Applications, challenges and new perspectives on the analysis of transcriptional regulation using epigenomic and transcriptomic data

A dissertation submitted towards the degree
Doctor of Natural Science
of the Faculty of Mathematics and Computer Science
of Saarland University

by

Florian Schmidt

Saarbrücken, April 2019

Tag des Kolloquiums:	26.08.2019
Dekan:	Prof. Dr. Sebastian Hack
Prüfungsausschuss:	
Vorsitzende:	Prof. Dr. Verena Wolf
Berichtserstatter:	Prof. Dr. Marcel Holger Schulz
	Prof. Dr. Hans-Peter Lenhof
	Prof. Dr. Jan Baumbach
Akademischer Mitarbeiter:	Dr. Peter Ebert

We shall not cease from exploration, and the end of all our exploring will be to arrive where we started and know the place for the first time.
- Thomas Stearns Eliot (Poet, 1888-1965) -

Acknowledgments

In alphabetical order, I have to thank the following people who supported me in various ways during the last couple of years:

Felipe Albrecht, Gus Augustus, Felicitas and Heinrich Baldauf, Linda Balzer, Nina Baumgarten, Stefan Barrois, Fatemeh Behjati, Lester Brewer, Benedikt Brors, Joachim Büch (with his dog Oli), Kailash Budhathoki, Cristina Cadenas, Michelle Carnell, Engin Cukuroglu, Mark Delegero, Deniz Demircioglu, Coskun Dev, Matthias Dietzen, Nadezhda Doncheva, Matthias Döring, Tanja Dorst, Dilip Durai, Frank Ebener, Peter Ebert, Christine Eng, Marieke Feis, Lars Feuerbach, Jonas Fischer, Jan Forster, Valentina Galata, Deborah Gérard, Nina and Gilles Gasparoni, Geeta and Ram Geetanjali, Jonathan Göke, Jan Grau, Anna Hake, Alf Hamann, Lisa Handl, Belinda Hanson, Ulrike Häuser, Jesko Hecking-Harbusch, Guruprasad Hegde, Jan G. Hengstler, Verena Heinrich, Karin Jostock, Adrian Kalb, Prabhav Kalaghatgi, Olga Kalinina, Elitza Kaloyanova, Sivarjan Karunanithi, Tim Kehl, Fabian Kern, Tariq Khaleeq, Christina Kiefer, Thorsten Klingen, Andreas Kolz, Avinash Kumar, Glenn Lawyer, Meike Lang, Thomas Lengauer, Hans-Peter Lenhof, Markus List, Jannik Luxenburger, Marcel Maltry, Panagiotis Mandros, Tobias Marschall, Alexander Marx, Jutta Meiser, Tobias Meiser, Marcel Meyerheim, Fabian Müller, Ralf Müller, Petra, Armin and Sofie Muno, Gerd Mutscheller, Manesh Narayan, Frank Nedwed, Stefan Neumann, Sarvesh Nikumbh, Karl Nordström, Ibrahim Özdemir, Klaus Pohl, Julia Polansky-Biskup, Polina Quaranta, Anna Ramisch, Max Schäffer, Gerhard Schaufert, Ruth Schneppen-Christmann, Pia Scherer-Geiss, Kristina Scherbaum, Michael Scherer, Lasse Sinkkonen, Sebastian Schirmer, Lara Schneider, Lisa Shama, Nora Speicher, Winfried Schuffert, Marcel H. Schulz, Julian Steil, Lukas Tost, Patrick Trampert, Elena Treiber, Jilles Vreeken, Jörn Walter, Mathias Wegner, Thorsten Will, Tobias Zender, and André Zenner.

For proofreading of my thesis, I thank Nina Baumgarten, Sabrina Hoppe, Sivarajan Karunanithi, Meike Lang, Markus List, Marcel Meyerheim, and Marcel H. Schulz.

Special thanks go to my group members Fatemeh Behjati, Dilip Durai and Sivarjan Karunanithi. We were much more than colleagues and I will miss our "family". Knowing to have such amazing company at work, gave me motivation and ensured that I was happy at work.

Our work would have been impossible without the administrative, and IT staff both at MPI-Inf and at the clusteroffice. All needs and requests that came up where solved and dealt with fast and smoothly.

I also have to thank Hans-Peter Lenhof, who is my mentor since I have started to study Bioinformatics. I thank him and Jan Baumbach for reviewing this thesis.

I am very grateful to Marcel Schulz who guided me in the past four years. Working with Marcel was a great experience, and I am happy that we will continue working together in the future. Marcel always supported all his students to grow and to

explore all possibilities of academia. Without him, I would have never spend time for a research visit in Singapore and my entire future would look very different. Also, he provided not only scientific inspiration, but also showed me that despite all efforts at work, one can still have a private life. Regarding that, I owe special gratitude to Mark Anthony Delegero, Sabrina Hoppe, Tobias Meiser, and Sivarajan Karunanithi who all made sure that I would not forget about my health, physically and mentally.

Finally, I have to thank my parents, Gaby and Lothar Schmidt for their constant support. Also my grandma, Ingeborg Schmidt, and my grandpa Wilhelm Joseph Hasenfratz, who unfortunately passed away towards the end of my doctoral studies, always had words of advice, although they never really understood what I was doing.

Abstract

The integrative analysis of epigenomics and transcriptomics data is an active research field in Bioinformatics. New methods are required to interpret and process large omics data sets, as generated within consortia such as the International Human Epigenomics Consortium. In this thesis, we present several approaches illustrating how combined epigenomics and transcriptomics datasets, e.g. for differential or time series analysis, can be used to derive new biological insights on transcriptional regulation. In this work we focus on regulatory proteins called transcription factors (TFs), which are essential for orchestrating cellular processes.

In our novel approaches, we combine epigenomics data, such as DNaseI-seq, predicted TF binding scores and gene-expression measurements in interpretable machine learning models. In joint work with our collaborators within and outside IHEC, we have shown that our methods lead to biological meaningful results, which could be validated with wet-lab experiments.

Aside from providing the community with new tools to perform integrative analysis of epigenomics and transcriptomics data, we have studied the characteristics of chromatin accessibility data and its relation to gene-expression in detail to better understand the implications of both computational processing and of different experimental methods on data interpretation.

Overall, we provide easy to use tools to enable researchers to benefit from the era of Biological Data Science.

Kurzfassung

Die integrative Analyse von Epigenetischen- und Genexpressionsdaten ist ein aktives Forschungsfeld in der Bioinformatik. Neue Methoden sind nötig, um die umfangreichen *omics* Datensätze, wie sie in Konsortien wie dem *International Human Epigenome Consortium* generiert werden, sinnvoll interpretieren zu können.

In dieser Dissertation stellen wir mehrere Ansätze vor, um die häufigsten *omics* Daten, wie beispielsweise differentielle Datensätze oder auch Zeitreihen zu verwenden, um neue Erkenntnisse über Genregulation auf transkriptioneller Ebene gewinnen zu können. Dabei haben wir uns insbesondere auf sogenannte Transkriptionsfaktoren konzentriert. Dies sind Proteine, die essentiell für die Steuerung regulatorischer Prozesse in der Zelle sind. In unseren neuen Methoden kombinieren wir epigenetische Daten, zum Beispiel DNaseI-seq oder ATAC-seq Daten, vorhergesagte Transkriptionsfaktorbindestellen und Genexpressionsdaten in interpretierbaren Modellen des maschinellen Lernens. Zusammen mit unseren Kooperationspartnern haben wir gezeigt, dass unsere Methoden zu biologisch bedeutsamen Ergebnissen führen, die exemplarisch im Labor validiert werden konnten.

Ferner haben wir im Detail Zusammenhänge zwischen der Struktur des Chromatins und der Genexpression untersucht. Dies ist von immenser Bedeutung, um den Einfluss von experimentellen Charakteristika aber auch von der Modellierung der Daten auf die biologische Interpretation zu vermeiden.

Contents

1 Introduction

The human body is composed of approximately 37.2 trillion cells [B+13a]. There are about 200 different *cell types* [A+05] that is cells with different purpose and morphology. Despite this diversity, almost all cells do share the same DNA sequence. This raises the question how cellular diversity is orchestrated and maintained on the molecular level.

To address this, Conrad Waddington introduced the term *epigenetics* which he derived from the Greek word *epigenesis* coined by Aristotle in his book *Generation of Animals* [AP15]. Waddington defines epigenetics as *"the branch of biology which studies the causal interactions between genes and their products which bring the phenotype into being"* [Wad08]. According to Robin Holiday, who has extended the definition, epigenetics accounts for gene-expression changes not only during development but also in the adult stage. His definition furthermore includes the possibility to transfer epigenetic information from one generation to another [Hol94]. Nowadays, epigenetics is defined as *"the study of changes in gene function that are mitotically and/or meiotically heritable and that do not entail a change in DNA sequence"* [WM01]. Mitotic and/or meiotic heritability refers to heritable traits between cells and between generations. The latter is known as transgenerational heritage. To describe the dynamics of cellular development, Waddington suggested the concept of the "epigenetic landscape" (Figure 1.1) [Wad57].

He compares cellular development to a marble rolling down from the top of a hill ending up in one of several valleys. Traversing down the landscape is a metaphor for a cell transitioning from a *pluripotent*, to a *multipotent*, to a fully differentiated state. The landscape itself can be interpreted as the result of the interplay of genes guiding cellular development. While the original theory of the landscape did not foresee that cells could leave a differentiated state, it was shown in 2006 that cells can be triggered to turn into pluripotent cells. Figuratively speaking, the cells are leaving the valley by travelling up the ridges [L+08]. Additionally, it was shown that cells are able to directly turn into another fully differentiated cell type, without a prior conversion into a pluripotent or multipotent state [K+14a]. For example, *fibroblasts* have been directly transformed to *cardiomyocytes* [I+10]. These transitions are induced by DNA-binding proteins, so called transcription factors (TF), which bind the DNA with high sequence specificity [V+09a]. Yamanaka *et al.* proposed a set of four TFs (Oct3/4, Sox2, c-Myc, Klf4) that are sufficient to convert fibroblasts to pluripotent cells, so called induced pluripotent stem cells (iPS) [L+08]. TFs are closely linked to the epigenetic landscape of a cell, i.e.

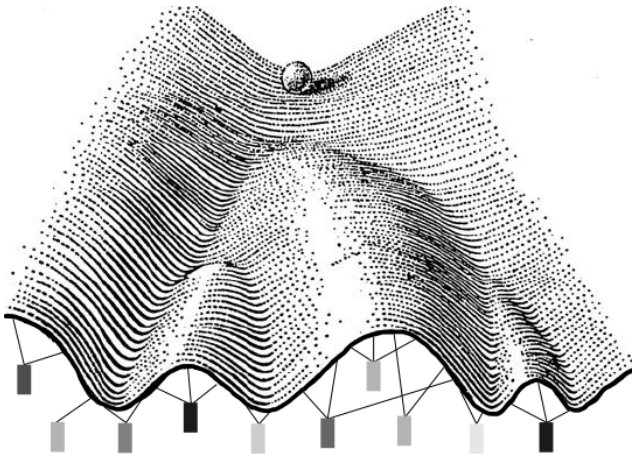


Figure 1.1: The epigenetic landscape as proposed by Waddington. He proposed that the landscape is shaped by interactions among genes, shown as strings and pegs below the landscape, respectively. The marble at the top represents a pluripotent cell. Traversing down the hill is a metaphor for the cells differentiation. Branching points indicate regulatory events affecting a cells faith. Figure adapted from Ghaffarizadeh *et al.* [G$^+$14a] obtained under licence CC BY 3.0.

the binding of TFs to the DNA is influenced by epigenetic mechanisms [HL17]. Concurrently, especially *pioneer TFs* have the potential to trigger alterations of a cell's epigenetic state [MD18].

To initiate transcription, TFs bind the DNA in distinct regions, mostly in so called *promoters*, which are located around a gene's Transcription Start Site (TSS) and *enhancers* [Y$^+$15]. Via epigenetic mechanisms, the accessibility of these sites is controlled and thereby gene-expression is regulated [HT17]. Especially the TF binding to enhancers has been associated with the regulation of cellular development [B$^+$14b]. However, epigenetics is not only essential for cellular development, but, as pointed out by Robin Holiday, also to account for gene-expression variation in adult cells, which can be influenced by external factors [Hol94]. For instance, it was shown that changes in a person's diet can influence several epigenetic marks at promoters and enhancers and thereby affect gene-expression [CF10, HT11]. In a more concrete example, chronic alcohol consumption can lead to a global *hypomethylation* of DNA by an inhibition of DNA methyltransferases (DNMTs) [Zak13, L$^+$00b].

Furthermore, alterations of enhancers on the molecular level have also been linked to several diseases. For example, genomic deletions of enhancers at the β-globin locus have been reported to be underlying *β-thalassemia* [K$^+$83]. Another example is the deletion of a TAD boundary enabling the LMNB1 promoter to interact with forebrain-specific enhancers. This is linked to the occurrence of ADLD, a neurological disorder [G$^+$15a].

In addition to effects caused by genomic rearrangements and deletions, also the occurrence of single nucleotide polymorphisms (SNPs) in enhancers is also linked to diseases. Emison *et al.* showed that a mutation in an intergenic enhancer of the *RET* gene has a significant contribution to the susceptibility of Hirschsprung's disease [E$^+$05]. In prostate cancer, several SNPs have been reported to be overlapping with known enhancer regions, thereby affecting TF binding [H$^+$14b]. Epigenetic changes in enhancers have also been reported for colon cancer [AZ$^+$12]. Maurano *et al.* performed a systematic analysis of disease-associated variations that hamper TF binding and showed that these are enriched in enhancer like sites identified using DNaseI Hypersensitive Sites (DHS) [M$^+$12a].

