Simulation studies #2 and #3. a, d) Heat maps illustrating the simulated features of 600 rows and 40 columns for simulations #2 and #3, respectively. b, e) Response variables for the data given in (a) and (d) for simulations #2 and #3, respectively. c, f) The heat maps of fused LASSO coefficients reflecting the informative regions by fusing the features together for simulations #2 and #3, respectively.

Simulation #3

We further investigated the fused LASSO optimization through designing the third experiment by keeping the distribution of data points more homogeneous. As shown in Figure 4.6d, the features consist of three distinct patterns that are consistent across the 600 samples. The corresponding response variable Y (Figure 4.6e) is generated based on the features X via the following equation:

$$Y_i = 2\sum_{j=1}^{5} X[i,j] + 3\sum_{j\in q_4} X[i,j] + 4 + \epsilon,$$

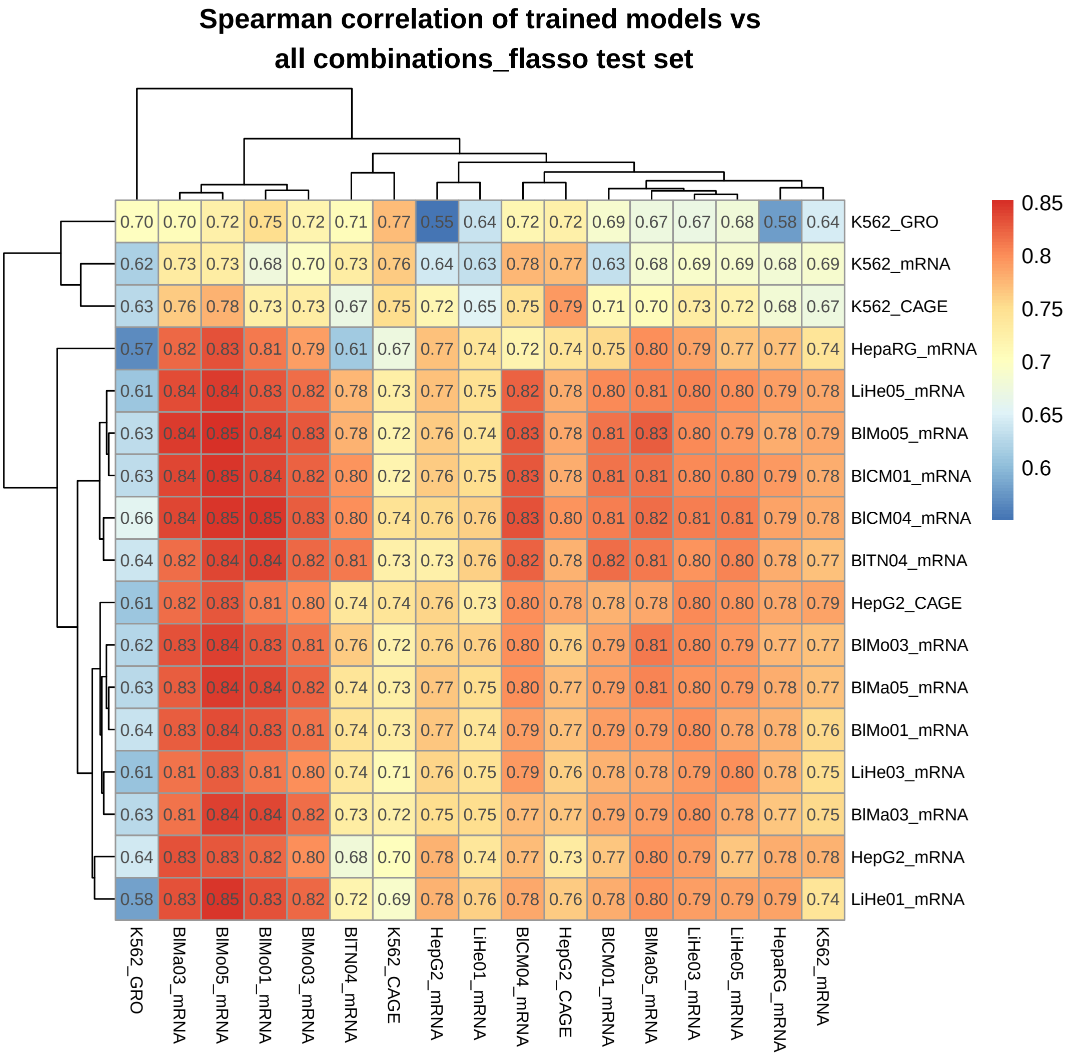
where ϵ follows a Gaussian distribution with 0 mean and 0.01 standard deviation denoting the additive noise. The variations in the colors are due to the slight variation of the response values that range only from 353.6 to 354.3. Because the response holds positive large values, the model decides to discard the middle block $(11, \cdots, 30)$ in the feature space as they possess smallest values. In addition, according to the formula above, these features do not contribute in generating the response variable Y, while the remaining features are included, with different magnitudes. Again, referring to the formula used to generate Y, the first 5 features are less impactful compared to the last 10 features (q_4), since their sum is multiplied by 2 and

3, respectively. This characteristic is remarkably reflected in the fused LASSO coefficient as well.

With the deeper insight gained from the above mentioned simulation experiments, we were convinced to carry out the task of predicting gene expression from histone modification in bidirectional genes using the fused LASSO models.

4.3.3 Fused LASSO suggests a unidirectional histone code in BPs

We trained individual fused LASSO models for each data set described in 4.2.1. For the response variable, in addition to RNA-seq measurement that we had for all data sets, we used CAGE for HepG2 and K562 and GRO-cap for K562 only. The performance of each model tested on their own allocated test data as well as other models (cross-comparison) is shown in Figure 4.7. It can be seen that, in general, the GRO-cap models do not generalize well on other data sets. This can be due to the difference in sensitivity of the sequencing assays.



	K562_GRO	BlMa03_mRNA	BlMo05_mRNA	BlMo01_mRNA	BlMo03_mRNA	BlTN04_mRNA	K562_CAGE	HepG2_mRNA	LiHe01_mRNA	BlCM04_mRNA	HepG2_CAGE	BlCM01_mRNA	BlMa05_mRNA	LiHe03_mRNA	LiHe05_mRNA	HepaRG_mRNA	K562_mRNA
K562_GRO	0.70	0.70	0.72	0.75	0.72	0.71	0.77	0.55	0.64	0.72	0.72	0.69	0.67	0.67	0.68	0.58	0.64
K562_mRNA	0.62	0.73	0.73	0.68	0.70	0.73	0.76	0.64	0.63	0.78	0.77	0.63	0.68	0.69	0.69	0.68	0.69
K562_CAGE	0.63	0.76	0.78	0.73	0.73	0.67	0.75	0.72	0.65	0.75	0.79	0.71	0.70	0.73	0.72	0.68	0.67
HepaRG_mRNA	0.57	0.82	0.83	0.81	0.79	0.61	0.67	0.77	0.74	0.72	0.74	0.75	0.80	0.79	0.77	0.77	0.74
LiHe05_mRNA	0.61	0.84	0.84	0.83	0.82	0.78	0.73	0.77	0.75	0.82	0.78	0.80	0.81	0.80	0.80	0.79	0.78
BlMo05_mRNA	0.63	0.84	0.85	0.84	0.83	0.78	0.72	0.76	0.74	0.83	0.78	0.81	0.83	0.80	0.79	0.78	0.79
BlCM01_mRNA	0.63	0.84	0.85	0.84	0.82	0.80	0.72	0.76	0.75	0.83	0.78	0.81	0.81	0.80	0.80	0.79	0.78
BlCM04_mRNA	0.66	0.84	0.85	0.85	0.83	0.80	0.74	0.76	0.76	0.83	0.80	0.81	0.82	0.81	0.81	0.79	0.78
BlTN04_mRNA	0.64	0.82	0.84	0.84	0.82	0.81	0.73	0.73	0.76	0.82	0.78	0.82	0.81	0.80	0.80	0.78	0.77
HepG2_CAGE	0.61	0.82	0.83	0.81	0.80	0.74	0.74	0.76	0.73	0.80	0.78	0.78	0.78	0.80	0.80	0.78	0.79
BlMo03_mRNA	0.62	0.83	0.84	0.83	0.81	0.76	0.72	0.76	0.76	0.80	0.76	0.79	0.81	0.80	0.79	0.77	0.77
BlMa05_mRNA	0.63	0.83	0.84	0.84	0.82	0.74	0.73	0.77	0.75	0.80	0.77	0.79	0.81	0.80	0.79	0.78	0.77
BlMo01_mRNA	0.64	0.83	0.84	0.83	0.81	0.74	0.73	0.77	0.74	0.79	0.77	0.79	0.79	0.80	0.78	0.78	0.76
LiHe03_mRNA	0.61	0.81	0.83	0.81	0.80	0.74	0.71	0.76	0.75	0.79	0.76	0.78	0.78	0.79	0.80	0.78	0.75
BlMa03_mRNA	0.63	0.81	0.84	0.84	0.82	0.73	0.72	0.75	0.75	0.77	0.77	0.79	0.79	0.80	0.78	0.77	0.75
HepG2_mRNA	0.64	0.83	0.83	0.82	0.80	0.68	0.70	0.78	0.74	0.77	0.73	0.77	0.80	0.79	0.77	0.78	0.78
LiHe01_mRNA	0.58	0.83	0.85	0.83	0.82	0.72	0.69	0.78	0.76	0.78	0.76	0.78	0.80	0.79	0.79	0.79	0.74

FIGURE 4.7: Model performance assessed in terms of Spearman correlation between predicted and measured response on test data. The rows of the heat map indicate the data the model was trained on and the columns reflect the test data the model was applied on. In other words, each entry in the heat map shows the Spearman correlation for the model trained on the data shown on the row when applied the test data shown on the column. For instance, the correlation between the predicted and measured mRNA expression of K562 cell line when the model was trained on the GRO-cap data is 0.64 (right top most corner of the heat map).

The feature heat maps of each model are compiled into one, embracing the coefficients for all models, as shown in Figure 4.8. What is striking in this heat map is the apparent preference of the model to select and assign higher values to coefficients downstream of the plus TSS; the region where the gene is transcribed. By inspecting the localization of each histone modification, one can see that, in general, the marks associated with gene activation hold positive values and marks associated with gene repression have negative values, except for H3K4me1. The results suggest a mutual exclusive relationship between the H3K4me1 and H3K4me3 marks. This can be justified by the fact that having three methyl groups on H3K4, masks the existence of one methyl group on the same amino acid. Given that all histone modifications are fed to the model as one feature vector, the model has been able to infer such mutual exclusivity.

Comparing the coefficients of H3K4me3 mark with respect to the sequencing protocol used, it becomes apparent that the model restricted the selection of coefficients to only a few 100 bp downstream of TSS, whereas for CAGE and RNA-seq this was stretched further to regions more distant from TSS.

Another interesting observation is the selection of coefficients for the H3K36me3 mark. As previously mentioned, this mark is associated with gene elongation and it often peaks more towards the gene termination site. This behavior is remarkably reflected in the histone map shown in Figure 4.8 for the RNA-seq models. The GRO-cap model, as expected, did not focus on this mark, as the corresponding response, which is the estimated nascent transcription, is independent of this mark. In other words, the nascent transcripts are mainly prevalent at the vicinity of the transcription start site rather than being distributed across the entire region downstream of the TSS. The CAGE model, however, spans the entire region, starting from TSS to all the way down to the last bin downstream. This can be due to the fact that the CAGE data represents the stable transcripts and therefore, both beginning and end of the transcribed region are included in the model.

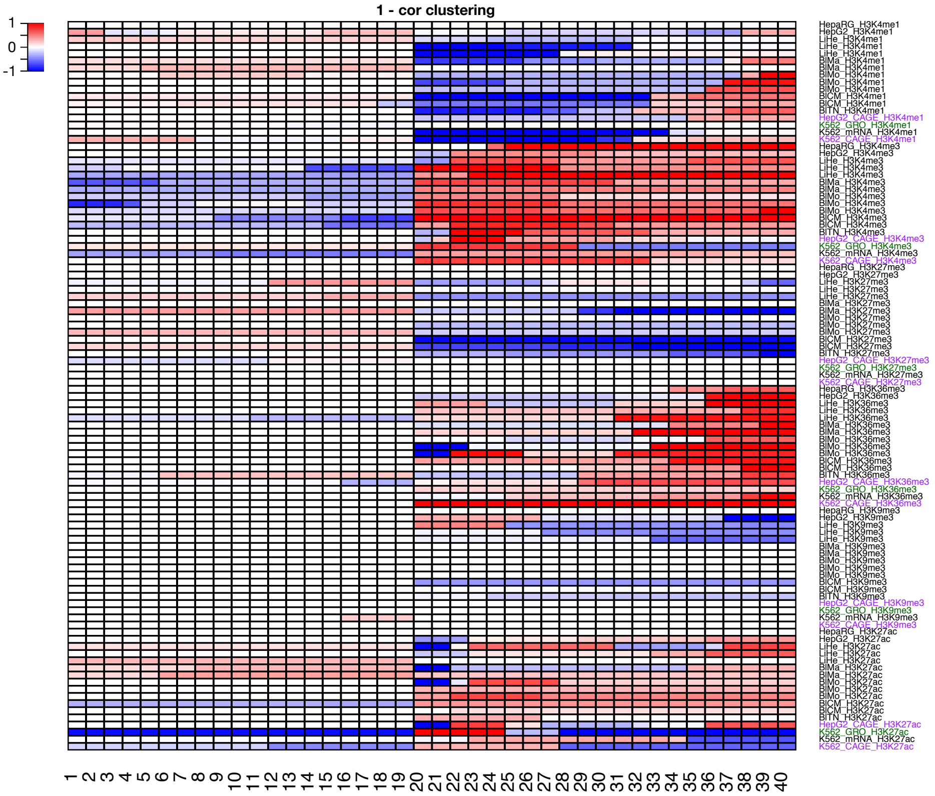


FIGURE 4.8: The histone map. Fused LASSO coefficients obtained from individual models learnt on the cell types described in 4.2.1 for the plus gene. The values are scaled between −1 (blue) and 1 (red) to ease the comparison across samples. The expression assays, RNA-seq, CAGE, and GRO-cap, are color coded by black, purple, and green, respectively. The heat map suggests a unidirectional localization of histone marks coinciding the direction of transcription.

Furthermore, we carried out the same prediction task using standard LASSO (only $|L_1|$ regularization). The heat map, illustrated in Figure 4.9, still suggests a unidirectional localization of histone marks coinciding the direction of transcription, but the sole sparsity regularization results in a scatter and thus unintuitive interpretation.

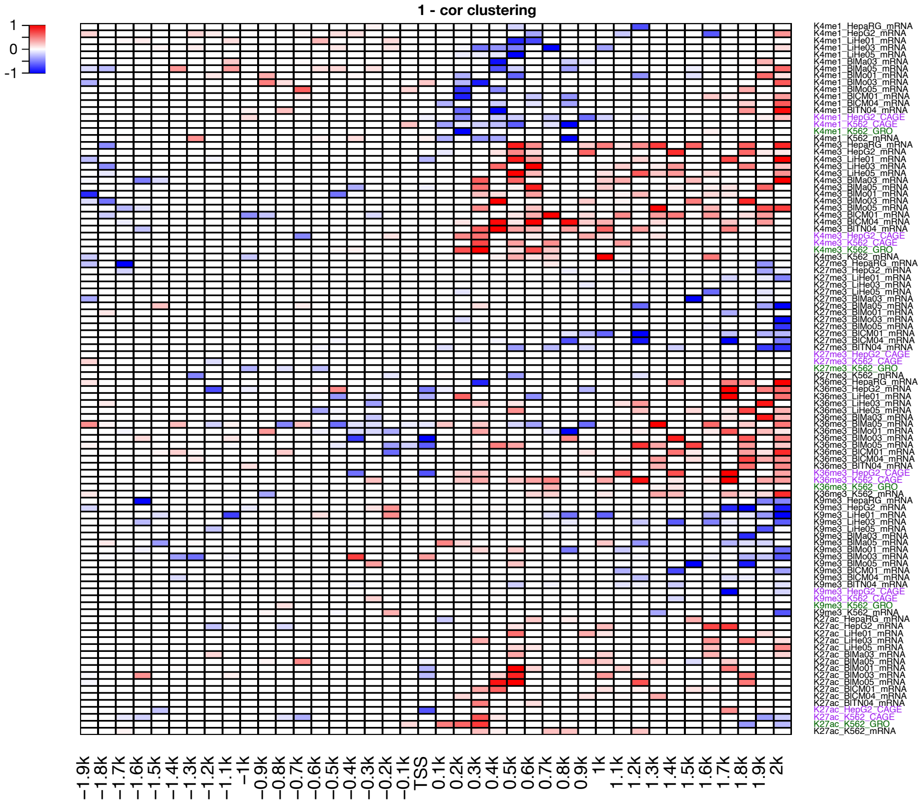


FIGURE 4.9: Standard LASSO coefficients obtained in the similar fashion as fused LASSO. The values are scaled between −1 (blue) and 1 (red) to ease the comparison across samples. The expression assays, RNA-seq, CAGE, and GRO-cap, are color coded by black, purple, and green, respectively.

We also were able to observe similar results for the expression of the minus gene as illustrated in Appendix Figure B.2.

4.3.4 Using a non-linear model results in similar performance as the linear fused LASSO

To assess how much of accuracy was sacrificed in favor of interpretability by using a linear model, we performed the same prediction task using support vector regression (SVR) with radial basis kernel. Figure 4.10 contrasts the performance of fused LASSO against SVR models. It can be observed that only for 7 out of 17 samples, SVR models outperforms the fused LASSO and except for the *K562_CAGE* sample, the differences are very marginal.

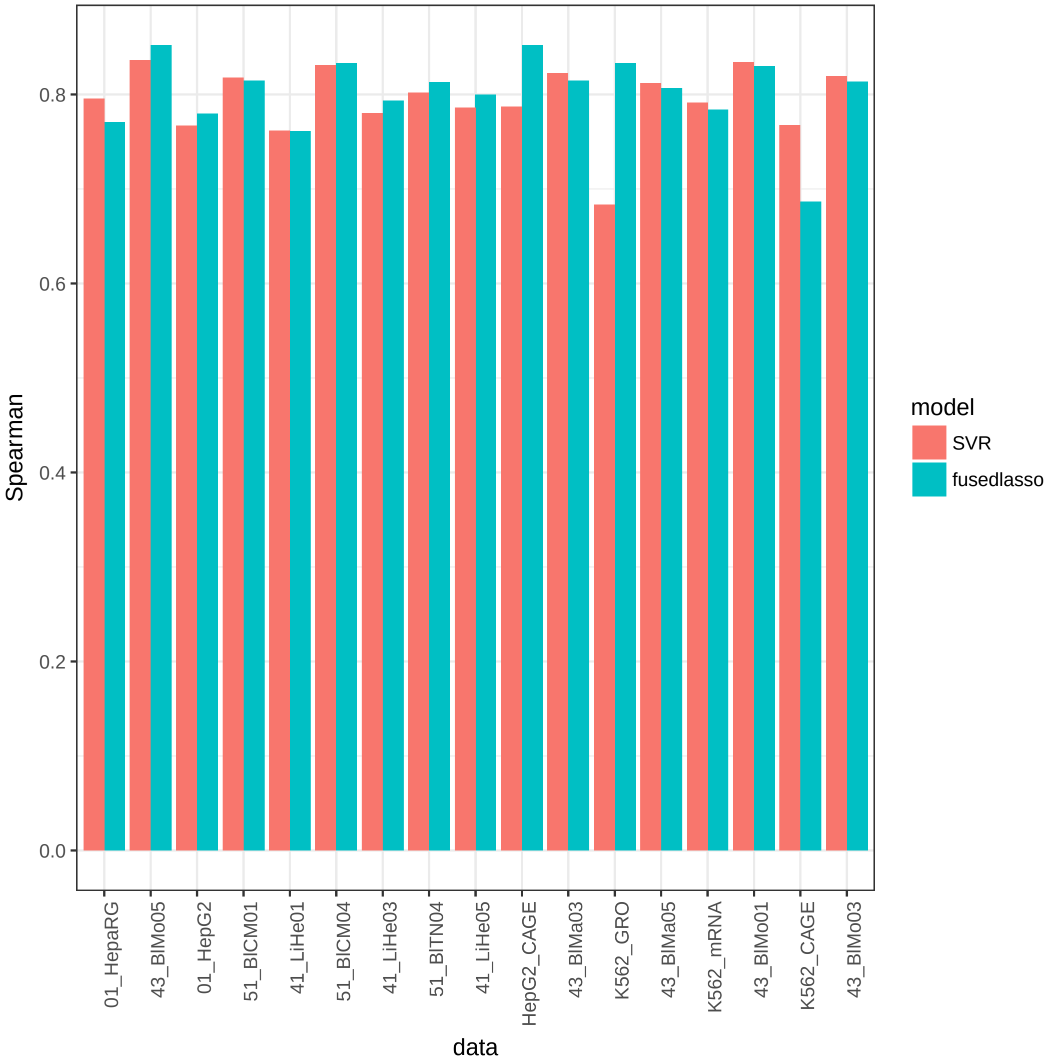


FIGURE 4.10: Comparison of performance, reported in terms of Spearman correlation, for linear (fused LASSO) and non-linear (SVR) models. Only in 7 samples out of 17, SVR outperforms the fused LASSO.

The peculiar difference in the performance of *K*562_*GRO* between the two types of models, fused LASSO and SVR, in Figure 4.10 inspired us to inspect the data points (genes) for which the prediction error is high. For this purpose, we computed the difference between predicted and measured GRO-cap values for the K562 cell line. We refer to cases where this difference is higher than 2 as outliers. By comparing the distribution of histone modification values between outlier and non-outlier samples, as illustrated in Figure 4.11, it can be noted that the outlier ones have higher values with smaller variance. Given that the SVR models were trained using the radial basis kernel function might have led to over-training caused by this additional layer of complexity (non-linear kernel).

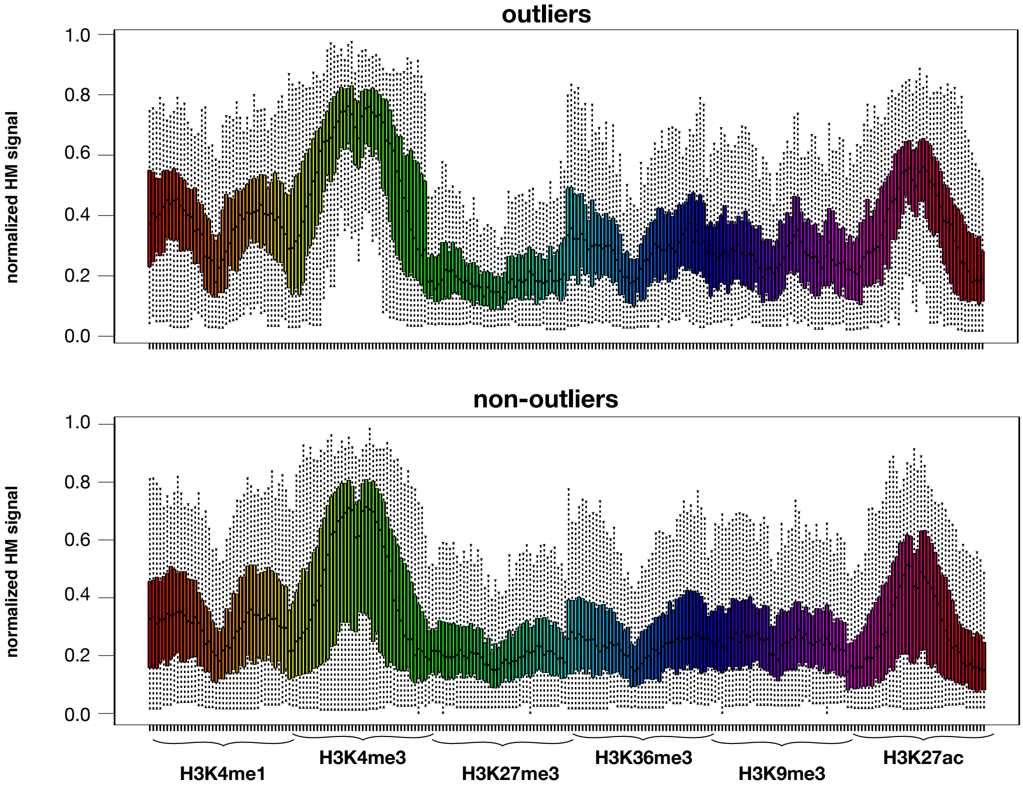


FIGURE 4.11: Comparison of histone modification signals among detected outlier and non-outlier genes.

4.4 Conclusion and discussion

We observed that the average ChIP-seq signal for the H3K4me3 histone modifications shows a bimodal distribution along the bidirectional promoters (Figure 4.1). This observation poses the question whether the expression regulation of one gene depends on the regulation of the other gene at a BP, therefore the bimodal pattern?

To address this question, we first inspected the correlations between the gene expression and genomic bins spread around the transcription start site of bidirectional genes, where we counted the HM ChIP-seq reads overlapping with those bins. As depicted in figures 4.4 and B.1 these correlation coefficients suggest that there exist strong associations between the abundance of HMs and direction of transcription.

The weaker correlations observed mostly at the upstream bins relative to TSS, inspired us to compute the partial correlation between gene expression and sum of HM abundance in upstream bins, when conditioned on sum of downstream bins, or vice versa. As presented in Table 4.1, the abundance of HMs at the upstream region plays an insignificant role in deriving the gene expression.

The correlation studies certainly are helpful in bringing insights into the data and the associations between variables. However, not only they are incapable of deriving such associations in an integrated manner, but also they have limited abilities in predicting the response variable for a given set of features. To overcome these limitations, we additionally, implemented a framework for associating HM ChIP-seq reads along a BP with gene expression. The purpose of this framework was to evaluate the performance of gene expression prediction and to understand at what location, relative to both TSSs, a histone modification is associated with gene expression. This allows us to estimate if HM abundance at the minus gene (upstream of the plus gene) depends on the expression of the plus gene. As opposed to previous studies, we included both TSSs in the model by partitioning the region +/- 2 kb for both TSSs in non-overlapping bins of size 100 bp.

We further investigated the transcriptional dependence of both TSSs by developing a general histone association map at BPs from learning associations of HMs, in various cell types, with gene expression measured by different assays (RNA-seq, CAGE, and GRO-cap), as shown in Figure 4.8. The GRO-cap models were trained on a data that was very narrow and sensitive around the TSS, where nascent transcripts were originated from, however the RNA-seq and CAGE models were more tailored to data with more stable transcripts. The heat map in Figure 4.8 demonstrates associations in direction of transcription in BPs, even though the overall average ChIP-seq signal of HMs shows a bimodal distribution as previously observed (Bornelöv, Komorowski, and Wadelius, 2015). Therefore, supporting the view that the chromatin modifying machinery is recruited to both sites at the BP in an independent manner.

Next, we found that the GRO-cap models differed in their selection of positive coefficients, neglecting the late-elongation mark H3K36me3, which was one of the strongest contributors in the RNA-seq and CAGE models. Although GRO-cap measures $5'$-capped nascent RNAs at the TSS, possibly including RNAs that are elongated over hundreds of bps, our results suggest that there is a strong enrichment for short, nascent TSS RNAs that are likely regulated at the pausing step of Pol2 (Henriques et al., 2013). Moreover, H3K27ac showed more importance in the GRO-cap models and thus probably involved in regulation of promoter-proximal pausing. This is in agreement with a previous observation of Chen et al., 2011 that showed strong positive association of H3K27ac enrichment downstream of the TSS for predicting stalled compared to elongating Pol2 signals.

In order to assess how much of performance was sacrificed in favor of interpretability, we trained non-linear regression models using the support vector regression method with radial basis kernel functions. By comparing the model performances in terms of the Spearman correlation between predicted and measured gene expression values, as illustrated in Figure 4.10, it can be concluded that the linear fused LASSO models were truly competing with the non-linear SVR models. Except seven out of 17 biological samples, the fused LASSO models were able to outperform the SVR models.

Taken together, we suggest that the histone modifications are, to the most part, assembled at the genomic region corresponding to the direction of transcription, even though the average HM ChIP-seq data exhibits bimodal patterns at the vicinity of BPs.

Chapter 5

Integrative analysis of single-cell expression data in bidirectional promoters

The work that is presented in this chapter has been published in the journal of Epigenetics & Chromatin in 2018 (Behjati Ardakani et al., 2018). This paper was selected as one of the top 10 "Reading Papers" in 2018 by the RECOMB/ISCB Regulatory systems Genomics group[1].

5.1 Introduction

As previously mentioned in Chapter 4, the regulation of bidirectional genes has not yet fully understood and several studies have attempted to address the question of how this regulation is carried out (Core, Waterfall, and Lis, 2008; Preker et al., 2008; Seila et al., 2008; Core et al., 2014b; Duttke et al., 2015a; Scruggs et al., 2015; Bagchi and Iyer, 2016; Duttke et al., 2015a; Lacadie et al., 2016).

BPs have been shown to harbor overrepresented TF binding sites such as GABPA, MYC, YY1, NRF-1, E2F1 and E2F4 (Lin et al., 2007). For instance, the introduction of GABPA binding sites into unidirectional promoters leads to bidirectional expression in 67% of the cases (Collins et al., 2007). Furthermore, the sequence elements at some BPs operate as indivisible units (Trinklein et al., 2004). Other TFs, however, avert bidirectional expression, for example, promoters that show elongation in only one direction often exhibit enrichment of CTCF binding sites (Core et al., 2014b; Bornelöv, Komorowski, and Wadelius, 2015). Nonetheless, more research is essential to investigate how TF binding determines directionality of initiation and elongation at BPs (Bagchi and Iyer, 2016).

It was recently shown that bidirectional promoters define a Nucleosome Free Region (NFR) between the two Transcription Start Sites (TSSs). The length of this region (NFR) might be an important structural element in the regulation of BPs, determining the accessibility of binding sites for various TFs at the promoter. This can, in turn, influence the intensity of gene expression as well as responsiveness to external stimuli (Duttke et al., 2015a; Scruggs et al., 2015). Recent studies (Core et al., 2014b; Duttke et al., 2015a; Scruggs et al., 2015) have pointed to a model, where an independent Pol2 complex assembles at each TSS and initiates transcription, such that accurate phasing of the +1 and -1 nucleosomes at these BPs allows epigenetic regulation through HMs. Comparisons between unidirectional and bidirectional promoters indicate that HMs associated with active gene expression exhibit

[1] https://www.iscb.org/recomb-regsysgen2019-submissions/recomb-regsysgen2019-reading

a bimodal distribution at BPs, and that upstream proximal enhancer marks may regulate downstream gene transcription (Bornelöv, Komorowski, and Wadelius, 2015; Scruggs et al., 2015).

In summary, previous studies rely on the comparison of unidirectional against bidirectional promoters to comprehend BP regulation. In this study, we take a different approach, by utilizing recent advances in single cell sequencing and investigating expression of genes at BPs in individual cells, to better understand their regulation. Recent developments in single cell genomics allow the measurement of RNA expression in individual cells with a similar accuracy as compared to bulk-sequencing of RNAs (Marinov et al., 2013; Wu et al., 2014). This advance has been used to identify previously unnoticed cell types and heterogeneous expression patterns, *e.g.*, (Pollen et al., 2014).

To investigate the expression behavior of bidirectional genes, we take advantage of novel and previously produced single cell RNA-seq (scRNA-seq) data for HepG2 and K562 cells. We discover four reproducible expression classes in BPs. These results also show that in a majority of cases, one gene at a BP shows considerably higher expression than the other. We also find novel associations of distinct structural and epigenetic features in these classes, using high resolution histone modification data sets produced at IHEC standards (Stunnenberg, Hirst, and Consortium, 2016) by the DEEP consortium or made available by ENCODE (ENCODEConsortium, 2012).

5.2 Methods

5.2.1 Single cell transcript expression

The TPM values for transcript isoforms of each Ensembl gene (GRCh37) were calculated using RSEM (Li and Dewey, 2011). Since we wanted to attribute the transcription expression, as opposed to gene expression, to each bidirectional gene, we summed the isoform TPM values of transcripts that had their annotated TSS overlapping within a 2 kb window downstream of the most 5′ TSS of that gene.

5.2.2 Bidirectional promoter (BP) gene set

The BP data set contained 1,242 divergent promoters with two core promoter elements, obtained from annotated ENSEMBL genes (GRCh37.75), such that the distance between TSSs of each BP does not exceed 500 bp. This set excludes loci overlapped by any other annotated gene region (±2 kb from the TSS).

5.2.3 Clustering BPs into four states

Hierarchical clustering using complete linkage method with Euclidean distance as distance metric was applied on the swapped BP matrix using R.

5.2.4 Constructing the single cell TPM matrix for BPs

For a particular BP, $BP_i = (g_{crick,i}, g_{watson,i})$, we evaluate the sum of TPM values across single cells as following:

$$Sum(g_{j,i}) = \Sigma_{c=1}^{N} TPM(g_{j,i}^{c}), \tag{5.1}$$

where N denotes the number of single cells, and $TPM(g^c_{j,i})$ returns the TPM value for gene $j \in \{crick, watson\}$ of BP_i in cell c.

The orientation of genes at a BP is not specific to the DNA strand, but the lower expressed gene of a BP is always swapped to the left and higher expressed gene to the right. In this way, without loss of generality, all analyses correctly adjust for differences of expression. In detail, we define $g_{H,i}$ denoting the gene of BP_i holding higher expression as follows:

$$g_{H,i} = \begin{cases} g_{watson,i}, & \text{if } Sum(g_{watson,i}) \geq Sum(g_{crick,i}) \\ g_{crick,i}, & \text{else}\,. \end{cases} \tag{5.2}$$

Similarly, we define $g_{L,i}$ denoting the gene of BP_i having lower expression:

$$g_{L,i} = \begin{cases} g_{watson,i}, & \text{if } Sum(g_{watson,i}) < Sum(g_{crick,i}) \\ g_{crick,i}, & \text{else}\,. \end{cases} \tag{5.3}$$

After defining $g_{H,.}$ and $g_{L,.}$ for each BP, we form the single cell matrix for BPs, scBP, as follows:

$$scBP = \begin{bmatrix} g^1_{L,1} & g^2_{L,1} & g^N_{L,1} & g^1_{H,1} & g^2_{H,1} & \cdots & g^N_{H,1} \\ g^1_{L,2} & g^2_{L,2} & g^N_{L,2} & g^1_{H,2} & g^2_{H,2} & \cdots & g^N_{H,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ g^1_{L,M} & g^2_{L,M} & g^N_{L,M} & g^1_{H,M} & g^2_{H,M} & \cdots & g^N_{h,M} \end{bmatrix}$$

5.2.5 Imputation of dropouts

To address the bias induced by dropouts, we performed the most, at the time of this study, recent and accurate dropout imputation method called scImpute (Li and Li, 2018b), which works toward improving the quality of single-cell data by eliminating the effects of dropouts without introducing new biases to the data. This tool takes two user-defined parameters, K and t. K denotes the number of existing cell types in the data, which we set to 1, as we work on cell lines. The second parameter, t, controls the dropout probabilities. The authors show that their results are robust to different values of t, therefore, we set the default value of 0.5 for this parameter. The comparison between raw and imputed read counts performed on the bidirectional genes is shown in Appendix Figure C.1 for both HepG2 and K562 samples. The Pearson correlation between the two quantities in both cell lines is $\sim$1.