Due to the apparent connection between enhancers, epigenetics and disease, researchers pursue the development of drugs targeting epigenetic mechanisms. Examples of FDA-approved drugs to treat cancer are Vorinostat and Vidaza [H$^+$14e]. Vorinostat is an histone deacetylase (HDAC) inhibitor approved for treatment of advanced primary cutaneous T-cell lymphoma [M$^+$07]. The drug leads to apoptosis, increases the sensitivity of tumors for cell death processes and makes them more susceptible for other drugs via an hyperacetylation of histones [H$^+$14e]. Vidaza reduces DNA methylation genome-wide by causing degradation of the enzyme DNA cytosine-5-methyltransferase 1 (DNMT1) and it is used to treat *myelodysplastic syndrome* (MDS) [K$^+$05].

These examples show that a better understanding of TFs, enhancers and epigenetics is of general importance to decipher not only cellular differentiation, but also to elucidate the molecular cause of diseases and to understand the impact of external factors on a beings health.

To advance in these fields of research, several national and international efforts such as DEEP [Con18], Blueprint [A$^+$12], Roadmap [K$^+$15] and ENCODE [D$^+$12b], pursue the epigenetic characterization of primary cells and tissues generating publicly available epigenomics data for hundreds of samples. While these consortia are dealing with bulk data, recent advances in single-cell profiling allow the joint profiling of chromatin accessibility and gene-expression in single-cells, providing detailed information on a cells regulatory landscape [C$^+$18a, B$^+$18b]. Those advances are enabling efforts like the *Human Cell Atlas* to fulfill the ambitious goal of characterizing all human cells [R$^+$17b] on the molecular level.

Computational biology is facing the challenge to process and to systematically analyze the generated data sets in an integrative way in order to derive new hypotheses on biological function and mechanisms. With our research, we want to provide not only new tools for data analysis and data exploration, but also try to gain new insights on characteristics of epigenomics data itself to help researchers in leveraging essential and truthful biological information. While data production has been addressed by many groups in international efforts, analyzing and understanding all data sets together is an ongoing endeavour that is still in its infancy. We aim at filling some gaps in understanding, integrating and utilizing various *omics* data types from the perspective of improving our knowledge on transcriptional regulation. Although many methods have been published to elucidate transcriptional

regulation, the wealth of epigenomics data that has been produced recently paired with advances in machine learning opens new possibilities, which have not been fully explored before.

Scope and outline of the thesis

In this thesis, we pursue the development of innovative bioinformatics approaches to facilitate the analysis and interpretation of large scale biological data sets as generated by the aforementioned consortia. We focus especially on integrative methods elucidating transcriptional regulation using gene-expression, chromatin accessibility and/or TF binding information. Essential background information on biology, maths and computer science is provided in Chapter 2.

In Chapter 3, we describe the Tepic framework which was designed in the scope of the DEEP project to help biologists in jointly interpreting chromatin accessibility and gene-expression data with predicted TF binding information, generated for various primary cell types such as primary human hepatocytes and T-cells. Tepic offers machine learning methods to infer key transcriptional regulators both within (Section 3.5) and between samples (Chapter 4). Additionally, as described in Section 3.4.3, we characterize potential confounding variables influencing such models, even if TF binding information is not predicted, but experimentally derived from TF ChIP-seq data.

The functionality of Tepic is complemented with the support for the analysis of time-series data, a sub module termed Epic-Drem, described in Chapter 5. Epic-Drem was used to infer novel, subsequently experimentally validated regulators of mesenchymal stem cell differentiation to both osteoblasts and adipocytes.

As Tepic is relying on chromatin accessibility data to predict TF binding, we systematically compared established experimental methods to assess chromatin accessibility: DNaseI-seq, ATAC-seq and NOMe-seq. Such a comparative study is essential to unravel potential assay specific biases that might affect further downstream analysis. As delineated in the 6^{th} chapter, such biases do exist and it is essential for the community to be aware of them because chromatin accessibility screens are among the default assays in epigenomic profiling. Particularly in the light of single-cell analysis it is important to understand potential confounders due to the sparsity of the data.

Another potential bias that can hamper integrative analysis are batch effects arising, for instance, from varying experimental protocols used by different labs. Although there are several existing methods to adjust for batch effects, we noticed a need for a measure that allows an objective comparison of batch effect adjustment methods. To this end, we suggest an ontology based similarity score, explained in Chapter 7 that can be applied even to heterogeneous data sets with low replicate numbers. In light of ongoing integrative analyses in the International Human Epigenomics Consortium (IHEC) such a method is needed to reliably merge data sets across different consortia.

Our contributions are concluded by a novel method called StitchIt which identifies candidate regulatory regions for a distinct gene utilizing a dynamic programming algorithm optimizing a score based on the Minimum Description Length (MDL) principle, applied to large epigenomic and transcriptomic data sets. We believe that this method, described in Chapter 8, can contribute to a better understanding of transcriptional regulation on the level of single genes. StitchIt provides a rich and unique resource of candidate regulatory regions for more than $30,000$ distinct genes.

In Chapter 9, we provide a discussion of our contributions, describe (future) challenges in the analysis of omics data with a focus on epigenomics and present an hypothesis how the field of research on transcriptional regulation might evolve in the future. At the end of the Chapters 3 to 8, we provide a section detailing the contributions of all researchers involved in the presented projects.

Throughout this thesis, terms written in *italic* font are briefly explained in the Glossary. Following the HGNC guidlines for gene and protein nomenclature, gene names of human genes are written with capital letters in italic font (*CTCF*), gene names of mouse genes are written in italic font with only the first letter in upper-case (*Ctcf*) and protein names are written in normal font in capital letters independent of the species (CTCF) [W+02].

2 Background

Biological (Section 2.1), mathematical and computational background knowledge (Section 2.2) that is of general relevance for the understanding of this thesis is provided within this chapter.

The data used throughout this work has been obtained from the International Human Epigenomics Consortium (IHEC). To adequately credit the contribution of the researchers generating the data, Section 2.3 contains an overview of the consortia providing epigenomics and transcriptomics data for the community.

2.1 Biological Background

This section provides a brief overview of the composition of DNA and RNA, on the structure and diversity of proteins, on the central dogma of molecular biology, on DNA sequencing and on transcription and translation. Further, we review the structure and composition of chromatin, provide an overview of the regulatory function of TFs, characterize established epigenetic modifications and introduce the notion of enhancers and repressors. For reasons of completeness, we also sketch regulatory mechanisms aside from transcriptional regulation through TFs. Note that this chapter is not meant to introduce the full spectrum of the covered topics, but should merely be seen as an introduction sufficient for a basic understanding of the following chapters.

2.1.1 DNA, RNA, Proteins and the definition of genes

Deoxyribonucleic acid (DNA)

DNA is a macromolecule containing the genetic information composing an organism. The basic building block of DNA is called nucleotide. A nucleotide is composed of the monosaccharide 2-deoxyribose, a phosphate group and a nitrogen containing base, which is either Adenine(A), Thymine(T), Cytosine(C), or Guanine(G). The phosphate is connected to the ribose with an *ester bond* to the C_5' atom, the base is connected to the C_1' atom via a *N-glycosidic bond* [Kni06, p.10].

An alternating sequence of monosaccharides and phosphate molecules form a chain, called DNA backbone. The chain is established by phosphate bridges between the C_5' atom of one and the C_3' atom of another nucleotide [Kni06, p.10]. The phosphates cause the DNA backbone to be negatively charged and hydrophilic.

There are two antiparallel DNA-strands. The backbone of one strand is oriented in $5' - 3'$ direction, while the other one is oriented in $3' - 5'$ direction. The two strands form a double helix structure, discovered by Watson and Crick in 1953 [WC53].

The connection between the strands in the helix is established by *hydrogen bonds* between the bases. Adenin interacts with thymin via 2 hydrogen bonds and cytosin interacts with guanin via 3 hydrogen bonds. The helix is further stabilized by hydrophobic bonds between neighbouring bases, a phenomenon referred to as base stacking [Kni06, p.12].

The not-symmetric arrangement of the DNA strands in the helix gives rise to two differently sized empty spaces between the strands, a larger region referred to as major groove and a smaller region, referred to as minor groove, shown in Figure 2.1a. The size of the grooves depends on the sequence context [O$^+$10]. It was shown that proteins which bind DNA at specific sequences predominantly bind in the major groove, while the minor groove is associated with non-specific DNA binding events and is especially bound by factors affecting DNA conformation [B$^+$98, S$^+$10].

X-ray experiments have revealed that in most living cells DNA exists as so called B-DNA, which is a right-handed helix with 10.4 to 10.5 base pairs per helix twist, an approximate distance of 0.34nm between two base pairs and a 90° angle between the bases and the main helix axis [Kni06, p.13]. Due to the flexibility of the glycosidic bonds, the bases are able to rotate, which gives rise to various conformations; namely Propeller twist, Twist, Roll and Tilt [Kni06, p.13], illustrated in Figure 2.1b. Knowledge of these local shapes has been shown to be important to describe the behaviour of DNA-binding proteins [A$^+$15a, M$^+$16b].

Aside from B-DNA, there is another right-handed helix structure called A-DNA. In A-DNA, there are about 11 base pairs per helix twist, the distance between base pairs is reduced to 0.26nm and the angle between the base pairs and the helix-axis changed from 90° in B-DNA to 71°-77° [Kni06, p.13]. While the biological relevance of A-DNA is being debated, a study in 2014 showed that B-DNA can be reversibly converted to A-DNA in prokaryotes [W$^+$14].

Yet another form of DNA is Z-DNA, a left-handed double helix with about 12 base pairs per helix twist and a distance of 0.38nm between base pairs [R$^+$84]. In Z-DNA, the backbone is oriented in a zigzag manner. The precise role of Z-DNA *in vivo* is still unclear, but Z-DNA has been observed together with B-DNA at the promoters of transcribed genes [W$^+$91, H$^+$05].

In Section 2.1.8 we detail how DNA can be compactly packed in the nucleus of a cell. Further information which is not required for the general understanding of this thesis, e.g. details on DNA-replication or DNA repair mechanisms are provided in the textbooks Molekulare Genetik [Kni06] and in The molecular biology of the cell [A$^+$05].

Ribonucleic acid (RNA)

Ribonucleic acid is a chain of ribonucleotides connected via phosphodiester bonds between the $3'$-hydroxy and $5'$-hydroxy group of ribose molecules. In RNA, thymin is replaced with uracil [Kni06, p.50]. There are several RNA subtypes, classified

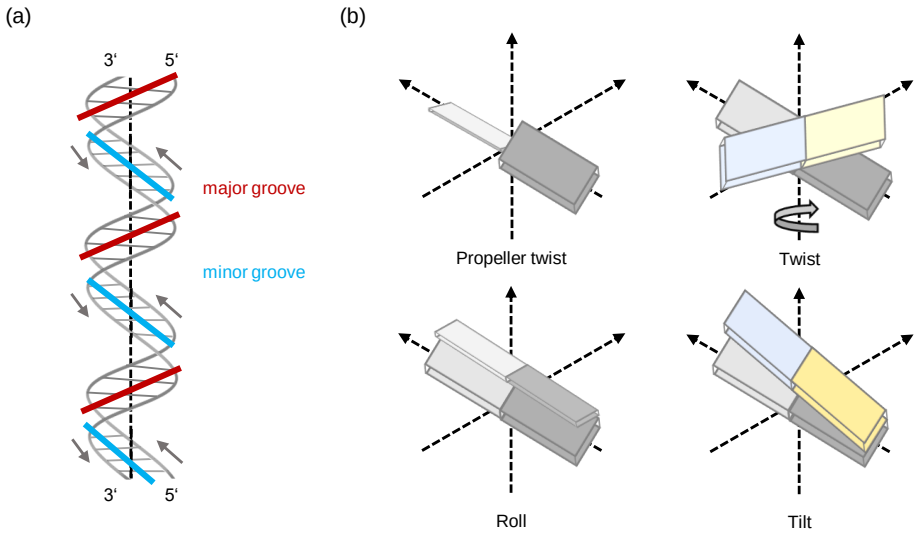


Figure 2.1: (a) Schematic of two antiparallel DNA strands forming a right-handed double helix with a major and minor groove between the DNA backbone. (b) Illustration of how nucleotides are oriented towards each other in terms of Propeller twist, Twist, Roll and Tilt. The figure is designed by the author of this thesis.

according to structure and function of the RNA molecules. An extensive overview is provided in Cech *et al.* [CS14]. In the scope of this thesis, only messenger RNAs (mRNAs) are relevant. They are described in detail below in the context of transcription.

Proteins

Proteins are macromolecules, just like DNA and RNA. Due to the diversity of functions carried out by proteins, Lodish *et al.* refer to them as the *"Molecules of life"* [L$^+$00a]. For example, collagen provides structure to cells [SR09]. Enzymes are catalyzing an uncountable number of reactions [Aga06]. For instance, the DNA-polymerase is an enzyme that is replicating DNA [Kni06, 177]. Furthermore, regulatory proteins such as TBP are inadmissible for gene regulation [Pug00]. Another example for protein function is their role in signaling cascades for inter- and intracellular communication [Hun00], for instance, insulin in the regulation of the glucose concentration in blood [Wil05]. Also, proteins are involved in transport, for instance, as hemoglobin in *erythrocytes*.

The function of a protein is reflected by a proteins three dimensional (3D) structure [Kni06, p.37]. One distinguishes several levels of structure in context of proteins. The primary structure of a protein refers to the sequence of amino acids composing a protein. Amino acids are small organic molecules with an amino group, a carboxyl group and a specific residue, also referred to as side chain, which

defines the type of amino acid (Figure 2.2a). The human genome codes for 20 different amino acids. These are connected via peptide bonds to form a directed chain of amino acids (Figure 2.2b) [Kni06, p.37,38]. The chain is directed from the N-terminal end that is the amino acid with an unbound amino group, to the C-terminal end that is an amino acid with an unbound carboxyl group. Amino acids are abbreviated using either a one or three letter code [A$^+$05, p.130].

The formation of hydrogen bonds between every fourth peptide gives rise to the so called α-helix, depicted in Figure 2.2c. The α-helix is a common secondary structure of proteins containing 3.6 amino acids within one helix twist [Kni06, pg.41]. Another secondary structure is the β-sheet, shown in Figure 2.2d. In contrast to the α-helix, in the *beta*-sheet hydrogen bonds are established between more distal parts of the entire amino acid chain that are oriented next to each other in close spatial proximity. If the chains are oriented in the same directions, one refers to the β-sheet as parallel, otherwise as antiparallel [A$^+$05, p.143].

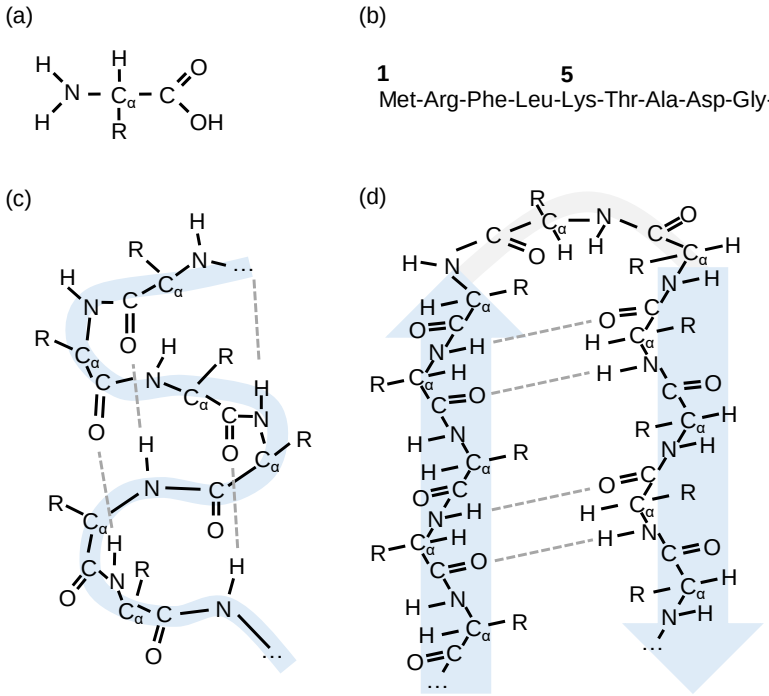


Figure 2.2: (a) Basic structure of an amino acid. The central carbon C_α is connected to the amino group, the carboxyl group, an amino acid specific side chain R and a hydrogen atom. (b) Primary structure of a protein with 11 amino acids. (c) Visualization of the secondary structure for the α helix. (d) Illustration of an antiparallel β sheet. Hydrogen bonds are shown by dotted lines in (c) and (d). Figures (c) and (d) after Figure 3 from [MK96] obtained under Springer Nature license 4501220300134.

The tertiary structure of a protein describes the entire 3D structure of one amino

acid chain, including all secondary structures. While the secondary structure is independent of interactions of the amino acid side-chains, these are essential for the formation of the tertiary structure. For example, side chains that are hydrophobic are typically oriented to the inner area of the structure, while hydrophilic chains are placed at the outside [Kni06, p.40]. The quaternary structure refers to protein complexes composed of several protein chains [A+05, p.144].

For a detailed overview of proteins we refer the reader to the textbook "Introduction to Proteins: Structure, Function and Motion" by Amid Kessel and Nir Ben-Tal [Dav11].

The definitions of the terms gene and genome

The term gene has been coined by Wilhelm Johannsen, a Danish botanist in 1909, who defined a gene as a unit of inheritance [Joh09]. The definition of gene has been evolving ever since. A detailed overview is provided for example by Portin and Wilkins [PW17].

In other words, a gene is a not necessarily continuous region of the DNA that is interacting with other genes. A gene's product, which is either a (non)coding RNA molecule or a protein, can affect other genes and may have a *phenotypic* effect.

The term genome is closely linked to gene. Following *NIHs* definition, *"a genome is an organism's complete set of DNA, including all of its genes. Each genome contains all of the information needed to build and maintain that organism"* [NIH18]. The human genome consists of about 3 billion base pairs [Kni06, p.19]. About 1% of those are coding for proteins, while 99% of DNA is non-coding [Zha12]. However, being non-coding does not imply being non-functional. About 10% of the genome is estimated to be functional, i.e. in terms of regulation [PH11].

2.1.2 The genetic code and DNA sequence alterations

Within a gene's open reading frame, the coding part of a protein coding gene, three consecutive base pairs, called triplets or codons, represent a distinct amino acid. As the genetic code is redundant, one amino acid can be referred to by several different triplets [Kni06, p.79-81]. The code is often explained in a circular table, c.f. Figure 2.3, which should be read from the inside out to determine the codons representing a distinct amino acid [BH72]. The circular representation nicely illustrates that the third base is often less important then the first two. Therefore, Francis Crick named the third base "*Wobble base*", as changes in this base do not necessarily have an influence on the resulting amino-acid [Cri66]. On the level of DNA, a protein coding gene always starts with the codon ATG, coding for the amino acid methionine. The coding sequence always terminates with one of three stop codons (TAA, TAG, or TGA) [Kni06, p.79-81].

Alterations of DNA such as *mutations*, *insertions*, or *deletions* occurring within the open reading frame of a protein coding gene can have several affects. Due to the Wobble base affect, the mutation might not translate to the resulting protein at all, in which case the mutation is a silent mutation. If the mutation transforms a codon

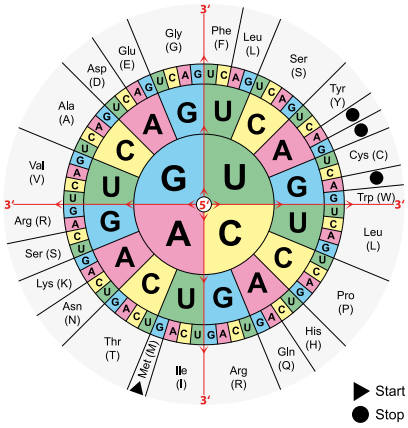


Figure 2.3: Circular representation of the genetic code. As indicated by the arrows, the representation should be read from the inside out to see which codon represents which amino acid. This figure is freely available in the public domain [Com18].

to a stop codon that will artificially terminate protein synthesis, the mutation is a nonsene mutation. If the amino acid represented by the triplet is changed, one refers to the mutation as a missense mutation [Kni06, p.249].

DNA sequence variations can also arise outside protein coding genes, in fact there are numerous known mutations in the non-coding part of the genome, as curated for example in the Cosmic database [F$^+$17a]. These mutations might affect the binding behaviour of DNA binding proteins by interfering with the sequence specificity of their binding profile [Z$^+$17]. An important class of variations at the resolution of a distinct base pair are single nucleotide polymorphisms (SNPs). These are heritable sequence variations existing in at least 1% of the population [Edu14]. As SNPs are heritable, they occur both in somatic and in germline cells. In contrast to that single nucleotide variations (SNV) occur only in somatic cells and are not subject to population based thresholding.

2.1.3 The central dogma of molecular biology

The central dogma of molecular biology has been postulated by Francis Crick in 1958 [Cri70]. It describes possible ways how information can be exchanged between DNA, RNA and proteins. In the version published in 1970, shown in Figure 2.4, the following directions are listed: DNA to RNA, a process we refer to as transcription, RNA to protein, known as translation and DNA to DNA, known as DNA replication. Nowadays, it is known that also the information flow from RNA to RNA [Ahl02], from RNA to DNA [Bal70, TM70] and from DNA to protein [U$^+$02], the latter albeit in an artificial system, do occur as well. Nevertheless, it is still true that information flow starting from the protein does not occur.

The central dogma is of importance, as it highlights that changes on the level of

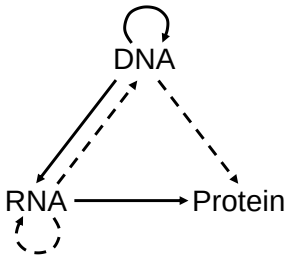


Figure 2.4: The central dogma of molecular biology as published by Francis Crick in 1970 illustrating the exchange of information between DNA, RNA and proteins. Solid arrows have been known to exist at the date of publication, dotted lines were hypothetical. Figure based on Figure 2 from [Cri70] obtained under Springer Nature licence 4487700933705.

DNA and RNA, as well as external effects during transcription and translation, can influence the resulting proteins. In the context of this thesis, epigenetic modifications of the DNA as well as effects on *transcription* are of special interest, therefore transcription is further detailed in the next section. For reasons of completeness, also translation is briefly described.

2.1.4 Transcription

In order for a gene to be expressed, i.e. to observe a phenotypic effect, the first step is to generate a RNA copy of a gene's DNA sequence, a so called transcript [A$^+$05, p.246]. If the expression of genes is assessed, this typically refers to a quantitative measurement of the abundance of their transcripts. A brief overview of the most common experimental methods for this task is provided in Section 2.1.7.

In eukaryotes, the enzyme RNA-polymerase II (RNA-PolII) is responsible for the synthesis of mRNA. In order to transcribe the mRNA molecule, the enzyme needs to bind to the promoter region of the respective gene. The promoter is positioned several base pairs upstream of the first coding nucleotid and among other molecular signatures is often characterized by the presence of CpG islands and presence of the TATA box [BK02]. To guide RNA-PolII to the promoter, a repertoire of TFs is required. These are called general transcription factors (TFIIA to TFIIH). Together, these factors give rise to the transcription preinitiation complex [Kni06, p.344]. Formation of the preinitation complex starts with TFIID, a factor containing TBP, the TATA binding protein, which binds to a gene's promoter. The binding of TFIID to the DNA is further stabilized by TFIIA. TFIIB supports the attachment of RNA-PolII and also pinpoints the enzyme to the TSS. TFIIF guides the polymerase to the promoter and also binds the preinitiation complex. Finally TFIIE guides TFIIH to the promoter. The latter is required to unwind the DNA double helix at the start site and to trigger RNA-PolII to leave the preinitiation complex via phosphorylation [Kni06, p.344ff]. However, TFIIH alone is not sufficient to reliably initiate transcription. The mediator complex is a large protein

complex that interacts with both the preinitation complex and other TFs bound to *enhancers*, regulatory regions controlling tissue specific gene-expression [AT15]. Therefore, understanding the interactions of TFs with enhancers is essential to unravel tissue specific regulation of transcription.

Upon successful transcription initiation, the RNA polymerase is supported by a set of elongation factors such as p-TEFb and ELL ensuring a steady generation of the RNA [Kni06, p.350]. The polymerase slides along the DNA in 3′-5′ direction. Thus the synthesized RNA is created in 5′-3′ orientation. The generated RNA sequence is exactly complementary to the unbound DNA strand, except for the replacement of thymine with uracil. Nucleotides are present in phosphorylated form to provide the energy necessary for the polymerization. Transcription terminates once the polymerase encounters a stop codon [A$^+$05, p.247ff].

The resulting RNA product is called precursor-mRNA (pre-mRNA). At this stage, both introns and exons are part of the sequence. Unlike exons, introns do not code for a protein, they are intervening sequences surrounding coding regions in eukaryotic genes [A$^+$05, p.252]. Introns are removed by a complex machinery, the spliceosome, in a process called splicing [W$^+$09]. Via alternative splicing that is the exclusion of exons from the final mRNA, various proteins can be encoded by the same gene. Alternative splicing is the reason why the transcript sequence can be different from the sequence obtained by linking all exons of the coding gene together. Different transcripts arising from the same gene are called isoforms [B$^+$10a].

At the 5′-end, the final mRNA molecule is attached to a five-prime cap, which is a 7-methylguanosine bound via a triphosphate bridge to the first nucleotide of the mRNA molecule [Sha76]. The cap structure is essential for the mRNA export from the nucleus [LI97], in avoiding degradation [B$^+$06a, G$^+$00] and in terms of translation initiation [R$^+$16]. Adjacent to the cap-structure, there is a chain of 50 to 100 non-coding nucleotides. The protein coding part of a mRNA, known as open reading frame (ORF), is framed by two distinct codons, the start codon, which is usually AUG and one of the three stop codons UAG, UGA, or UAA. Following the stop codon, another sequence of non-coding nucleotides is attached. A mRNA is completed at the 3′-end by the so called Poly-A tail, which is a sequence of up to 200 adenin nucleotides [Kni06, p.431]. The Poly-A tail is linked to mRNA stability [And05]. The mRNA molecule is exported outside the nucleus to act as a blueprint for protein synthesizes during translation [A$^+$05, p.256].

2.1.5 Translation

Translation refers to the synthesis of a protein by ribosomes according to the blueprint provided by the mRNA. Ribosomes consist of two major subunits, a small subunit known as 40S subunit and a larger one known as 60S subunit. In simple terms, the 40S unit slides along the codons on the mRNA while the 60S unit actively joins the amino acids to form the peptide [Kni06, p.441ff]. During initiation of transcription, the 40S subunit is placed together with the amino acid methionine (provided by a transferRNA (tRNA)) at the start codon of the mRNA via specific TFs, so called eukaryotic initiation factors (eIF). Upon the correct positioning of

the 40S subunit the 60S subunit is being attached as well, thereby complementing the ribosome [Kni06, p.441ff]. A set of elongation factors is required to maintain translation of the entire mRNA sequence into a protein. Translation terminates once the ribosome encounters a stop codon which triggers the release of the peptid chain and the disassembling of the ribosome [Kni06, p.441ff].

Further details on translation can be found in a Nature review series on *Translation and protein quality control* [Bio18]. Genuth and Barna provide an overview of regulatory mechanisms of translation in [GB18]. As the projects introduced in this thesis are focusing on transcriptional regulation, translational mechanisms are not introduced.