5.2.6 Quality of scRNA-seq

Imputed expression of bidirectional genes averaged over single cells was compared with their corresponding bulk RNA-seq expression. For both, HepG2 and K562, the single cell expression agrees well with bulk measurements (Spearman correlation coefficient of $\sim$0.8, Appendix Figure C.2). Additionally, the imputed TPM values were divided into three intervals, $1 < \text{TPM} < 10$, $10 \leq \text{TPM} \leq 100$, $\text{TPM} > 100$ to account for the number of genes falling in those intervals per cell (Appendix Figure C.3a, and similarly for the imputed read counts in C.3b).

5.2.7 Prediction of RNA stability from histone data

To examine the possible effects of post-transcriptional regulation on our transcription states, we adopted the approach proposed by (Wang et al., 2012), where, using different histone modification ChIP-seq data sets, they predict the expression level of RNA. Similar to their approach, we trained an ordinary least squares (OLS) model on our bidirectional genes to predict the average single cell RNA expression values from six histone modifications, H3K4me1, H3K4me3, H3K27me3, H3K36me3, H3K9me3, and H3K27ac. We created input features for the regression task as follows. The data for each histone modification is represented by two bins, one holds the sum of read counts at 2 kb upstream of the TSS and the other holds the same for the region 2 kb downstream (12 features in total). Next, we fit a linear regression model to our data set with feature matrix of size $2{,}484 \times 12$ and response vector of size 2,484 using the lm function in R, where 2,484 is the number of bidirectional genes considered in this study. The studentized residuals were calculated between the measured average transcript expression values and the predicted values. As hinted by (Wang et al., 2012), a gene is marked as stable if the corresponding studentized residual for that gene is above 1, conversely a gene is unstable if the studentized residual is below -1 If none of the above is the case, the gene is annotated as neutral (Appendix Figure C.4a). For each state, we assessed the percentage of genes being classified into stable, unstable, or neutral. To compute the enrichment of the stable and unstable mRNAs in each state, we performed the hypergeometric test on these three categories, separately for each state, and used the $p \leq 0.05$ as significance cutoff.

5.2.8 Bidirectional gene signature: concordant or discordant

We highlight two signatures describing the changes in bidirectional gene expression. The first signature pertains to consistent changes of expression of the two genes across single cells in a consistent manner. For instance, when the expression of one gene of a BP is always higher than the other across the cells. Patterns as such create a signature that we call *concordant* signature. On the other hand, there exist cases where the expression of these two genes flips across cells, i.e., the expression of one gene is higher in one cell and in another cell the expression of the counterpart gene is higher. Since the signature these BPs can be attributed to reflects a discordant behavior we call this signature as *discordant*. To analytically distinguish these two signatures for the gene pairs in a BP, we computed the Wilcoxon signed rank test on their imputed single cell expression values (BPs whose both genes had zero expression in all cells were discarded for the test). We consider a gene pair being *concordant* if the p-value after using Benjamini-Hochberg multiple testing correction is smaller than or equal to 0.05. We, additionally, define concordant ratio as the number of *concordant* BPs normalized by the total number of BPs in a given cluster.

5.2.9 Data for HepG2 and K562 cell lines

Epigenomic data for the HepG2 cell lines were produced by the DEEP consortium and are publicly available on the DEEP data portal[2] as well as the IHEC data grid[3], including metadata information providing further details on sample sorting and preprocessing.

[2] http://deep.dkfz.de/#/home

[3] http://epigenomesportal.ca/ihec/grid.html

The rest of the data, K562 (HM ChIP-Seq, TF ChIP-seq, CAGE), and HepG2 (TF ChIP-seq, CAGE) were obtained from ENCODE.

5.2.10 GRO-cap and CAGE expression estimation

GRO-cap data for K562 was downloaded from GEO via the accession code GSE60456 provided by Core et al., 2014b and CAGE data for both K562 and HepG2 samples from ENCODEConsortium, 2012. The count of the reads overlapping with a window in the region [0,+100] bp downstream of a gene's TSS is used to define the expression derived from the GRO-cap or CAGE assays.

5.2.11 Measuring average methylation in BPs

WGBS-seq data for HepG2 was produced by DEEP and for K562 was obtained from ENCODEConsortium, 2012. Both files were processed using the RnBeads package in R (Assenov et al., 2014) to compute the average methylation levels around the TSSs. In other words, for each TSS, the methylation level was computed in a 2 kb window (partitioned into bins of 100 bp) downstream of the L and the H gene, respectively (40 bins in total). Additionally, the methylation level was measured within the region between the TSSs of L and H genes. Finally, the computed methylation levels were concatenated in genomic order, resulting a vector of 41 elements in total.

5.2.12 Measuring G-C content in BPs

GC-content profiles were determined based on the human GRCh37 reference genome. For each TSS, the GC-content was measured in a 2 kb window (partitioned into bins of 100 bp) downstream of each of the L and the H genes, separately (40 bins in total). Additionally, the GC-content was computed within the region between the TSSs of L and H genes. For visualization, the results were concatenated in genomic order forming a vector of size 41 per BP.

5.2.13 Measuring small RNA abundance in BPs

The BAM alignment files for small RNA data measured at the nuclear fraction of the HepG2 and K562 cell lines were obtained from ENCODEConsortium, 2012. Next, bamCoverage from bedtools was used to produce the bedgraph files, to which the binning approach explained in 5.2.12 was applied, in order to derive the small RNA profiles around the BPs. For clearer illustration, values larger than 200 were capped to 200.

5.2.14 Enrichment of gene products partitioned according to transcription states

We partitioned the gene product annotations into two general groups, protein-coding (PC) and the rest as non-coding (NC). In the scope of BPs, we established a new notation, $gp \in \{NC \rightarrow NC, NC \rightarrow PC, PC \rightarrow NC, PC \rightarrow PC\}$, denoting the gene products for a pair of genes. We counted the occurrences of each of these four gene product categories for the gene pairs of our transcription states, as shown in tables 5.2 and C.1. To assess the enrichment of such occurrences, we performed a hypergeometirc test to their contingency table, $C \in \mathbb{Z}^{4\times4}$, where $C_{i,j}$ represents the frequency of the j^{th} gene product category in the i^{th} state. In detail, let $h(x; N, n, k)$ be the hypergeometirc distribution, where N and n are the population and sample size, respectively. k is the frequency of successes in the population, and x denotes the frequency of successes in the sample. We used the following assignment of parameters of this distribution for each entry $C_{i,j}$ of the contingency matrix C:

$$h(C_{i,j}; \Sigma_{r=1}^{4}\Sigma_{s=1}^{4}C_{r,s}, \Sigma_{r=1}^{4}C_{r,j}, \Sigma_{r=1}^{4}C_{i,r}) \ . \tag{5.4}$$

The p-value obtained from this test is used to assess the significance of enrichment of a gene product category in a particular state.

5.2.15 Enrichment of TF ChIP-seq data

To capture and preserve the spatial distribution of the TF ChIP-seq signal around the promoter, the ChIP-seq reads are counted in bins of size 100 bp spanning a window beginning from the TSS of each bidirectional gene and extending up to 2000 bp downstream of the TSS. An additional bin with variable length is dedicated to count for the reads falling within the region defining the distance between the two TSSs of a BP. The 20 bins from the L gene, the bin for region between both TSSs, together with the 20 bins from the H gene are all combined into a vector of size 41, which represents the binned ChIP-seq signal per BP for a particular TF. To evaluate the enrichment score of the i^{th} TF at a particular BP, we define:

$$Enrich(TF^i) = \Sigma_{j=1}^{41} log_2(\frac{TF_j^i + 1}{BG_j^i + 1}), \tag{5.5}$$

where TF^i is the signal measured for i^{th} TF (for HepG2, $i \in \{1, \ldots, 44\}$ and for K562, $i \in \{1, \ldots, 50\}$) at the given BP, TF_j^i holds the read counts computed at the j^{th} bin of TF^i signal and BG_j^i holds the median of TF^i signal computed at the j^{th} bin across all BPs.

5.2.16 Definition of transcript length

For each gene, we consider all the annotated transcripts that start within 2 kb downstream of the most 5′ TSS of the gene. Then we measured the length of the exonic region encompassed by these transcripts, which we refer to as *transcript length*. Note that this is not the whole gene's *transcript length* as other transcripts of the gene that would start outside of the 2 kb region are not included.

5.2.17 Definition of transcripts span

For each gene, we consider all the annotated transcripts that start within 2 kb downstream of the most 5′ TSS of the gene. The region spanned by those transcripts is referred to as *transcripts span*. For instance, if the following transcripts start downstream within 2 kb of the most 5′ TSS, $T_1 = (start : 0, end : 1000)$, $T_2 = (start : 200, end : 3000)$, $T_3 = (start : 200, end : 2000)$, then the *transcripts span* would be equal to $(start : 0, end : 3000)$, where *start* and *end* are relative coordinates to the most 5′ TSS. Note that all regions in this interval are considered, regardless of their exonic or intronic annotations. Also note that other transcripts of the gene that would start outside of the 2 kb region are not considered for the definition of *transcripts span*.

5.2.18 Chromatin state segmentation score

We obtained the 18-states ChromHMM (Ernst and Kellis, 2012) annotation for both cell lines, HepG2 was generated by DEEP, and K562 was downloaded from Roadmap (Consortium et al., 2015). For the sake of simplicity, we collapsed all states related to TSS to one state called, TSS. Similarly, we defined Enhancer and Repressed states and assigned all the remaining states to Others, resulting in four summarized states in general. Later, for each gene g we defined a window, W_g, starting at the TSS of the gene and extending up to the size of the *transcripts span* (see 5.2.17 for definition of *transcript span*). We then computed the average number of bases having a particular chromatin state, s, overlapping in that window. $ChromScore_g^s$ holds this value and is described in the following:

$$ChromScore_g^s = \frac{\Sigma\{|R| : R \subseteq W_g \; and \; state(R) = s\}}{W_g}, \tag{5.6}$$

where R represents a region in the genome, $|R|$ denotes the size of this region, and $state(R)$ holds the chromatin state assigned by ChromHMM to region R. It is worth mentioning that since the ChromHMM state annotation is continuous across the genome, the following equation holds:

$$\Sigma_{s\in\{TSS,Enhancer,Repressed,Others\}} ChromScore_g^s = 1\,, \tag{5.7}$$

and thus $ChromScore$ is appropriately normalized to account for variable lengths of *transcripts span* per gene. $ChromScore$ can also be attributed to a cluster of genes, C, via the following:

$$ChromScore_C^s = \Sigma_{g\in C} ChromScore_g^s\,, \tag{5.8}$$

As the last step, we convert the $ChromScore_C^s$ into percentages to make the score comparable across different clusters of genes with different gene sizes:

$$percent(ChromScore_C^s) = \frac{ChromScore_C^s}{\Sigma_{s\in\{TSS,Enhancer,Repressed,Others\}} ChromScore_C^s}\,. \tag{5.9}$$

Figure 5.1 illustrates the desired segmented regions for the computation of $ChromScore$ given the explanation above.

5.3 Results

5.3.1 Four states of transcription with distinct bidirectional characteristics

We propose a novel approach to study the relationship between the regulation of the two genes at a bidirectional promoter by exploiting RNA-seq data at the single cell level. Our approach is in contrary to the existing studies that rely on bulk RNA-seq data. Bulk RNA-seq averages gene expression across individual cells, and thus obscures interesting patterns of bidirectional gene expression (Figure 5.2a).

As previously described in 5.2.2, our BP data set contains 1,242 divergent promoters with two core promoter elements, obtained from annotated ENSEMBL genes (GRCh37.75). This BP set meets the criteria of having the distance between TSSs of each BP not exceeding 500 bp. Loci overlapped by any other annotated gene region (±2 kb from the TSS) are excluded from our BP set. From two scRNA-seq data

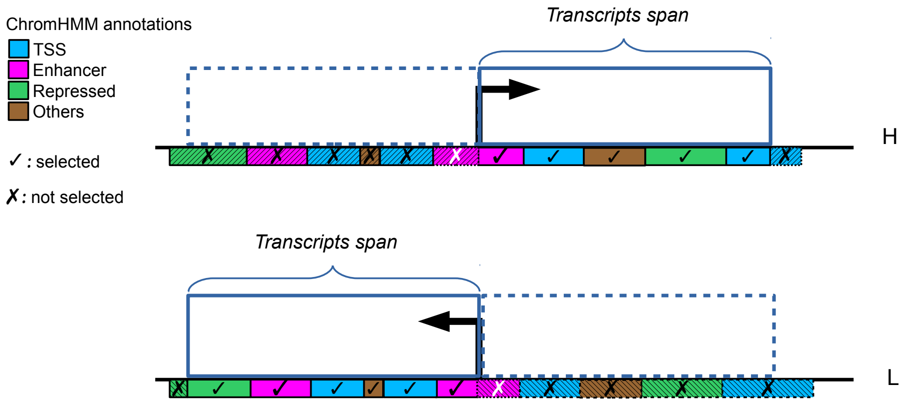


FIGURE 5.1: A schematic representation for computing the segmented genomic regions using ChromHMM for a region defined at the *transcript span* of either genes (L and H). The overlapping regions are taken into account for computing the *ChromScore*.

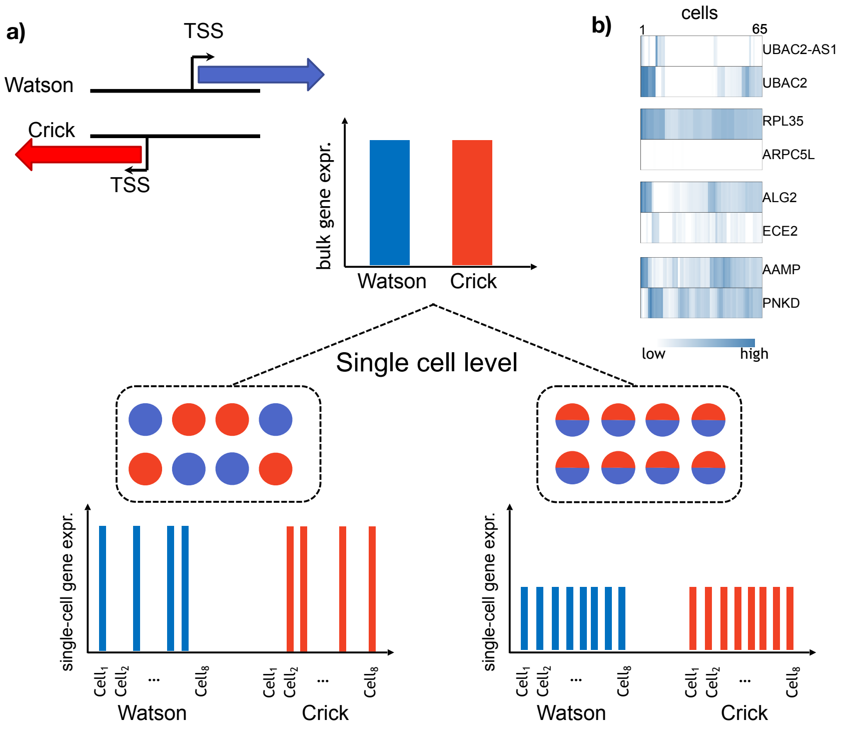


FIGURE 5.2: Advantages of studying BPs at single cell level. a) An illustration of a BP, defined based on two genes located on opposing strands of DNA (Watson and Crick). Bulk RNA measurements at the BP may hide complexity of BP gene regulation. This is shown in the left single cell expression scenario, where one of the genes is expressed and the other is silent in the same cell compared to the other scenario where single cell expression agrees with bulk measurements. b) Heat maps of 65 HepG2 single cell RNA-seq expression measured in four bidirectional promoters (TPM).

sets, 65 cells from HepG2 (see C.1) and 42 K562 cells (Pollen et al., 2014), we determined the single cell expression of genes associated to these BPs. To perform a sanity check on the single cell data, we contrasted the expression of bidirectional genes averaged over single cells with their corresponding bulk expression. As illustrated in Appendix Figure C.2, the average single cell expression concurs with bulk measurements with Spearman correlation coefficient of $\sim$0.8, for both HepG2 and K562 samples.

Figure 5.2b provides examples of single cell expression patterns in HepG2 cells for a few chosen BPs. All these four BPs exhibit distinct patterns. For instance, considering the *ALG2/ECE2* gene pair, the magnitude of expression alternates across the cells, i.e., in some cells *ALG2* is higher expressed than *ECE2* and vice versa. The *AAMP* and *PNKD* genes also demonstrate this alternation, but in a more frequent manner. These observations triggered us to systematically investigate these diversities through building a matrix holding expression values specific to BPs for clustering analysis (see 5.2.4).

As displayed in Figure 5.3, we first construct two matrices representing the single cell expression in BPs, one for the gene located on the Watson strand (Watson matrix) and the other for the gene located on the Crick strand (Crick matrix). Next, we swap a row of the Watson matrix with the corresponding row of Crick matrix, if the average single cell expression of the former is lower than the latter. In this way, for a given BP, we always keep the higher expressed gene (H) on the right side and the lower expressed one (L) on the left. It is worth mentioning that this step merely facilitates the follow-up analyses and does not destroy any desirable information in the data. The final BP matrix is the result of concatenating the swapped Watson and Crick matrices. Each row of this matrix represents a BP and each column represents a cells holding the expression of genes. This means that there are as many rows as the number of BPs (N=1,242) and as many columns as twice the number of single cells; the first half of the columns represent single cell expression of L genes and the second half represent the same for H genes. Given that the combined matrix contains the joint expression information for both genes of a BP in each row, we used hierarchical clustering to identify groups of BPs based on the similarity in their single cell expression patterns.

The result of this clustering led to discovering four distinct transcription states in both cell lines (Figure 5.4 HepG2, and Figure C.5a K562) with the following characteristics: 1) *Bidirectional Lowly Expressed* (*BLE*), where both genes of a BP are lowly expressed, 2) *Bidirectional Weak Difference* (*BWD*), where the H gene is higher expressed than the L gene with a weak difference between the two, 3) *Bidirectional Strong Difference* (*BSD*), where the H gene is much higher expressed than the L gene and higher than in *BWD*, 4) *Bidirectional No Difference* (*BND*), where both genes of a BP are expressed relatively at the same rate.

Table 5.1 compares the number of BPs in each transcription states for both HepG2 and K562 cell lines and reveals that most the BPs are common between the two samples (1090 out of 1242 in total). Next, we investigated the association between the transcription state and the type of gene products encoded in a BP (see 5.2.14). We discovered that for both cell lines the *BWD* and *BND* states are enriched with BPs (hyper-geometric test, $p \leq 0.05$) where both bidirectional genes are protein-coding (PC→PC, tables 5.2 andC.1). On the other hand, the BPs in *BLE* state are enriched either with two non-coding genes ($NC \rightarrow NC$) or with the L gene being protein-coding and the H gene being non-coding ($PC \rightarrow NC$).

Using the single cell data, we identified and counted the *concordant* and *discordant* BPs in all states for both cell lines (see 5.2.8 and Figure 5.5 and Table 5.3). The *BLE*

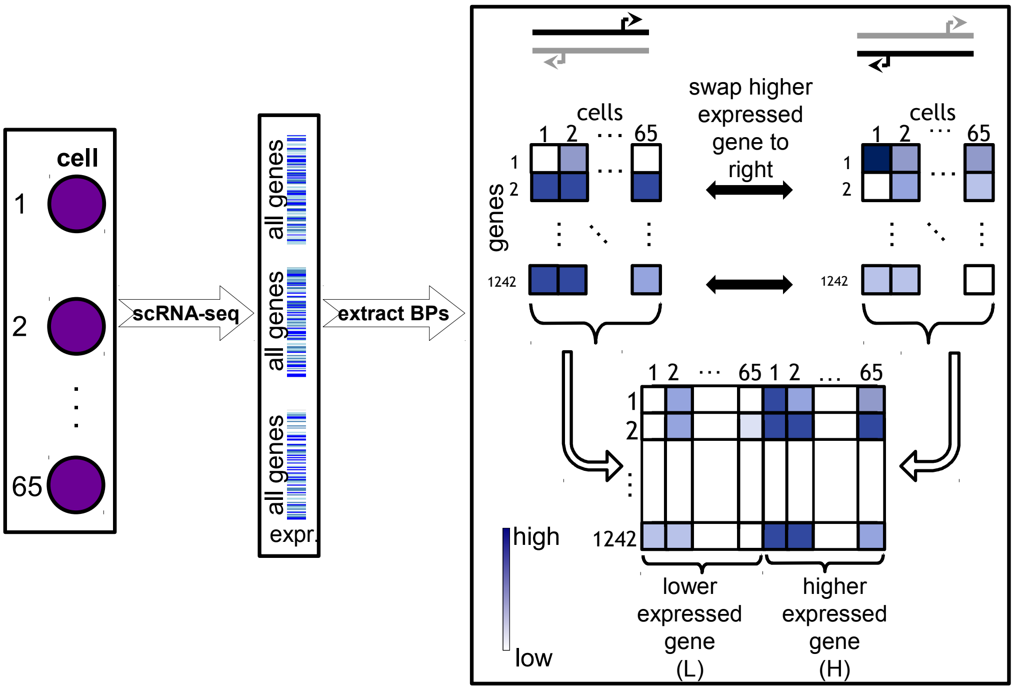


FIGURE 5.3: After single cell sequencing and estimating the gene expression of all genes in a cell, a set of 1,242 BPs was extracted. Single cell expression of either genes of a BP was arranged in two separate matrices for which the rows represent the BPs and columns the cells. Next, we swap the higher expressed gene to the matrix on the right and lower expressed one to the left. The resulting matrices are combined into one joint BP single cell expression matrix.

cluster	*BLE*	*BSD*	*BWD*	*BND*	total
HepG2	900	94	208	40	1242
K562	870	65	272	35	1242
overlap	804	50	128	18	1090

TABLE 5.1: Number of BPs falling into each transcription state in HepG2 and K562 cells and their overlap.

	BLE	*BSD*	*BWD*	*BND*
NC→NC	78*	2	12	2
NC→PC	273	32	55	2
PC→NC	142*	7	26	3
PC→PC	407	53	115*	33*

TABLE 5.2: Number of BPs falling into the gene product categories (NC→NC, NC→PC, etc.) in HepG2. Statistically enriched values are marked with $*$ (Hypergeometric test $p<0.05$).

state is, in general, lowly expressed and the stochasticity in the expression makes it difficult to observe a consistent pattern. On the other hand, the *BSD* state includes BPs where one gene's expression is always higher than the other, resulting in a concordant ratio of 1. As expected, the *BND* state shows the smallest concordant ratios, i.e, highest discordance, which is due to the frequent alternations taking place in the expression profile of the genes in this state.

The CAGE expressions in the transcription states are displayed in Figure 5.6. It can be observed that the distributions shown in Figure 5.6 agree with the expression characteristics of each state (similarly for the bulk RNA-seq and CAGE in K562 cell line, Appendix Figure C.5b and C.6a).

Even though the representation used in Figure 5.4 is concise, it does not deliver a suitable visualization for exploring the changes in the expression of L and H genes within the same cell. Therefore, it is reasonable to compute the correlation between the expression of L and H genes of a BP across the single cells. This allows us to quantitatively assess the relation between single cell expression of bidirectional genes in these states, as depicted in Figure 5.7a for both K562 and HepG2 cell lines. This analysis revealed that the state with the highest correlation is *BND*. On the contrary, the *BSD* state showed smaller correlation values, indicating a more independent regulation of its bidirectional genes.

To address which mechanism(s) are involved in driving such differences in regulation of BPs, we explored the following aspects: 1) structural features, 2) epigenetic signals, and 3) transcriptional regulatory elements.

concordant (%)	*BLE*	*BSD*	*BWD*	*BND*
HepG2	0.80	1.00	0.95	0.72
K562	0.58	1.00	0.88	0.88
overlap	0.59	1.00	0.81	0.72

TABLE 5.3: Ratio of *concordant* BPs shown separately in each transcription state for both cell lines as well as their overlap.

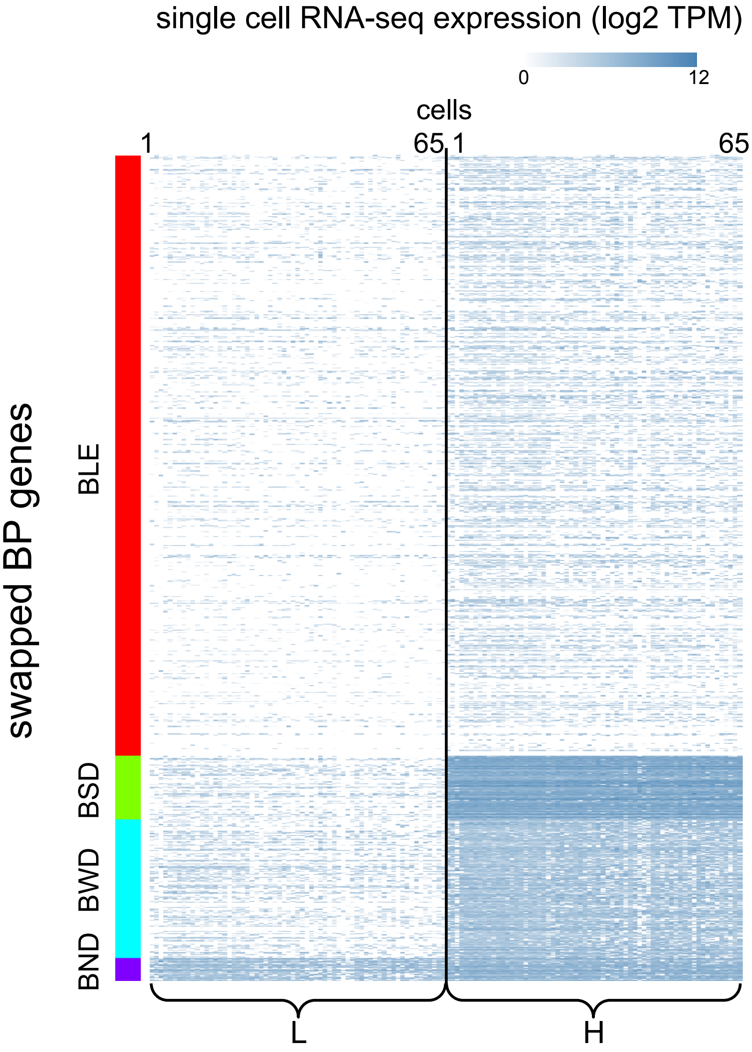


FIGURE 5.4: Single cell RNA-seq expression in bidirectional promoters. Hierarchical clustering of the HepG2 single cell transcript expression matrix visualized as a heat map (log2, TPM). The four distinct clusters (*BLE*, *BSD*, *BWD*, *BND*) are referred to as *transcription state* in this manuscript.

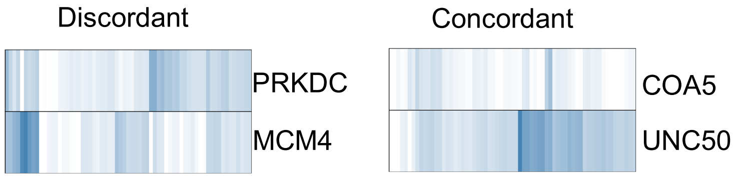


FIGURE 5.5: Examples of *concordant* and *discordant* BPs in HepG2.

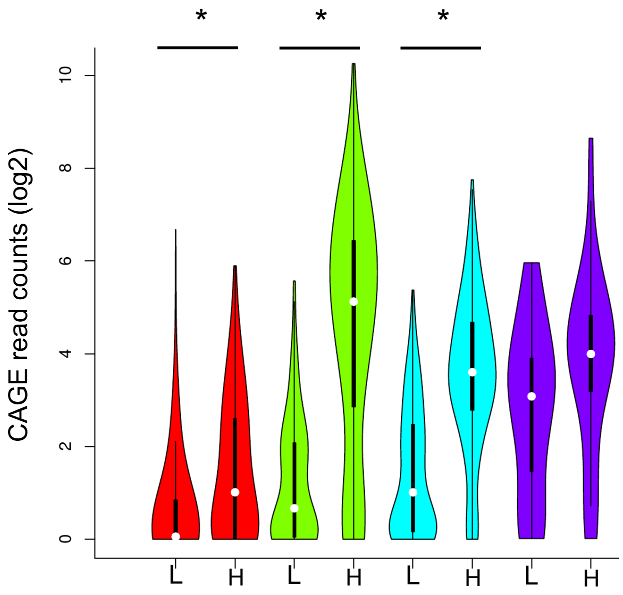


FIGURE 5.6: CAGE read counts, measured for each bidirectional gene (L and H), shown for each transcription state. Color code as in A. Significant differences are marked with $*$ (paired and two-sided Mann-Whitney test, p <0.05).

5.3.2 Structural features associated with transcription states

We first tested whether the distance between TSSs of bidirectional genes was associated with the transcription states. Figure 5.7b provides the distributions of TSS distances in each state for both cell lines. We observed that the *BLE* state shows significantly larger TSS distances compared to the other states (t-test, p $\leq$ 0.001). On the contrary, *BND* had the smallest median distance (significant for HepG2, t-test p $\leq$ 0.05). This observation together with the correlation analysis in Figure 5.7a suggests that the smaller distance may influence recruitment of a common regulatory complex that facilitates the simultaneous regulation of both genes.

Given that the scRNA-seq protocol measures steady-state fully elongated mRNAs, we wondered whether the length of the transcribed region differs in the genes associated to the BPs. For this, we inspected the region spanned by all transcripts originating from transcription start sites within 2 kb from the most 5′ TSS of a BP gene, a region we refer to as *transcripts span* (see 5.2.17). Surprisingly, this length was significantly smaller (Mann-Whitney test, p $\leq$ 3.6e-05) for the H genes of states *BSD* and *BWD* compared to their counterpart L genes. Linking this observation to the actual transcription expression depicted in Figure 5.7c for these two states suggests that the expressions of L and H genes are inversely related to their *transcripts spans* in BPs. To elucidate whether this association holds for all genes or only BPs, we measured the *transcripts span* for all 63,678 annotated genes in the human genome. We found no association of *transcripts span* with gene expression for all genes (Figure S5B), indicating that such structural configuration might be specific to BPs. Since the estimated TPM values are derived from the exonic regions only, we further examined the *transcript length* by measuring the exonic region of all transcripts initiating within the 2 kb from the most 5′ TSS of a BP gene (Figure C.9, also see 5.2.17 and 5.2.16). Despite the evident differences observed in the single cell expression profiles of the *BSD* and *BWD* states, their *transcript length* dose not show a relevance in deriving those differences.

We further investigated whether the difference in GC-content was involved in yielding the variations observed on the expression patterns, but we found no apparent differences (Appendix Figure C.8 and 5.2.12).

5.3.3 Histone modification and DNaseI patterns reflect the characteristics observed in transcription states

To explore the role of epigenetics in transcription states observed in Figure 5.2D, we produced seven histone modifications (H3K4me1, H3K4me3, H3K36me3, H3K27me3, H3K9me3, H3K27ac, and H3K122ac) and DNaseI-seq data for HepG2 cells within the DEEP consortium. Further we obtained data for DNaseI-seq and all modifications, except H3K122ac, for K562 cells from ENCODEConsortium, 2012. Figure 5.8 depicts the normalized read counts measured around the TSSs of bidirectional genes stratified according to the transcription states for all HepG2 data sets (similarly, for K562 in Appendix Figure C.10a). Generally, we observed that the epigenetics data show specific patterns related to these states. For instance, it is notable that the *BLE* state had the lowest abundance for HMs associated with active promoters (H3K4me1, H3K4me3, H3K36me3, H3K27ac, and H3K122ac) and highest for H3K27me3 and H3K9me3 that are mostly associated with repressed promoters (Bannister and Kouzarides, 2011). On the other hand, the *BND* state exhibited the very opposite behavior compared to *BLE*, reflecting their expression characteristics observed in Figure 5.4.

Another interesting observation is the agreement of the elongation mark profiles, H3K36me3, with the *transcripts span* distribution shown in Figure 5.7c. In general, the larger the increase of the H3K36me3 mark the shorter the *transcripts span* for the gene. For instance, the *BSD* state that has the shortest *transcripts span* exhibits the sharpest increase in its H3K36me3 profile only downstream of the TSS. This is compatible to the previous observation that the H3k36me3 mark increases gradually and peaks at the end of genes (Sein, Värv, and Kristjuhan, 2015) and we can observe that general trend for the *transcripts span* on our data as well (Appendix Figure C.10b).

In a recent study by Wang et al., 2012, it has been shown that mRNA stability can be estimated using HM data at promoters. Previous research on genome-wide measurements of RNA half-lives suggested that lncRNAs exhibit a wide range of stabilities similar to that of protein-coding transcripts (Clark et al., 2012). Therefore, we used the approach by Wang et al., 2012 to estimate which genes appear to be stable and unstable, with the idea that this could also explain differences in the gene expression behavior we observe in the different states. Briefly, this method uses HM signals at promoters, as features, and gene expression measurements, as response, to learn a linear model that predicts gene expression. Using outlier analysis, genes that show lower (higher) expression as predicted are marked as unstable (stable) (see 5.2.7 and Appendix Figure C.4a). The results reveal that the putative stable genes are significantly (hyper-geometric test, $p \leq 0.05$) enriched in all the states except *BLE* (consistent across both HepG2 and K562 samples). On the other hand, the *BSD* state was significantly enriched in the putative unstable category, with $\sim$21% and 30% of its genes being inferred as unstable in HepG2 and K562 samples, respectively (Appendix Figure C.4b).