2.1.6 An introduction to DNA sequencing

DNA sequencing technology underwent an extraordinary advancement. The initial sequencing of the human genome was carried out by two competitors, the privately funded *Celera Genomics* [V$^+$01], with a budget of 300 million USD and the publicly funded *Human Genome Project* [L$^+$01] that finished with a total cost of 2.7 billion USD. Both attempts took several years for completion.

In contrast to that, a current sequencing machine, like the Illumina HiSeq X Ten, is able to sequence the entire human genome in less than three days for less than 1000 USD [Inc17]. While the initial projects used shotgun sequencing [And81] or bacterial artificial chromosomes (BACs) [SKM01] together with Sanger sequencing machines [HC16] which deliver only low throughput, rapid advancements in sequencing technology in the early 2000s led to the development of new sequencing technologies, now known as second generation sequencing. Meanwhile, third generation sequencing has been introduced, which provide longer reads compared to the second generation sequencing technologies, albeit at lower coverage [M$^+$14].

Note that for reasons of brevity, the variety of sequencing technologies including second and third generation sequencing are not introduced in detail here. We refer the reader to "The sequence of sequencers: The history of sequencing DNA" by Heather and Jain [HC16] and to "DNA sequencing at 40: past, present and future" by Shendure *et al.* [S$^+$17b] for a detailed overview. This section focuses on short read sequencing using Illumina sequencers as such data is predominantly used in this thesis. The sequencing process is depicted in Figure 2.5.

Similar to Sanger sequencing, the core concept of NGS is sequencing by synthesis. As the name suggests a DNA double strand is synthesized and during the synthesis, it's sequence is captured using, for instance, fluorescent dyes attached to the newly incorporated nucleotides. High throughput is achieved by conducting this reaction for millions of fragments in parallel [Inc16].

To sequence a sample, its DNA (or cDNA, e.g. for RNA-seq) is fragmented and adapters are ligated to the 5′ and 3′ ends. These fragments are subsequently amplified using a *polymerase chain reaction* (PCR) and purified in a gel electrophoresis, e.g. to remove not ligated adapters or fragments that are too large to be sequenced [Sci19, Inc16]. The resulting set of fragments is called a library [Inc16].

In the next step, the library is brought onto a flow cell, a glass slide, where the

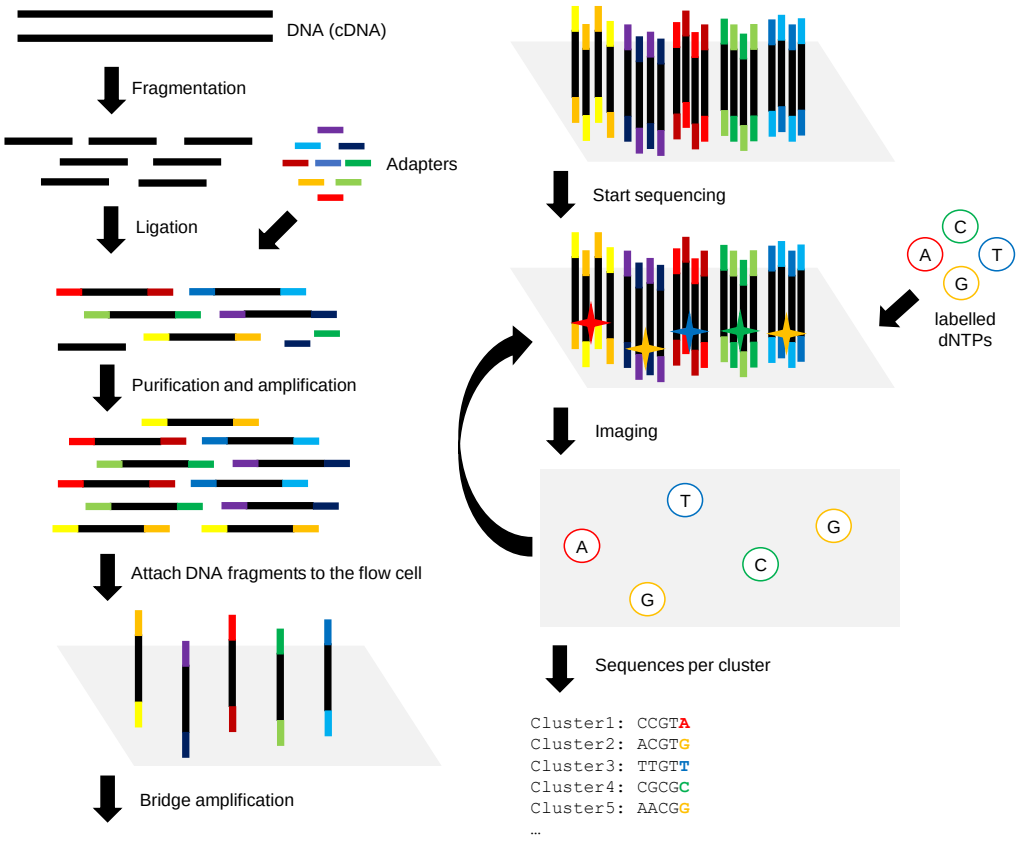


Figure 2.5: Schematics of a next generation sequencing run as performed by Illumina sequencers. Adapters are ligated to fragmented (c)DNA. The resulting fragments are both amplified and purified. Subsequently, they are attached to a flow cell and subject to bridge amplification forming clusters of reads. The actual sequencing is performed following a sequencing by synthesis principle using fluorescent dNTPS. These are added to the sequence in repeated synthesis cycles, inferring the sequence composition nucleotide per nucleotide. The figure is designed by the author of this thesis.

adapters attached to the DNA fragments are captured by complementary oligos that is short DNA sequences between 15 and 30 base pairs in length, bound to the surface of the flow-cell. In case of Illumina sequencing, the bound DNA fragments are amplified using bridge amplification producing local clusters harboring identical, single stranded DNA sequences.

Upon successful amplification, the sequences can be sequenced by adding *"reversible terminator-bound dNTPs"* [Inc16]. These bind to the present DNA fragments one nucleotide per synthesis cycle. Via imaging, the fluorescent dye is captured and the added nucleotide is inferred per DNA fragment. This process is

repeated until the double stranded DNA sequences are completed. The obtained reads can either be aligned against a reference genome, or can be assembled *de novo*.

NGS not only paved the way for affordable and fast whole genome sequencing, it is also the foundation for other applications, e.g. RNA-seq, analysis of DNA methylation, chromatin accessibility and for ChIP-seq analysis of histone modifications as well as TF binding. These are detailed below. Further information on the sketched sequencing method is available in [Inc16].

2.1.7 Experimental methods to measure gene-expression

As mentioned in the context of transcription, gene-expression is usually measured in terms of mRNA abundance. This section provides a brief overview of the two most common high-throughput methods to experimentally quantify *in vivo* gene-expression: DNA microarrays and the more recent RNA-sequencing which is used throughout the work in this thesis.

DNA microarrays

A DNA microarray is a surface containing single stranded DNA sequences that are attached to the array in a grid like structure, where the position and sequence of each DNA sequence is known. These so called probes bind to complementary, labelled DNA sequences obtained from the sample to be analyzed. In case of gene-expression analysis, cDNA (complementary DNA) is used, which is *reversely transcribed* from extracted mRNA. These target sequences are tagged with fluorescent dyes. Upon hybridization with complementary probes, the markers can be evaluated using laser microscopy and the visual readout is transformed into an abundance measure. By up-scaling the number of probes on a microarray, a large number of genes can be screened at the same time [Bum13]. Nowadays, a single microarray, e.g. the Affymetrix *GeneChip*^{TM} *Human Transcriptome Pico Assay 2.0* is sufficient to assess the expression of all known human transcript isoforms [Sci18]. In addition to the assessment of steady state gene-expression levels, also differential analysis can be conducted by using two different dyes for two distinct samples representing different conditions, e.g. healthy and diseased [T^{+}07].

Roger Bumgarner provides a thorough overview of the history and the applications of DNA microarrays, which is not limited to gene-expression estimation but also allows, for instance, protein binding analysis (c.f. Section 2.1.12) or the screening for the presence of SNPs in target sequences [Bum13].

RNA-seq

With the advancement of (NGS) technologies an alternative way of quantifying gene-expression emerged that is to sequence cDNA reversely transcribed from the mRNA extracted from a sample. Compared to DNA microarray based quantification, RNA-seq delivers *"a more detailed and quantitative view of gene-expression,*

alternative splicing and allele-specific expression" [KM15]. Here, we briefly sketch a RNA-seq workflow to quantify gene-expression from bulk data, visualized in Figure 2.6. In this thesis, single-cell gene-expression data has not been used, therefore single-cell RNA-seq is not discussed. Hwang *et al.* provide an introduction into the details of single-cell RNA-seq analysis [H$^+$18c].

The first step of a RNA-seq experiment is to isolate RNA from the sample of interest. Typically this is done for both technical and biological replicates to reduce technical noise and biological variance, respectively. Secondly, a RNA-seq library is created for sequencing. In gene-expression experiments, this involves an enrichment for mRNA, e.g. via a selection of molecules with a poly-A tail, or by a depletion of ribosomal RNA (rRNA). After purification, the RNA is reverse transcribed to cDNA which is sequenced using NGS [KM15]. To adjust for a variety of technical biases and confounders, spike-ins, which are DNA plasmids of varying length, are added to the library and can be used to normalize the transcriptomics data [C$^+$15a, KM15].

Upon the completion of the sequencing, the raw reads are aligned to the reference genome, for instance, using TopHat [T$^+$09]. Alternatively, transcripts can be assembled *de novo*, which is necessary if no reference genome is available. This can be performed e.g. with Oases [S$^+$12c]. Using tools like Cufflinks [T$^+$10], gene-expression can be quantified from RNA-seq bam files, which contain the aligned reads. Recently, methods such as Salmon [P$^+$17a] and Kallisto [B$^+$16d] have emerged, which are avoiding a full alignment of the raw reads and use k-mer based hashing against a reference index instead. While achieving a significant speedup compared to alignment based approaches, these novel methods still achieve accurate quantification results.

The aforementioned tools quantify gene or transcript expression in terms of reads per kilobase of transcripts per million mapped reads (RPKM) (formula 2.1), fragments per kilobase of transcript per million mapped reads (FPKM), or transcripts per million (TPM) (formula 2.2). These metrics attempt to adjust for affects caused by gene length and sequencing depth [C$^+$16c].

The RPKM value for a distinct gene/transcript i can be computed as

$$RPKM_i = \frac{R_i}{l(i)} \frac{10^6}{TR}, \tag{2.1}$$

where R_i is the number of reads aligned to gene/transcript i, TR is the total number of obtained reads and $l(i)$ denotes the length of gene/transcript i in kilobases. FPKM is computed in the same way except that FPKM is designed for paired-end data and ensures that paired reads are not counted twice.

TPM is computed differently as

$$TPM_i = \frac{R_i}{l(i)} \frac{10^6}{TRPK}, \tag{2.2}$$

$$TRPK = \sum_{i \in \mathcal{I}} \frac{R_i}{l(i)}, \tag{2.3}$$

where $\mathcal{I}$ is the set of all considered genes/transcripts. By construction, the sum of TPM values between samples is identical, unlike the sum of RPKM and FPKM values. This makes it easier to compare gene-expression across multiple samples [Blo15].

Further details on RNA-seq can be found in A SURVEY OF BEST PRACTICES FOR RNA-SEQ DATA ANALYSIS by Ana Conesa *et al.* [C+16c] as well as in RNA SEQUENCING AND ANALYSIS by Kimberly Kukurba and Stephen Montgomery [KM15]. RNA-seq has been the workhorse for gene-expression quantification in many projects such as ENCODE [D+12b] and the Genotype-Tissue Expression (GTEx) project [L+13d].

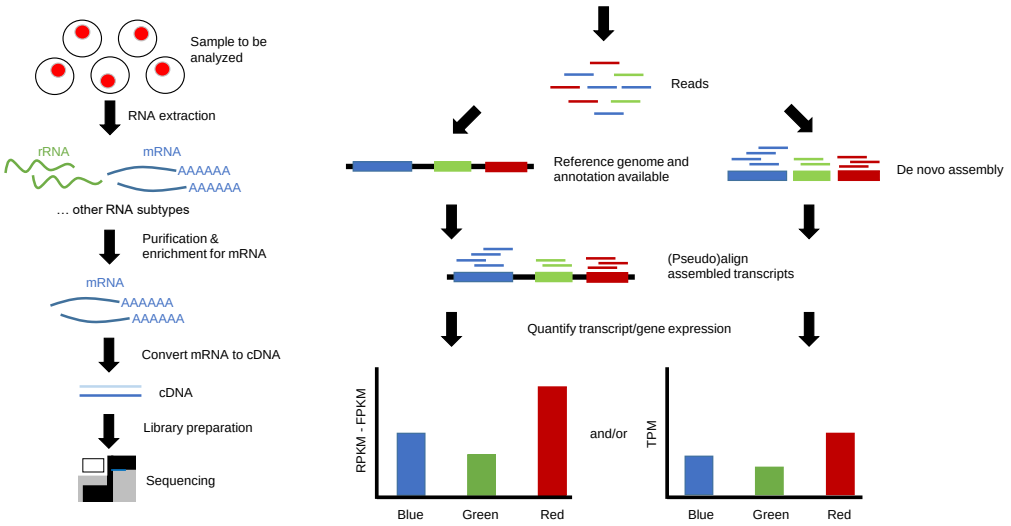


Figure 2.6: Workflow of a RNA-seq experiment: RNA is extracted from a sample of interest and, in case of gene-expression quantification, enriched for mRNA. The latter is reversely transcribed to cDNA, which is being sequenced using NGS. The obtained reads are either aligned against a known reference or are assembled *de novo*. Gene-expression is typically quantified in terms of RPKM, FPKM, or TPM. Figure designed after Figure 1 and 2 from [KM15] under permission and copyright of Cold Spring Harbor Laboratory Press.

2.1.8 Chromatin organization

Eukaryotic DNA is systematically organized and packed. Essential for DNA packing are histone proteins. The complex of DNA, histone proteins and other non-histone DNA binding proteins is called *chromatin*.

In the cell, there are five different kinds of histone proteins: H1, H2A, H2B, H3 and H4. Histone proteins are the building blocks of nucleosomes, the basic unit of chromatin organization. A nucleosome consists of eight histone proteins, two H2A, two H2B, two H3 and two H4 proteins forming the histone octamer, as well as of

a 146bp long stretch of double stranded DNA that is wrapped around the histone octamer in 1.65 left-handed turns [A$^+$05, p.195ff]. Each histone protein is composed of three connected alpha helix structures and possesses a flexible N-terminal tail, attached to the core protein. These tails are composed of mainly positively charged amino acids. Therefore, they are attracted to the negatively charged DNA backbone (Figure 2.7a). Histone tails are highly important for gene-expression regulation, as detailed in Section 2.1.9 [E$^+$14b, BK11].

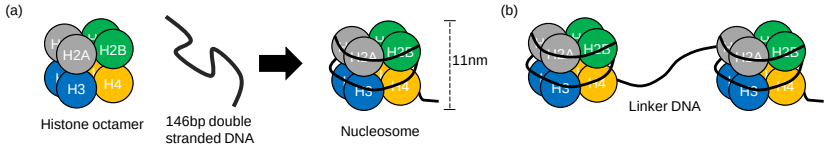


Figure 2.7: (a) Blueprint of nucleosomes. A nucleosome consists of four different kinds of histone proteins forming the histone octamer; specifically two H2A, two H2B, two H3, two H4 proteins as well as a stretch of double stranded DNA of size 146bp. (b) Multiple nucleosomes can form a chain, connected via so called linker DNA. The linker DNA can have varying length, typically at most 80bp [A$^+$05, p.195].

Pairs of nucleosomes can be connected by so called linker-DNA, which can be up to 80bp in length, thereby forming a chain (Figure 2.7b). Literature refers to a series of chained nucleosomes as beads-on-a-string [H$^+$13b], shown in Figure 2.8. At this stage of chromatin compaction, yet another histone protein, the H1 protein, is getting involved. It is attached to the linker DNA between nucleosomes [Z$^+$13a]. H1 is essential for further compaction of chromatin as well as for regulating accessibility of the linker DNA [Z$^+$13a]. As depicted in Figure 2.8, the 11nm structure is further compacted into a 30nm fibre often described as either a solenoid or a zigzag structure. However, we note that the occurrence of the 30nm structure *in vivo* is still being debated [H$^+$13b]. Looping of the 30nm structure leads to further compaction, the 300nm structure. This is further compressed and tightly coiled, thereby forming a section of the final chromosome. The length of the totally compacted chromosome with a width of 1400nm is only $\frac{1}{10.000}$ of the length of the not compacted DNA [A$^+$05, p.196f][Ann08].

A systematic organization of the chromatin is not only indispensable for chromatin packaging but also for gene regulation. Depending on the level of chromatin condensation, chromatin can be divided into two classes.

Heterochromatin describes a fully condensed form, where the chromatin is highly condensed. At this stage, transcription is generally not possible and due to the presence of nucleosomes surrounding the DNA, the latter is inaccessible for DNA binding proteins. Chromatin staying in heterochromatic state all the time forming functional structures e.g. *centromers* and *telomers* is called constitutive heterochromatin. Chromatin that might change its heterochromatic state is named facultative heterochromatin [Kni06, p.161].

Transcriptionally active chromatin that is accessible for other non-histone pro-

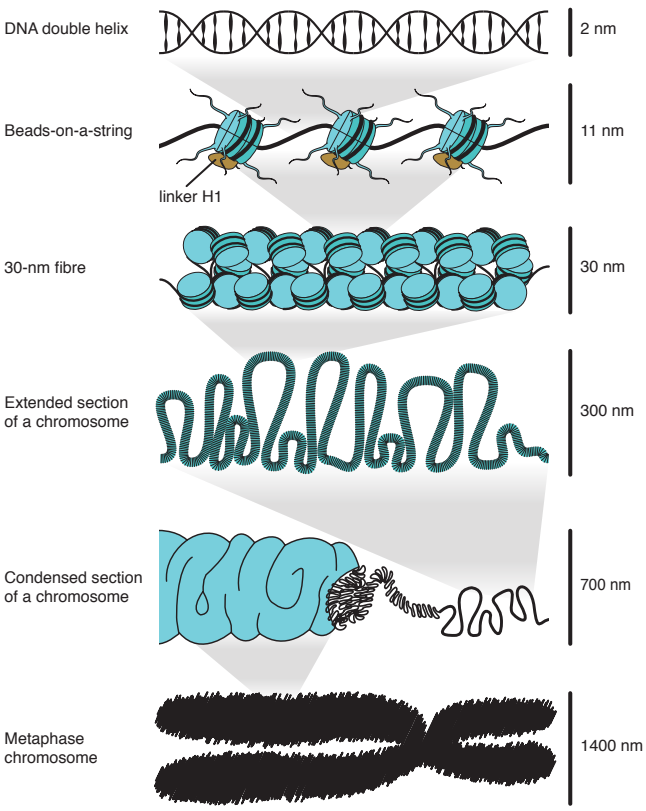


Figure 2.8: Hierarchical overview of chromatin packaging. The DNA double helix is wrapped around the histone octamer, forming nucleosomes. These are connected forming the beads-on-a-string structure. The latter is forming a 30nm fibre, mediated by the H1 protein. The 30nm fibre is further condensed in two stages yielding the final compaction of the chromatin. Figure provided by Fabian Mueller [Mue17].

teins due to the absence of tightly bound nuclesomes, is called *euchromatin*. It was shown that euchromatin is not only closely linked to transcriptionally active genomic regions, but also to regions of regulatory influence, e.g. TF binding sites [Kni06, p.143]. Therefore, the genome-wide identification of nucleosome free regions (NFRs) (also known as nucleosome depleted regions (NDRs)) is important for the elucidation of transcriptional regulation. Several experimental methods that are commonly used for this task are described in Section 2.1.10.

Whether chromatin is present as hetero or euchromatin is largely dependent on the epigenome, specifically on modifications and cross-talk of the histone tails of nucleosomes [BK11]. Especially modifications on the H4 tail have been shown to be highly associated with chromatin state [D$^+$03]. Depending on the nature of the histone modifications (HMs), chromatin remodelling enzymes alter the position of nucleosomes on a stretch of DNA or even disassemble the histone octamer, thereby making the DNA accessible for other DNA binding factors [Kni06, p.401f].

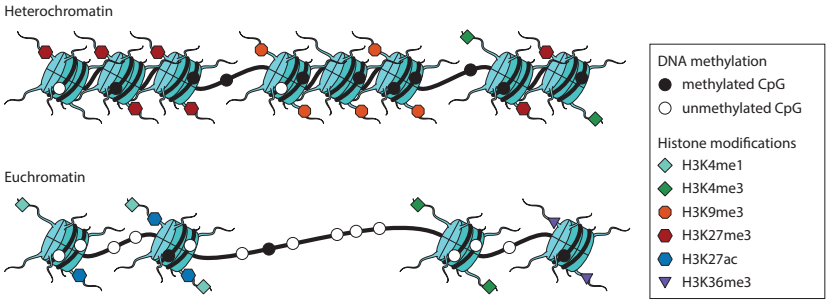


Figure 2.9: Chromatin can be either in a densely packed state, called heterochromatin, or in a lose state, known as euchromatin. The latter is generally accessible for DNA binding proteins. Both states exhibit distinct histone modification and DNA methylation signatures. Figure provided by Fabian Mueller [Mue17].

In addition to HMs, methylation of cytosines in a CpG context has been associated with chromatin condensation as well [M+05]. These observations are illustrated in Figure 2.9, further details on histone modifications and DNA methylation are provided in Section 2.1.9.

Aside from chromatin condensation, chromatin structure fulfills yet another regulatory role. As mentioned above in the context of transcription, the mediator complex allows interactions between the transcriptional machinery bound to the promoter of a gene and DNA binding proteins bound to potentially distant regulatory elements, so called enhancers. These enhancers are brought into close physical proximity by 3D loops of the chromatin, mediated by the cohesin complex [H+17, HL13]. Such loops allow for a crosstalk between genomic regions that are up to several mega bases away. Methods to assess 3D chromatin organization are briefly described in Section 2.1.11.

Such approaches allow the cell type specific characterization of chromatin contacts and revealed the presence of *topologically associating domains (TADs)* [S+16a, D+16c]. TADs can be characterized as *"two regions within a TAD associate on average more frequently with each other than with regions outside of the TAD"* [D+16c]. TADs have been shown to be fairly cell type invariant, thus providing a scaffold for potential cell type specific enhancer-promoter loops [D+16c].

2.1.9 Epigenetic modifications

As mentioned in Chapter 1, epigenetics defines *"the study of changes in gene function that are mitotically and/or meiotically heritable and that do not entail a change in DNA sequence"* [WM01]. In the following, we describe two well established epigenetic marks: Histone Modifications (HMs) and the methylation of cytosines in a CpG context, often referred to as DNA methylation. For reasons of brevity, other epigenetic mechanisms, e.g. regulation through small RNA molecules, are not detailed and are also not analyzed in the remainder of the thesis.

Histone Modifications

As explained above in Section 2.1.8, histone proteins are composed of a conserved inner core made up of three coiled alpha helices as well as of flexible amino acid chains, known as histone tails, anchored at the inner core interacting with the DNA wrapped around the histone octamer. These histone tails can be modified e.g. by methylation, acetylation and/or phosphorylation [ZG15].

HMs are described using a distinct nomenclature: Initially, the histone protein whose tail is modified is stated, followed by the one letter abbreviation of the modified amino acid, its position in the histone tail and an abbreviation of the modification. For instance, H3K18ac refers to an acetylation of the lysine at position 18 on the tail of histone H3 [Kni06, p.396].

The association of HMs to gene-expression and regulation are manifold. For example, H3K4me3 has been associated with active promoters, whereas H3K36me3 has been associated with elongation of transcription. These effects are achieved by various means. For instance, modifying the lysines with acetyl groups neutralizes the attractive force of the normally positively charged tails to the negatively charged DNA backbone. Therefore, H3K27ac induces a looser connection between histone tail and DNA, allowing DNA binding proteins to access the DNA enabling chromatin remodellers to displace the entire nucleosome [BK11]. However, not all HMs carry out their function via structural changes, e.g. H3K4me3 is detected by TFIID mediating its binding to DNA [vI$^+$08]. In Table 2.1, an overview of the most commonly analyzed six HMs is provided. The occurrence of HMs *in vivo* can be analyzed using ChIP-seq (chromatin immuno precipitation followed by sequencing) experiments. The ChIP-seq assay is explained in the next section.

A machinery of enzymes is involved in maintaining and altering HMs. For instance, acetylation marks are set by Histone-Acetyl-Transferases (HATs) and can be removed by Histone-Deacetylases (HDACs) [LT03]. Histone Methyltransferases (HMTase) on the other hand are responsible for adding methyl groups to amino acids of the histone tails [Tri04]. HMs influence and depend on each other, a phenomenon known as histone crosstalk [L$^+$10a].

Abbreviation	Type of modification	Relation to gene-expression	Reference
H3K4me3	Tri-methylation	Active promoter	[H$^+$07]
H3K27ac	Acetylation	Active enhancer	[C$^+$10b]
H3K4me1	Mono-methylation	Active and poised enhancers	[C$^+$10b]
H3K27me3	Tri-methylation	Facultative heterochromatin	[J$^+$16b]
H3K36me3	Tri-methylation	Elongation mark	[WC12]
H3K9me3	Tri-methylation	Constitutive heterochromatin	[GS17]

Table 2.1: Overview of the six commonly analyzed HMs in genome-wide studies

During DNA replication, HMs are maintained by several means that are still

under investigation and are not yet fully understood [B+13c]. Nevertheless several insights have been gained already. For example, the repressive mark H3K9me3, which is associated with heterochromatin, is re-established at the DNA replication fork by a complex interplay of the TF heterochromatin protein 1 (HP1), the *chaperone* Chromatin Assembly Factor 1 (CAF-1), the protein proliferating cell nuclear antigen (PCNA) and several small RNAs [Q+04, L+11].

In contrast to that, the active promoter mark H3K4me3 as well as the repressive H3K27me3 mark have been reported to be enriched prior to *S-phase* and are maintained by dilution during DNA replication rather than by being set *de novo* [L+12a].

The transgenerational inheritance of HMs is subject to ongoing research and the means are highly debated although several studies provide evidence that epigenetic information is forwarded through generations [LB13, BM+18]. An overview of the current state of research is provided in [C+14a].

ChIP-seq analysis

ChIP-seq is a genome-wide profiling technique utilizing antibodies for the localization of specific proteins of interest such as TFs, RNA polymerases, or histones bound to DNA [Mar07].

A ChIP-seq experiment consists of four major steps, depicted in Figure 2.10: Firstly, DNA-binding proteins are crosslinked to the DNA using formaldehyde. Secondly, the DNA is fragmented via sonication or digestions using an endonuclease. Next, the protein of interest is subject to immunoprecipitation using a specific antibody that binds the assayed protein. The sample is purified and marked proteins are pulled-down. The last experimental step is to reverse the crosslinks and to sequence the DNA-fragments bound by the protein of interest.