The DNaseI-seq profile of the *BND* state revealed not only the highest signal, but also the widest spread around the TSS compared to the other states. This fits to the observation that there is similar amount of single cell transcription for both genes. Due to recent reports about small promoter-associated RNAs (Zamudio,

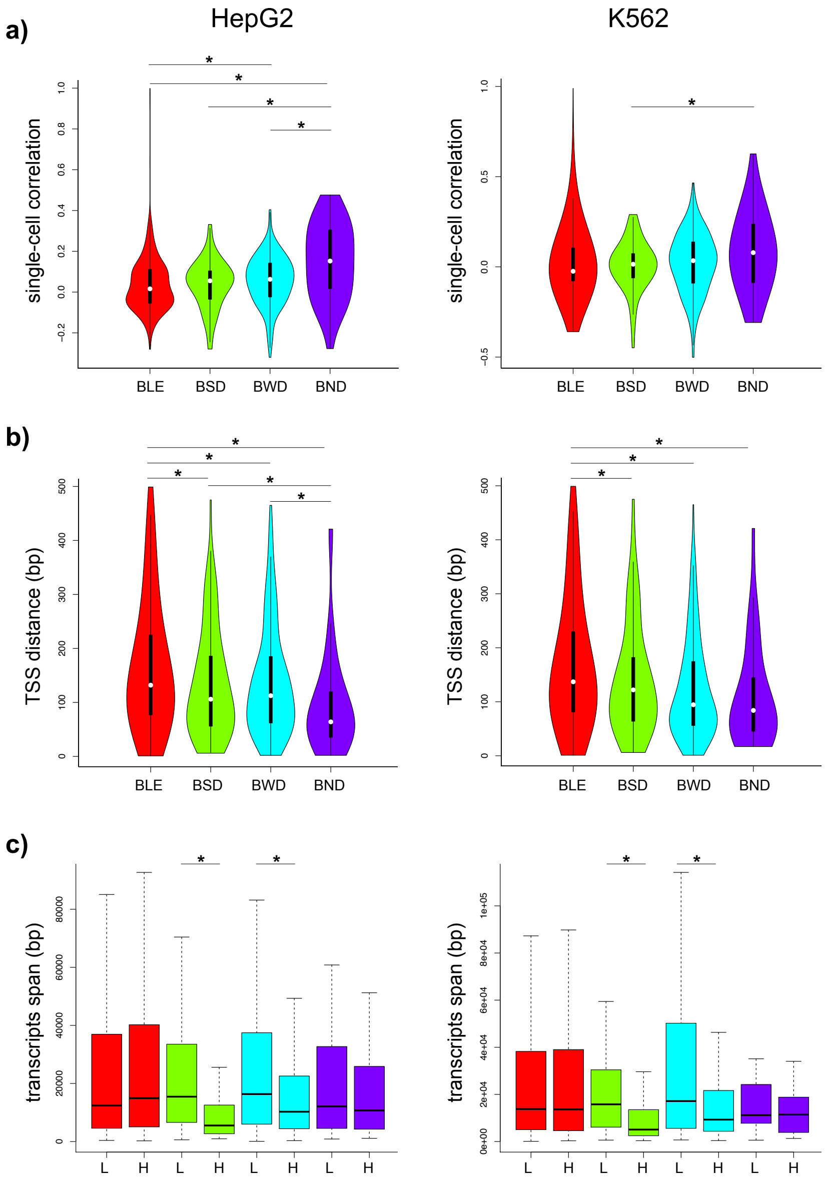


FIGURE 5.7: Structural features of BPs for HepG2 (left column) and K562 cells (right column). a) Distributions of Pearson correlation coefficients (y-axis) calculated from all single cell measurements for each BP in one of the states (x-axis). b) Distributions of TSS distance of BPs in each state. c) Length distributions of *transcripts span* for L and H genes of BPs shown in each state. Significant differences are marked with $*$ (paired and two-sided Mann-Whitney test, p <0.05). For all subfigures the color-coding is consistent with Figure 5.4

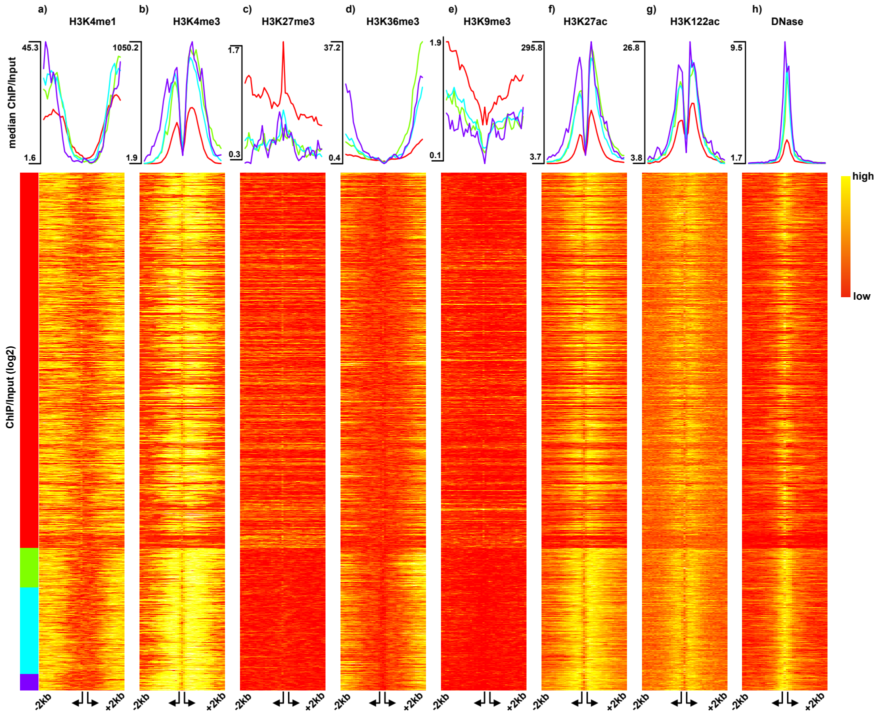


FIGURE 5.8: Epigenetic characteristics of transcription states in HepG2 cells. a-g) Histone modification (ChIP/Input) shown as median profiles (top panel) and log-transformed values as heatmap (bottom panel). h) DNase1-seq median profiles (top panel) and log-transformed raw counts (bottom panel). Arrangement of genes as in Figure 5.4. The reads are measured in 40 bins of size 100 bp spanning a window of size 4000 bp centered around the TSSs, with an additional variable bin between the TSSs.

Kelly, and Sharp, 2014; Wang et al., 2016), we obtained small RNA data (ENCODEConsortium, 2012) for HepG2 and K562 samples (see Experimental Procedures) and grouped them according to the defined transcription states. Although we observed residual small RNA expression in the vicinity of the bidirectional TSSs, we found no consistent patterns associated with the transcription states (Appendix Figure C.10c).

We also examined the average methylation profiles obtained in the four transcription states (see 5.2.11) due to the previously reported associations with gene expression (Siegfried and Simon, 2010; Schübeler, 2015). The results were consistent with other studies where higher level of DNA methylation coincided mostly with silent genes (*BLE*). Consistent with the enrichment of HMs, genes in the *BND* state showed the least amount of DNA methylation (Appendix Figure C.10d).

5.3.4 The BND state coincides with strongest regulatory activity

It was shown that specific TFs preferentially bind to bidirectional promoters (Trinklein et al., 2004; Bornelöv, Komorowski, and Wadelius, 2015). As we observed that the DNA accessibility profiles differed among the transcription states (Figure 5.8h), we were inspired to investigate binding of transcription factors. We obtained ChIP-seq data for several transcription factors (ENCODEConsortium, 2012) (44 for HepG2 and 50 for K562). One hypothesis was that there may exist TFs that bind in the proximal region of a BP and influence gene expression as was observed in our transcription states.

To test this, we established a novel enrichment score tailored to BPs (Appendix Figure C.11 and 5.2.15), which preserves the spatial distribution of the ChIP-seq signal along a BP. We applied the enrichment analysis for both cell lines (HepG2 in Figure 5.9a and K562 in Appendix Figure C.11b). As expected, states with higher expression showed more TF binding in general. However, we were not able to single out distinct TF subsets that could be associated with a particular state. In fact, the states *BSD*, *BWD* and *BND* showed enrichment for many of the TFs that we analyzed. This triggered us to investigate whether the number of TFs that are regulating a BP differed in those states. Figures 5.9b,c provide violin plots demonstrating the number of positively enriched TFs per BP for each state in HepG2 and K562 cell lines, respectively. The *BND* state showed the highest percentage of positively enriched TFs (t-test, $p \leq 0.05$) suggesting that more TFs are required to regulate gene expression in this state.

Next, we tested whether specific genomic regions, such as enhancers, are associated with these four transcription states. For this, we inspected the genome-wide segmentation of HepG2 and K562 cells using an 18-state ChromHMM model (Ernst and Kellis, 2012) (see 5.2.18). For simplification we collapsed all TSS-related, enhancer-related, and repression-related ChromHMM states into *TSS*, *Enhancer*, and *Repressed*, respectively. We assigned all the remaining chromatin states to *Others* (data not shown). The results provided in Figures 5.9d, e suggest that the enhancer-related regions are the most frequent amongst the *BSD* and *BND* states, reflecting their stronger expression profiles. In the case of HepG2 (Figure 5.9d), this quantity is even higher than the number of *TSS* regions. Concurrent with (Figure 5.8) most of the repressed regions belong to the *BLE* state, where genes were lowly expressed.

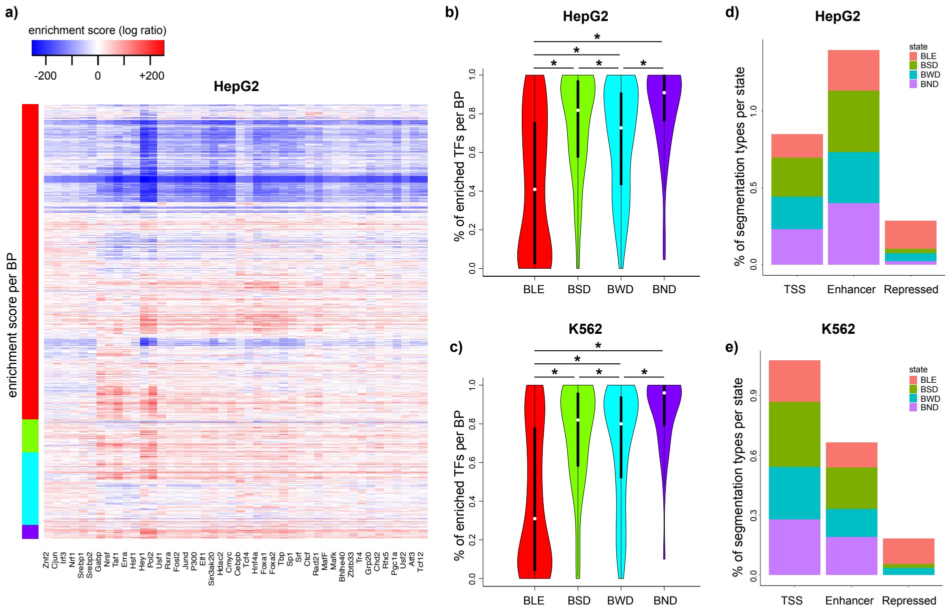


FIGURE 5.9: Heat map of TF (columns) enrichment scores (log ratio against background) for each BP (rows) in HepG2 cells. BPs are sorted as in Figure 5.4. b, c) Distribution of percentages of TFs per BP (enrichment score in (a) > 0) in each state for HepG2 (top panel) and K562 (bottom panel). d, e) ChromHMM annotation, summarized into the categories: *TSS*, *Enhancer*, and *Repressed*, are shown as percentages in a bar plot per state (see 5.2.15)

Conclusion and discussion

In this work we compared single cell expression of genes at BPs. Currently, we only have access to single cell protocols for RNA-seq, and other techniques for quantification of transcription start sites cannot be used (Scruggs et al., 2015; Core et al., 2014b; Shiraki et al., 2003b). Thus, other effects on the mRNA steady state level, *e.g.* post-transcriptional regulation, may influence the gene clustering produced. Here, we have used two high quality single cell data sets for ENCODE cell lines allowing us to benefit from a plethora of epigenomic data sets, which are available or have been produced in this work. We found that 88% of the BPs have the same transcription state in scRNA-seq data despite the difference in origin of HepG2 and K562 cells, which suggests that the majority of these configurations may be stable for many cell types.

In previous work that has analyzed BP regulation, analyses were often limited to a certain configuration at the BP, *e.g.* a non-coding gene upstream of a coding gene, therefore care has to be taken when comparing to previous studies. Here, we have limited our results to annotated protein- or non-coding genes that originate from a bidirectional promoter. We found that the BPs that show similar expression for both genes are mostly restricted to a configuration with two protein-coding genes. It was shown previously that core promoter strength differs for genes with bidirectional expression and unidirectional promoters (Duttke et al., 2015a). Here, we show that, beyond differences in the strength of the core promoter, the number of TF regulators that bind to BPs with high bidirectional expression is largest compared to all other expression configurations we observed. In this analysis we used several ChIP-seq data sets for TFs and developed a BP-specific enrichment analysis approach that measures spatial differences in read coverage along the BP regions compared to the median background, unified in a single quantity for each BP and TF. This is different to other studies that have compared TF ChIP-seq data at BPs, *e.g.* (Bornelöv, Komorowski, and Wadelius, 2015), where the background often was defined as a set of unidirectional promoters rather than all BPs. Thus, to find enrichment in the observed states we properly adjust for the fact that there are two genes, which are regulated by TF binding.

We observed that the *BND* state shows the largest (although not strong) single cell correlation values and that there is a trend with correlation at BP genes being inversely proportional to TSS distance (Figures 5.7a,b). A similar observation was recently made for BPs in the rice genome with correlation measured over several bulk RNA-seq data sets (Fang et al., 2016). Small distance between the two TSSs may ease the coupled regulation of transcription from both, for example through a shared or co-regulated mediator complex (Allen and Taatjes, 2015).

We also found that the *transcripts span*, the genomic region covered by all transcripts that start in the vicinity of the TSS, was imbalanced for the *BSD* and *BWD* states, with the shorter span linked to the highly expressed gene at the BPs. One possibility is that shorter regions of elongation lead to faster transition cycles for Pol2, assuming similar elongation rate of both genes at a BP. This could be a mechanism by the cell to create imbalanced expression output from a shared regulatory region of two BP genes. We also showed that these two states have the highest percentage of stable and unstable genes inferred by our outlier detection approach. We found out that in these two states only the lowly expressed genes were inferred as unstable. As 3′UTR length is found to be associated with regulation of mRNA stability (Mayr, 2017), we investigated the 3′UTR length between the lowly and highly expressed genes in the stable and unstable categories (Appendix Figure C.4c). However, the results showed no apparent significant trend. This probably means that different sets of post-transcriptional regulators are involved in individual bidirectional gene regulation.

Anecdotally, we investigated bulk GRO-cap data for K562 cells (Core et al., 2014b), and found that the amount of capped nascent transcripts is more similar for both genes at a BP in our states (Appendix Figure C.6b), compared to the amount of stable RNAs expressed (CAGE and RNA-seq). Even though the nascent RNA amount is similar we get significantly different steady-state transcript expression, which could be explained by the difference in length of the genomic region to be elongated, here referred to as *transcripts span*. Once single cell measurements of nascent transcription are available, one could investigate the difference in elongation and transcriptional initiation in these BPs.

Taken together, we observed three different genomic and epigenomic architectures underlying single cell transcription states in BPs. We propose a model depicted in Figure 5.10 to describe these architectures. This model supports distinct characteristics of the *BLE* state, where the bidirectional genes were lowly expressed. They mostly exhibited large TSS distance and more prevalence of repression associated HMs, fewer regions of accessible DNA, and less TF binding. The *BSD* and *BWD* states, on the other hand, had reduced TSS distance in comparison with *BLE* and more abundance of activation associated HMs as well as higher rate of TF binding. Interestingly, the *transcripts span* associated to the H gene of BPs in these states was observed to be shorter than the L one. Lastly, *BND* showed strongest single cell co-expression and smallest TSS distance among the states. Furthermore, we observed the widest accessible regions of DNA, the largest number of binding TFs, and highest amount of activation related HMs.

Although the transcription state definition was based on the single cell data, several bulk data sets showed consistent and matching patterns for those states. Our results suggest that novel statistical methods can be developed to deconvolute masked subpopulations of cells measured with different bulk epigenomic assays with the help of single cell RNA-seq data. Further advances in single cell sequencing technologies (Schwartzman and Tanay, 2015) are necessary such that we can measure not only RNA expression, but also TF binding and histone modifications in single cells to understand the hidden complexity, in particular, in BP regulation.

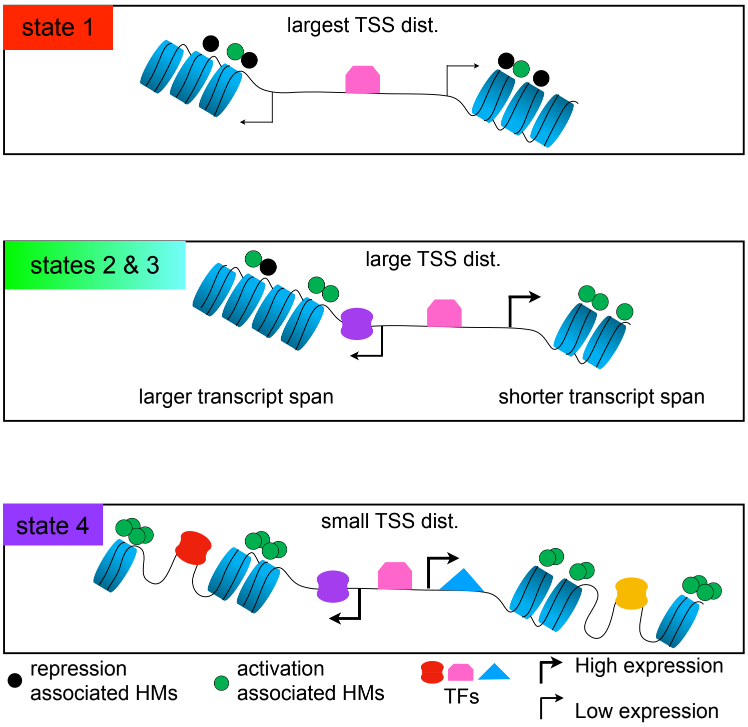


FIGURE 5.10: Hypothetical model for three different genomic architectures underlying epigenetic regulations of BPs. BPs that drive single cell expression patterns observed in the *BLE* state show large TSS distance and higher abundance of repression associated histone marks and depletion of most TFs. *BSD* and *BWD*, on the other hand, exhibit smaller TSS distance and more TF binding compared to *BLE*. In addition, the *transcript span* of the H gene is observed to be significantly smaller compared to the L gene. BPs categorized in *BND* show the smallest TSS distance with the most TF binding events that require more accessible DNA to regulate both the L and H genes.

Chapter 6

Tree-based multi-tasking to predict gene regulation in single cells

6.1 Introduction

In the previous chapter 5, we emphasized on the advantage of using single cell sequencing to unravel interesting expression patterns among the bidirectional genes. The promising results we obtained from that project, motivated us to explore a more global picture of transcriptional regulation by considering all genes. However, the main objective here is to understand whether the single cell data can be used to infer cell-specific transcriptional regulatory components. There have been several related studies exploring the answer to this question.

ACTION (Mohammadi et al., 2018) is one of the several existing methods that was developed to discover new cell types using single cell data (Kotliar et al., 2019; Kanter et al., 2019; Baron et al., 2016). It not only reveals novel cell types, but also seeks to identify underlying regulatory factors from single cell expression data sets. Mohammadi et al., 2018 devised a cell similarity metric that intensifies the effects of preferentially expressed markers. They further used a geometry-based approach to identify the primary functions of cells.

Another tool for inferring regulatory networks, called SCENIC, was proposed by Aibar et al., 2017. SCENIC is a computational approach to infer gene regulatory networks specific to identified states obtained from single cell data. The result of their computations creates a binary matrix, called activity matrix, through which identifying single cells that have significantly higher subnetwork activity is feasible. In the first step of their approach, they train random forest regression models to learn non-linear associations between the expression of TFs, as features, and the expression of target genes, as response. The output of this step yields potential regulators as well as their importance measure, which is used to interpret the influence of a TF on predicting its target gene. In addition, they perform enrichment analysis on the motifs of TFs of interest. These motifs are significantly over-represented in the surroundings of the TSS (10 kb around the TSS or 500 bp upstream the TSS) of the target genes. The TFs with strong motif enrichment are then selected for building the regulons. However, their framework cannot find negative associations between a TF and its target gene, due to low motif enrichment they observed for these cases in the data sets they studied. The positively associated regulons are then used to be incorporated with the single cell data. Through this step, the activity of each regulon in each cell is evaluated by calculating the AUC scores, integrating the expression ranks across all genes in a regulon. Finally, these scores are used to create the desired activity matrix as the output of their workflow.

Later, Suo et al., 2018 exploited SCENIC and modified it by defining a Jensen-Shannon divergence based score to assess the specificity of the regulons with respect to the cell types. By considering the regulons having high values of such customized score, they were able to infer both known and novel regulatory elements in the mapped mouse cell atlas derived from comprehensive single cell transcriptomic analysis.

Despite the interesting studies conducted regarding cell type identification and/or regulatory network inference, the single cell data suffers from the inherent technical noise of so called, dropouts. Dropouts are referred to genes that are falsely identified as zero-expressed. In simpler words, any zero that is observed in the expression count matrix of single cell data, can be viewed as either a correctly identified silent gene or incorrectly identified silent gene due to the dropout effect. There have been several methods (Gong et al., 2018; Li and Li, 2018a; Tracy, Yuan, and Dries, 2019) that attempted to address and solve this problem by imputing the missing expression values, but there is no clear way to achieve this nor be certain that the imputed values are correctly inferred.

The work we present here is conceptually similar to SCENIC (Aibar et al., 2017), but it is methodologically different. Similar to SCENIC, we study the associations between the single cell gene expression and transcription factors. However, in contrast to SCENIC, we compute the binding affinities of many TFs instead of relying on their gene expression. This allows us to study the cis-regulatory associations with gene expression independent of cell types or the corresponding expression data. In other words, it provides a generic profile for cell-specific cis-regulatory activities. In addition, we include other types of features obtained from epigenetic data or TF ChIP-seq to infer more cell type dependent associations. We train statistical models, where the expression measurements of genes across single cells are considered as the tasks in a multi-task-learning (MTL) setup. This intuitive analogy motivated us to build MTL regression models with group LASSO regularization predicting the single cell RNA-seq data from the aforementioned features.

We trained our models on two single cell gene expression data sets, induced pluripotent stem cells (iPCs) and human skeletal muscle myoblasts (HSMM), and inspected the coefficients of these models to identify interesting set of features that best explain the gene expression in single cells. In addition, we compared the MTL results with standard univariate response regression models. These results indicate that the MTL models that integrate the information among all single cell gene expressions not only produce more interpretable models but also often lead to higher accuracy.

6.2 Methods

6.2.1 Data preparation

In this section, we explain how the feature and response matrices were generated for our statistical models. It should be noted that Florian Schmidt generated the following feature data (*static*, *epigenetic*, and *dynamic*) using his pipelines. I exploited these data for the remaining of the work.

Static features

We ran TRAP (Roider, 2007) to quantify the binding affinities of 726 TFs at the promoter area defined by a window of size 2 kb centered at the transcription start site

(TSS). These affinity values were later log-transformed to be fed to the statistical models. More precisely, we define $F_S \in \mathbb{R}^{n \times p}$ be the feature matrix representing the affinity values measured for n genes, arranged at the rows, and p TFs, arranged at the columns. This matrix contains the data for what we call *static* features.

Epigenetic features

Using TEPIC (Schmidt et al., 2017), the binding affinities of 726 TFs were measured in peaks defined based on the DNase-seq data within the 50 kb window around the TSSs of HepG2 cells produced by DEEP and mapped against human genome hg19. Similar to *static* features described above, we define the feature matrix $F_E \in \mathbb{R}^{n \times p}$, where n is the number of genes and the p number of TFs for which the binding affinities were computed. In this setup, we also include three extra features representing the number of peaks (*Peak_Counts*), the length of the open region (*Peak_Length*), and the aggregated signal (*Peak_Signal*) computed withing the 50 kb window around the TSS. Because this particular type of feature is derived from the peaks in the DNase-seq data, we refer to this feature setup as *epigenetic* features.

Dynamic features

ChIP-seq data for 123 TFs of the HepG2 cell line were downloaded from ENCODE and the read counts were measured in a 3 kb window defined around the gene's TSS (mapped against genome hg38) to be combined with the iPSCs single cell data for the model training. Precisely, for n genes and p TFs, the feature matrix, $F_D \in \mathbb{R}^{n \times p}$ contains the ChIP-seq read counts measured at the TSSs of the genes. We refer to these features as *dynamic* features.

Responses (tasks)

The expression values of all genes (TPM normalization) measured for a single cell, are considered as a task for the multi-task learning framework. We acquired induced Pluripotent Stem Cells (iPSCs) data generated and provided by Kathrin Kattler from the lab of Prof. Dr. Jörn Walter (Saarland University) in cooperation with Prof. Dr. Jan Hengstler's group (IfADo Dortmund) within the BMBF funded project StemNet. These cells contain two annotated cell types, Primary Human Hepatocyte (PHH) and Hepatocyte Like Cells (HLC). After discarding the low quality values from the iPSCs data with 238 cells, there were 14142, 4827, and 14188 genes for *static*, *epigenetic*, and *dynamic* features, respectively.

We obtained the Human Skeletal Muscle Myoblasts (HSMM) data from Trapnell et al., 2014 and applied the filtering approach suggested by the monocle's tutorial (Trapnell and Cacchiarelli, 2014). At first, the detected genes were defined using the *detectGenes* function by setting the *min_expr* argument to 0.1. A gene is kept if there are at least 10 cells in where the gene was detected (based on the aforementioned definition of detected genes), otherwise discarded. The filtered data (kept genes), contained 19,566 genes and 306 cells for the *static* features. It is worth mentioning that we only generated the *static* features for this data, as there was no valid annotation of the cells that we could rely on for the downstream analysis in our study. Therefore, this data set was only used to demonstrate the results based on the different choices of tree structures required for the tree-guided MTL models.

Filtering

We further reduced the gene set, by completely removing all the affinities computed for the genes that the variance in their feature space (TF affinities) was less than the third quartile of the variances measured for each gene. More precisely, given the F_S matrix, we compute the variance over the TF affinities across all genes, as follows:

$$var_i = variance(F_S[i,]),\ i \in \{1, \cdots, n\} \tag{6.1}$$

where $F_S[i,]$ is a vector of size p, holding the affinity values in the i^{th} row. Next, we define a threshold t based on the third quartile computed over var_i's $\forall i \in \{1, \cdots, n\}$, as a cutoff to decide whether the $gene_i$ should be kept or not:

$$gene_i : \begin{cases} kept & \text{if } var_i \geq t \\ discarded & \text{else}\,. \end{cases} \forall i \in \{1, \cdots, n\} \tag{6.2}$$

Similarly, we applied this filtering procedure on the other two feature setups, *epigenetic* and *dynamic*.

6.2.2 Single-task learning method

We trained individual models with elastic net regularization with 5-fold cross-validation model selection exploring the α parameter within the range of 0 and 1 with step size of 0.1 from the glmnet package in R.

6.2.3 Multi-task learning methods

In the multi-task learning regression setup, where the response variable is a multivariate vector (matrix) a slightly different objective function is considered. Let $X \in \mathbb{R}^{n \times p}$ denote the input matrix for n observations (samples) and p features. Let $Y \in \mathbb{R}^{n \times k}$ denote the response matrix, whose columns are vectors of observations for those k tasks. We look for an appropriate coefficient matrix, $B \in \mathbb{R}^{p \times k}$ that establishes the linear relation between X and Y with the error term ϵ as described in the following formula:

$$Y = XB + \epsilon\,. \tag{6.3}$$

There are various ways to obtain the optimal values for the B coefficient matrix. In this section, we describe several multi-task learning setups used in this study to understand the performance of different formulations and also downstream interpretation of the obtained results.

Ordinary MTL

To optimize a multi-task regression model with elastic-net regularization, the following objective function is used:

$$B^* = arg\min_B(\Sigma_{i=1}^k (y_i - X\beta_i)^T.(y_i - X\beta_i) + \alpha\Sigma_{j=1}^p \|\beta^j\|_2), \tag{6.4}$$

where B^* denotes the optimal coefficient matrix, α is a tuning parameter that controls the magnitude of the coefficients through the L_2 norm regularization, and y_i is a vector of size n holding the response values of the i^{th} task. Similarly, β^j is a vector of size p denoting the coefficients corresponding to the i^{th} task. The $(.,.)$ operator denotes the inner product between two vectors.

Given this optimization formula, we trained an MTL model with elastic net regularization using the glmnet package in R, where the family argument was set to *mgaussian* to account for the multi-tasking nature of the setup. We used 5-fold cross validation to optimize over the α search grid defined within the range of 0 and 1 with the resolution of 0.05. The models generated using the aforementioned formulation are referred to as ordinary MTL (OMTL) throughout the remaining of the text.

6.2.4 Tree-guided group-lasso MTL model fitting

In the ordinary MTL scenario all tasks share the same relevant features. However, it is possible that a subset of highly related tasks may share a common set of relevant features, whereas weakly related tasks are less likely to be affected by the same features (Kim and Xing, 2010). An improvement was proposed by Kim and Xing, 2010 to address this shortcoming of OMTL models. Through their proposed method, which they refer to as tree-guided MTL, the relationship among the tasks is represented as a tree T with V vertices. Each leaf node of T is associated with a task and the internal nodes reflect the groupings of the tasks. This tree structure can be inferred directly from the data or even may be available as prior knowledge beforehand. Within this tree, each node $v \in V$ is associated with a weight w_v, typically representing the depth of the subtree rooted at node v.

$$B^* = arg \min_B (\Sigma_{i=1}^{k} (y_i - X\beta_i)^T . (y_i - X\beta_i) + \lambda \Sigma_{j=1}^{p} \Sigma_{v \in V} \| w_v \beta^j_{G_v} \|_2), \qquad (6.5)$$

where λ is the regularization parameter and $\beta^j_{G_v}$ is a group of regression coefficients $\{\beta^j_i : i \in G_v\}$. We used the LinearMTL package in R, implemented by Tobias Heinen, to train the tree-guided MTL models. We first partitioned 60% of the data for training and 40% for test. Then, we normalized the data to have zero mean and unit variance. For the sake of model selection, we performed a 5-fold cross validation, through which 21 distinct values of λ, defined within the range of 0 and 1 with the resolution of 0.05, were explored. Finally, we trained the models by setting the maximum number of iterations to 200.

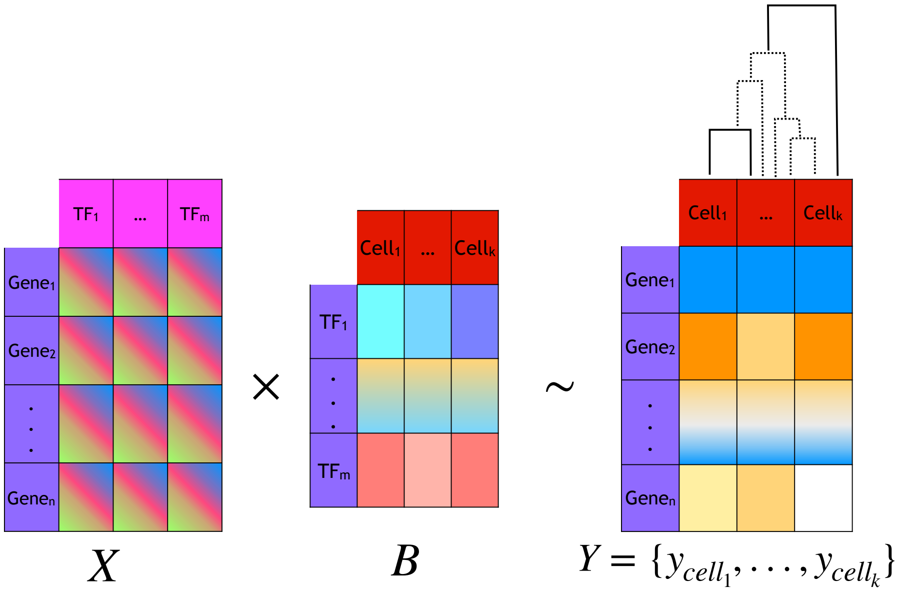


FIGURE 6.1: Schematic illustration of the tree-guided multi-task learning in the context of single cells. The rows of the feature matrix, X, are the genes for which one of the feature setups described previously would be used. The response matrix, Y, consists of the gene expression values measured in single cells. And finally, the coefficients matrix, B, establishes a linear association between the X and Y, where the rows indicate the features and columns the cells.

6.2.5 Construction of the trees used for the tree-guided MTL models

Pseudo-time ordering tree

We ran the monocle package implemented by Trapnell et al., 2014, in order to obtain the pseudo-time ordering of the iPS cells. We retrieved the minimum spanning tree output by the tool to construct the required tree structure for the tree-guided MTL models.

Real data trees

We generated several trees on the true gene expression data that are listed below:

- **HC**: Hierarchical clustering was applied on the real single cell expression data. The clustering tree is then used to guide the tree-guided MTL models.
- **twoDmonocleHC**: Coordinates of the single cells in the reduced dimension space (a matrix of size $2 \times \#cells$) derived from the monocle model trained on the real data were fed to the BuildTreeHC function of the LinearMTL package. This function, as the name suggests, builds the hierarchical tree based on those coordinates in the reduced dimension space.
- **monocle**: In order to generate the groups suggested by the monocle pseudo-time ordering tree, $T = (V, E)$, we defined clusters of cells based on the connected components remained after removing the edges attributed to the nodes

with the highest degree. More precisely, let m be the set of nodes with the maximum degree in T, we define a new tree, T' as follows:

$$T' = (V, E')E' = E \backslash \{e : (v, v') | v \neq v' \text{and } \forall v \in V, \text{and } \forall v' \in m\} . \tag{6.6}$$

This procedure converts the MST generated by monocle to a tree required by the tree-guided MTL models.

Baseline tree

We constructed another tree that serves as a baseline for our tree-guided MTL models. The tree structure forms a star shape, with a root and *#cells* children. More precisely, let k be the number of cells. Then, the baseline tree has $k+1$ nodes, labeled by $0, 1, \cdots, k$, where 0 represents the root and the remaining nodes represent the k cells. Every non-root node has one and only one edge connecting it to the root. Clearly, the root has immediate links to other nodes, i.e., degree of k. This tree structure is considered baseline, because it does not suggest any particular grouping of the cells, as they all are uniformly connected to the root.

Random data tree

In order to generate appropriate random data, we shuffled the expression values for a given gene across the cells and then regenerated the monocle tree for this randomized shuffled data. Given that the genes are arranged in rows and cell in columns, for each row, we shuffled the gene expression values across the cells and then gave it to the monocle tool. Finally, we trained the tree-guided MTL model using the *twoDmonocleHC* tree structure described above.

6.2.6 Selection of top features

Since for the *static* and *epigenetic* features, several hundreds of TFs were included in the set and visualizing this many TFs makes the interpretation infeasible, we decided to shrink this set by selecting those that pass a certain criteria. Essentially, for a given TF arranged in the rows of the coefficient matrix, we compute the sum of absolute values for that TF across all cells. If this value is higher than our predefined threshold of 0.5, we keep that TF, or discard it otherwise.

6.3 Results

6.3.1 Hierarchical clustering based trees result in better performing models

The tree-guided MTL models expect a tree structure to guide the model on how the tasks should be grouped when optimizing the objective function. However, the choice of the tree for the tree-guided MTL is on user's shoulder. Therefore, we explored several cases for which we thought they can convey the structure existing in the single cell gene expression data.

An intuitive choice was to derive the structure from a pseudo-time ordering applied on the single cells. Through traversing the trajectory obtained from monocle, we built a tree representing the pseudo-time ordering of the cells (see 6.2.5). Since the transformation from the pseudo-time ordering to a tree can be arbitrary, we applied hierarchical clustering on the matrix holding the data for pseudo-time ordering and used the resulting tree for our tree-guided MTL models (see 6.2.5).

In addition, we applied hierarchical clustering on the gene expression data directly, to compare the results when other trees have been used (6.2a-c). The scatter plots in Figure 6.2 show the Pearson correlation coefficients computed between the predicted and measured values of gene expression in each individual cell. Interestingly, we see that applying the hierarchical clustering either on the gene expression values (*HC*) or on the reduced dimension data obtained from monocle software (*twoDmonocleHC*) leads to better results than the tree structure inferred from the pseudo-time ordering (*monocle*).

We further examined the performance of the tree-guided models with two other types of tree structures, *random* and *baseline* (see 6.2.5). Figures 6.2d-f compare the performance of *random* and *baseline* with the *HC* and *twoDmonocleHC* models. These results suggest that the choice of hierarchical tree, performed on either the full gene expression data or the reduced space, are valid and reliable as they outperformed the models with *random* and *baseline* trees.

Apart from the tree-guided models, we also generated the ordinary multi-task learning (*OMTL*) models to examine the efficiency of the tree-guided over *OMTL* models. Scatter plots provided in figures 6.2g-i allow us to compare the performance of *OMTL* models with the tree-guided MTL models, where different trees are used. In general, it can be seen that the tree-guided models are superior to the *OMTL* ones, especially when the hierarchical tree is used.

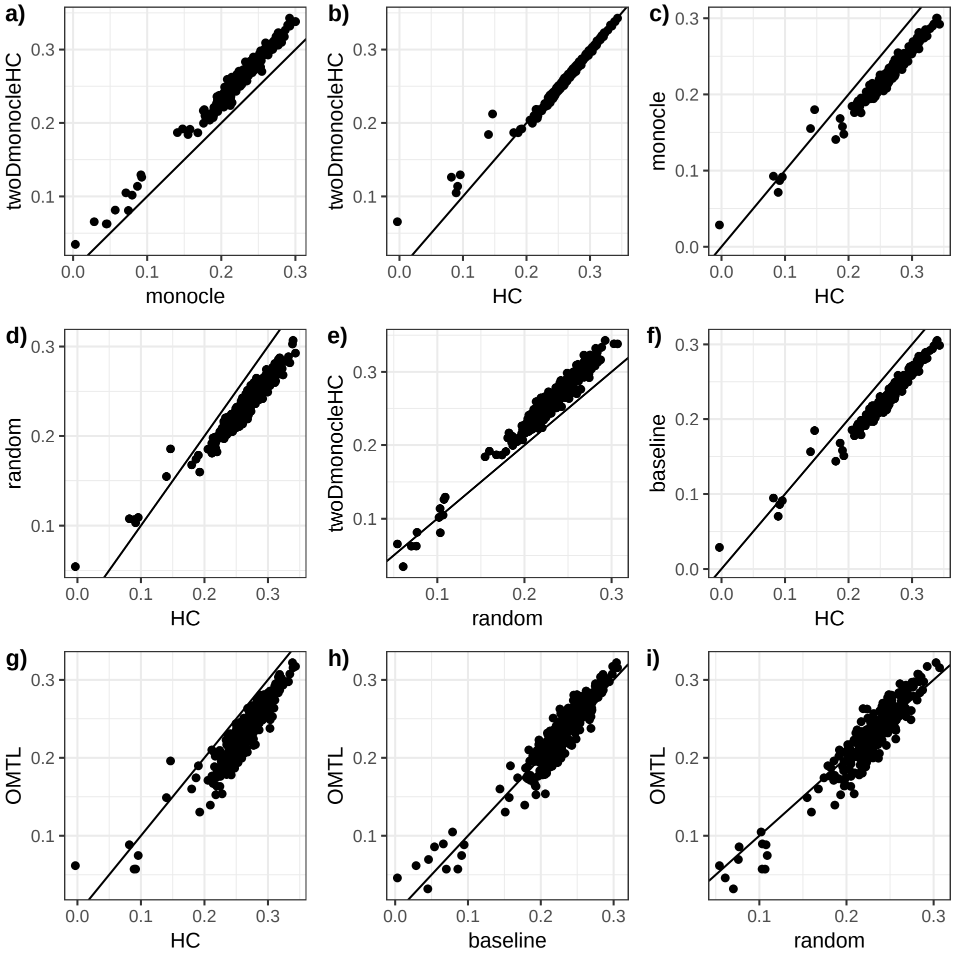


FIGURE 6.2: Comparison of MTL models on HSMM data. Each point in the scatter plot represents the Pearson correlation coefficient computed between the predicted and measured values of gene expression per cell. The diagonal line indicates the identity line to ease the comparison between the models placed on x and y axes.

In addition to the *OMTL* models, we also trained individual single-task models, by providing the gene expression profile per cell as the response variable of each model. Figure 6.3 illustrates the single-task learning (*STL*) framework. The predictions obtained from each individual model were later used to compute the correlation values between the prediction and actual measurements of gene expression.

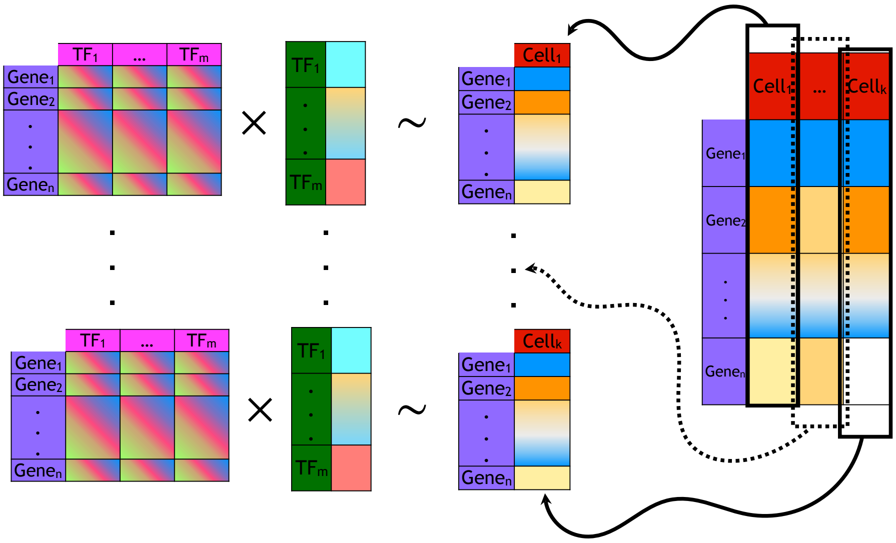


FIGURE 6.3: Schematic illustration of the single-task learning in the context of single cells. The vector of gene expression values measured in each single cell plays the role of the response variable in a distinct learning task. Ultimately, there are as many models as the number of single cells.

Finally, we designed another experiment that contrasts the performance of various tree-guided MTL models with single-task models obtained from individual builds of elastic net regularized univariate response models generated for each cell. Figure 6.4 illustrates the distribution of Pearson correlation coefficients obtained between the predicted and measured values of gene expression for each of the single-task and tree-guided multi-task learning with *HC*, *baseline*, and *monocle* provided as the guiding trees. These results further support the advantage of tree-guided MTL, in particular the *HC* setup, over the single-task models, where no information is shared among the tasks.

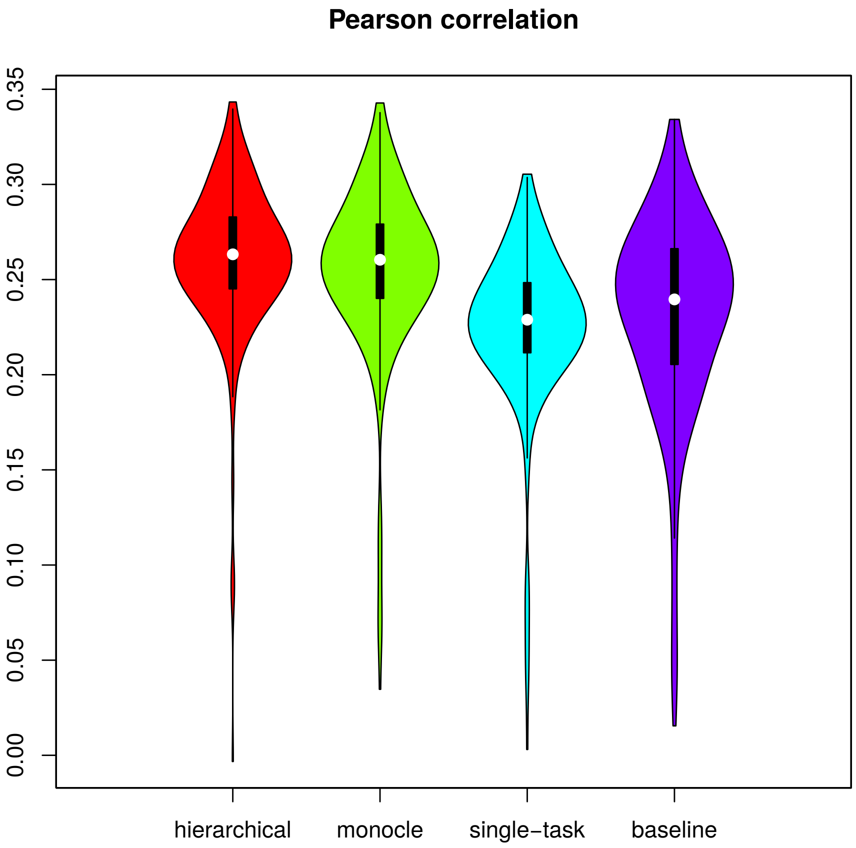


FIGURE 6.4: Comparison of single-task and tree-guided MTL models on HSMM data. The x-axis refers to three tree-guided MTL models (hierarchical (*HC*), *monocle*, *baseline*), as well as a single-task learning model. The y-axis shows the Pearson correlation coefficients between predicted and measured values of single cell gene expression on test data.

In conclusion, the tree-guided MTL indeed provides advantages compared to *OMTL* and the best choice for the tree is through conducting hierarchical clustering on the entire gene expression data (*HC*). Thus, we decided to generate the downstream tree-guided MTL models using *HC* as the guiding tree.

6.3.2 The impact of feature types on the prediction results

We wanted to explore the associations of gene expression in single cell to features that are independent of the cell content or configuration. Therefore, we designed a feature setup, which we named *static*, to link the cis-regulatory characteristics of ∼ 700 transcription factors with the gene expressions measured in single cell (see 6.2.1).

Figure 6.5 schematically illustrates the genomic area, where the *static*, *epigenetic*, and *dynamic* features are generated from. In *static* features, for each transcription start site of a gene, the TF binding affinities are measured within the 2 kb window around the TSS. These affinity scores are used to form the feature matrix for the *static* setup (Figure 6.5a). Figure 6.5b, illustrates the peaks obtained from the DNase-seq data used to identify the open chromatin regions in a 50 kb window around the TSS. The TF binding affinities are computed in the segments of this 50 kb window that correspond to the peaks. Finally, Figure 6.5c, shows the region in where the reads of ChIP-seq data of 123 TFs are counted. The resulting measurements form the third feature setup, *dynamic* features.

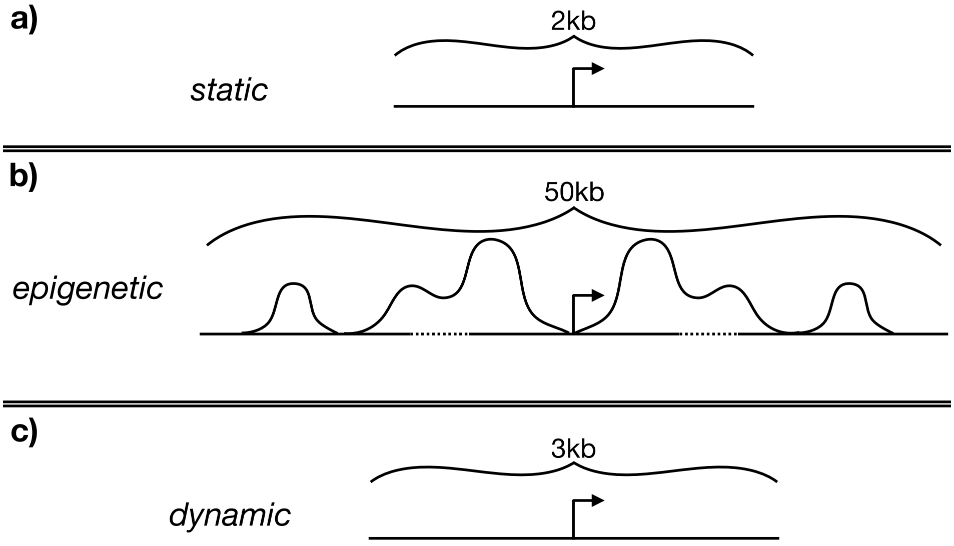


FIGURE 6.5: Genomic regions where the a) *static*, b) *epigenetic*, and c) *dynamic* features are generated from.

Figure 6.6 shows the Pearson correlation coefficients between predicted and measured values of gene expression obtained from the tree-guided MTL model trained on this setup. The results indicate that predicting the expression of Primary Human Hepatocyte (PHH) cells using the *static* features is more difficult than the expression of Hepatocyte Like Cells (HLC).

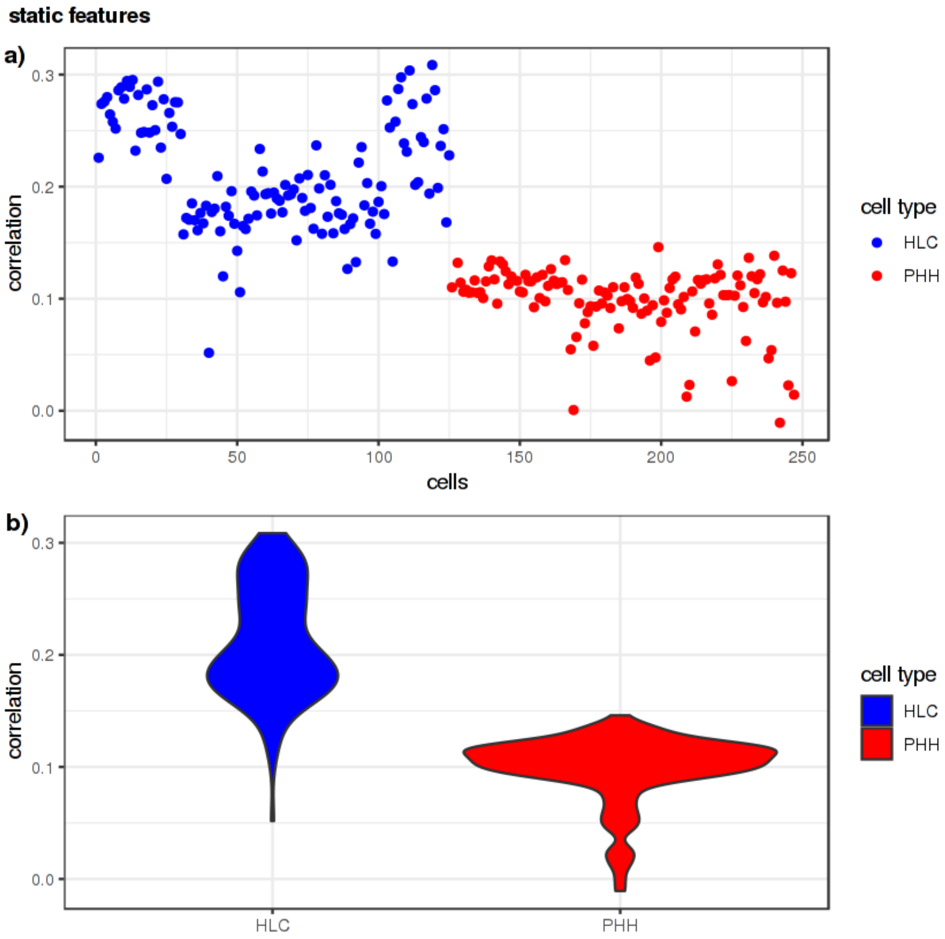


FIGURE 6.6: a) Scatter and b) violin plots illustrating the Pearson correlation coefficients obtained on the test partition of the iPSCs gene expression regarded as response and *static* features. The two subpopulations, HLC and PHH, are colored with blue and red, respectively.

Next, we deployed the *epigenetic* features (see 6.2.1) to see how the models perform when this type of feature is used. The *epigenetic* features represent the affinity binding of hundreds of TFs measured in specific genomic segments. These genomic segments are defined through the epigenetic signal of DNase-seq peaks. Therefore, they potentially should capture more specific associations between the TF binding affinity and chromatin openness. We built the tree-guided MTL models to predict the gene expression values in single cells using the *epigenetic* features. Similar to the *static* setup, we generated the correlation values to evaluate the model performance as shown in Appendix Figure D.1. It can be seen that, overall, the *epigenetic* setup improves over the *static* setup, especially for the PHH cells. In other words, there is a notable rise in the correlation values of both cell types, in particular the PHH cells.

We stretched our analysis by building another tree-guided MTL model that takes the *dynamic* features (see 6.2.1) to predict the single cell gene expression. The dynamic features hold the data for over 100 TF ChIP-seq reads that their corresponding genomic location overlapped with the 3 kb window centered at the TSS of the genes. This setup, interestingly, resulted in the best performance accuracy in comparison to the formerly mentioned setups (Appendix Figure D.2).

Even though the top performing model was obtained from the *dynamic* features, we favored to present the results for the *static* features in the main text of this thesis. Because, the *static* features are independent of any cell type or tissue and can be coupled with any available single cell expression data. However, on the contrary, the other two feature types (*epigenemic* and *dynamic*) are specific to a given cell type, preferably the one that matches best the gene expression data.

6.3.3 Imputation generally improves the accuracy

The results shown in the previous section, 6.3.2, are performed on the original unaltered expression data. However, we were curious to find out how the results will change when we impute the missing values, which were introduced by the dropout effect. Therefore, we imputed the data using the scImpute tool (Li and Li, 2018a) and repeated the experiments described in 6.3.2 with the difference of using the imputed expression values as the response matrix. Appendix figures D.3, D.4, and D.5 exhibit the prediction accuracy, measured in terms of Pearson correlation coefficients between the imputed gene expression and the predicted values for the *static*, *epigenetic*, and *dynamic* feature setups, respectively. Figure 6.7 provides an overview of the performance of the tree-guided MTL models on the three feature setups as well as the imputation status of the expression data, imputed or not imputed. These results reveal that, not surprisingly, the imputation enhances the prediction accuracy, regardless of the feature setup. It is interesting to observe that for the *epigenetic* setup, not only the correlation values are increased, but also the distribution of these values is changed in favor of having a smaller variance across the cells. The change of distribution is notable for the other two setups as well, but that does not necessarily lead to a smaller variance.

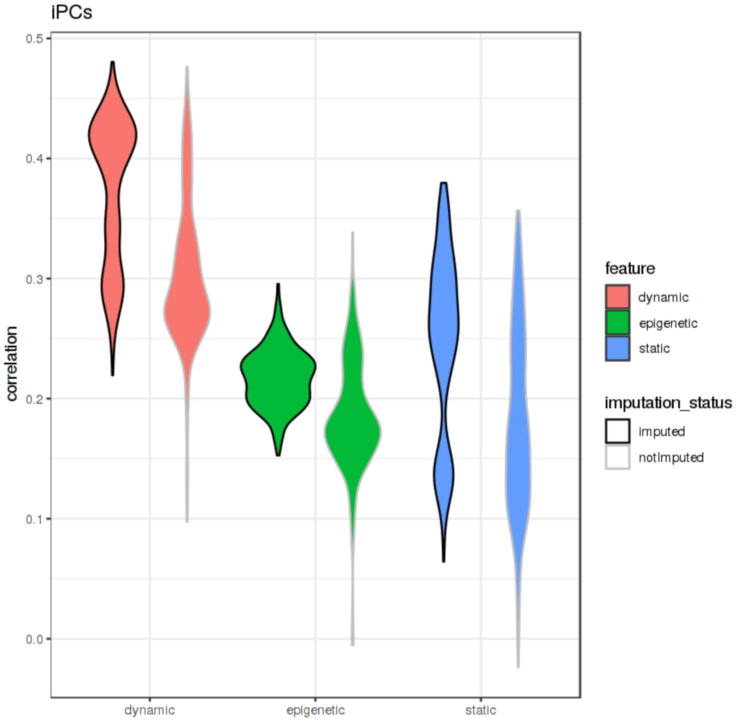


FIGURE 6.7: Comparison of all tree-guided MTL models on the test partition of the iPSCs data. *Dynamic*, *epigenetic*, and *static* setups are colored with red, green, and blue, respectively. The imputation status is indicated by different border colors, black for imputed and gray for not imputed.

6.3.4 Distinct regulatory elements are attributed to the HLC and PHH cell types

Observing such difference in the prediction accuracy, triggered our curiosity in inspecting the model coefficients that correspond to the TFs in cells. The heat map in Figure 6.8 depicts the coefficients of the top features (see 6.2.6) derived from the tree-guided MTL model trained on the *static* features to predict the gene expression in iPS cells.

In this heat map, it can be noted that, firstly, the cells are very well clustered according to the model coefficients, as the built-in hierarchical clustering in the heat map function nicely arranged the HLC and PHH cell types separately from each other. Secondly, these results show certain blocks of TFs playing distinct roles in regulating the gene expression in single cell. For instance, the transcription factor YY2 holds positive coefficient values for the HLC cells, whereas its coefficient values for the PHH cells are negative. This is interesting, since YY2 has a dual affect on gene expression, i.e., it can both repress and activate transcription (Nguyen et al., 2004).

On the other hand, HNF1A, which is essential for the expression of various liver-specific genes, was considered irrelevant for the HLC cells by the model, as it has assigned zero to coefficients corresponded to this particular TF on these cells. However, HNF1A holds positive coefficient values for the PHH cells.

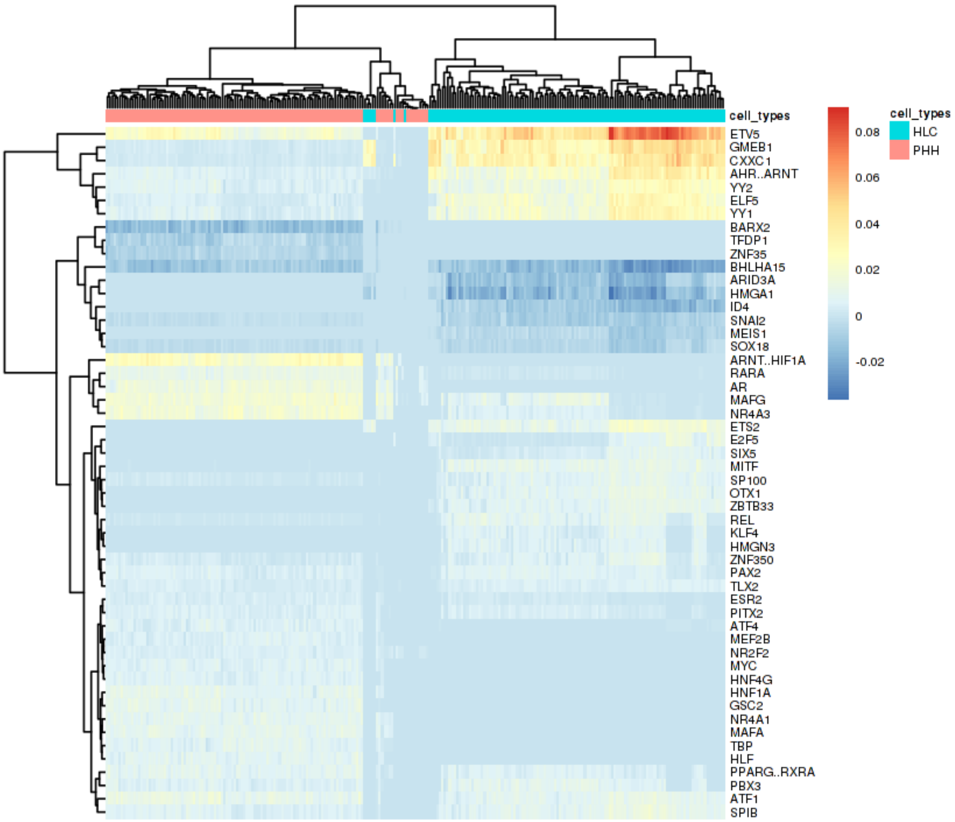


FIGURE 6.8: Heat map illustrating the top features (see 6.2.6) derived from the tree-guided MTL trained on the *static* features to predict the unimputed gene expression in iPS cells.

Similar trends, in terms of spotting distinct cell type specific TFs among the model coefficients, was also observed in other feature setups, as well as models trained on the imputed gene expression data. The heat maps representing the coefficients of top features for the *epigenetic* and *dynamic* features linking the unimputed data are provided in Appendix figures D.6 and D.7, respectively. The results obtained on the imputed data are shown in Appendix figures D.8, D.9, and D.10.

6.3.5 Tree-guided MTL outperforms single-task learning models

Our last experiment is devoted to investigate the two distinguishable learning approaches, multi-tasking versus single-tasking, trained based on the three feature types. Similar to before, the tree-guided MTL models guided by the *HC* tree structure are trained on the *static*, *epigenetic*, and *dynamic* features, separately, to predict the imputed and unimputed gene expression measurement in single cell. Simultaneously, the single-task models are trained using the elastic net regularization, where gene expression measurements per cell form a univariate response vector, resulting in as many regression models as the number of cells. Figure 6.9 provides the scatter plots representing the Pearson correlation coefficients between the measured and predicted gene expressions in single cells.

We can analyze the results shown in these plots in two different ways. Firstly, it is apparent that the tree-guided MTL models most often outperform the single-task learning elastic net models for all features setups of the unimputed (notImputed) scenarios. We additionally calculated the ratio of cells, where the tree-guided MTL models outperformed single-task models for each feature setup on the unimputed

Feature	Ratio
static	0.87
epigenetic	0.70
dynamic	0.56

TABLE 6.1: Ratio of cells for which the corresponding tree-guided MTL outperformed the competing single-task models on the unimputed data.

data, as displayed in Table 6.1. This further supports the advantage of using a learner that is able to share the information across different tasks (single cells in this particular problem).

Secondly, when the imputed data is considered, this advantage becomes strikingly prominent as, except for very few cells, the MTL models undoubtedly outperform the *STL* ones for all feature setups. This observation hints at the better generalizablity of the models when less noisy data is supplied. The imputation indeed improved the performance of the *STL* models compared to the unimputed scenario, by having higher correlation values and less variance across the cells, but the MTL models could better exploit this improvement in the data quality, and enhance their prediction accuracy.

These results conclude that irrespective of the feature type and imputation status of the gene expression data as the response, the MTL models are superior to the *STL* models.

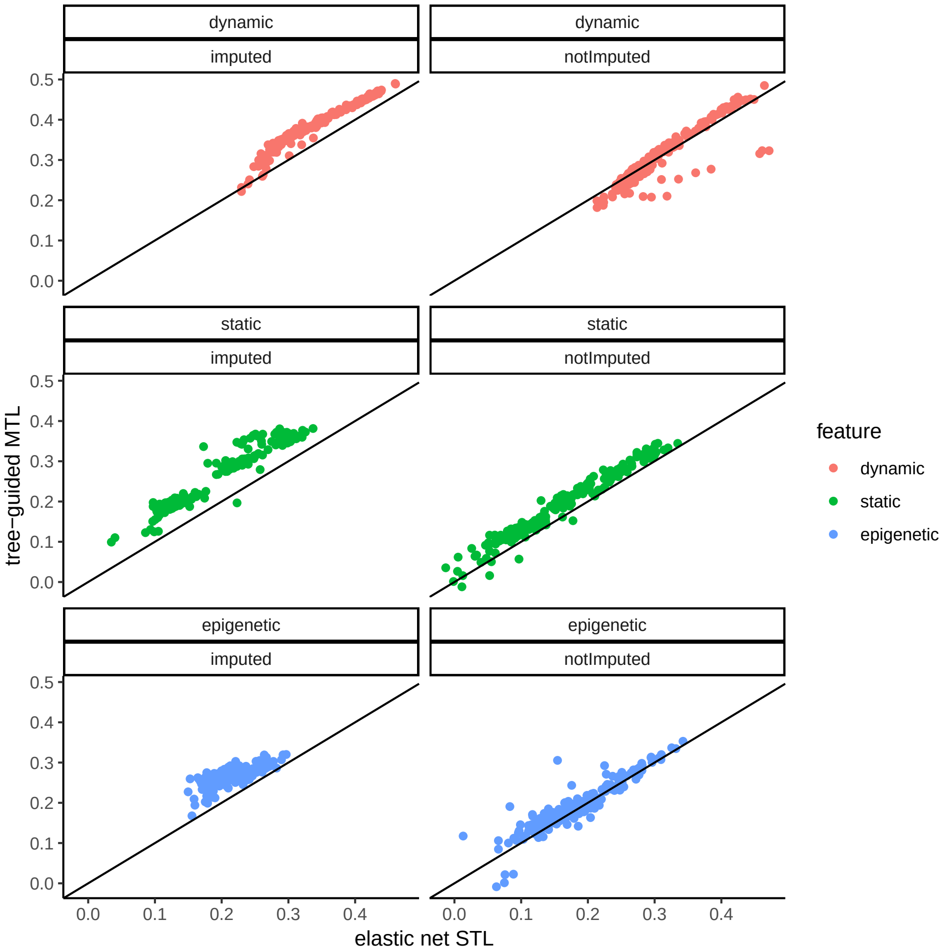


FIGURE 6.9: Comparison between single- and multi-task learning models trained on either of *static*, *epigenetic*, or *dynamic* feature setups. The response variables are the single cell gene expressions with or without imputation procedure indicated by *imputed* and *notImputed*, respectively.

Conclusion and discussion

In this work, we utilized the single cell RNA-seq data partly from the HSMM cells obtained from Trapnell et al., 2014, and mainly from the induced pluripotent stem cells data set provided by Kathrin Kattler from the lab of Prof. Dr. Jörn Walter (Saarland University) in cooperation with Prof. Dr. Jan Hengstler's group (IfADo Dortmund). We built several statistical models that intrinsically address the dropout issue and simultaneously find associations between the single cell RNA-seq expression measurements and various transcriptional characteristics, such as transcription factor binding affinities and ChIP-seq signals.

The discrepancies observed among the gene expression profiles in single cells, trivially, hints at the existence of specific differences in the transcriptional regulatory mechanism. This mechanism can involve the static cis-regulatory characteristics all the way up to the more elaborated features such as the chromatin accessibility and transcription factor binding. Devising computational methods that are able to infer associations between gene expression in single cells and cis-regulatory motifs as well as epigenetic characteristics has attracted the attention of researchers in the field.

There have been studies focusing on linking the transcriptional regulatory elements in single cells (Mohammadi et al., 2018; Aibar et al., 2017; Suo et al., 2018).

Aibar et al., 2017 relies on the gene expression of several transcription factors to infer cell-specific associations between the expression of other genes and those TFs. In contrast to their approach, we computed the binding affinities of hundreds of TFs, in addition to exploiting TF ChIP-seq data to infer such associations. We found the multi-task learning approach a suitable machine learning candidate to conduct our computational experiments. We employed the tree-guided MTL method to benefit both from the information sharing delivered by multi-tasking and also grouping the cells according to the tree structure provided as an additional input.

We designed three different feature setups, *static*, *epigenetic*, and *dynamic* to inspect the association of cis-regulatory and epigenetic features with cell-specific transcription. The connection was established between these feature setups and either the imputed or unimputed gene expression values in single cells through various statistical learning models. Even though the main focus was on the tree-guided MTL framework, we explored the prediction accuracy on ordinary MTL (without a tree structure) and single-task learning using the elastic net regularization.

In order to decide what tree structure to choose for training the tree-guided MTL models, we tested various relevant choices of which we settled on the *HC* case. The results were compared with the models trained on the same data using *baseline* and *random* trees for further validation of the goodness of our choice.

The tree-guided MTL models using the *HC* tree structure trained on each of the *static*, *epigenetic*, and *dynamic* feature setups predicting the iPS single cell expression led to multifaceted results. Firstly, we noticed that the prediction accuracy was in general higher for the HLC cells compared to PHH cells. To further investigate this peculiar finding, we computed the Pearson correlation between the individual features of each feature setup and the gene expression separated by the cell type (PHH or HLC) as shown in the Appendix figures D.12, D.13, and, D.11. These correlation values appear to be in favor of the HLC cells, as regardless of which feature setup considered, the distributions tend to exhibit larger values for the HLC cells.

Not only we strove for obtaining improvements in model accuracy by using the tree-guided MTL approach, but also were interested in acquiring more interpretable results. By inspecting the coefficients of our models, we were able to pinpoint distinct transcription factors that show cell-type specific regulation in iPS cells.

In view of the fact that the MTL shares the information among the tasks to facilitate the learning procedure, we speculated that the concern regarding the dropout issue in single cell data would be intrinsically addressed. To evaluate whether this speculation was justified, we compared the prediction accuracy with the counterpart learning approach of MTL, which is the single-task learning approach. The results, in fact, support our speculation by revealing the superiority of MTL models over the STL ones. Therefore, even though our approach is different to imputation methods, but because it leverages the idea of sharing the cell similarity, it can actually improve the prediction of gene expression. However, the run-time complexity depends heavily on the number of cells and as this number grows (> 2000), the current implementation becomes prohibitively slow.

As a future work we think that we can extend the model to include single ATAC-seq data, upon availability, to attain *epigenetic* features that are estimated in accessible chromatin regions defined through higher resolution assessments. In addition, the feature setup can be further extended to embrace more diverse transcriptional regulatory characteristics, such as annotated enhancers as opposed to promoter regions. Likely, our proposed approach could be even more powerful if we had modeled the noise in the single cell data into our objective function. In the end, we believe that the single cell sequencing technology has paved the way for excelling our understanding of gene regulation at more fine grained levels.