The genomic locations of the assayed protein can be determined computationally using an enrichment analysis. To account for potential sequencing biases, a control experiment without an antibody is performed. The result of the control experiment is also known as *Input*. It can be considered as a baseline in the enrichment analysis of the protein of interest.

The results of a ChIP-seq experiment reflect a snapshot of the binding behavior of a protein of interest across many cells at a specific time point. In recent years, the ChIP-seq technique became prevalent in biology and has been used on a large scale especially by ENCODE to characterize the binding behavior of TFs and to characterize the landscape of several histone modifications in many diverse cell types [D+12b]. In the meantime, the ChIP-seq protocol has been applied to the single-cell level [R+15b].

However, the technique has some important limitations. First of all, not for every target protein a reliable antibody is available. Also, antibody efficiency can differ between different batches of antibodies compromising reproducibility [Vos14]. Secondly, the experiments are still expensive, labor-intensive and time consuming, so it is impossible to chip all known proteins of interest in various species and cell types. Another drawback is that, especially in case of TFs, it is not certain whether

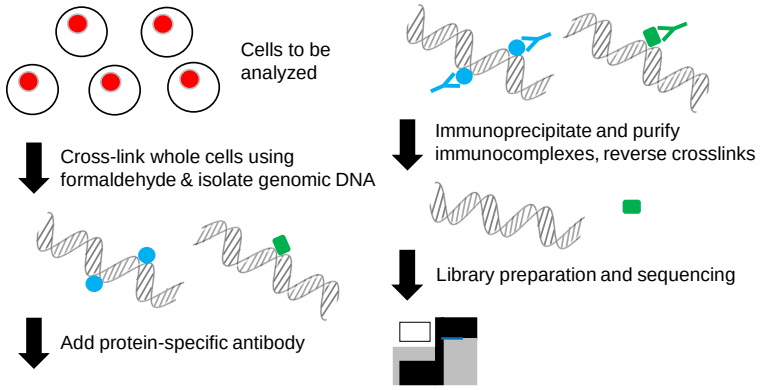


Figure 2.10: Schematic overview of a ChIP-seq experiment, after Figure 1 from Mardis E.R. [Mar07] obtained under Springer Nature license 4499340794413.

the ChIP-seq experiment identifies direct DNA-binding events or indirect binding events, e.g. a TF is bound to another protein but not directly to the DNA [Fur12].

Methylation of cytosines in CpG dinucleotides

A well characterized epigenetic mark is the naturally occurring methylation of cytosines in CpG dinucleotides leading to 5-methylcytosine (Figure 2.11).

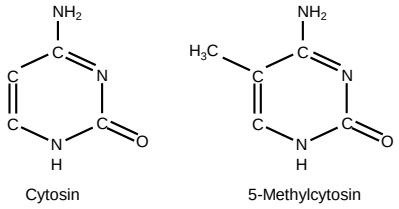


Figure 2.11: Molecular structure of cytosin and 5-methylcytosine.

Genome wide studies suggest that DNA methylation is repressing retrotransposons, is regulating *monoallelic gene-expression* of imprinted genes, is responsible for X chromosome inactivation and is regulating the binding behaviour of TFs to promoters, although 75% of human promoters are located within unmethylated CpG islands [E+17].

DNA methylation is established and maintained by a group of specialized enzymes, so called DNA methyltransferases (DNMTs). A review on their regulation and function is provided in `The DNA methyltransferase family: a versatile toolkit for epigenetic regulation` [Lyk18]. By interactions of the DNMTs with DNA during replication, DNA methylation can be efficiently maintained during cell division [H+12c]. The removal of a methyl group requires a series of oxidation reactions, delineated in [I+11, WZ14]. Importantly, 5-methylcytosines are

less stable than cytosines and can undergo deamination, which converts them to a thymin. This is a key process for the acquisition of mutations in eukaryotes [Kni06, p.265f].

The presence of DNA methylation is profiled genome-wide using bisulfite conversion of the genome, a process detailed in the next section. Although DNA methylation in a CpG context is not considered for further analysis in this thesis, it is important for another assay, NOMe-seq, introduced below.

Bisulfite conversion to characterize DNA methylation genome-wide

The bisulfite conversion was introduced in 1992 by Frommer *et al.* to analyse 5-methyl-
cytosines in DNA sequences [F$^+$92]. The bisulfite conversion affects non-methylated cytosines such that they are converted to uracils, a base normally occurring in RNA only. Methylated cytosines are not affected by the bisulfit treatment.

While bisulfit treated sample have been analyzed using PCR experiments in the beginning, the development of NGS technologies allows a high-throughput characterization of bisulfit converted reads [Z$^+$13b]. Upon reverse transcription, the unmethylated cytosines that have been replaced with uracils are amplified as thymines.

Note that the bisulfite conversion is not depending on the DNA sequence surrounding (5-methyl)cytosines. Aligning the bisulfit reads against a reference genome allows a quantitative characterization of DNA methylation. Further, we point out that standard aligners should not be used to align bisulfit reads due to the reduced alphabet and special nature of this bisulfit data [G$^+$13].

2.1.10 Experimental Methods used to characterize the chromatin accessibility landscape of a cell

There are several experimental approaches to study nucleosome occupancy, which is in the remainder of this thesis also termed chromatin accessibility. The most prevalent methods to monitor chromatin organization are DNaseI-seq [W$^+$79], ATAC-seq [B$^+$13d] and NOMe-seq [K$^+$12]. The conceptual idea of these assays is detailed in Figure 2.12 as well as in the following three sections.

DNaseI-seq

DNaseI is a endonuclease, which is a class of enzymes cleaving internal phosphodiester bridges of both single and double stranded DNA [Kni06, p.24]. DNaseI has already been used in 1979 by Wu *et al.* to analyze chromatin activity of the heat shock protein in Drosophila melanogaster [W$^+$79]. Due to advancements of the DNaseI protocol [S$^+$04] and the development of NGS technologies, analyzing chromatin accessibility using DNaseI at the genome wide level became possible [B$^+$08].

The DNaseI enzyme cuts freely accessible DNA releasing small DNA fragments. After digestion of a sample with DNaseI, these fragments are extracted via size-selection and sequenced. DNaseI Hypersensitive Sites (DHSs) can be obtained

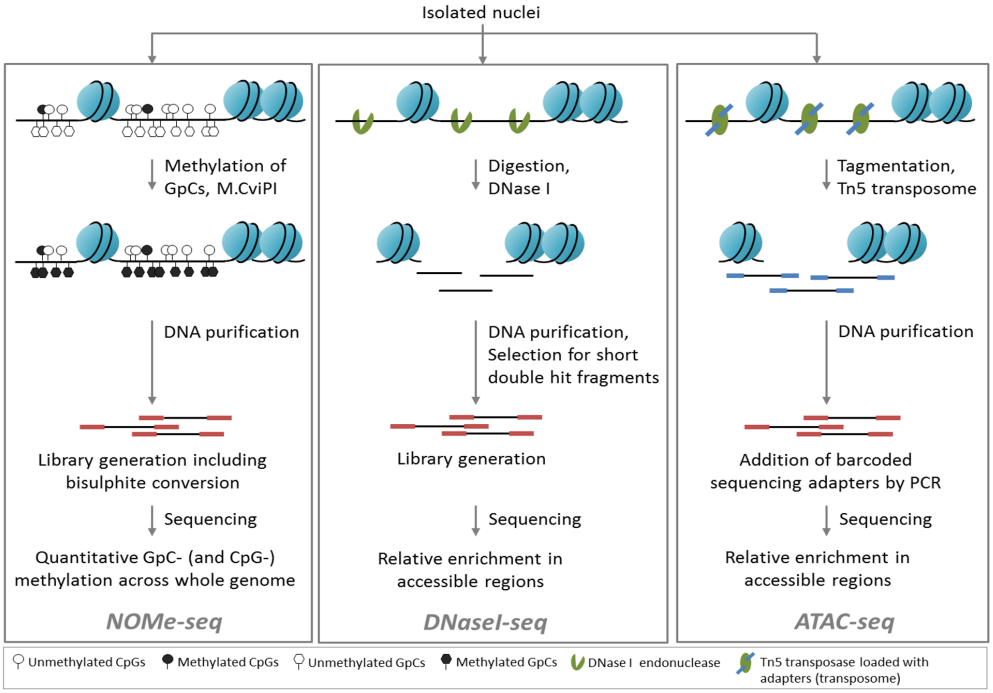


Figure 2.12: Schematic workflow of three common assays to assess chromatin accessibility. In NOME-seq, the methyltransferase M.CviPI methylates all accessible GpCs in the genome. DNA is extracted, purified and bisulphite converted. Upon sequencing NOMe-seq provides information on natural CpG methylation as well as on chromatin accessibility deduced from GpC methylation. In DNaseI-seq, the endonuclease DNaseI cleaves the DNA at nucleosome free regions. Upon digestion, the resulting DNA fragments are purified and enriched for short DNA fragments, which are subject to sequencing. The relative enrichment of the fragments along the genome provides insights on the accessibility of the chromatin. In ATAC-seq, the Tn5 transposase is adding a transposable element to accessible parts of the genome. After DNA purification and sequencing, the relative enrichment of these elements provides details on chromatin accessibility on the whole genome scale. Figure provided by Nina Gasparoni.

using peak calling algorithms such as MACS2 [Z+08a] or JAMM [I+15] applied to the genome-wide DNaseI signal.

Large scale DNaseI studies by Blueprint, Roadmap and ENCODE [A+12, K+15, D+12b] showed that this approach boosts the identification of regulatory elements in different cell and tissue types [T+12]. Moreover, it was shown that numerous genetic variants identified in genome-wide association studies (GWAS) overlap with DHSs [S+12a] and DNaseI footprinting has been shown to be highly informative for TF binding [G+14d] (see Section 2.1.12 for further details). These examples stress

the importance of DHSs for gene regulation.

Although DNaseI-seq is a fairly old and established assay, it is still challenging in several ways. First of all, the experiments require pre-testing to identify the right incubation conditions, because both under- and over-digestion would greatly influence the obtained results and lead to incorrect conclusions [H$^+$14c]. This pre-testing requires additional biological material, which, especially for primary cells, can be limited. Nevertheless, DNaseI-seq protocol has been scaled down to the level of single cells [J$^+$15].

ATAC-seq

Another method to assess chromatin accessibility is ATAC-Seq, where ATAC is the abbreviation for Assay for Transposase-Accessible Chromatin. ATAC-Seq has been introduced in 2013 with the goal of finding a faster and less demanding method in terms of cell material than DNaseI-seq [B$^+$13d]. The assay uses a hyperactive Tn5 transposase, a prokaryotic transposase that cuts accessible DNA and simultaneously inserts pre-loaded oligonucleotides into the sequence [B$^+$15b]. Upon DNA isolation and PCR-amplification, libraries can be prepared and sequenced. As for DNaseI, peak calling can be used to determine regions of enrichment from the aligned reads. There are several advantages of ATAC-seq over DNaseI-seq. The over-digestion problem mentioned for DNaseI-seq is not a severe issue with ATAC-seq, as the Tn5 reaction is an end-point reaction. Furthermore, ATAC-seq can be carried out in less than 3 hours at low costs, making the method of choice for large scale studies [C$^+$18b]. Moreover, ATAC-seq can be more easily applied to the single-cell level than DNaseI-seq [M$^+$16c], therefore ATAC-seq might be gaining even more importance in the future.

NOMe-seq

NOMe-seq is an alternative method to analyze chromatin accessibility pursuing a completely different strategy than both DNaseI-seq and ATAC-seq. Instead of using an endonuclease, NOMe-seq is based on a methyltransferase, specifically the enzyme M.CviPI methylating cytosines in a GpC (not a CpG) context. The enzyme M.CviPI was identified in 1998 in Chlorella virus NYs-1 and has been cloned into both Escherichia coli and Saccharomyces cerevisiae [X$^+$98].

NOMe-seq is a method to characterize both chromatin accessibility and DNA methylation at CpGs and has been introduced by Kelly *et al.* in 2012 [K$^+$12]. As GpC methylation, unlike CpG methylation, does not occur *in vivo*, incubating cells or isolated nuclei with the M.CviPI GpC methyltransferase will result in methylated GpCs, if the corresponding sequence is not occupied by nucleosomes or other DNA binding proteins. Upon incubation, the DNA is extracted, Bisulfit converted and sequenced. Thereby, one obtains a readout of standard CpG methylation as well as of GpC methylation from the same DNA molecules. GpC methylation can be used to infer information on chromatin accessibility, for example using the findNDR tool [T$^+$14].

Like ATAC-seq, NOMe-seq does not require a lot of input material and the methylation is an endpoint reaction, therefore there is only a little risk of over-exposure. A draw-back of NOME-seq is that the accessibility information depends on the presence of GpCs in the genome, therefore the data will be sparse in GpC depleted areas. Just like ATAC-seq, also NOMe-seq can be fairly easily applied to single-cells [Pot17].

2.1.11 Experimental Methods used to characterize long range chromatin contacts

There are several different techniques capturing the three-dimensional structure of chromatin. In this thesis, only a subset of them have been used, specifically Hi-C, Capture Hi-C and ChIA-PET. These are detailed in the following subsections. For a general overview of all available methods, we refer the reader to "C-ing the genome: A compendium of chromosome conformation capture methods to study higher-order chromatin organization" by Barutcu *et al.* [B$^+$16a].

3C based methods

The 3C (chromosome conformation capture) protocol was the first high-throughput protocol to assess chromatin interactions *in vivo*. Hi-C and capture Hi-C used in this thesis are based on the 3C technique, which is therefore briefly explained as well.