Chapter 7

Summary and conclusion

Many phenotypic and genotypic disorders can be traced back to aberrant expression of certain genes. What goes wrong in the cell that results in such abnormalities has always concerned the biologists and physicians. Through generating novel hypotheses for gene regulation, the path will be paved to develop appropriate strategies for prevention or treatment of diseases. Vast amount of research has been gone into unraveling the mystery of gene regulation (Cramer, 2019; Smith and Flodman, 2018). However, each study can only address a few pieces of this expansive puzzle of gene regulation. Throughout this dissertation, we presented our contribution by tackling interesting problems related to understanding the gene regulation mechanism.

In Chapter 3, we explained the problem of TF binding site (TFBS) prediction, which was pursued in the context of the DREAM challenge competition (ENCODE-DREAM, 2017). We described a random forest (RF) based ensemble learning framework that allowed us to predict the TFBS in predefined genomic bins of several human cell types. Our approach demonstrated its strength, in particular, in the multi tissue settings, by determining the cofactors associated to the TF of interest (TF for which its binding sites were to be predicted). In addition, our proposed method was capable of better generalizing across samples, in comparison to models that were trained only on a single sample. We further explored the relation between the performance of our RF classifiers and the number of tissues used for training. The main obstacle on the way of building our computational models for this specific task was the imbalanced nature of the classes, *Bound* and *Unbound*. Therefore, precautions were required to be taken into consideration to prevent the models being biased towards the major class. In summary, we showed that modeling cofactors can be instrumental in predicting TFBS and that ensemble learning is a promising approach to gain generalizability across tissues.

In the subsequent two chapters, 4 and 5, we targeted gene regulation in bidirectional promoters. These are a class of gene pairs that are located in proximity to each other but on the opposite strands of DNA (plus and minus strands). These genes are, in particular, interesting because the promoter architecture embracing these genes allows miscellaneous patterns of gene expression resulting from a disjoint or coupled regulation. As it was previously shown that the histone modification data can be used to build accurate models for predicting gene expression (Karlić et al., 2010), we decided to leverage such data for studying the gene regulation in bidirectional promoters.

In Chapter 4, we discussed our feature design strategy tailored for addressing this particular problem. Unlike Karlić et al., 2010, we measured the histone modification read counts in bins of 100 bp spanning the promoter region in order to capture the spatial distribution of the signal to be later used in our machine learning framework. We exploited the fused LASSO algorithm that is able to provide interpretable models when correlated features exists in the data. Through a series of simulation case studies, we investigated the behavior of the models by varying the associations between the feature and response variables fed to the models for training. As the results of simulation analyses were convincing and promising, we applied our approach on the real histone modification and RNA-seq data provided by the DEEP consortium. Through generating multiple fused LASSO models, each trained on a different cell types, we were able to build a histone map extracted from the model coefficients, for all these cell types. The prominent trend observed in this map suggested a unidirectional association of histone marks to the expression measured at the bidirectional promoters.

We continued studying this peculiar gene regulation mechanism at bidirectional promoters in Chapter 5 by utilizing the single cell RNA-seq data. Using the single cell gene expression profiles in bidirectional promoters, we derived three different genomic and epigenomic architectures that were specific to these promoters. We concluded our findings by proposing a hypothetical model that describes these architectures. This model supports distinct characteristics of HM abundance, DNA accessibility, TF abundance, TSS distance between gene pairs, as well as the *transcripts span* of these genes pairs (Behjati Ardakani et al., 2018).

Thrilled by the potentials single cell sequencing has, we stretched our interest in understanding the gene regulation mechanism by developing a novel statistical framework tailored for single cell data. As provided in Chapter 6, we described our work, which is a general framework for establishing cell specific associations through employing a tree-guided multi-task learning (MTL) algorithm (Kim and Xing, 2010). Despite other related approaches (Mohammadi et al., 2018; Aibar et al., 2017; Suo et al., 2018), we incorporated the information shared among the single cell gene profiles using our multi-tasking approach. We examined the results obtained from different tree structures used for grouping the cells. The tree structure that provided the most promising results was selected to be used for the rest of the analysis. We designed three different feature setups, *static*, *epigenetic*, and *dynamic* reflecting genetic and epigenetic characteristics. We then established the connection between these feature types and either the imputed or unimputed gene expressions in single cell. In order to showcase the power of our tree-guided MTL design, we additionally compared the prediction accuracy on ordinary MTL (without a tree structure) and single-task learning using the elastic net regularization. By inspecting the coefficients of our models trained on induced pluripotent stem cells with two annotated cell types, HLC and PHH, we were able to identify distinct transcription factors that appear to be cell type specific in these cells.

The approaches presented herein provide new integrative means for studying the gene regulation mechanism. Through exploiting or tailoring already existing computational methods, we were able to achieve new insights into the genetic and epigenetic characteristics involving gene expression. In order to grasp a comprehensive overview of this complex mechanism of gene expression, it is essential to combine various sources of data reflecting the genetic and epigenetic markup of a cell. In particular, with the advances in single cell sequencing, we believe that integrating different single cell sequencing assays can result in revolutionary leaps in understanding the gene regulation at the single cell resolution.

Appendix A

Supplementary materials for Chapter 3

A.1 Supplementary Figures

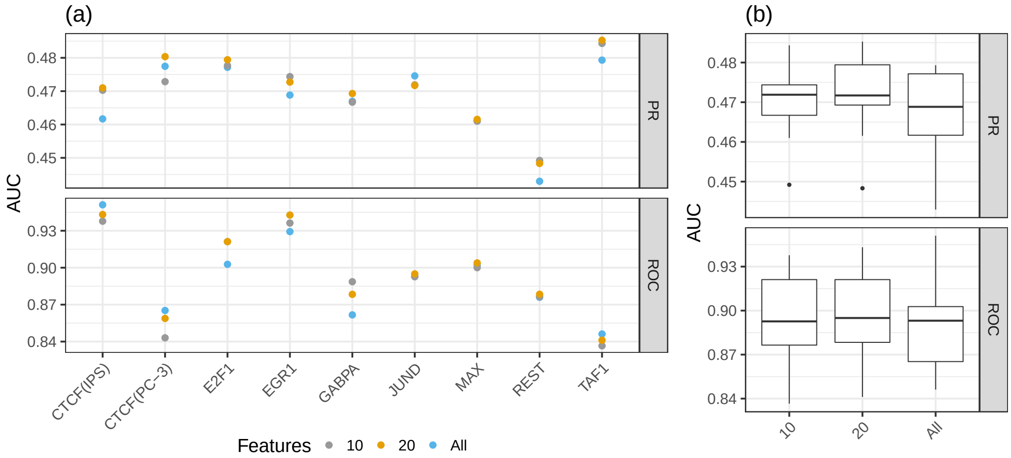


FIGURE A.1: PR-AUC and ROC-AUC for different sets of features: considering *all* features, the top 10, and the top 20 features on several test tissues. One can see that the there is a slight advantage for the top20 and top10 model over the full model in these scenarios. The performance is shown for individual tissues in (a) and separately for the size of the feature matrices in (b).

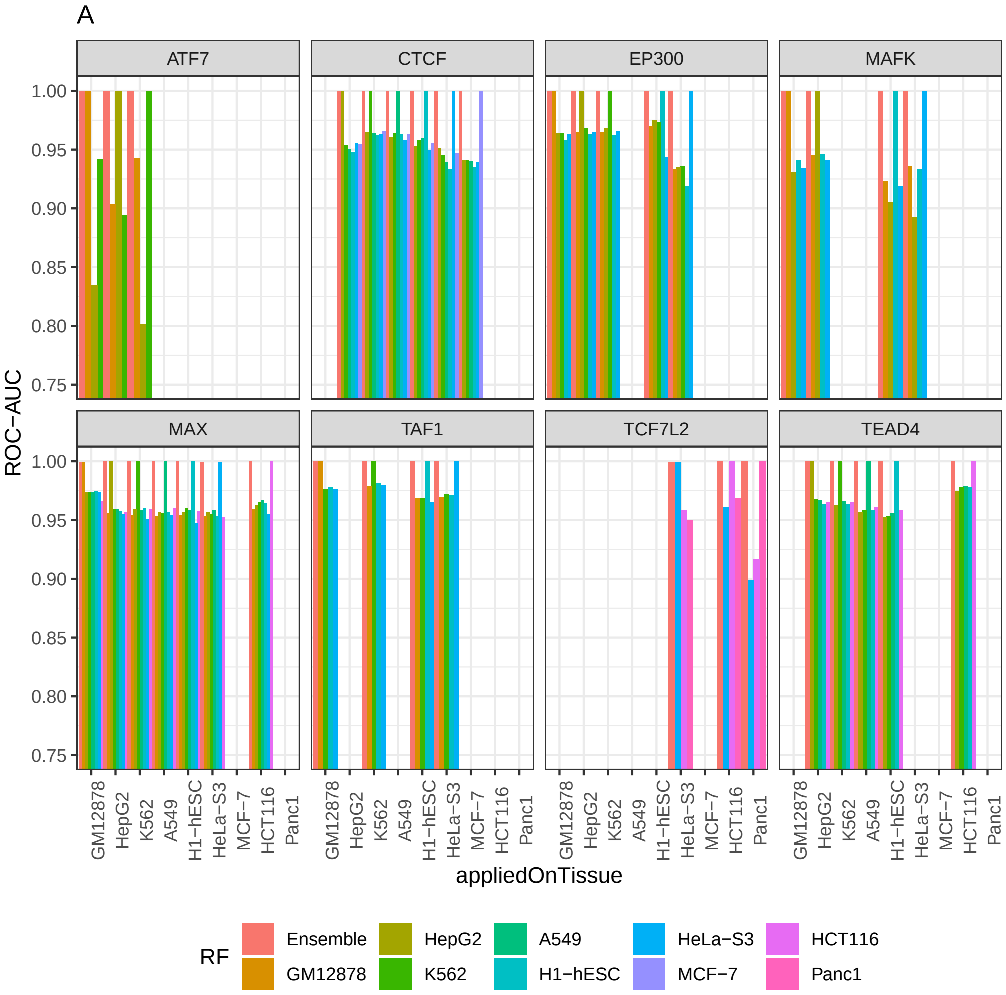


FIGURE A.2: Within and cross tissue comparisons for ensemble and tissue specific RFs. Model performance is assessed in terms of (a) ROC-AUC and (b) PR-AUC.

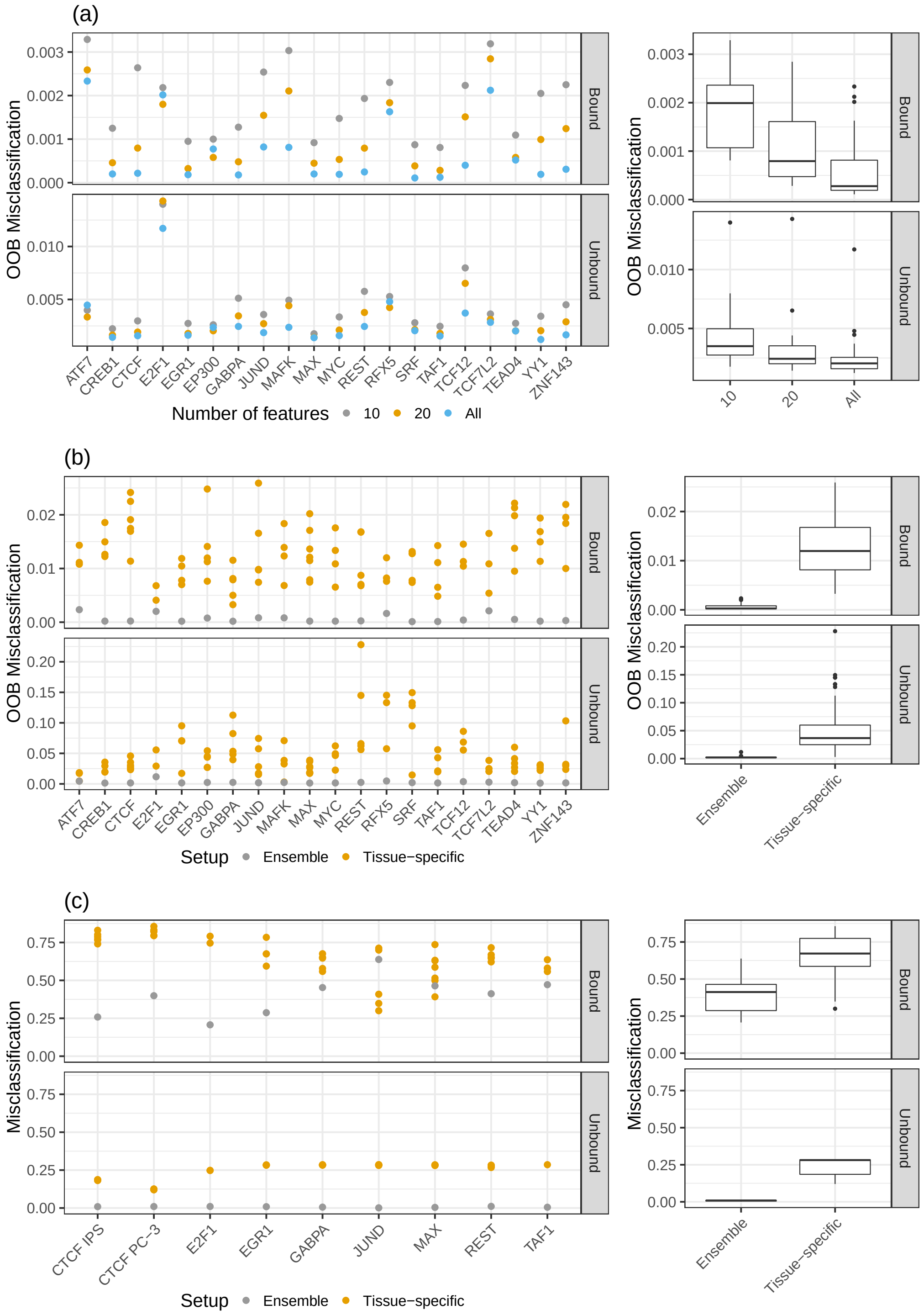


FIGURE A.3: a) Classification error for the *Bound* and *Unbound* classes for different sets of features: considering *all* features, the top 10, and the top 20 features. One can see that the difference in model performance between the top 20 and *all* feature cases is only marginal. b) Comparison of the out of bag (OOB) error between ensemble models and tissue-specific random forest (RF) classifiers. Especially in the *Unbound* case, the ensemble models show superior performance compared to the tissue-specific RF classifiers. c) Misclassification rate computed on unseen test data for ensemble and tissue-specific RF classifiers. As in b) we see that the ensemble models generally outperform the tissue-specific ones. Note that the scale of the y-axis is different for the *Bound* and *Unbound* classes in (a) and (b).

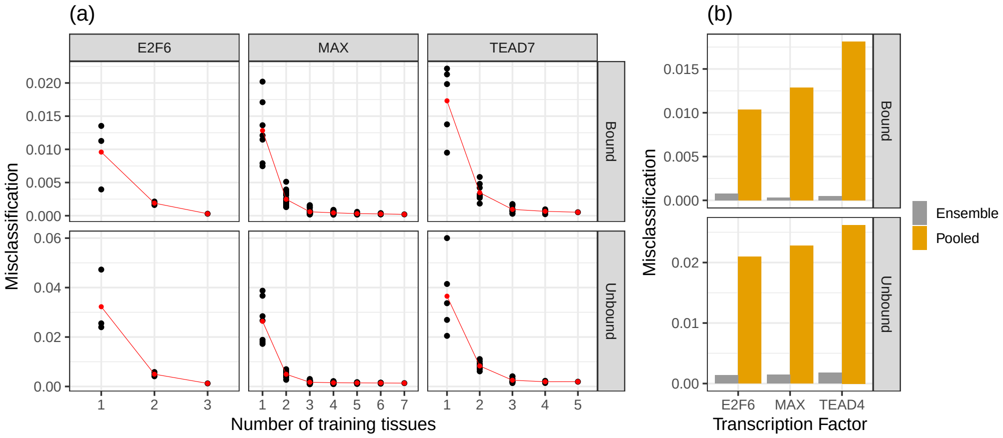


FIGURE A.4: a) Relation of the OOB error for three TFs (E2F6, MAX, and TEAD4) to the number of tissues used for training. The OOB reduces if more tissues are included in the ensemble learning. Red dots represent the mean classification error across all tissue-specific classifiers. Individual models are represented by the black points. b) Comparison between true ensemble models for E2F6, MAX, and TEAD4 and RF classifiers trained on pooled data sets comprised of training data for all available tissues. The ensemble models perform better than the models based on aggregated data.

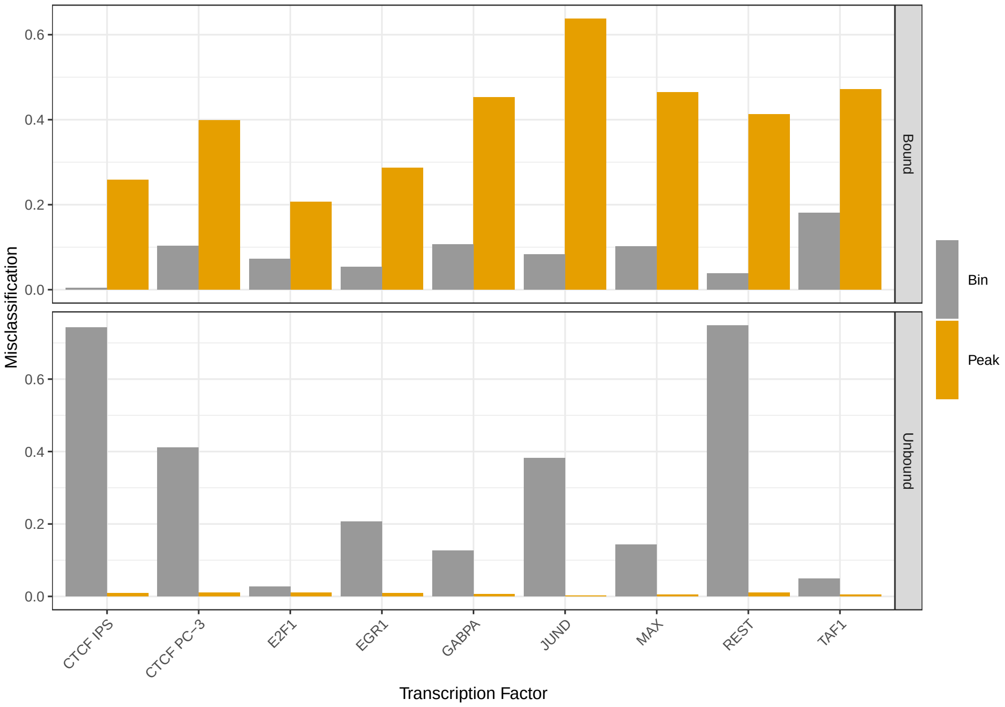


FIGURE A.5: Comparison of misclassification rate depending on the feature design computed on test data.

Appendix B

Supplementary materials for Chapter 4

B.1 Supplementary Figures

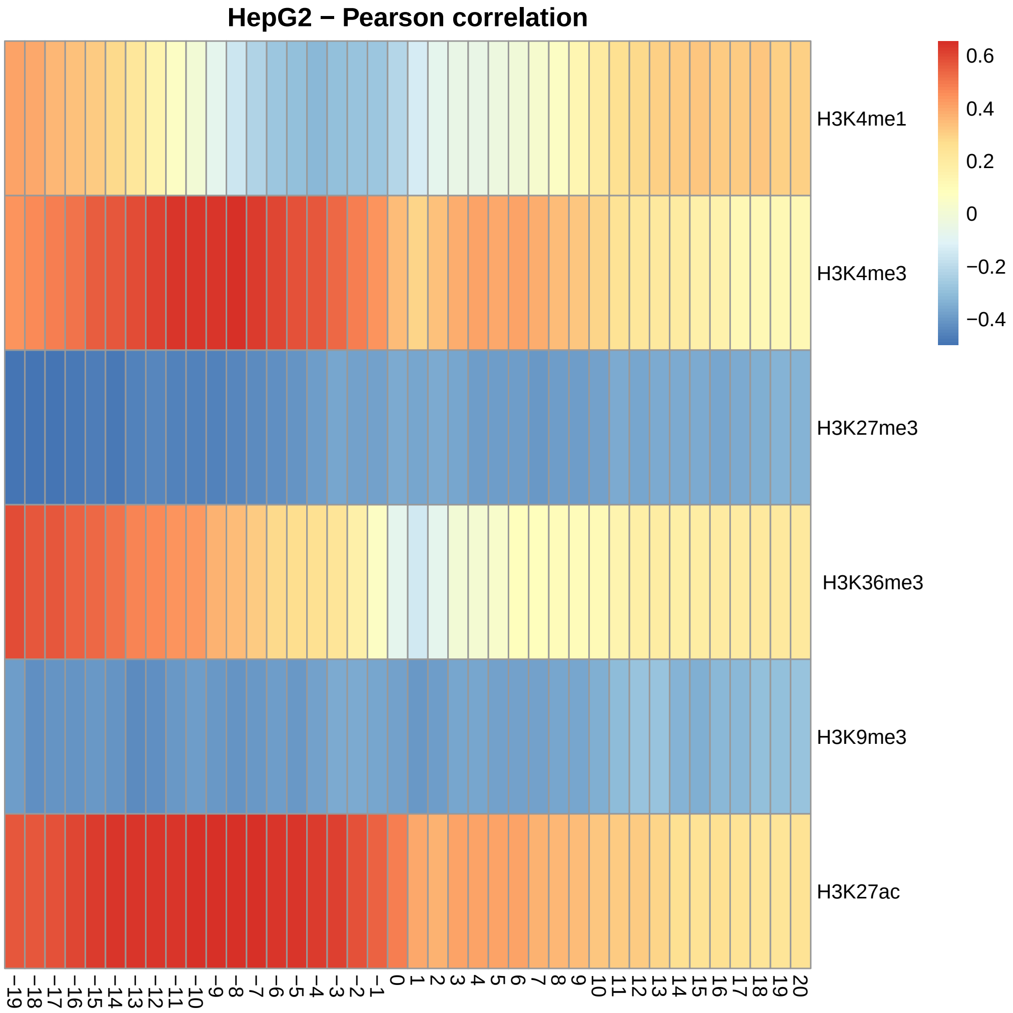


FIGURE B.1: Heat map showing the Pearson correlation coefficients between bins of six histone modification data and minus gene expression for cell line HepG2. Both HM and expression data are log-transformed. Each cell of the heat map correspond to the bins described in 4.3. 0 indicates the TSS of the minus gene, where the window of size 4 kb is anchored on it.

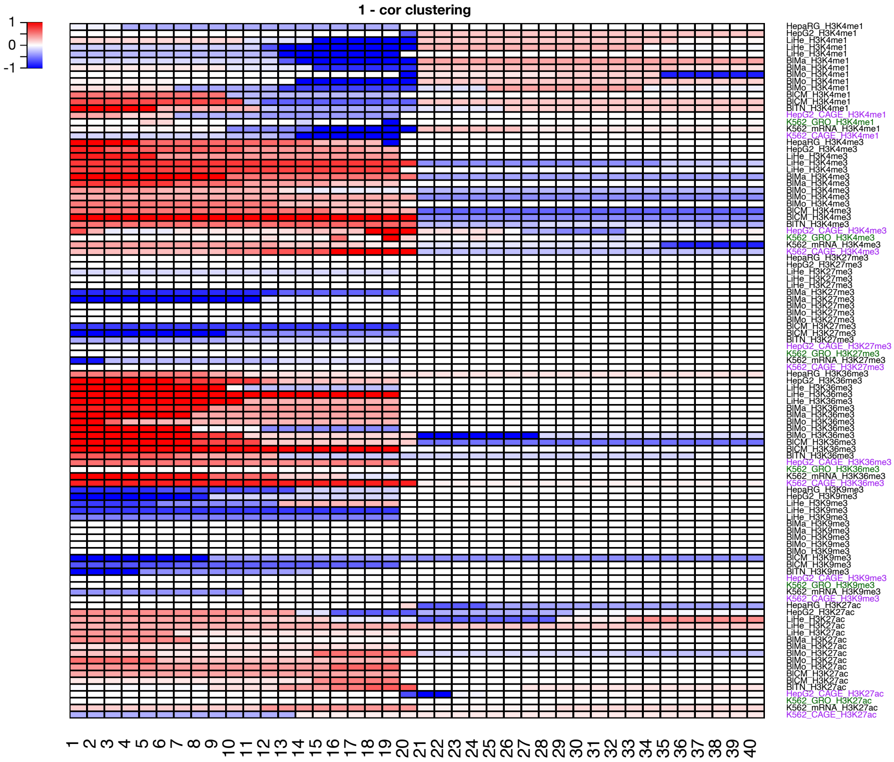


FIGURE B.2: The histone map. Fused LASSO coefficients obtained from individual models learnt on the cell types described in 4.2.1 for the minus gene. The values are scaled between -1 (blue) and 1 (red) to ease the comparison across samples. The expression assays, RNA-seq, CAGE, and GRO-cap, are color coded by black, purple, and green, respectively. The heat map suggests a unidirectional localization of histone marks coinciding the direction of transcription.

Appendix C

Supplementary materials for Chapter 5

C.1 Methods

Single cell RNA-seq

Single HepG2 cells were manually picked to prepare poly-A enriched cDNA libraries using Smart-seq2 as described by Picelli et al., 2014 with modifications. Briefly, 65 single cell samples were supplemented with 0.5 μl of a 1:40,000 dilution of the Ambion ERCC RNA Spike-In Mix 1 (Thermo Sientific, #4456740). After cell lysis polyadenylated mRNA was reverse transcribed using a biotinylated template switch oligo (5′-Biotin-AAGCAGTGGTATCAACGCAGAGTACATrGrG+G-3′) with two riboguanosines (rG) and one LNA-modified guanosine (+G) at the 3′ end. Preamplified cDNA (18 PCR cycles) was purified with Agencourt Ampure XP Beads (Beckman Coulter, #A 63881) in a 1:1 ratio. cDNA quality of 8 random samples was assessed on the Agilent 2100 Bioanalyzer (Agilent Technologies, #G2938C) using the Agilent high-sensitivity DNA kit (Agilent Technologies, # 5067- 4626). Sequencing libraries were prepared using the Nextera XT DNA Sample Preparation Kit (Illumina, #FC-131- 1024) with approximately 480 pg of cDNA in a 4 μl tagmentation reaction followed by a dual indexing PCR with 9 cycles. Individual single cell libraries were pooled and purified with 0.8 X Agencourt Ampure XP Beads. The library pool was sequenced on a HiSeq 2500 (Illumina) using the TruSeq SBS Kit v3-HS (Illumina, #FC-401- 3001) in a single read run with 90 bp read length.

Bulk RNA expression quantification

BAM files of RNA-seq reads for HepG2 were produced using TopHat 2.0.11 (Kim et al., 2013), with Bowtie 2.2.1 (Langmead and Salzberg, 2012) and NCBI build 37.1 with parameters: *–library-type fr-firststrand* and *–b2-very-sensitive*. Cufflinks was used for gene expression computation (Trapnell et al., 2012) using GENCODE release 19 (GRCh37.p13).

Mapping of ChIP-seq data

Reads were mapped to the 1000 genomes phase 2 assembly of the human reference genome (NCBI build 37.1, downloaded from `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/`) with a hardware-accelerated implementation of Burrows-Wheeler Aligner BWA aln version 0.6.2 (Liu et al., 2012) with -q 20, and BWA 0.6.2 sampe with -a 1000. Merging and duplicate marking was performed with Picard version 1.125[1].

Imputation of dropouts

The scRNA-seq expression data were imputed using the scImpute tool. To observe how the values before and after imputation changed, we plotted the expression of genes in each cell for raw read counts (y-axis) and imputed ones (x-axis), as illustrated in Figure C.1

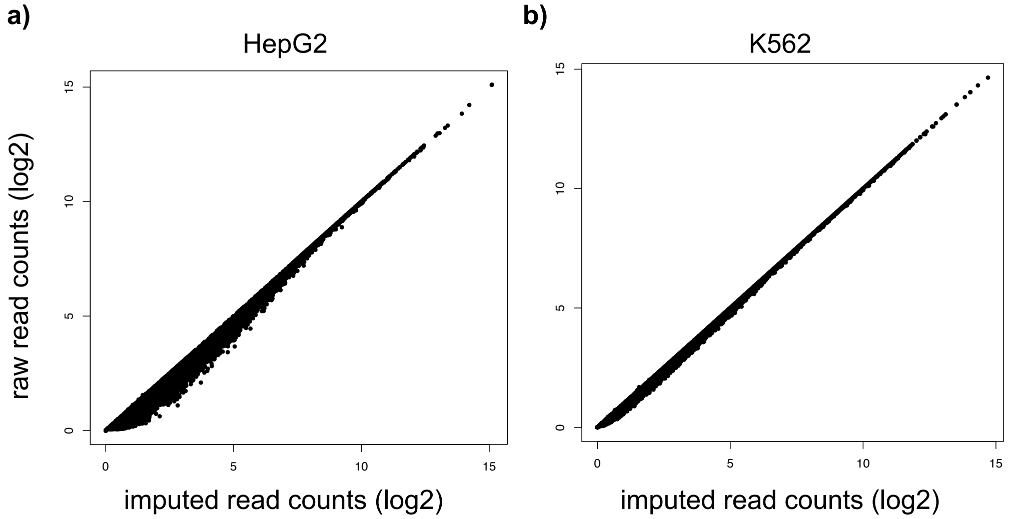


FIGURE C.1: Comparison between raw and imputed read counts for both cell lines, a) HepG2 and b) K562.

Quality of scRNA-seq

As part of our quality control procedure, we measured the correlation between the average imputed single cell expression and bulk measurements for both HepG2 and K562, (Spearman correlation coefficient of ∼0.8, Appendix Figure C.2)

In order to account for the number of genes falling into three intervals, $1 < \text{TPM} < 10$, $10 \leq \text{TPM} \leq 100$, $\text{TPM} > 100$, the imputed TPM values were divided into those intervals per cell, as shown in Figure C.3a. We also performed this partitioning based on the imputed read counts and provided the results in Figure C.3b.

Measuring 3′UTR length in BPs

3′UTR coordinates for our BPs were retrieved from annotated ENSEMBL genes (GRCh37.75) to show the 3′UTR length of the highly and lowly expressed genes, particularly, in the stable and unstable categories (see 5.2.7) as illustrated in Appendix Figure C.4c. The Mann-Whitney test was used between the highly and lowly expressed genes within each category to compute the p-values with the 0.05 for significance calling.

[1] http://broadinstitute.github.io/picard

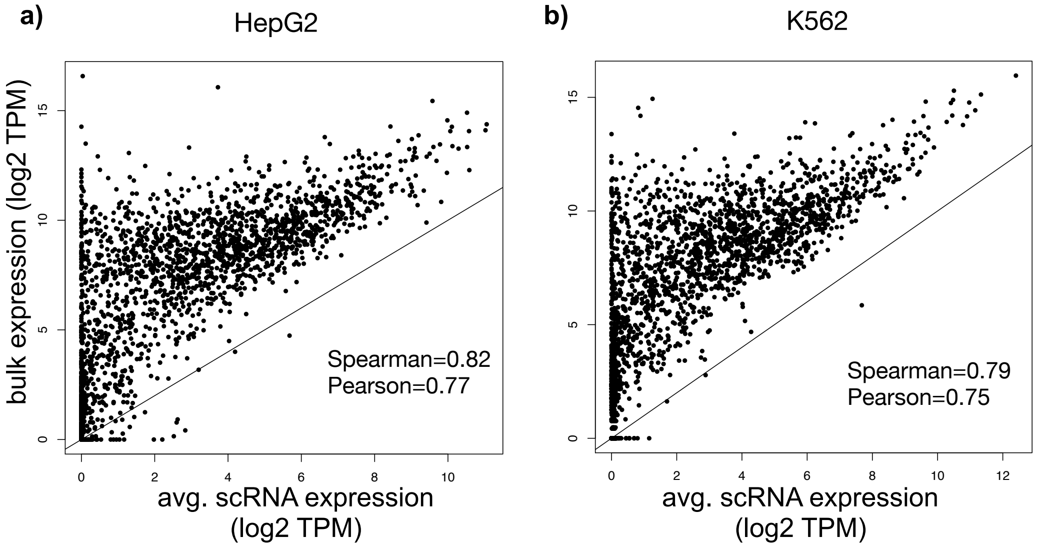


FIGURE C.2: Contrasting average imputed scRNA-seq with the corresponding bulk expression for both cell lines, a) HepG2 and b) K562.

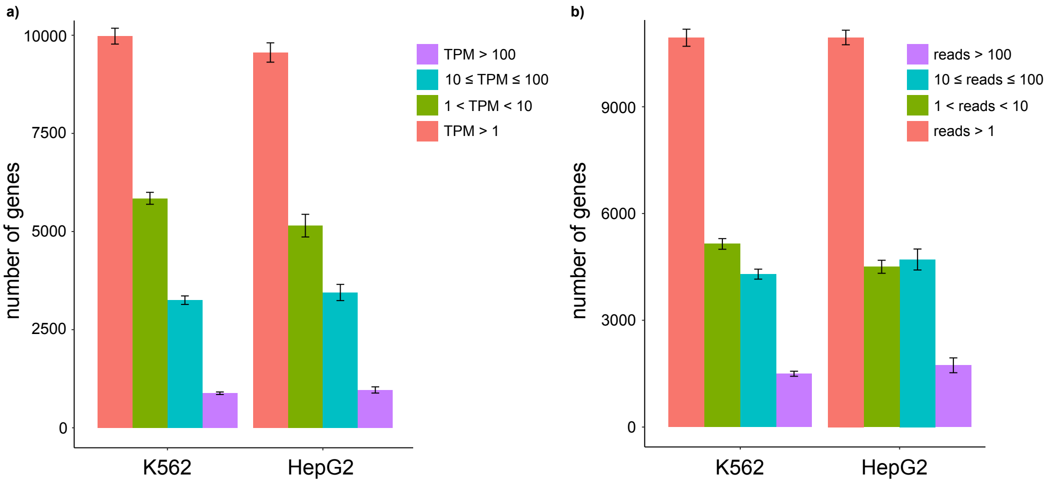


FIGURE C.3: a) Number of expressed genes according to the four intervals defined on TPM values per cell. TPM > 1 shows the overall number of genes detected to have expression higher than 1 TPM. On average, there are over 7,000 genes expressed for both cell lines. b) Similar to (a) but for read counts.

Measuring H3K36me3 in *transcript span* of BPs

The H3K36me3 ChIP-seq reads are counted in the region starting from the TSS of a bidirectional gene extending down to the *transcript span* partitioned into 10 bins. It is worth noting that the bin sizes might vary between genes as they have variable *transcript span* lengths. Therefore, read counts are normalized according to the bin size.

C.2 Supplementary Tables

	BLE	*BSD*	*BWD*	*BND*
NC→NC	79*	3	9	3
NC→PC	255	21	75	5
PC→NC	141*	6	35	2
PC→PC	395	35	153*	25*

TABLE C.1: Number of BPs falling into the gene product categories (NC→NC, NC→PC, etc.) in K562. Statistically enriched values are marked with $*$ (Hypergeometric test $p<0.05$).

C.3 Supplementary Figures

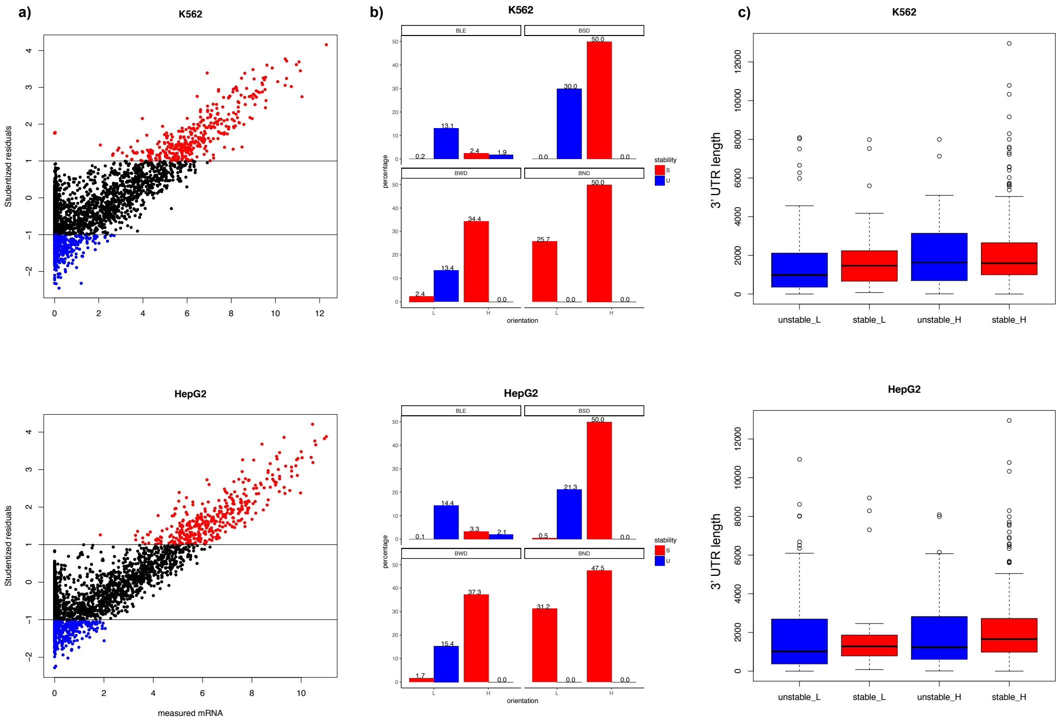


FIGURE C.4: Prediction of RNA stability. a) Scatter plot showing the studentized residulas versus the measured mRNA as average single cell transcript expression for both K562 (top panel) and HepG2 (bottom panel) samples. b) Percentage of L and H genes inferred as stable or unstable per state. c) The 3′ UTR length distribution shown for L and H genes per each stability category.

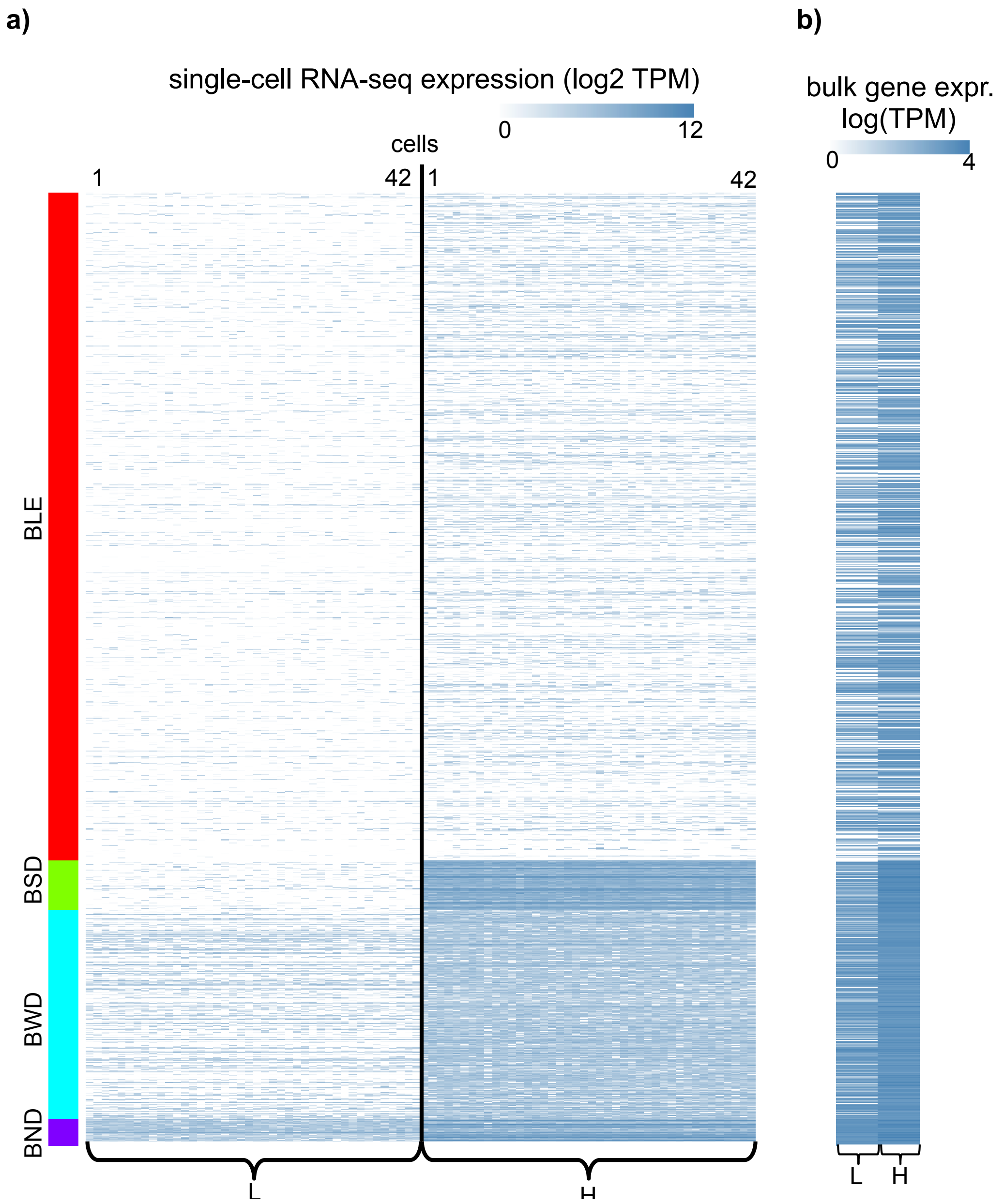


FIGURE C.5: a) Hierarchical clustering of the K562 single cell transcript expression matrix visualized as heat map (log2, TPM) and grouped into four distinct clusters (*BLE*, *BSD*,*BWD*,*BND*). b) Heat map of bulk RNA-seq expression in K562 cells (log2, TPM), arranged according to (a).

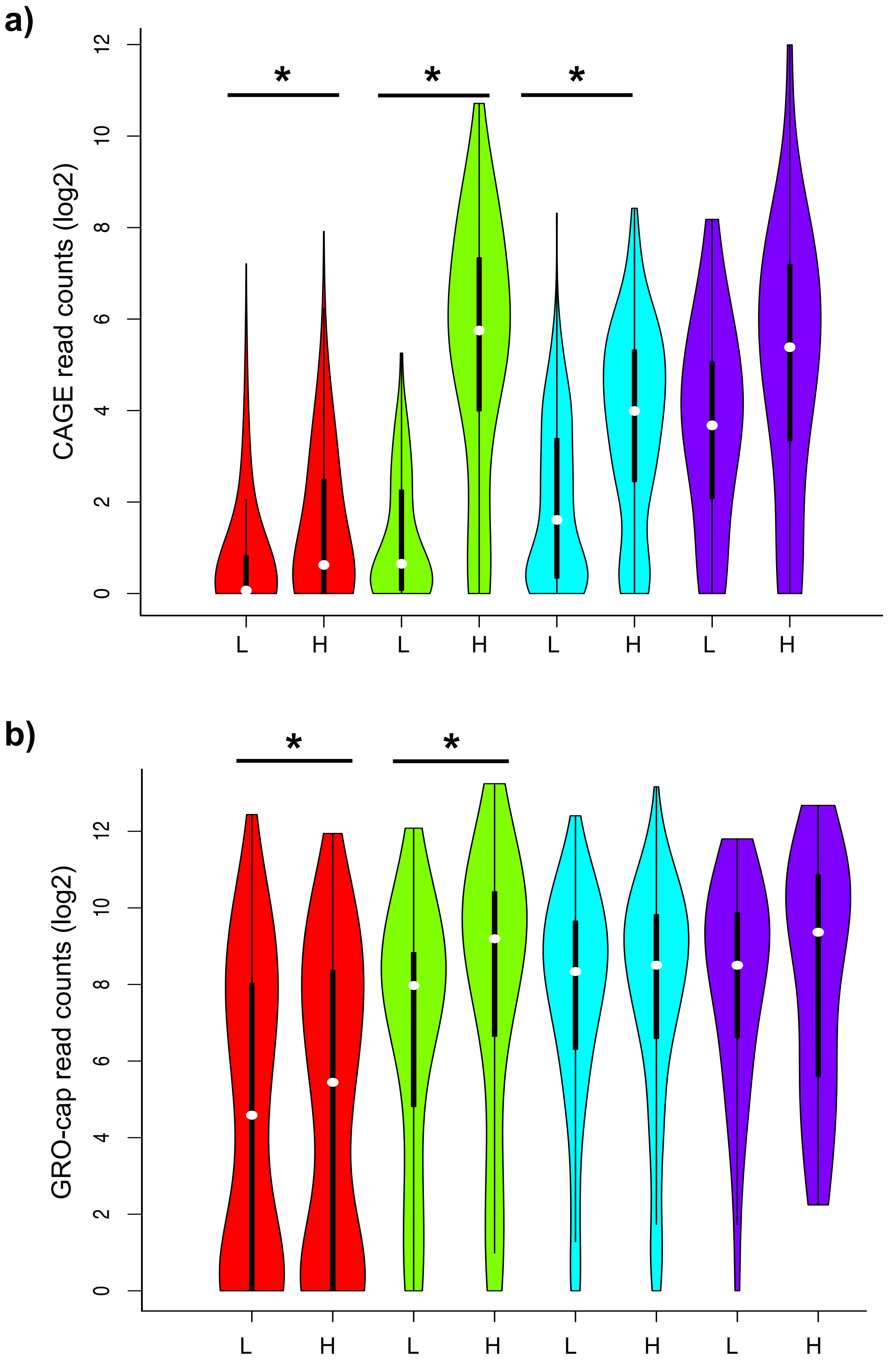


FIGURE C.6: a) CAGE read counts measured for each bidirectional gene (L and H), shown for each transcript state. Color code similar to Appendix Figure C.5a. Significant differences are marked with $*$ (paired and two-sided Mann-Whitney test, $p \leq 0.05$). b) Similar to (a) except that the results are shown for GRO-cap read counts.

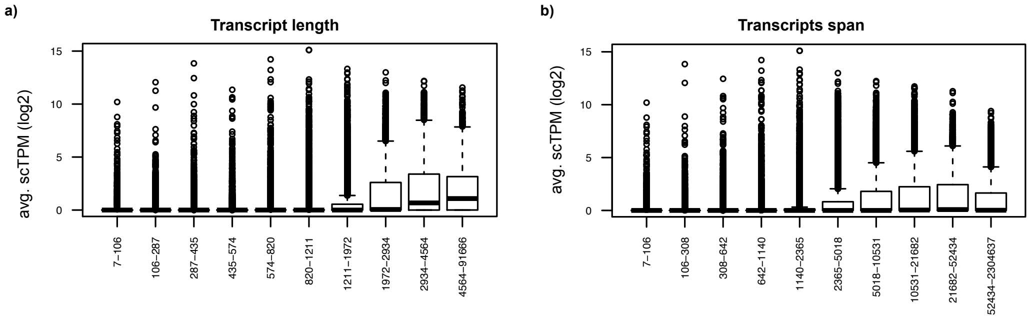


FIGURE C.7: Effects of *transcript length* (a) and *transcript span* (b) on average single cell TPM expression in all genes.

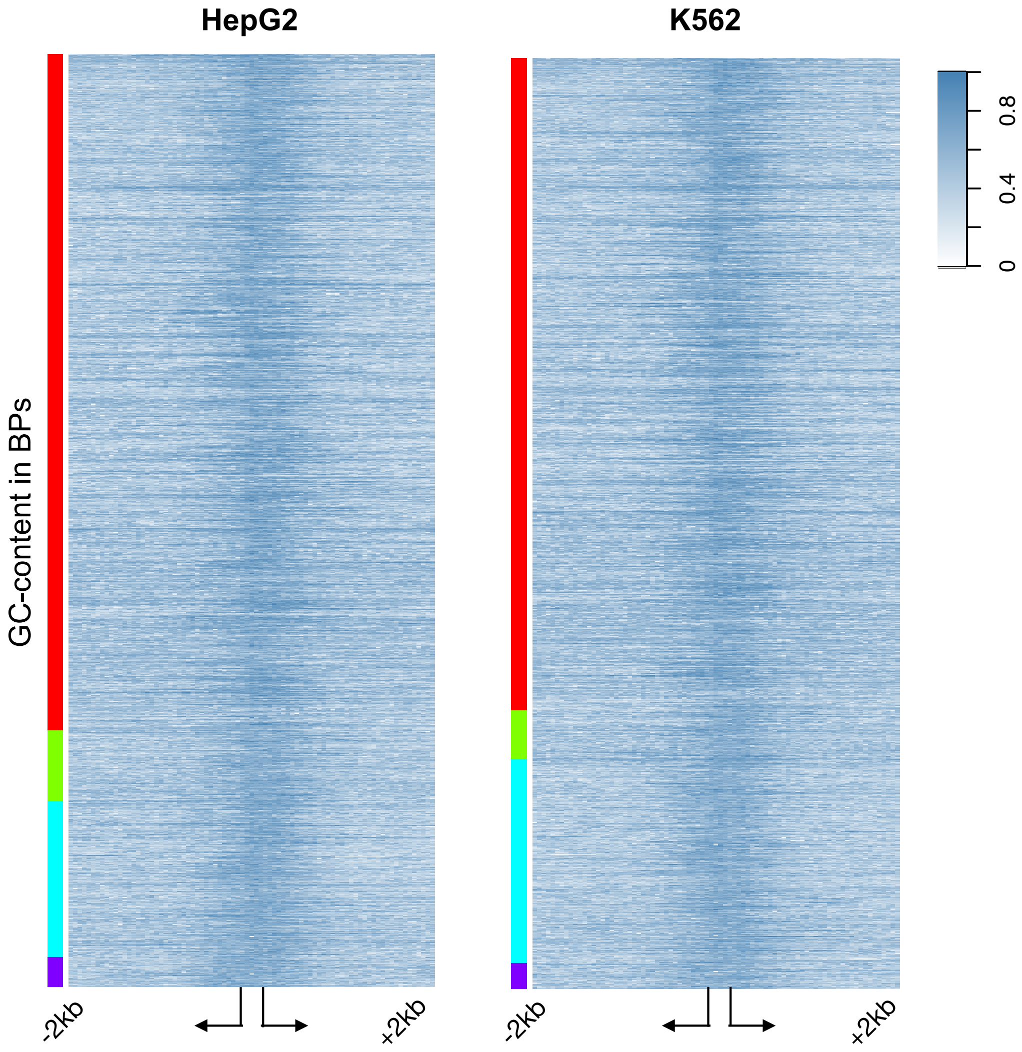


FIGURE C.8: Heat maps of the G-C content measured around the TSSs of bidirectional genes (see 5.2.11) for HepG2 (left panel, color coded as in Figure 5.4) and K562 (right panel, color coded as in Appendix Figure C.5a).

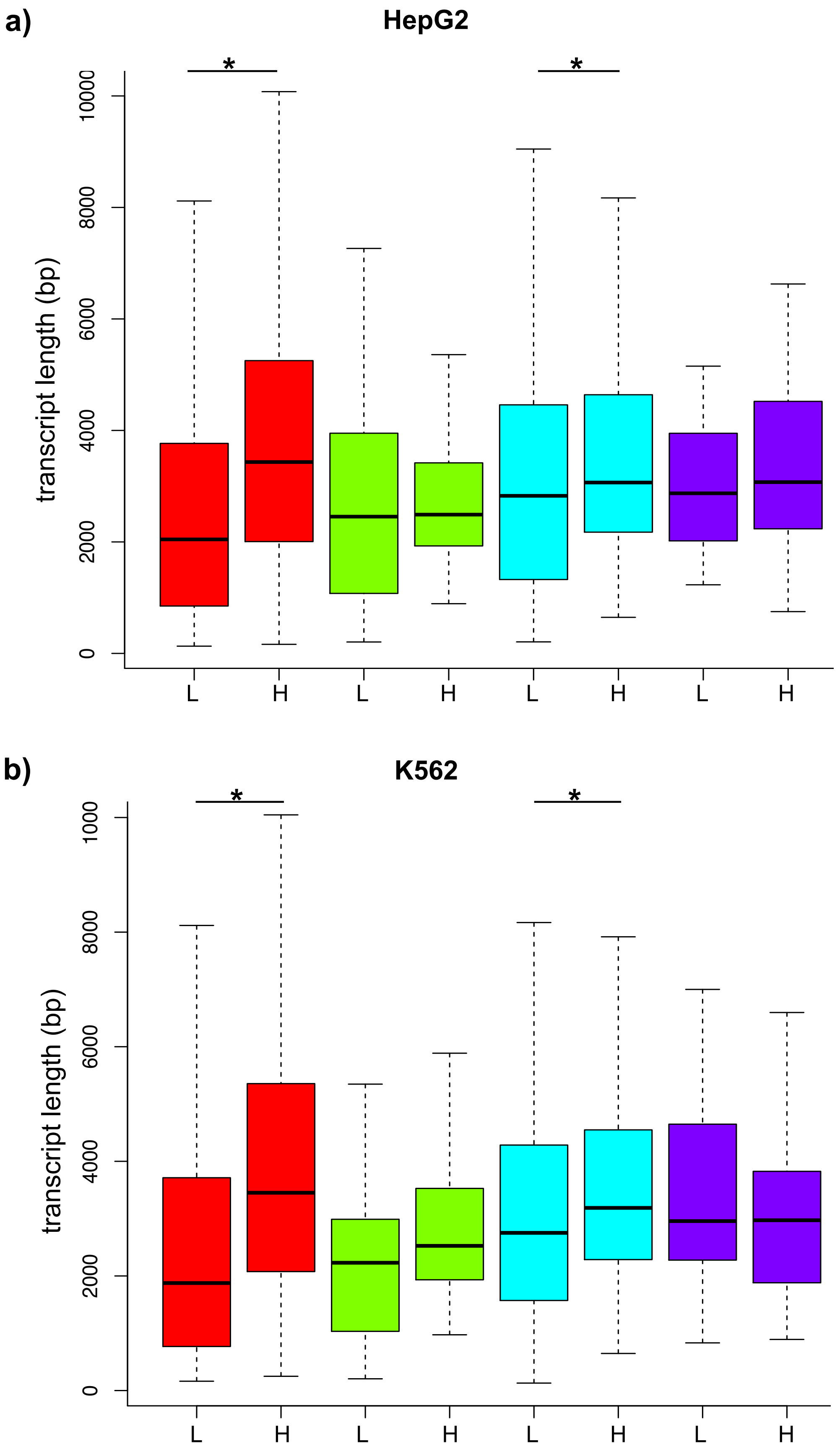


FIGURE C.9: Distributions of transcript length of L and H genes in a) HepG2 (color coded as in Figure 5.4) and b) K562 (color coded as in Appendix Figure C.5a). Significant differences are marked with $*$ (paired and two-sided Mann-Whitney test, $p \leq 0.05$)

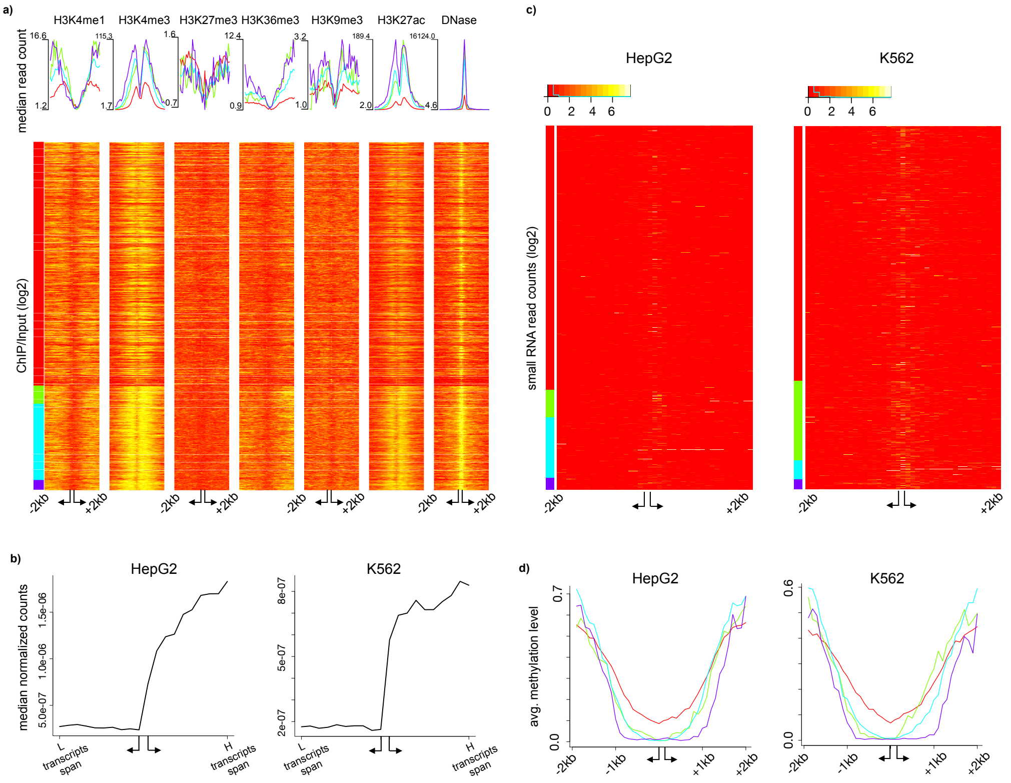


FIGURE C.10: Epigenetic characteristics in transcription states. a) Histone modification (ChIP/Input) and DNA-seq1 (raw read counts) shown as median profiles (top panel) and log-transformed values as heat map (bottom panel). Arrangement of genes as in Figure C.5. The reads are measured in 40 bins of size 100 bp spanning a window of size 4000 bp centered around the TSSs, with an additional variable bin accounting for the region between the TSSs. b) Measured H3K36me3 ChIP-seq counts in bins of variable size covering the region starting from the TSS of L and H genes extending down to the *transcript span*. c) Small RNA abundance heat maps measured similar to (a) for HepG2 (left panel) and K562 (right panel). For better visibility, bins holding values more than 200 were set to 200. d) Average methylation profiles measured by WGBS-seq read counts in bins of 100 bp following the approach explained in (a), shown for both HepG2 (left panel) and K562 (right panel).

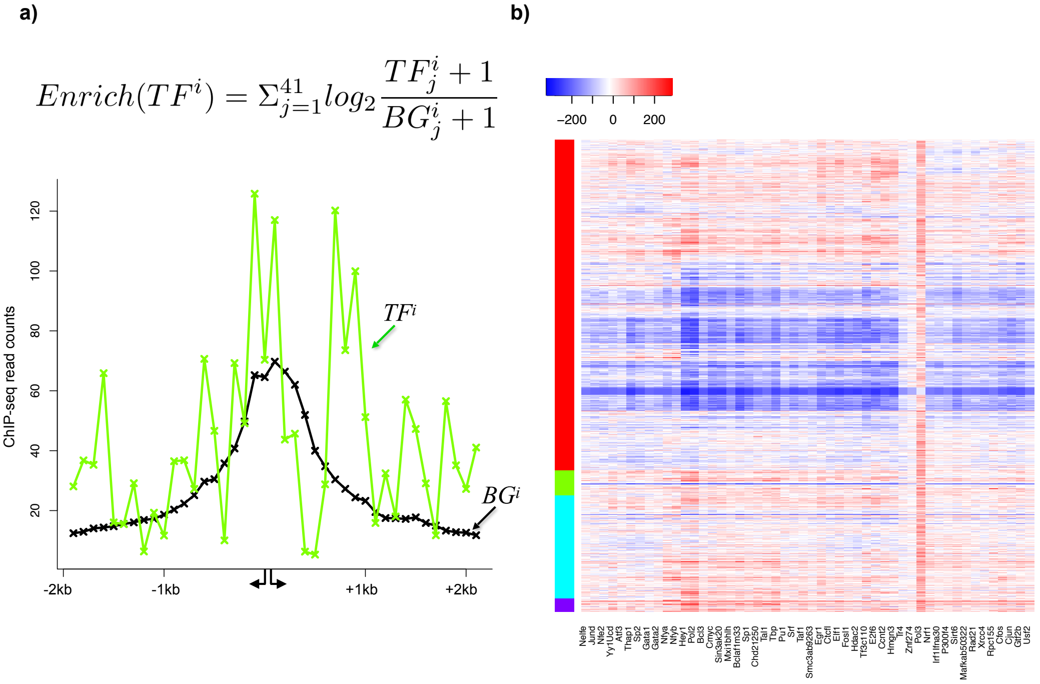


FIGURE C.11: Transcription factor enrichment. a) Enrichment score, $Enrich(TF^i)$, formula to compute the enrichment of TFs tailored for BPs. At the bottom, the binding profile of TF^i for a BP is shown. The curve shown in black represents the background defined based on the bin-wise median of TF^i binding across all BPs. The example demonstrates the effectiveness of the $Enrich(TF^i)$ score in capturing the spatial differences between true TF signal and background. b) Heat map of 50 TF (columns) enrichment scores (log ratio against background) for each BP (row) in K562 cells. Row color annotations are consistent with the clustering shown in Appendix Figure C.5.

Appendix D

Supplementary materials for Chapter 6

D.1 Supplementary Figures

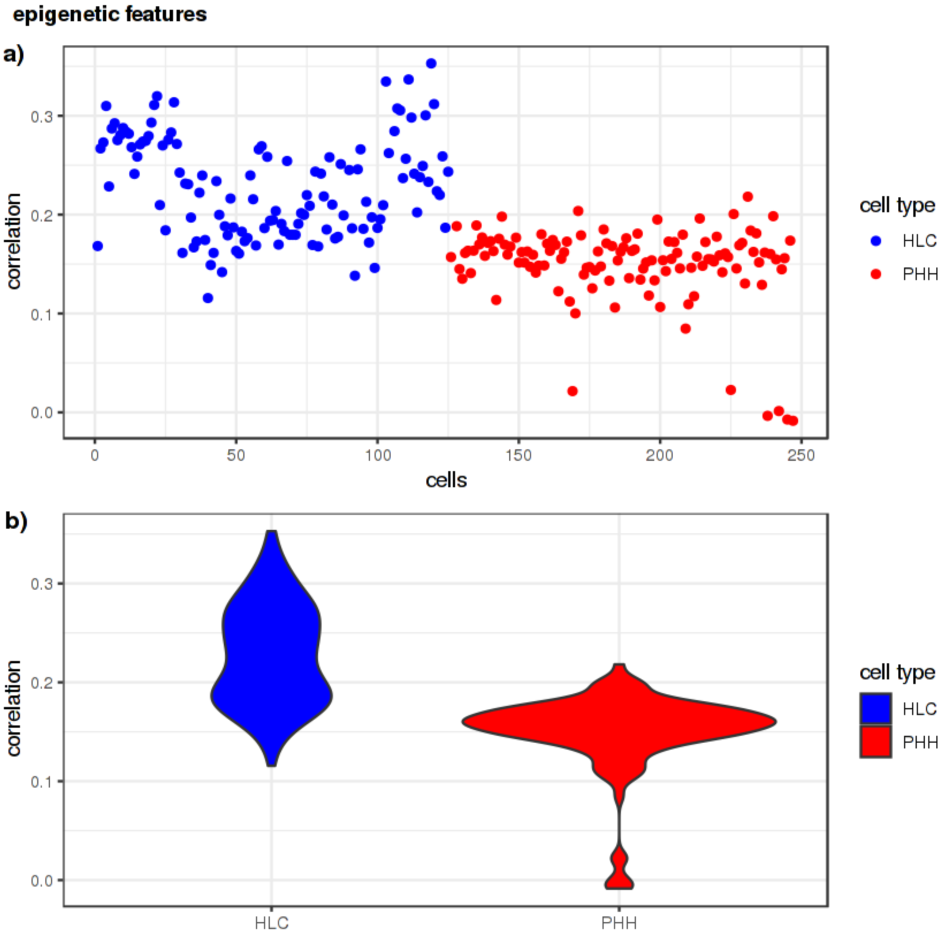


FIGURE D.1: a) Scatter and b) violin plots illustrating the Pearson correlation coefficients obtained on the test partition of the unimputed iPSCs gene expression data regarded as response and *epigenetic* features. The two sub-populations, HLC and PHH, are colored with blue and red, respectively.

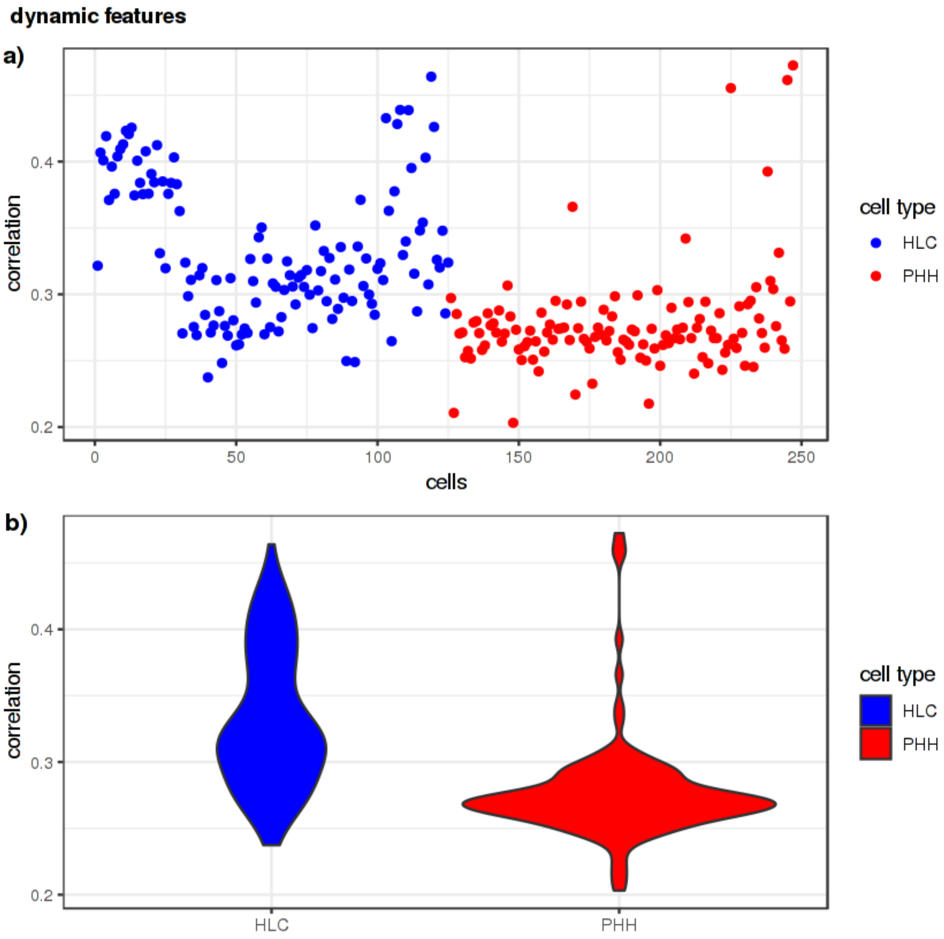


FIGURE D.2: a) Scatter and b) violin plots illustrating the Pearson correlation coefficients obtained on the test partition of the unimputed iPSCs gene expression data regarded as response and *dynamic* features. The two sub-populations, HLC and PHH, are colored with blue and red, respectively.

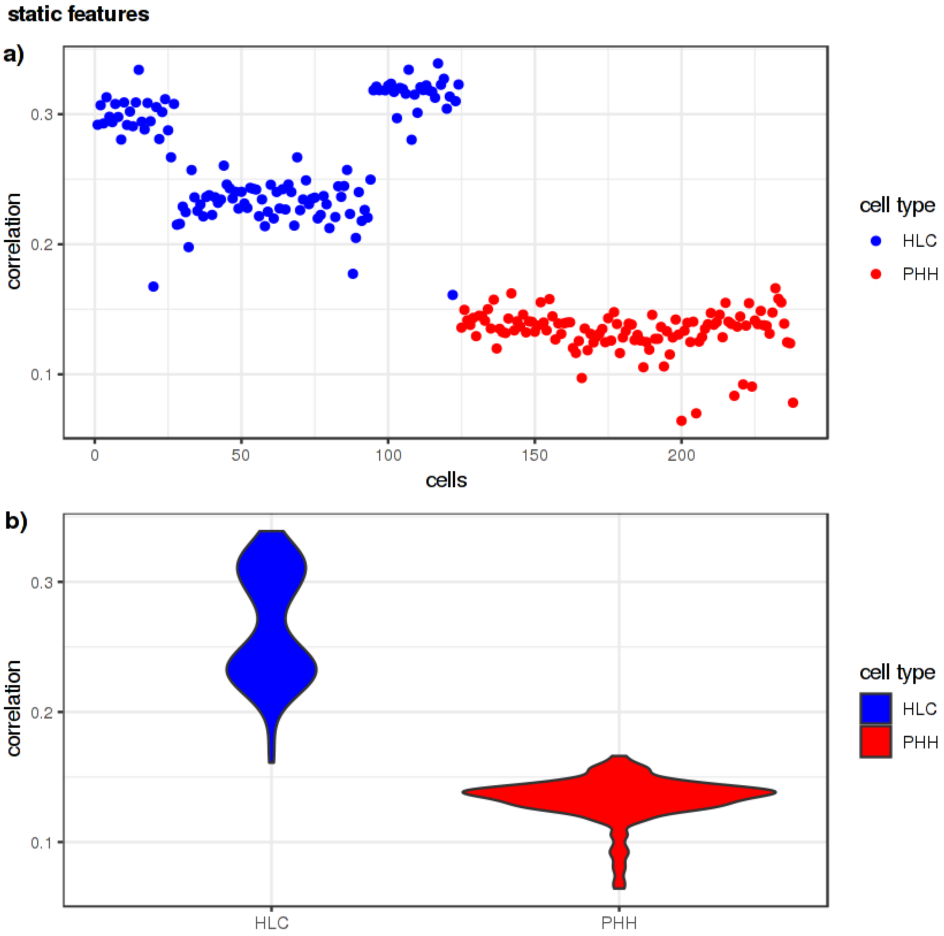


FIGURE D.3: a) Scatter and b) violin plots illustrating the Pearson correlation coefficients obtained on the test partition of the imputed iPSCs gene expression data regarded as response and *static* features. The two sub-populations, HLC and PHH, are colored with blue and red, respectively.