The goal of a 3C experiment is to asses the interaction frequency of two loci based on their spatial proximity in three-dimensional space, averaging out the effects of many samples within a population. This is achieved by cross-linking the nucleus with formaldehyde and subsequently fragmenting the chromatin using restriction enzymes. The digested ends are ligated, purified and analyzed either via PCR or sequencing. Note that the PCR requires primers designed for a set of regions of interest. Thus 3C is often described as a "hypothesis driven" technique, i.e. it can be used to validate hypothesized contacts, but it is not designed to discover genome-wide contacts *de novo* [D$^+$02, B$^+$16a]. Hi-C can be seen as an up-scaled version of the original 3C approach. It allows to measure interactions between any pair of genomic regions resulting in genome-wide contact matrices. By amplifying ligation products of the entire genome a broad coverage of all genomic regions can be obtained [LA$^+$09].

The entire Hi-C workflow is sketched in Figure 2.13. The major difference to the 3C approach is that during ligation of the sticky ends produced by the restriction enzyme, biotin is added to mark the ligation sides. The generated products are purified and again subjected to DNA shearing. Sequences marked with biotin are pulled down using streptavidin beads and are analyzed using NGS. The resulting reads can be aligned to a reference genome. Local signal enrichment's provide insights on chromatin interactions [LA$^+$09]. These can be identified, e.g. with the tool HiCCUPS [LA$^+$09].

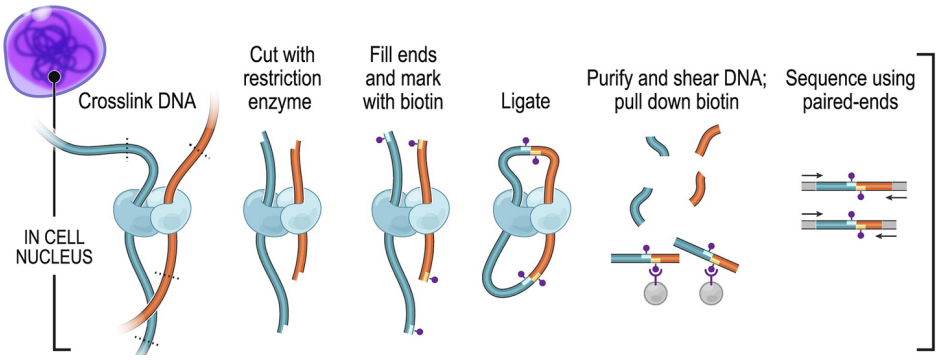


Figure 2.13: Schematic workflow of a Hi-C experiment. DNA contacts are fixed and crosslinked. DNA-strands are cut using a restriction enzyme. Ends of remaining DNA complexes are filled up and marked with biotin and ligated. After purifying and shearing DNA, biotin marked fragments are pulled down. Paired-end sequencing produces reads of pulled down fragments, which are aligned to the reference genome. Figure provided by Fabian Kern, designed after Figure 1A from [LA$^+$09] under Science license number 4498821444625.

While Hi-C provides genome-wide contact information, the resolution of this information is at best in the range of 1kb to 40kb. Resolution refers to the size of the interval in genomic space were an interaction was found. Although the achieved resolution is sufficient to unravel sub-TAD structures, it might not be sufficient to pinpoint the precise location of loops. To circumvent this limitation, researchers developed capture Hi-C. Capture Hi-C follows the same experimental workflow as a standard Hi-C experiment, with the difference that the library is enriched for a distinct set of sequences prior to sequencing, e.g. promoter sequences or a set of distinct genes of interest. This allows to sequence these regions at much greater depth, thereby improving resolution [D$^+$14].

ChIA-PET

While 3C based methods are solely based on deciphering interactions of DNA with DNA, another goal is pursued by Fullwood *et al.* In 2009, they presented an approach to describe genome-wide interactions between DNA and a distinct protein, called chromatin interaction analysis by paired-end tag sequencing (ChIA-PET). In ChIA-PET, chromatin interactions are fixed by formaldehyde cross-linking. The cross-linked molecules are subject to sonication and the DNA-protein complexes are extracted using ChIP-seq. The DNA sequences contained in the extracted complexes are ligated using proximity ligation and are sequenced using paired-end tags (PET). Aligning the resulting reads to a reference genome provides insights about (long range) chromatin interactions with the screened protein [F$^+$09].

2.1.12 Transcription Factors

Transcription Factors (TFs) are proteins essential for orchestrating gene function and response to internal and external stimuli. TFs either bind directly to the DNA according to a TF specific sequence preference or interact with already bound factors via protein-protein interactions. Although the importance of TFs is acknowledged by the field, only for about 30% of TFs, their regulatory function is known [V$^+$09a]. At the same time, deregulated TFs have been related to developmental disorders [V$^+$09a], TFs are also related to cancer [Cle04] and to autoimmune diseases [vdVN07].

As mentioned in the context of transcription, a complex of so called general transcription factors binds to the promoter of a gene and is essential to facilitate binding of the RNA polymerase [Kni06, 344ff.]. While the assembly and interplay of these TFs is highly conserved and thus similar between different genes and cell types, the mediator complex links the transcriptional initiation machinery to distal regulatory regions called enhancers and repressors acting as binding sites for additional TFs required to drive cell type specific transcription [K$^+$10a]. This interplay is graphically shown in Figure 2.14, the importance of enhancers and repressors is further detailed in Section 2.1.13.

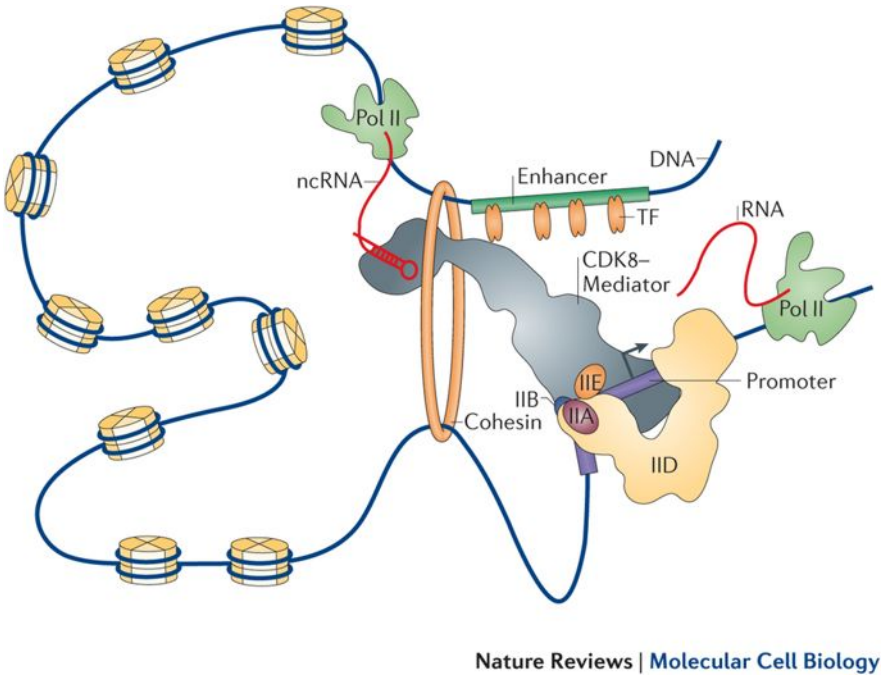


Figure 2.14: Regulation of transcriptional initiation through an interplay of TFs bound at the promoter and an enhancer. The enhancer is brought into proximity of the Promoter via DNA looping, stabilized by Cohesin. Figure obtained from [AT15] under Springer Nature license 4500250091468.

Aside from the role in initiating transcription, the fine tuning of cell type specific expression is an important task, especially in light of cellular differentiation, cell cycle control, intracellular signaling and response to external factors [F$^+$17b, L$^+$14a, Fra08, O$^+$13].

In any of these tasks, TFs can either function as an activator or as a repressor, i.e. they influence gene-expression either positively or negatively. Activators influence gene-expression e.g. by recruiting HATs or chromatin remodellers to make the chromatin more accessible for the transcriptional machinery. Alternatively, they might exert their function by modifying proteins, e.g. via phosphorylation. An important class of TFs with activating function are pioneering TFs. These have been shown to be able to bind even to heterochromatin and initiate the transformation to euchromatin. Thereby, they are indispensable for cell fate decisions [IDZ16]. The opposing function is carried out by repressors. These are interacting, for instance, with HDACs to reduce chromatin accessibility.

We have illustrated that TFs are essential regulatory proteins, however, the question arises how the activity of TFs can be influenced. Up till date, several different mechanisms are known. For example, TFs are regulated by regulation of their own transcription, e.g. TFs can regulate themselves via negative feedback loops. In a negative feedback loop a TF binds to a regulatory region of its own gene, in this case a repressor, thereby reducing its own transcription [B$^+$14c]. As for other proteins, TFs can also be post translationally modified, e.g. by phosphorylation [P$^+$13d]. Also, they might depend on other TFs to function, as they act together in a protein complex [WH14]. We have already mentioned the importance of chromatin accessibility for TF binding, so we stress again that regulating chromatin accessibility is another way regulating TF activity by controlling their potential binding sites [H$^+$07].

Earlier in this chapter, we mentioned that translation occurs at ribosomes outside the nucleus. Consequently, once the mRNA of TF has been translated into the actual TF protein, the protein needs to be guided back into the nucleus to carry out its regulatory function. As this transport is also relying on interactions with other proteins, regulating a TF's transit into the nucleus is yet another way of regulating TF activity [KO00]. Similarly, membrane-bound transcription factors are TFs anchored at the cellular membrane in a stand-by mode. They are being activated and transported into the nucleus upon external stimuli [L$^+$18b].

As mentioned above, TFs bind the DNA at distinct sequences. The preferred binding sequences as well as the functionality of a TF is depending on its three dimensional structure. The 3D structures involved in DNA binding are categorized into several DNA-binding domains. The most common DNA-binding domains in eukaryotes are the homeobox, the helix-loop-helix domain and the zinc finger domain. TFs sharing the same DNA-binding domain often exhibit very similar DNA sequence preferences. This is a drawback for their accurate computational binding prediction [P$^+$08].

The homeobox domain consist of roughly 60 amino acids forming three α-helices. The first two helices are oriented antiparallel to each other, connected with a loop.

The third helix is placed at a right angle to the other two helices. Further, the third helix is binding to the DNA by interacting with the DNAs major groove. The family of Oct-factors involved for example in embryonic development contain a homeo domain [BM17].

The helix-loop-helix domain is composed of two α-helices. One helix binds the DNA at the major groove and is linked to the second, shorter helix via a loop. Helix-loop-helix-domains tend to build dimers with proteins from the same protein family. Often, these bind to the DNA at loci with a palindromic sequence. The E12/E47 proteins regulating immunoglobin genes are known to hold a helix-loop-helix domain [MM00].

Zinc finger domains are assembled of one or more copies of a roughly 30 amino acid long sequence forming a loop. This loop is stabilized by a zinc ion, which is fixed to cysteine- and histidine side chains. The participating amino acids form two distinct secondary structures, a α-helix and a β-sheet. The α-helix is attached to the major groove of the DNA and the β-sheet interacts with the DNAs backbone to stabilize the binding of the entire factor. The TF Sp1, which is involved in cell growth, apoptosis and chromatin remodeling, contains three zinc finger domains and binds to TATA-box free promoters [N$^+$97].

To elucidate the function and importance of TFs, their DNA binding preference must be learned and their potential cell type specific binding sides throughout the genome must be identified. In the next section, we sketch experimental approaches how to determine the binding preference of TFs. The previously introduced ChIP-seq experiments can be used to pinpoint the binding sites of TFs in a genome wide manner, which has been done by ENCODE on a large scale for many TFs and other proteins involved in transcriptional regulation such as RNA-Pol II [D$^+$12b]. However due to experimental limitations, CHIP-seq experiments can not be carried out for all TFs in all cell types. Computational methods predicting TF binding sites attempt to fill this gap. An overview of such methods is provided in Section 3.1. Due to the extraordinary importance of TFs in orchestrating cell type specific gene-expression, a precise knowledge base of their binding, functionality and interplay is essential for a better understanding of gene regulation.

Determining the sequence specificity of TFs

Several experimental methods have been suggested to infer the binding preference of TFs. In addition to the already introduced ChIP-seq method [Bai11], also protein binding microarrays (PBMs) [Bul07] and the SELEX assay (systematic evolution of ligands by exponential enrichment) [ES90, J$^+$10] can be used for this purpose.

TF binding motifs can be derived from TF ChIP-seq data using computational methods, e.g. DREME or Dimond [Bai11, GPGK13].

PBMs are similar to DNA microarrays mentioned before. With PBMs a tagged protein is bound to a microarray containing a set of predefined sequences. The readout of the array provides information on the composition of the sequences that were bound by the assayed TF [Bul07].

The SELEX experiment is an iterative procedure, designed to identify oligonucleotides binding to a tested protein with high affinity. The factor of interest is brought together with a pool of candidate single stranded DNA oligos. After incubation and washing of the unbound oligos, the bound ones are amplified and slightly modified during subsequent PCR. After several SELEX cycles, the best matching oligos are extracted and the binding preference of the factor can be determined [ES90].

2.1.13 Enhancers and repressors

We have mentioned that TFs bind both to the promoter of genes which are in close proximity to a gene's transcription start site (TSS) and to enhancers, regulatory regions that can be located far way from the regulated gene in genomic space [Y$^+$15]. Since enhancers have been described for the first time in 1981 by Banerji *et al.* [B$^+$81], numerous studies shed light on their functional relevance.

For example, enhancers were shown to be essential in cell differentiation [LA$^+$14]. Also, it has been reported that mutations occurring in enhancer regions, can not only lead to changes in gene-expression [K$^+$83] but can also increase the probability to contract certain diseases, for instance *Hirschsprung's disease* [E$^+$05], *Type 2 Diabetes* [Lys08], or acute myeloid leukemia (AML) [G$^+$14c]. These effects are likely to be caused by an altered binding of TFs due to mutations occurring in enhancer sequences [H$^+$14f]. Ongoing research suggested that enhancers might be therapeutic targets, e.g in case of sickle cell disease by reducing the expression of the TF BCL11A via remodeling of one of its enhancers [Y$^+$15, S$^+$15].