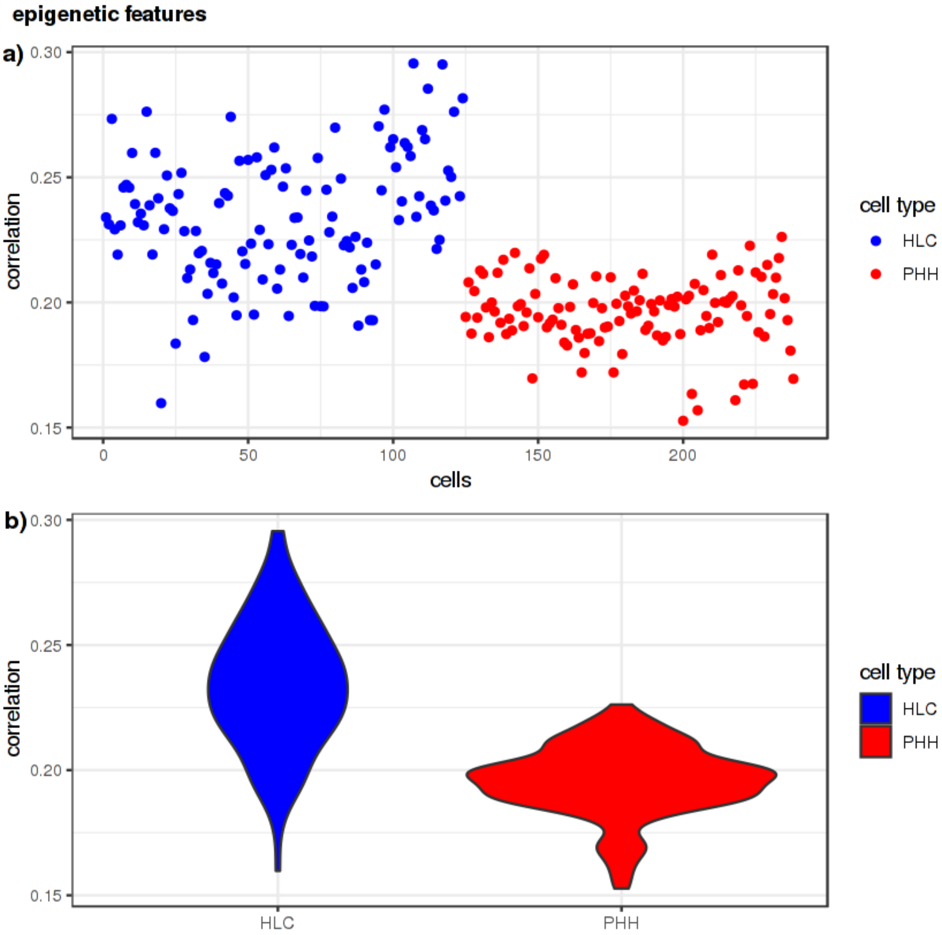


FIGURE D.4: a) Scatter and b) violin plots illustrating the Pearson correlation coefficients obtained on the test partition of the imputed iPSCs gene expression data regarded as response and *epigenetic* features. The two sub-populations, HLC and PHH, are colored with blue and red, respectively.

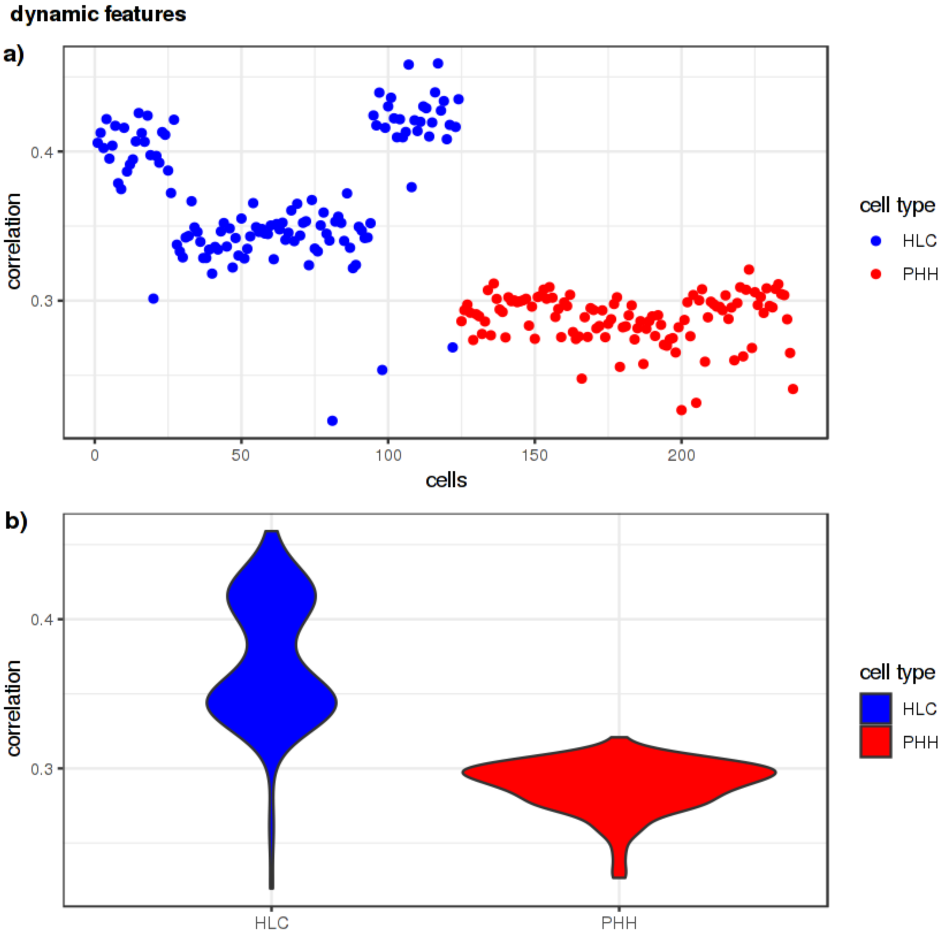


FIGURE D.5: a) Scatter and b) violin plots illustrating the Pearson correlation coefficients obtained on the test partition of the imputed iPSCs gene expression data regarded as response and *dynamic* features. The two sub-populations, HLC and PHH, are colored with blue and red, respectively.

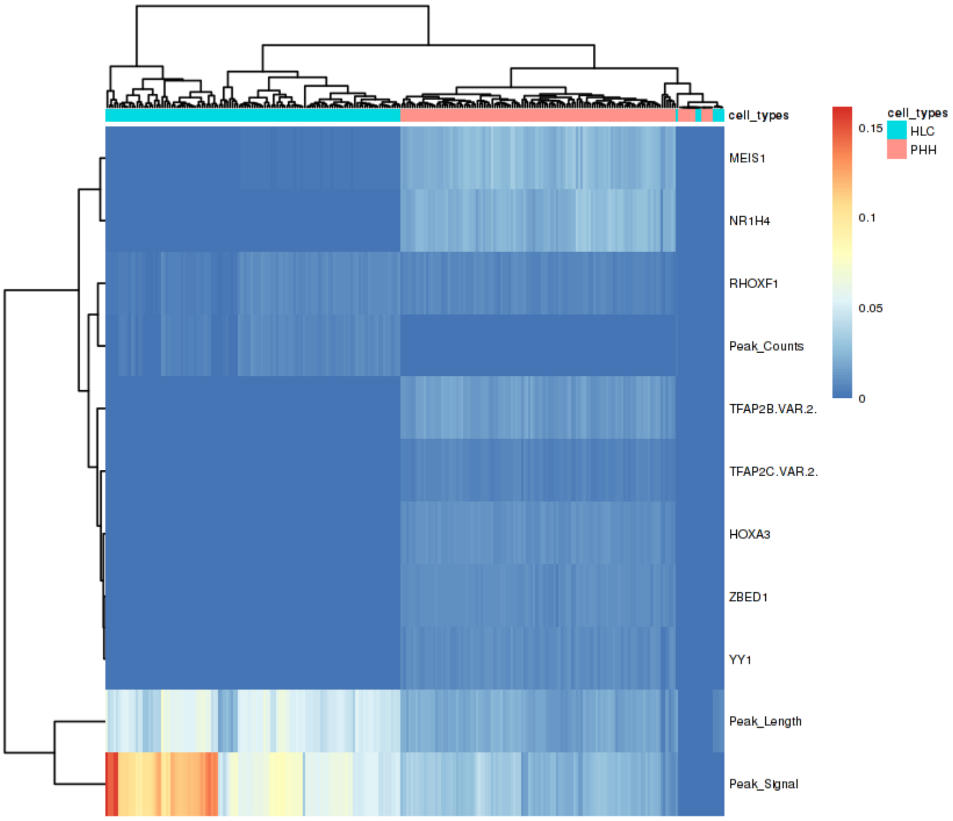


FIGURE D.6: Heat map illustrating the top features (see 6.2.6) derived from the tree-guided MTL trained on the *epigenetic* features to predict the unimputed gene expression in iPSCs cells.

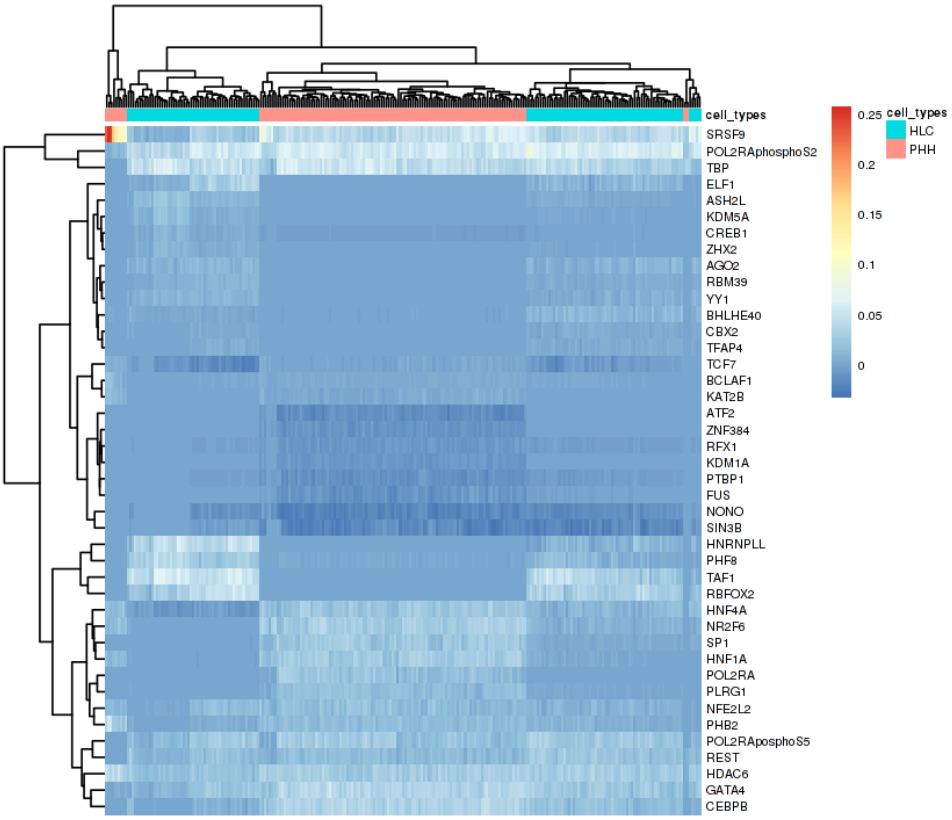


FIGURE D.7: Heat map illustrating the top features (see 6.2.6) derived from the tree-guided MTL trained on the *dynamic* features to predict the unimputed gene expression in iPSCs cells.

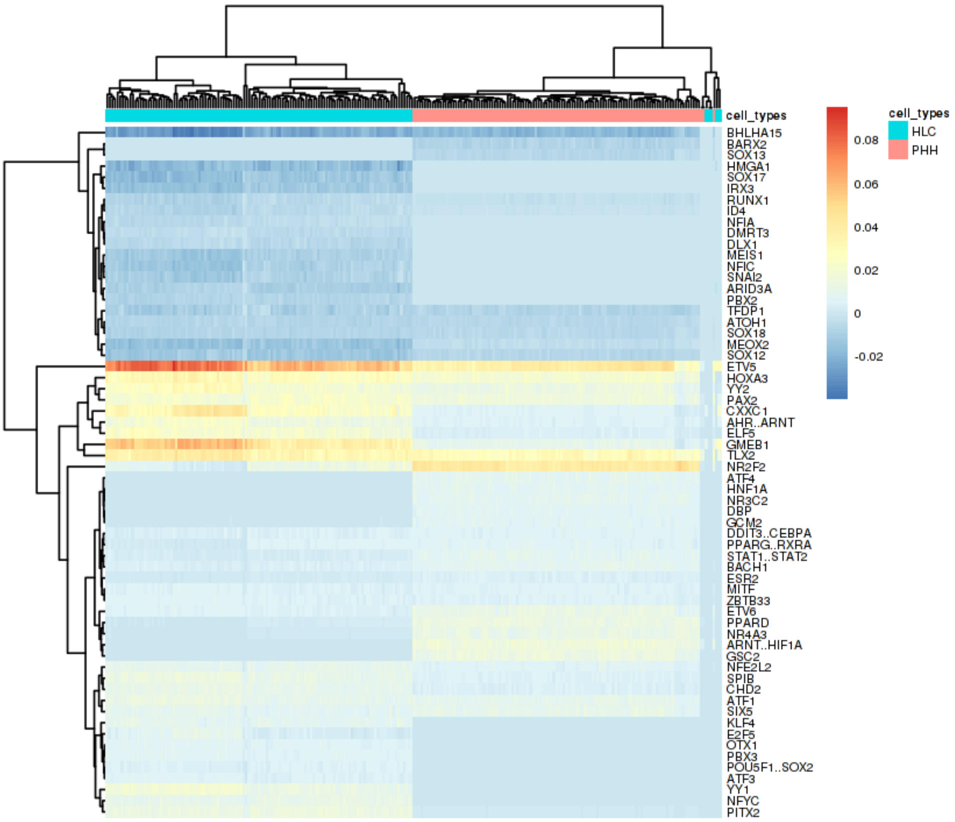


FIGURE D.8: Heat map illustrating the top features (see 6.2.6) derived from the tree-guided MTL trained on the *static* features to predict the imputed gene expression in iPSCs cells.

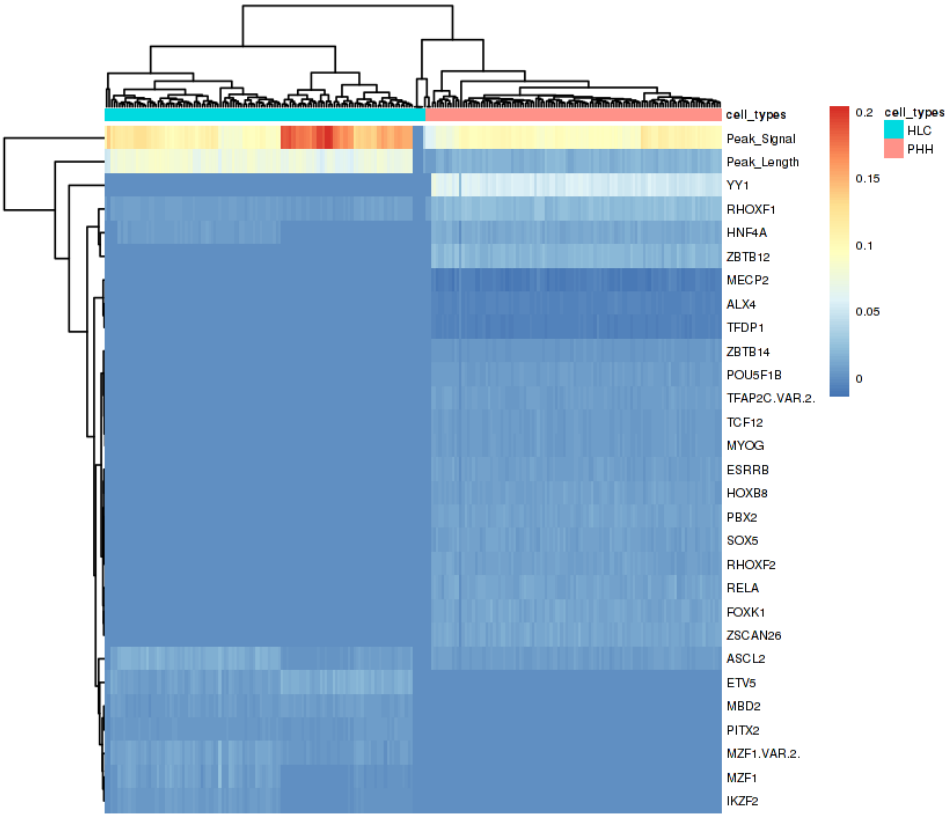


FIGURE D.9: Heat map illustrating the top features (see 6.2.6) derived from the tree-guided MTL trained on the *epigenetic* features to predict the imputed gene expression in iPCs cells.

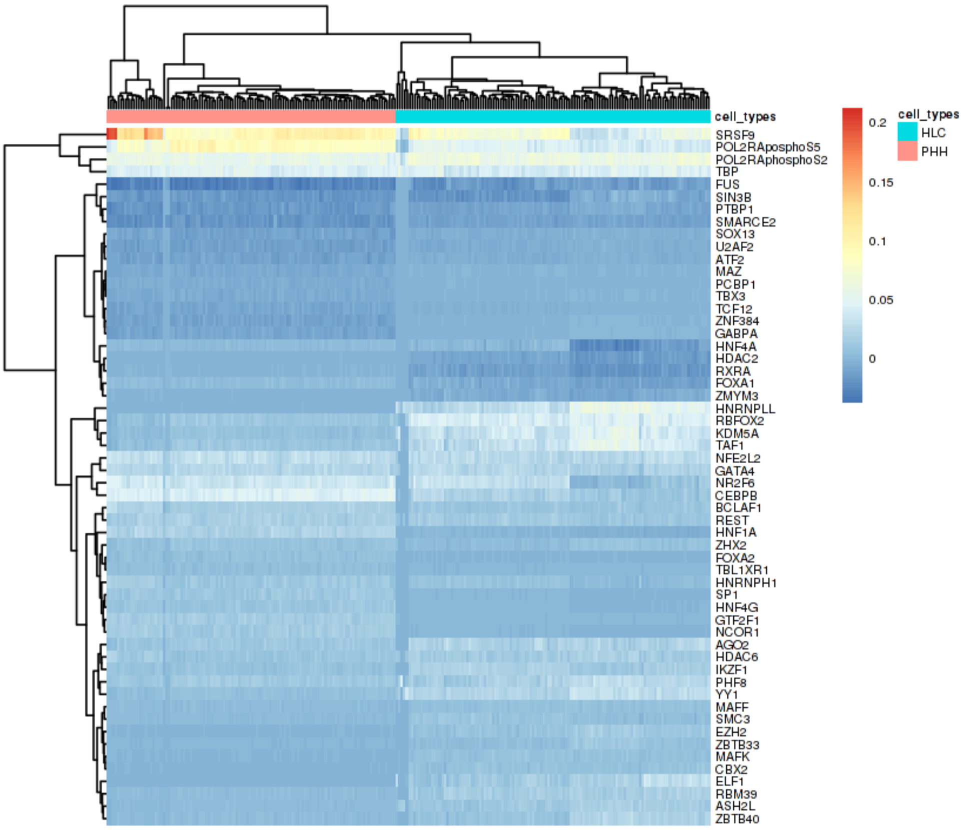


FIGURE D.10: Heat map illustrating the top features (see 6.2.6) derived from the tree-guided MTL trained on the *dynamic* features to predict the imputed gene expression in iPSCs cells.

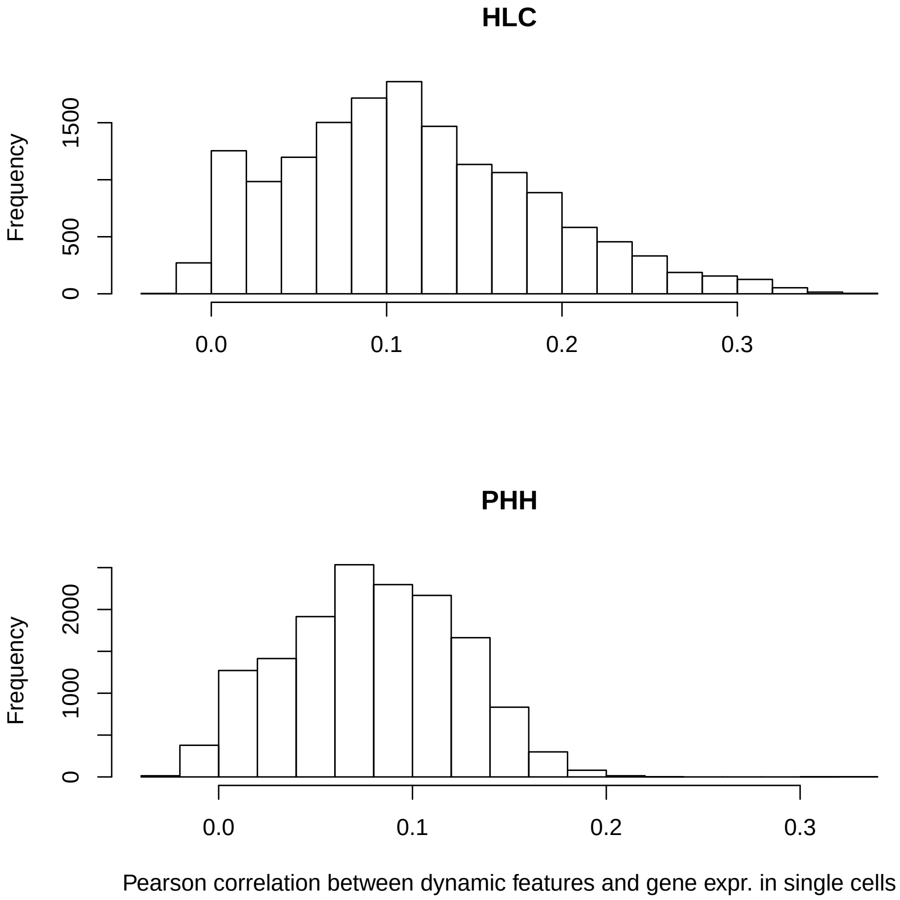


FIGURE D.11: Histograms depicting the Pearson correlation between the *dynamic* features and gene expression in PHH and HLC cells (data not imputed).

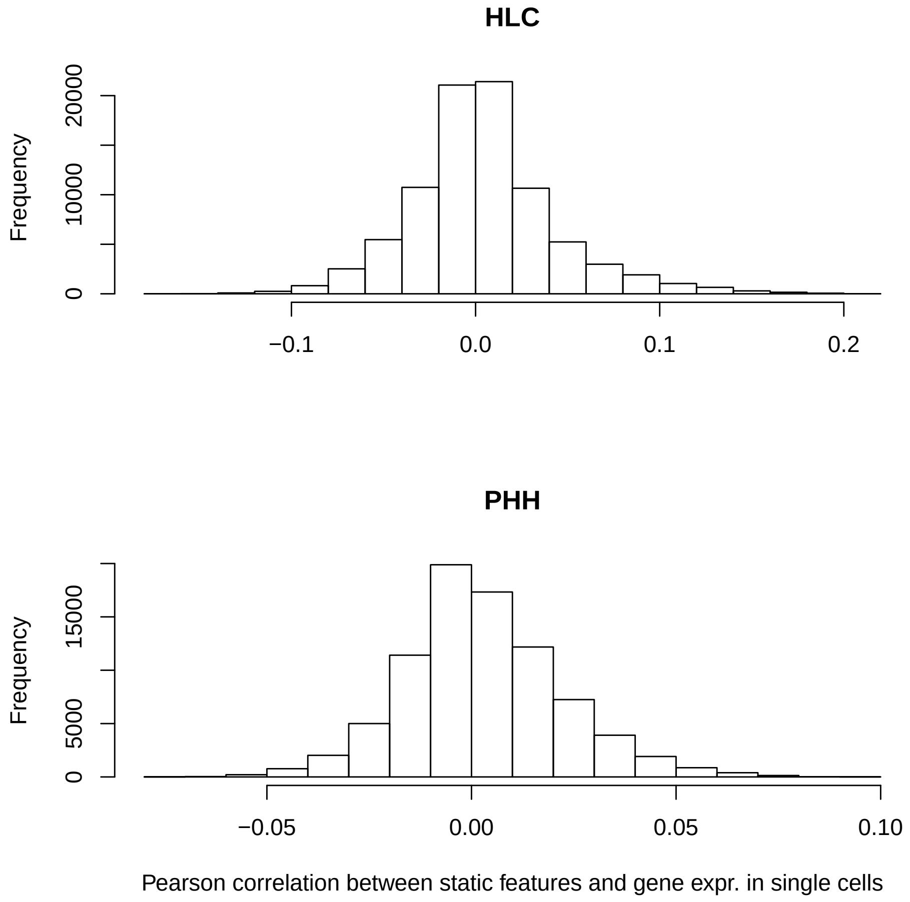


FIGURE D.12: Histograms depicting the Pearson correlation between the *static* features and gene expression in PHH and HLC cells (data not imputed).

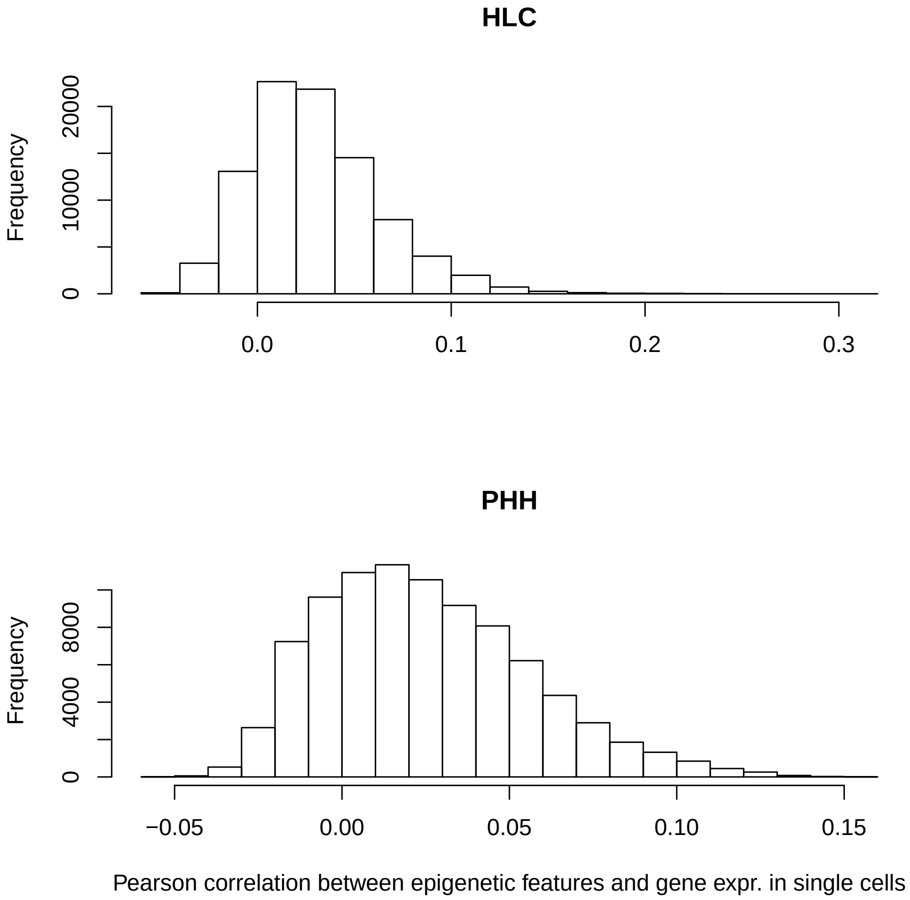


FIGURE D.13: Histograms depicting the Pearson correlation between the *epigenetic* features and gene expression in PHH and HLC cells (data not imputed).

Bibliography

Adachi, Noritaka and Michael R. Lieber (2002). “Bidirectional Gene Organization: A Common Architectural Feature of the Human Genome”. In: *Cell* 109.7, pp. 807–809. DOI: 10.1016/S0092-8674(02)00758-4. URL: https://doi.org/10.1016/S0092-8674(02)00758-4.

Aibar, Sara et al. (2017). “SCENIC: single-cell regulatory network inference and clustering”. In: *Nature Methods* 14. DOI: 10.1038/nmeth.4463. URL: https://doi.org/10.1038/nmeth.4463.

Alipanahi, B. et al. (2015). “Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning”. In: *Nat. Biotechnol.* 33.8, pp. 831–838.

Allen, Benjamin L and Dylan J Taatjes (2015). “The Mediator complex: a central integrator of transcription”. In: *Nature reviews. Molecular cell biology* 16.3, pp. 155–166. DOI: 10.1038/nrm3951. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4963239/.

Amendola, Mario et al. (2005). “Coordinate dual-gene transgenesis by lentiviral vectors carrying synthetic bidirectional promoters”. In: *Nature Biotechnology* 23.1, pp. 108–116. DOI: 10.1038/nbt1049. URL: https://doi.org/10.1038/nbt1049.

Andersson, Robin et al. (2015). “Human Gene Promoters Are Intrinsically Bidirectional”. In: *Molecular Cell* 60.3, pp. 346–347. DOI: 10.1016/j.molcel.2015.10.015.

Arnold, T. and R. Tibshirani (2014). “Efficient Implementations of the Generalized Lasso Dual Path Algorithm”. In: *Journal of Computational and Graphical Statistics* 5. eprint: 1405.3222 (stat.CO).

Ashburner, M. et al. (2000). “Gene ontology: tool for the unification of biology. The Gene Ontology Consortium”. In: *Nature genetics* 25.1, pp. 25–29. ISSN: 1061-4036. DOI: 10.1038/75556. URL: https://www.ncbi.nlm.nih.gov/pubmed/10802651.

Assenov, Yassen et al. (2014). “Comprehensive analysis of DNA methylation data with RnBeads”. In: *Nat Meth* 11.11, pp. 1138–1140. URL: http://dx.doi.org/10.1038/nmeth.3115.

Bagchi, Dia N. and Vishwanath R. Iyer (2016). “The Determinants of Directionality in Transcriptional Initiation”. In: *Trends in Genetics* 32.6, pp. 322–333. DOI: 10.1016/j.tig.2016.03.005. URL: http://dx.doi.org/10.1016/j.tig.2016.03.005.

Bainbridge, Matthew N. et al. (2006). “Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach”. In: *BMC Genomics* 7.1, p. 246. ISSN: 1471-2164. DOI: 10.1186/1471-2164-7-246. URL: https://doi.org/10.1186/1471-2164-7-246.

Bannister, Andrew J and Tony Kouzarides (2011). “Regulation of chromatin by histone modifications”. In: *Cell Research* 21.3, pp. 381–395. DOI: 10.1038/cr.2011.22. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3193420/.

Bannister, Andrew J. et al. (2005). “Spatial Distribution of Di- and Tri-methyl Lysine 36 of Histone H3 at Active Genes”. In: *Journal of Biological Chemistry* 280.18, pp. 17732–17736. DOI: 10.1074/jbc.M500796200. eprint: http://www.jbc.org/content/280/18/17732.full.pdf+html. URL: http://www.jbc.org/content/280/18/17732.abstract.

Baron, Maayan et al. (2016). “A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure”. In: *Cell systems* 3.4. 27667365[pmid], 346–360.e4. ISSN: 2405-4712. DOI: 10.1016/j.cels.2016.08.011. URL: https://www.ncbi.nlm.nih.gov/pubmed/27667365.

Barski, Artem et al. (2007). “High-Resolution Profiling of Histone Methylations in the Human Genome”. In: *Cell* 129.4, pp. 823 –837. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2007.05.009. URL: http://www.sciencedirect.com/science/article/pii/S0092867407006009.

Bartel, David P. (2004). “MicroRNAs: Genomics, Biogenesis, Mechanism, and Function”. In: *Cell* 116.2, pp. 281–297. ISSN: 0092-8674. URL: http://www.sciencedirect.com/science/article/pii/S0092867404000455.

Barth, Teresa K. and Axel Imhof (2010). “Fast signals and slow marks: the dynamics of histone modifications”. In: *Trends in Biochemical Sciences* 35.11, pp. 618–626. ISSN: 0968-0004. URL: http://www.sciencedirect.com/science/article/pii/S0968000410000940.

Behjati Ardakani, F, F Schmidt, and MH Schulz (2019). “Predicting transcription factor binding using ensemble random forest models [version 2; peer review: 2 approved]”. In: *F1000Research* 7.1603. DOI: 10.12688/f1000research.16200.2.

Behjati Ardakani, Fatemeh et al. (2018). “Integrative analysis of single-cell expression data reveals distinct regulatory states in bidirectional promoters”. In: *Epigenetics & Chromatin* 11.1, p. 66. ISSN: 1756-8935. DOI: 10.1186/s13072-018-0236-7. URL: https://doi.org/10.1186/s13072-018-0236-7.

Bornelöv, Susanne, Jan Komorowski, and Claes Wadelius (2015). “Different distribution of histone modifications in genes with unidirectional and bidirectional transcription and a role of CTCF and cohesin in directing transcription.” en. In: *BMC genomics* 16.1, p. 300. ISSN: 1471-2164.

Breiman, Leo (2001). “Random Forests”. In: *Machine Learning* 45.1, pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. URL: https://doi.org/10.1023/A:1010933404324.

Budden, D. M. et al. (2014). “Predicting expression: the complementary power of histone modification and transcription factor binding data”. In: *Epigenetics Chromatin* 7, p. 36.

Buenrostro, Jason D et al. (2013). “Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position”. In: *Nature methods* 10. DOI: 10.1038/nmeth.2688. URL: https://www.ncbi.nlm.nih.gov/pubmed/24097267.

Burley, S. K. and R. G. Roeder (1996). “BIOCHEMISTRY AND STRUCTURAL BIOLOGY OF TRANSCRIPTION FACTOR IID (TFIID)”. In: *Annual Review of Biochemistry* 65.1. PMID: 8811195, pp. 769–799. DOI: 10.1146/annurev.bi.65.070196.004005. eprint: https://doi.org/10.1146/annurev.bi.65.070196.004005. URL: https://doi.org/10.1146/annurev.bi.65.070196.004005.

Calo, Eliezer and Joanna Wysocka (2013). “Modification of enhancer chromatin: what, how, and why?” In: *Molecular cell* 49.5. 23473601[pmid], pp. 825–837. ISSN: 1097-4164. DOI: 10.1016/j.molcel.2013.01.038. URL: https://www.ncbi.nlm.nih.gov/pubmed/23473601.

Cannell, Ian. G., Yi Wen Kong, and Martin Bushell (2008). “How do microRNAs regulate gene expression?” In: *Biochemical Society Transactions* 36.6, pp. 1224–1231. ISSN: 0300-5127. DOI: 10.1042/BST0361224. eprint: https://portlandpress.com/biochemsoctrans/article-pdf/36/6/1224/852809/bst0361224.pdf. URL: https://doi.org/10.1042/BST0361224.

Caruana, Rich (1997). "Multitask Learning". In: *Mach. Learn.* 28.1, pp. 41–75. ISSN: 0885-6125. DOI: 10.1023/A:1007379606734. URL: https://doi.org/10.1023/A:1007379606734.

Chen, Yun et al. (2011). "Prediction of RNA Polymerase II recruitment, elongation and stalling from histone modification data." In: *BMC genomics* 12.1, p. 544. ISSN: 1471-2164.

Clark, Michael B et al. (2012). "Genome-wide analysis of long noncoding RNA stability". In: *Genome Research* 22.5, pp. 885–898. DOI: 10.1101/gr.131037.111. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3337434/.

Collas, Philippe (2010). "The Current State of Chromatin Immunoprecipitation". In: *Molecular Biotechnology* 45.1, pp. 87–100. ISSN: 1559-0305. DOI: 10.1007/s12033-009-9239-8. URL: https://doi.org/10.1007/s12033-009-9239-8.

Collins, Patrick J et al. (2007). "The ets-related transcription factor GABP directs bidirectional transcription." In: *PLoS genetics* 3.11. Ed. by Lisa Stubbs, e208. ISSN: 1553-7404.

Consortium, Roadmap Epigenomics et al. (2015). "Integrative analysis of 111 reference human epigenomes Open". In: *Nature* 518.7539, pp. 317–330. DOI: 10.1038/nature14248. URL: https://doi.org/10.1038/nature14248.

Consortium, The Gene Ontology (2018). "The Gene Ontology Resource: 20 years and still GOing strong". In: *Nucleic Acids Research* 47.D1, pp. D330–D338. ISSN: 0305-1048. DOI: 10.1093/nar/gky1055. eprint: http://oup.prod.sis.lan/nar/article-pdf/47/D1/D330/27437640/gky1055.pdf. URL: https://doi.org/10.1093/nar/gky1055.

Core, Leighton et al. (2014a). "Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers". In: *Nature genetics* 46. DOI: 10.1038/ng.3142.

Core, Leighton J., Joshua J. Waterfall, and John T. Lis (2008). "Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters". In: *Science* 322.5909, pp. 1845–1848. ISSN: 0036-8075. DOI: 10.1126/science.1162228. eprint: http://science.sciencemag.org/content/322/5909/1845.full.pdf.

Core, Leighton J et al. (2014b). "Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers". In: *Nature Genetics* 46.12, pp. 1311–1320. ISSN: 1061-4036.

Cramer, Patrick (2019). "Organization and regulation of gene transcription". In: *Nature* 573.7772, pp. 45–54. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1517-4. URL: https://doi.org/10.1038/s41586-019-1517-4.

Creyghton, Menno P. et al. (2010). "Histone H3K27ac separates active from poised enhancers and predicts developmental state". In: *Proceedings of the National Academy of Sciences* 107.50, pp. 21931–21936. ISSN: 0027-8424. DOI: 10.1073/pnas.1016071107. eprint: https://www.pnas.org/content/107/50/21931.full.pdf. URL: https://www.pnas.org/content/107/50/21931.

Crocker, J. et al. (2015). "Low affinity binding site clusters confer hox specificity and regulatory robustness". In: *Cell* 160.1-2, pp. 191–203.

Cuellar-Partida, G. et al. (2012). "Epigenetic priors for identifying active transcription factor binding sites". In: *Bioinformatics* 28.1, pp. 56–62.

Dahm, Ralf (2008). "Discovering DNA: Friedrich Miescher and the early years of nucleic acid research". In: *Human Genetics* 122.6, pp. 565–581. ISSN: 1432-1203. DOI: 10.1007/s00439-007-0433-0. URL: https://doi.org/10.1007/s00439-007-0433-0.

Duttke, Sascha H.C. et al. (2015a). "Human Promoters Are Intrinsically Directional". In: *Molecular Cell* 57.4, pp. 674–84. ISSN: 10972765.

— (2015b). "Perspectives on Unidirectional versus Divergent Transcription". In: *Molecular Cell* 60.3, pp. 348–349. ISSN: 10972765.

Efron, B. (1979). "Bootstrap Methods: Another Look at the Jackknife". In: *Ann. Statist.* 7.1, pp. 1–26. DOI: 10.1214/aos/1176344552. URL: https://doi.org/10.1214/aos/1176344552.

Elison, Gregory L., Ruijie Song, and Murat Acar (2017). "A Precise Genome Editing Method Reveals Insights into the Activity of Eukaryotic Promoters". In: *Cell Reports* 18.1, pp. 275 –286. ISSN: 2211-1247. DOI: https://doi.org/10.1016/j.celrep.2016.12.014. URL: http://www.sciencedirect.com/science/article/pii/S2211124716316874.

Elison, Gregory L. et al. (2018). "Insights into Bidirectional Gene Expression Control Using the Canonical GAL1/GAL10 Promoter". In: *Cell Reports* 25.3, 737 –748.e4. ISSN: 2211-1247. DOI: https://doi.org/10.1016/j.celrep.2018.09.050. URL: http://www.sciencedirect.com/science/article/pii/S2211124718314955.

ENCODE-DREAM (2017). *ENCODE-DREAM in vivo transcritpion factor binding site prediction challenge*. Accessed: 2018-02-03. URL: {https://www.synapse.org/#!Synapse:syn6131484/wiki/402034}.

ENCODEConsortium (2012). "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414, pp. 57–74.

Ernst, Jason and Manolis Kellis (2012). "ChromHMM: automating chromatin-state discovery and characterization". In: *Nature Methods* 9.3. DOI: 10.1093/nar/gkv1495. eprint: http://www.nature.com/nmeth/journal/v9/n3/full/nmeth.1906.html.

Fang, Yuan et al. (2016). "Histone modifications facilitate the coexpression of bidirectional promoters in rice". In: *BMC Genomics* 17.1, p. 768. ISSN: 1471-2164. DOI: 10.1186/s12864-016-3125-0. URL: http://dx.doi.org/10.1186/s12864-016-3125-0.

Ferdous, Mohsina M. et al. (2018). "Predicting gene expression from genome wide protein binding profiles". In: *Neurocomputing* 275, pp. 1490 –1499. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2017.09.094. URL: http://www.sciencedirect.com/science/article/pii/S0925231217316235.

Fux, Cornelia and Martin Fussenegger (2003). "Bidirectional expression units enable streptogramin-adjustable gene expression in mammalian cells". In: *Biotechnology and Bioengineering* 83.5, pp. 618–625. DOI: 10.1002/bit.10713. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bit.10713. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bit.10713.

Gong, Wuming et al. (2018). "DrImpute: imputing dropout events in single cell RNA sequencing data". In: *BMC Bioinformatics* 19.1, p. 220. DOI: 10.1186/s12859-018-2226-y. URL: https://doi.org/10.1186/s12859-018-2226-y.

Grau, Jan, Ivo Grosse, and Jens Keilwagen (2015). "PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R". In: *Bioinformatics (Oxford, England)* 31.15. 25810428[pmid], pp. 2595–2597. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btv153. URL: https://www.ncbi.nlm.nih.gov/pubmed/25810428.

Gusmao, E. G. et al. (2014). "Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications". In: *Bioinformatics* 30.22, pp. 3143–3151.

Harrow, Jennifer et al. (2012). “GENCODE: the reference human genome annotation for The ENCODE Project”. In: *Genome research* 22.9, pp. 1760–1774. ISSN: 1549-5469. DOI: 10.1101/gr.135350.111. URL: https://www.ncbi.nlm.nih.gov/pubmed/22955987.

Hashimshony, Tamar et al. (2012). “CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification”. In: *Cell Reports* 2.3, pp. 666 –673. ISSN: 2211-1247. DOI: https://doi.org/10.1016/j.celrep.2012.08.003. URL: http://www.sciencedirect.com/science/article/pii/S2211124712002288.

Heintzman, Nathaniel D. et al. (2007). “Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome”. In: *Nature Genetics* 39. Article, 311 EP –. URL: https://doi.org/10.1038/ng1966.

Henriques, Telmo et al. (2013). “Stable pausing by RNA polymerase II provides an opportunity to target and integrate regulatory signals”. In: *Molecular cell* 52.4, 10.1016/j.molcel.2013.10.001. DOI: 10.1016/j.molcel.2013.10.001.

Hume, M. A. et al. (2015). “UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions”. In: *Nucleic Acids Res.* 43.Database issue, pp. D117–122.

Ibrahim, M. M. et al. (2015). “JAMM: a peak finder for joint analysis of NGS replicates”. In: *Bioinformatics* 31.1, pp. 48–55.

Igarashi, K. et al. (1994). “Regulation of transcription by dimerization of erythroid factor NF-E2 p45 with small Maf proteins”. In: *Nature* 367.6463, pp. 568–572.

Jayaram, N. et al. (2016). “Evaluating tools for transcription factor binding site prediction”. In: *BMC Bioinformatics*.

Kahara, J. et al. (2015). “BinDNase: a discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data”. In: *Bioinformatics* 31.17, pp. 2852–2859.

Kannan, M. B. et al. (2012). “The small MAF transcription factors MAFF, MAFG and MAFK: current knowledge and perspectives”. In: *Biochim. Biophys. Acta* 1823.10, pp. 1841–1846.

Kanter, Jurrian K de et al. (2019). “CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing”. In: *Nucleic Acids Research* 47.16, e95–e95. ISSN: 0305-1048. DOI: 10.1093/nar/gkz543. eprint: http://oup.prod.sis.lan/nar/article-pdf/47/16/e95/30023943/gkz543.pdf. URL: https://doi.org/10.1093/nar/gkz543.

Karlić, Rosa et al. (2010). “Histone modification levels are predictive for gene expression”. In: *Proceedings of the National Academy of Sciences* 107.7, pp. 2926–2931. ISSN: 0027-8424. DOI: 10.1073/pnas.0909344107. eprint: https://www.pnas.org/content/107/7/2926.full.pdf. URL: https://www.pnas.org/content/107/7/2926.

Keilwagen, J. and J. Grau (2015). “Varying levels of complexity in transcription factor binding motifs”. In: *Nucleic Acids Res.* 43.18, e119.

Kersey, Paul Julian et al. (2018). “Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species”. In: *Nucleic acids research* 46. DOI: 10.1093/nar/gkx1011. URL: https://www.ncbi.nlm.nih.gov/pubmed/29092050.

Kim, Daehwan et al. (2013). “TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions”. In: *Genome Biology* 14.4, R36. ISSN: 1465-6906. DOI: 10.1186/gb-2013-14-4-r36.

Kim, Joomyeong and Hana Kim (2012). “Recruitment and Biological Consequences of Histone Modification of H3K27me3 and H3K9me3”. In: *ILAR Journal* 53.3-4, pp. 232–239. ISSN: 1084-2020. DOI: 10.1093/ilar.53.3-4.232. URL: https://doi.org/10.1093/ilar.53.3-4.232.

Kim, Seyoung and Eric P. Xing (2010). "Tree-Guided Group Lasso for Multi-Task Regression with StructuredSparsity". In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pp. 543–550. URL: https://icml.cc/Conferences/2010/papers/352.pdf.

Köhler, Sebastian et al. (2008). "Walking the Interactome for Prioritization of Candidate Disease Genes". In: *The American Journal of Human Genetics* 82.4, pp. 949–958. DOI: 10.1016/j.ajhg.2008.02.013.

Kotliar, Dylan et al. (2019). "Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq". In: *eLife* 8. Ed. by Alfonso Valencia et al., e43803. ISSN: 2050-084X. DOI: 10.7554/eLife.43803. URL: https://doi.org/10.7554/eLife.43803.

Kulakovskiy, I. V. et al. (2016). "HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models". In: *Nucleic Acids Res.* 44.D1, pp. D116–125.

Lacadie, Scott A. et al. (2016). "Divergent transcription and epigenetic directionality of human promoters". In: *FEBS Journal*, n/a–n/a. ISSN: 1742-4658. DOI: 10.1111/febs.13747. URL: http://dx.doi.org/10.1111/febs.13747.

Langmead, Ben and Steven L Salzberg (2012). "Fast gapped-read alignment with Bowtie 2". In: *Nat Meth* 9.4, pp. 357–359.

Lee, Jeong-Heon and David G. Skalnik (2005). "CpG-binding Protein (CXXC Finger Protein 1) Is a Component of the Mammalian Set1 Histone H3-Lys4 Methyltransferase Complex, the Analogue of the Yeast Set1/COMPASS Complex". In: *Journal of Biological Chemistry* 280.50, pp. 41725–41731. DOI: 10.1074/jbc.M508312200. eprint: http://www.jbc.org/content/280/50/41725.full.pdf+html. URL: http://www.jbc.org/content/280/50/41725.abstract.

Lee, Jeong-Heon et al. (2007). "Identification and Characterization of the Human Set1B Histone H3-Lys4 Methyltransferase Complex". In: *Journal of Biological Chemistry* 282.18, pp. 13419–13428. DOI: 10.1074/jbc.M609809200. eprint: http://www.jbc.org/content/282/18/13419.full.pdf+html. URL: http://www.jbc.org/content/282/18/13419.abstract.

Li, Bo and Colin N. Dewey (2011). "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome". In: *BMC Bioinformatics* 12.1, p. 323. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-323. URL: https://doi.org/10.1186/1471-2105-12-323.

Li, Wei Vivian and Jingyi Jessica Li (2018a). "An accurate and robust imputation method scImpute for single-cell RNA-seq data". In: *Nature Communications* 9.1, p. 997. DOI: 10.1038/s41467-018-03405-7. URL: https://doi.org/10.1038/s41467-018-03405-7.

— (2018b). "An accurate and robust imputation method scImpute for single-cell RNA-seq data". In: *Nature Communications* 9.1, p. 997. DOI: 10.1038/s41467-018-03405-7. URL: https://doi.org/10.1038/s41467-018-03405-7.

Liaw, Andy and Matthew Wiener (2002). "Classification and Regression by randomForest". In: *R News* 2.3, pp. 18–22. URL: http://CRAN.R-project.org/doc/Rnews/.

Lin, Jane M et al. (2007). "Transcription factor binding and modified histones in human bidirectional promoters." In: *Genome research* 17.6, pp. 818–27. ISSN: 1088-9051.

Liu, S. et al. (2017). "Assessing the model transferability for prediction of transcription factor binding sites based on chromatin accessibility". In: *BMC Bioinformatics* 18.1, p. 355.

Liu, Shutong et al. (2019). “Single cell sequencing reveals gene expression signatures associated with bone marrow stromal cell subpopulations and time in culture”. In: *Journal of Translational Medicine* 17.1, p. 23. ISSN: 1479-5876. DOI: 10.1186/s12967-018-1766-2. URL: https://doi.org/10.1186/s12967-018-1766-2.

Liu, Yaping et al. (2012). “Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data.” In: *Genome biology* 13.7, R61. ISSN: 1465-6914.

Luo, K. et al. (2013). “Using DNase digestion data to accurately identify transcription factor binding sites”. In: *Pac Symp Biocomput*, pp. 80–91.

Marinov, G. K. et al. (2013). “From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing”. In: *Genome Research* 24.3, pp. 496–510. ISSN: 1088-9051.

Mathelier, A. et al. (2016). “JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles”. In: *Nucleic Acids Res.* 44.D1, pp. D110–115.

Mayr, Christine (2017). “Regulation by 3′-Untranslated Regions”. In: *Annual Review of Genetics* 51.1. PMID: 28853924, pp. 171–194. DOI: 10.1146/annurev-genet-120116-024704. eprint: https://doi.org/10.1146/annurev-genet-120116-024704. URL: https://doi.org/10.1146/annurev-genet-120116-024704.

Mohammadi, Shahin et al. (2018). “A geometric approach to characterize the functional identity of single cells”. In: *Nature Communications* 9.1, p. 1516. ISSN: 2041-1723. DOI: 10.1038/s41467-018-03933-2. URL: https://doi.org/10.1038/s41467-018-03933-2.

Mongia, Aanchal, Debarka Sengupta, and Angshul Majumdar (2019). “McImpute: Matrix Completion Based Imputation for Single Cell RNA-seq Data”. In: *Frontiers in Genetics* 10, p. 9. ISSN: 1664-8021. DOI: 10.3389/fgene.2019.00009. URL: https://www.frontiersin.org/article/10.3389/fgene.2019.00009.

Mortazavi, Ali et al. (2008). “Mapping and quantifying mammalian transcriptomes by RNA-Seq”. In: *Nature Methods* 5. DOI: 10.1038/nmeth.1226. URL: https://www.nature.com/articles/nmeth.1226#supplementary-information.

Natarajan, A. et al. (2012). “Predicting cell-type-specific gene expression from regions of open chromatin”. In: *Genome Res.* 22.9, pp. 1711–1722.

Neil, Helen et al. (2009). “Widespread bidirectional promoters are the major source of cryptic transcripts in yeast”. In: *Nature* 457, 1038 EP –. URL: https://doi.org/10.1038/nature07747.

Nguyen, Nang et al. (2004). “Molecular Cloning and Functional Characterization of the Transcription Factor YY2”. In: *Journal of Biological Chemistry* 279.24, pp. 25927–25934. DOI: 10.1074/jbc.M402525200. eprint: http://www.jbc.org/content/279/24/25927.full.pdf+html. URL: http://www.jbc.org/content/279/24/25927.abstract.

O’Connor, T. R. and T. L. Bailey (2014). “Creating and validating cis-regulatory maps of tissue-specific gene expression regulation”. In: *Nucleic Acids Res.* 42.17, pp. 11000–11010.

Ouyang, Zhengqing, Qing Zhou, and Wing Hung Wong (2009). “ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells”. In: *Proceedings of the National Academy of Sciences* 106.51, pp. 21521–21526. ISSN: 0027-8424. DOI: 10.1073/pnas.0904863106. eprint: https://www.pnas.org/content/106/51/21521.full.pdf. URL: https://www.pnas.org/content/106/51/21521.

Park, Daechan et al. (2014). “Simultaneous mapping of transcript ends at single-nucleotide resolution and identification of widespread promoter-associated non-coding RNA governed by TATA elements”. In: *Nucleic Acids Research* 42.6, pp. 3736–

3749. ISSN: 0305-1048. DOI: 10.1093/nar/gkt1366. eprint: http://oup.prod.sis.lan/nar/article-pdf/42/6/3736/25035500/gkt1366.pdf. URL: https://doi.org/10.1093/nar/gkt1366.

Patikoglou, G. and S. K. Burley (1997). "EUKARYOTIC TRANSCRIPTION FACTOR-DNA COMPLEXES". In: *Annual Review of Biophysics and Biomolecular Structure* 26.1. PMID: 9241421, pp. 289–325. DOI: 10.1146/annurev.biophys.26.1.289. eprint: https://doi.org/10.1146/annurev.biophys.26.1.289. URL: https://doi.org/10.1146/annurev.biophys.26.1.289.

Picelli, Simone et al. (2014). "Full-length RNA-seq from single cells using Smart-seq2". In: *Nature Protocols* 9, pp. 171–181. DOI: 10.1038/nprot.2014.006. URL: http://www.pnas.org/content/103/5/1412.long.

Pique-Regi, R. et al. (2011). "Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data". In: *Genome Res.* 21.3, pp. 447–455.

Pjanic, Milos et al. (2011). "Nuclear factor I revealed as family of promoter binding transcription activators". In: *BMC genomics* 12, p. 181. DOI: 10.1186/1471-2164-12-181.

Pollen, Alex A et al. (2014). "Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex". In: *Nature Biotechnology* 32.10, pp. 1053–1058. ISSN: 1087-0156.

Pradeepa, Madapura M. (2017). "Causal role of histone acetylations in enhancer function". In: *Transcription* 8.1, pp. 40–47. ISSN: 2154-1272. DOI: 10.1080/21541264.2016.1253529. URL: https://www.ncbi.nlm.nih.gov/pubmed/27792455.

Preker, Pascal et al. (2008). "RNA Exosome Depletion Reveals Transcription Upstream of Active Human Promoters". In: *Science* 322.5909, pp. 1851–1854. ISSN: 0036-8075. DOI: 10.1126/science.1164096. eprint: http://science.sciencemag.org/content/322/5909/1851.full.pdf.

Pruitt, Kim D., Tatiana Tatusova, and Donna R. Maglott (2005). "NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins". In: *Nucleic acids research* 33. DOI: 10.1093/nar/gki025. URL: https://www.ncbi.nlm.nih.gov/pubmed/15608248.

Quinlan, A. R. and I. M. Hall (2010). "BEDTools: a flexible suite of utilities for comparing genomic features". In: *Bioinformatics* 26.6, pp. 841–842.

Ramsköld, Daniel et al. (2012). "Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells". In: *Nature Biotechnology* 30. DOI: 10.1038/nbt.2282. URL: https://www.nature.com/articles/nbt.2282#supplementary-information.

Reznikoff, W. S. (1993). "THE TN5 TRANSPOSON". In: *Annual Review of Microbiology* 47.1. PMID: 7504907, pp. 945–964. DOI: 10.1146/annurev.mi.47.100193.004501. eprint: https://doi.org/10.1146/annurev.mi.47.100193.004501. URL: https://doi.org/10.1146/annurev.mi.47.100193.004501.

Roider, H. G. et al. (2007). "Predicting trancription factor affinities to DNA from a biophysical model". In: *Bioinformatics* 23.2, pp. 134–141.

Rokach, Lior (2010). "Ensemble-based classifiers". In: *Artificial Intelligence Review* 33.1, pp. 1–39. ISSN: 1573-7462. DOI: 10.1007/s10462-009-9124-7. URL: https://doi.org/10.1007/s10462-009-9124-7.

Schmidt, Florian et al. (2017). "Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction". In: *Nucleic acids research* 45.1. 27899623[pmid], pp. 54–66. ISSN: 1362-4962. DOI: 10.1093/nar/gkw1061. URL: https://www.ncbi.nlm.nih.gov/pubmed/27899623.

Schübeler, Dirk (2015). “Function and information content of DNA methylation”. English. In: *Nature* 517.7534, pp. 321–326.

Schwartzman, Omer and Amos Tanay (2015). “Single-cell epigenomics: techniques and emerging applications”. In: *Nat Rev Genet* 16.12, pp. 716–726.

Scruggs, Benjamin S et al. (2015). “Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin.” In: *Molecular cell* 58.6, pp. 1101–12. ISSN: 1097-4164.

Seila, Amy C. et al. (2008). “Divergent Transcription from Active Promoters”. In: *Science* 322.5909, pp. 1849–1851. ISSN: 0036-8075. DOI: `10.1126/science.1162253`. eprint: `http://science.sciencemag.org/content/322/5909/1849.full.pdf`.

Sein, Henel, Signe Värv, and Arnold Kristjuhan (2015). “Distribution and Maintenance of Histone H3 Lysine 36 Trimethylation in Transcribed Locus”. In: *PLOS ONE* 10.3, pp. 1–10. DOI: `10.1371/journal.pone.0120200`. URL: `https://doi.org/10.1371/journal.pone.0120200`.

Shiraki, Toshiyuki et al. (2003a). “Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage”. In: *Proceedings of the National Academy of Sciences* 100.26, pp. 15776–15781. ISSN: 0027-8424. DOI: `10.1073/pnas.2136655100`. eprint: `https://www.pnas.org/content/100/26/15776.full.pdf`. URL: `https://www.pnas.org/content/100/26/15776`.

— (2003b). “Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage”. In: *Proceedings of the National Academy of Sciences of the United States of America* 100.26, pp. 15776–15781. DOI: `doi:10.1073/pnas.2136655100`. URL: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC307644/`.

Siegfried, Zahava and Itamar Simon (2010). “DNA methylation and gene expression”. In: *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 2.3, pp. 362–371. ISSN: 1939-005X. DOI: `10.1002/wsbm.64`. URL: `http://dx.doi.org/10.1002/wsbm.64`.

Simon, Jeremy M. et al. (2012). “Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA”. In: *Nature protocols* 7. DOI: `10.1038/nprot.2011.444`. URL: `https://www.ncbi.nlm.nih.gov/pubmed/22262007`.

Smith, Moyra and Pamela L. Flodman (2018). “Expanded Insights Into Mechanisms of Gene Expression and Disease Related Disruptions”. In: *Frontiers in molecular biosciences* 5. 30542652[pmid], pp. 101–101. ISSN: 2296-889X. DOI: `10.3389/fmolb.2018.00101`. URL: `https://www.ncbi.nlm.nih.gov/pubmed/30542652`.

Song, Lingyun and Gregory E. Crawford (2010). “DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells”. In: *Cold Spring Harbor protocols* 2010. DOI: `10.1101/pdb.prot5384`. URL: `https://www.ncbi.nlm.nih.gov/pubmed/20150147`.

Stunnenberg, Hendrik G, Martin Hirst, International Human Epigenome Consortium, et al. (2016). “The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery”. In: *Cell* 167.5, pp. 1145–1149.

Suo, Shengbao et al. (2018). “Revealing the Critical Regulators of Cell Identity in the Mouse Cell Atlas”. In: *Cell Reports* 25.6, 1436–1445.e3. ISSN: 2211-1247. DOI: `10.1016/j.celrep.2018.10.045`. URL: `https://doi.org/10.1016/j.celrep.2018.10.045`.

Szklarczyk, D. et al. (2017). “The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible”. In: *Nucleic Acids Res.* 45.D1, pp. D362–D368.

Tanay, A. (2006). "Extensive low-affinity transcriptional interactions in the yeast genome". In: *Genome Res.* 16.8, pp. 962–972.

Tang, Fuchou et al. (2009). "mRNA-Seq whole-transcriptome analysis of a single cell". In: *Nature Methods* 6. DOI: `10.1038/nmeth.1315`. URL: `https://doi.org/10.1038/nmeth.1315`.

Tibshirani, Robert et al. (2005). "Sparsity and smoothness via the fused lasso". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.1, pp. 91–108. ISSN: 1369-7412.

Tiklová, Katarína et al. (2019). "Single-cell RNA sequencing reveals midbrain dopamine neuron diversity emerging during mouse brain development". In: *Nature Communications* 10.1, p. 581. DOI: `10.1038/s41467-019-08453-1`. URL: `https://doi.org/10.1038/s41467-019-08453-1`.

Tracy, Sam, Guo-Cheng Yuan, and Ruben Dries (2019). "RESCUE: imputing dropout events in single-cell RNA-sequencing data". In: *BMC Bioinformatics* 20.1, p. 388. DOI: `10.1186/s12859-019-2977-0`. URL: `https://doi.org/10.1186/s12859-019-2977-0`.

Trapnell, Cole and Davide Cacchiarelli (2014). *Monocle: Differential expression and time-series analysis for single-cell RNA-Seq and qPCR experiments*. URL: `http://monocle-bio.sourceforge.net/monocle-vignette.pdf`.

Trapnell, Cole et al. (2012). "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks". In: *Nat. Protocols* 7.3, pp. 562–578.

Trapnell, Cole et al. (2014). "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells". In: *Nature Biotechnology* 32, 381 EP –.

Trinklein, Nathan D et al. (2004). "An Abundance of Bidirectional Promoters in the Human Genome". In: *Genome Research* 14.1, pp. 62–66. DOI: `10.1101/gr.1982804`. URL: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC314279/`.

Vaquerizas, J. M. et al. (2009). "A census of human transcription factors: function, expression and evolution". In: *Nat. Rev. Genet.* 10.4, pp. 252–263.

Waardenberg, A. J. et al. (2016). "Prediction and validation of protein-protein interactors from genome-wide DNA-binding data using a knowledge-based machine-learning approach". In: *Open Biol* 6.9.

Wang, Chenghe et al. (2016). "Promoter-associated endogenous and exogenous small RNAs suppress human bladder cancer cell metastasis by activating p21 CIP1/WAF1 expression". In: *Tumor Biology* 37.5, pp. 6589–6598. ISSN: 1423-0380. DOI: `10.1007/s13277-015-4571-z`. URL: `https://doi.org/10.1007/s13277-015-4571-z`.

Wang, Chengyang et al. (2012). "Computational inference of mRNA stability from histone modification and transcriptome profiles". In: *Nucleic Acids Research* 40.14, pp. 6414–6423. DOI: `10.1093/nar/gks304`. URL: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3413115/`.

Wang, Zhong, Mark Gerstein, and Michael Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nature Reviews Genetics* 10.1, pp. 57–63. ISSN: 1471-0064. DOI: `10.1038/nrg2484`. URL: `https://doi.org/10.1038/nrg2484`.

Watson, J. D. and F. H. C Crick (1953). "A structure for deoxyribose nucleic acid". In: *Nature* 171. URL: `https://www.nature.com/scitable/content/Molecular-Structure-of-Nucleic-Acids-16331`.

Weber, Andreas P.M. (2015). "Discovering New Biology through Sequencing of RNA". In: *Plant Physiology* 169.3, pp. 1524–1531. ISSN: 0032-0889. DOI: `10.1104/pp.15.01081`. eprint: `http://www.plantphysiol.org/content/169/3/1524.full.pdf`. URL: `http://www.plantphysiol.org/content/169/3/1524`.

Wei, Wu et al. (2011). "Functional consequences of bidirectional promoters". In: *Trends in Genetics* 27.7, pp. 267 –276. ISSN: 0168-9525. DOI: `https://doi.org/10.1016/j.tig.2011.04.002`. URL: `http://www.sciencedirect.com/science/article/pii/S0168952511000588`.

Wolpert, David H. (1992). "Stacked generalization". In: *Neural Networks* 5.2, pp. 241 –259. ISSN: 0893-6080. DOI: `https://doi.org/10.1016/S0893-6080(05)80023-1`. URL: `http://www.sciencedirect.com/science/article/pii/S0893608005800231`.

Wu, Angela R et al. (2014). "Quantitative assessment of single-cell RNA-sequencing methods". In: *Nat Meth* 11.1, pp. 41–46. URL: `http://dx.doi.org/10.1038/nmeth.2694`.

Xiao, Yu et al. (2011). "Caenorhabditis elegans chromatin-associated proteins SET-2 and ASH-2 are differentially required for histone H3 Lys 4 methylation in embryos and adult germ cells". In: *Proceedings of the National Academy of Sciences* 108.20, pp. 8305–8310. ISSN: 0027-8424. DOI: `10.1073/pnas.1019290108`. eprint: `https://www.pnas.org/content/108/20/8305.full.pdf`. URL: `https://www.pnas.org/content/108/20/8305`.

Xu, Zhenyu et al. (2009). "Bidirectional promoters generate pervasive transcription in yeast". In: *Nature* 457, 1033 EP –. URL: `https://doi.org/10.1038/nature07728`.

Yan, Chao et al. (2015). "Decoupling of divergent gene regulation by sequence-specific DNA binding factors". In: *Nucleic Acids Research* 43.15, pp. 7292–7305. ISSN: 0305-1048. DOI: `10.1093/nar/gkv618`. eprint: `http://oup.prod.sis.lan/nar/article-pdf/43/15/7292/17434051/gkv618.pdf`. URL: `https://doi.org/10.1093/nar/gkv618`.

Yan, J. et al. (2013). "Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites". In: *Cell* 154.4, pp. 801–813.

Yardımcı, G. G. et al. (2014). "Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection". In: *Nucleic Acids Res.* 42.19, pp. 11865–11878.

Yuan, Ming and Yi Lin (2006). "Model selection and estimation in regression with grouped variables". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.1, pp. 49–67. DOI: `10.1111/j.1467-9868.2005.00532.x`. eprint: `https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2005.00532.x`. URL: `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2005.00532.x`.

Zamudio, Jesse R, Timothy J Kelly, and Phillip A Sharp (2014). "Argonaute-bound small RNAs from promoter-proximal RNA Polymerase II". In: *Cell* 156.5, pp. 920–934. DOI: `10.1016/j.cell.2014.01.041`. URL: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4111103/`.

Zawel, Leigh and Danny Reinberg (1995). "COMMON THEMES IN ASSEMBLY AND FUNCTION OF EUKARYOTIC TRANSCRIPTION COMPLEXES". In: *Annual Review of Biochemistry* 64.1. PMID: 7574492, pp. 533–561. DOI: `10.1146/annurev.bi.64.070195.002533`. eprint: `https://doi.org/10.1146/annurev.bi.64.070195.002533`. URL: `https://doi.org/10.1146/annurev.bi.64.070195.002533`.

Zhao, Yingming and Benjamin A. Garcia (2015). "Comprehensive Catalog of Currently Documented Histone Modifications". In: *Cold Spring Harbor Perspectives in Biology* 7.9. DOI: `10.1101/cshperspect.a025064`. eprint: `http://cshperspectives.cshlp.org/content/7/9/a025064.full.pdf+html`. URL: `http://cshperspectives.cshlp.org/content/7/9/a025064.abstract`.

Zhou, Vicky W., Alon Goren, and Bradley E. Bernstein (2010). "Charting histone modifications and the functional organization of mammalian genomes". In: *Nature Reviews Genetics* 12. Review Article, 7 EP –. URL: `https://doi.org/10.1038/nrg2905`.

Zou, Hui and Trevor Hastie (2005). "Regularization and variable selection via the Elastic Net". In: *Journal of the Royal Statistical Society, Series B* 67, pp. 301–320.