While it is established in the community that enhancers drive transcription by interactions with the transcripitional machinery, several other means of enhancer function have been suggested. For example, it was discovered in 2010 that parts of non-coding enhancer regions are actually being transcribed into long non-coding RNAs (ncRNAs) so called enhancer RNAs (eRNA). Interestingly, eRNA expression has been shown to be correlated to the expression of the enhancers' target gene [K$^+$10b]. However, a recent study using single cell RNA-seq data suggests that eRNAs accumulation is not required to maintain the transcription of the enhancers' target gene, illustrating that still only little is known about the role of eRNAs [R$^+$17a]. Another class of long ncRNAs has been described by Oron *et al.* It is hypothesized that these ncRNAs serve as a scaffold for the assembly of TFs required for transcription [OC11]. Yet another way of enhancer function was described for several genes that require the presence of H3S10ph to initiate transcription by releasing RNA-polymerase II from promoter-proximal pausing. The release of RNA-Pol II is mediated via histone crosstalk between the respective enhancers and promoters [OC11].

To understand the function of enhancers further, putative enhancer regions need to be identified and linked to their target genes. Recently, considerable progress has been made in identifying putative enhancer regions utilizing epigenetics data, especially *H3K27ac* and *H3K4me1*. These two marks have been used in several computational approaches to suggest putative enhancer regions. These are introduced

in Section 3.4.5. Furthermore, DHSs not located at promoter regions are also good candidates for putative enhancer regions due to their enrichment for TF-binding sites and their positive correlation to gene-expression [PR$^+$11, R$^+$15a].

However, it is still not fully understood how enhancers interact with their potentially distantly located targets. The most prevalent hypothesis is that enhancers are brought to close proximity to their target genes by chromosomal re-organization and DNA-looping. This hypothesis is known as the looping model. It is opposing the so-called scanning model stating that an enhancer is usually influencing only the active promoter located most closely to the enhancer in genomic space [BK98]. Experimental evidence could be found for both models, hence it is likely that both mechanisms are occurring in nature [Y$^+$15].

For instance, long range enhancer-gene interactions, as proposed by the *looping* model, have been experimentally determined using fluorescence in situ hybridization (FISH), via enhancer RNAs (eRNAs) and their correlation to target genes, or 3C-based high-throughput methods such as Hi-C and Capture-Hi-C as introduced before in Section 2.1.11 [M$^+$15b]. Detailed analyses of individual genes, e.g. the *β-globin gene* showed that multiple chromatin contacts occur simultaneously at one genomic loci and also overlap with DHSs [dLG03]. Because of the tissue specificity of enhancers and their importance for tissue specific gene-expression the exact knowledge of their location and their accurate linking to target genes is inadmissible for a clear understanding of gene regulation.

2.1.14 CRISPR/Cas9 and viability screens

The Cas9 enzyme is a DNA-endonuclease. Unlike the already introduced DNaseI endonuclease, which is cutting the DNA wherever it is accessible, the Cas9 enzyme is a RNA-guided endonuclease. It cuts DNA at a sequence complementary to that of the guide RNA (gRNA). Cas9 has been found in bacteria, where it is guided by clustered regularly interspaced short palindromic repeats (CRISPR), DNA fragments derived from viruses that have infected the bacteria before. The CRISPR/Cas9 machinery is playing a vital role in the immune response in prokaryotes by the targeted degradation of pathogens [BM14].

Due to the high specificity and the relatively simple way of controlling Cas9, the enzyme has been used for research purposes in several ways. Using the Cas9 enzyme in its wild-type form to induce breaks of the DNA-double strand are known as CRISPR knockout screens (CRISPR-ko). The name arises from the observation that the repair of the DNA double strand, causes insertions or deletions that might affect TF binding sites and thereby influence gene-expression [S$^+$14b]. Other applications include the usage of Cas9 to specifically alter DNA methylation or histone modifications [C$^+$16b, H$^+$15a]. An overview of various other applications of Cas9 is presented in a review by John G. Doench [Doe18].

CRISPR/Cas9 allows to perform high throughput perturbation analysis of large populations, e.g. viability screens. Viability screens attempt to highlight genes affecting cell fitness as well as their regulatory elements. Millions of cells are transduced with a distinct gRNA. In a positive selection, all cells are challenged, for

instance, using a drug. Cells that could nevertheless proliferate and are thus enriched over time are subject to DNA sequencing. By comparing the gRNA to the genomic sequence, the most likely target(s) of the used gRNAs can be determined and thereby genes or putative regulatory sites involved in mediating resistance can be determined [Doe18].

2.1.15 Looking beyond transcriptional regulation through TFs

In this thesis, we focus especially on the already introduced regulation through TFs at promoter and enhancer regions. However, there are numerous other regulatory mechanisms beyond those holding an important role in gene regulation. This section attempts to provide an overview based on review articles on transcriptional and translational regulation, as well as on post-translational protein modifications [LY13, H$^+$12b, Spo18].

In addition to the binding of TFs regulated by chromatin accessibility, DNA methylation, HMs and several non coding RNA molecules are known to have regulatory function. For instance, long non coding RNAs (lncRNA) are associated with several regulatory functions [B$^+$16b]. They are involved in recruiting chromatin remodellers [FT$^+$09]. Also, they can interact with TFs or directly bind RNA-PolII and thereby directly influence transcription [Z$^+$08b, WB08]. Additionally, lncRNAs have been implicated in monoallelic gene-expression and in affecting the spliceosome [B$^+$90, H$^+$14d].

Diverse regulatory functions are carried out by antisense RNAs (asRNA). An asRNA is a RNA molecule that is complementary to a mRNA. As delineated by Pelechano *et al.*, asRNAs can induce DNA methylation, lead to changes in HMs, can slow down or terminate transcription. Besides, asRNAs can bind their mRNA complements and thereby inhibit translation [PS13].

MicroRNAs (miRNA) have been shown to lead to transcript degradation and are involved in translational repression [Shi06]. Similarly, short interfering RNAs (siRNA) degrade mRNA by a process known as RNA interference (RNAi) [D$^+$17]. We also remind the reader of the not yet fully understood function of eRNAs, non-coding RNAs transcribed at active enhancers [K$^+$10b].

Besides, the translated proteins can be subject to several post-translational modifications (PTMs), altering a proteins activity and live span in the cell [Spo18]. Proteins can undergo reversible chemical modifications, e.g. phosphorylation, acetylation, or methylation [Spo18]. Proteins can also be exposed to redox reactions, for instance, via S-sulfenation that is the addition of an SOH group [Spo18].

Protein phosphorylation is a key PTM. It describes the addition of a phosphate group to an amino acid, most frequently to serines. Phosphorylation is inadmissible in regulating *"protein synthesis, cell division, signal transduction, cell growth, development and aging"* [A$^+$17b]. The well known tumor suppressor P53 is activated via phosphorylation [A$^+$17b].

Acetylation typically effects the N-terminal end of a protein or of lysine residues. Acetylation has been implicated for example in hormone regulation, regulation of blood pressure, stress response, cardiac rhythms, *osteogenesis* and *haematopoiesis*

[D+16d]. Protein methylation, aside from the already discussed HMs, has also been shown to affect other proteins, e.g. the methylation of HSP70 has been shown to foster cancer proliferation. DNA methyltransferases themselves can be methylated, which influences their activity [L+14b]. Protein sulfenation has been linked to oxidative stress and aging as well as to neurodegeneration in the rat brain [Y+18]. In addition to such chemical modifications, the protein can be modified by the reversible addition of polypeptides, e.g. ubiquitin. The addition of ubiquitin is called ubiquitination [Spo18]. An ubiquitinated protein is targeted for degradation by the 26S *proteasome*, therefore ubiquitin is essential to adjust the amount of protein present in a cell [BS04]. Besides, a protein can be modified by the addition of other complex molecules e.g. via ADP-ribosylation, which is implicated for example in DNA repair and in chromatin decondensation [LY15, Spo18].

2.2 Mathematical and Computational Background

Within this section, we provide a basic understanding of linear and logistic regression, introduce the concept of hypothesis testing, explain the methodology of peak calling algorithms that are prevalent in the analysis of epigenetics data, introduce dynamic programming and provide a brief introduction into information theory, more specifically the minimum description length principle (MDL). We start with a brief description of the concepts of statistical learning used in later chapters of this thesis. For a thorough introduction, we refer to the book The Elements of Statistical Learning [H+06].

2.2.1 Regression

In a regression problem, we consider an output variable Y with $Y \in \mathbb{R}^n$ and an input matrix $X \in \mathbb{R}^{n \times m}$. The number of samples (sometimes referred to as observations), is denoted by n, the number of predictors (also known as features) is denoted by m. All features for a specific sample i are denoted with x_i. A specific feature j for a specific sample i is denoted with $x_{i,j}$. The output variable for sample i is denoted with y_i. Note that the index i is often omitted in practice, that is x refers to an individual sample contained in X and y refers to an individual observation in Y. The joint probability distribution of X and Y is denoted with $Pr(X, Y)$. The aim of regression is to find a function $f(X)$ predicting Y using the input matrix X. To find $f(X)$, we define a loss function $L(Y, f(X))$ penalizing prediction errors, for example the squared error loss $L(Y, f(X)) = (Y - f(X))^2$. Thus, we get for the estimated prediction error (EPE):

$$
\begin{aligned}
EPE(f) &= E(Y - f(X))^2 && (2.4)\\
&= \int [y - f(x)]^2 p(x, y) dx dy && (2.5)\\
&= \int [y - f(x)]^2 p(y|x) p(x) dx dy && (2.6)\\
&= \int_x (\int_y [y - f(x)]^2 p(y|x) dy) p(x) dx && (2.7)\\
&= \int_x (E_{Y|X}([y - f(x)]^2 | X) dx) && (2.8)\\
&= E_X E_{Y|X}([y - f(x)]^2 | X). && (2.9)
\end{aligned}
$$

This form of the EPE shows that $f(x)$ can be obtained by minimizing the EPE point by point, i.e. per sample, yielding:

$$
\begin{aligned}
f(x) &= argmin_c E_{Y|X}([Y - c]^2 | X = x), && (2.10)\\
f(x) &= E(Y | X = x), && (2.11)
\end{aligned}
$$

where equation (2.11) is known as the regression function, stating that, using squared error loss, the best prediction of Y is obtained by the conditional mean

across all data points [H+06, p.18f]. This is utilized in the linear regression estimator explained in the next section.

Linear regression

In linear regression, the regression function (2.11) is either linear in the input matrix X or assumes that linearity is an adequate approximation. In a linear model, the response Y is approximated by a linear combination of the feature vectors denoted by X_j as:

$$f(X) = \beta_0 + \sum_{j=1}^{m}(X_j\beta_j), \tag{2.12}$$

where β_0 is the intercept and the β_j's are the model coefficients. Using a quadratic loss function as in (2.4), the least squares approach is used to determine the regression coefficients β by minimizing the residual sum of squares (RSS) as

$$RSS(\beta) = \sum_{i}^{n}(y_i - f(x_i)) \tag{2.13}$$

$$= \sum_{i}^{n}(y_i - \beta_0 - \sum_{j=1}^{m}(x_{i,j}\beta_j)) \quad \text{using (2.12).} \tag{2.14}$$

Recall that $x_{i,j}$ denotes the value of feature j for sample i. Intuitively, RSS is the sum of quadratic errors made by the linear model. With $X \in \mathbb{R}^{n\times 2}$, this is often visualized as fitting a hyper plane in the three dimensional space spanned by X_1, X_2, and the response Y. Formula (2.14) can be more comprehensively written in matrix notation as

$$RSS(\beta) = (Y - X\beta)^T(Y - X\beta), \tag{2.15}$$

with X being a $n \times (m + 1)$ matrix. The addition of one column holding $1's$ is required to account for the intercept captured by β_0. As (2.15) is a differentiable function, global extrema can be obtained by setting its first derivative $RSS(\beta)'$ to zero:

$$RSS(\beta)' = -2X^T(Y - X\beta). \tag{2.16}$$

Solving (2.16) for β yields

$$\hat{\beta} = (X^TX)^{-1}X^TY, \tag{2.17}$$

which is a global minima if X^TX is positive definite as the second derivative of (2.15) is positive in this case too [H+06, p.44ff]. The inferred model coefficients $\hat{\beta}$ can be used to compute the estimated response, denoted by $\hat{Y}$.

Model performance

Predicting the response $\hat{Y}$ from the input X used for training as $\hat{Y} = X\hat{\beta}$ allows to quantify the training error in terms of the mean squared error measure (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{2.18}$$

However, the training error does not provide good insights on the generalizability of a model that is its performance on unseen data. Therefore, one typically applies the model on unseen data, also known as test data, where X' denotes the features of the test data and Y' denotes the response vector [H$^+$06, p.24,46,219]. In addition to the MSE, Pearson and Spearman correlation are used to assess the performance of linear models in this thesis. The Pearson correlation between two variables A and B is defined as

$$\rho(A, B) = \frac{cov(A, B)}{\sigma_A \sigma_B}, \tag{2.19}$$

where $cov(A, B)$ is the covariance between A and B and σ is the standard deviation. The value of ρ is in $[-1, 1]$, where -1 and 1 are the extreme cases. The closer the value of ρ is to 1, the more correlated A and B are, the closer the value of ρ is to -1 the more anti-correlated A and B are. Values around 0 indicate that there is no correlation between A and B. Pearson correlation has been shown to be susceptible to outliers. Therefore, we also use the rank based Spearman correlation r_S, which is more robust in these cases [D$^+$75].

$$r_S(A, B) = \rho(g(A), g(B)) = \frac{cov(g(A), g(B))}{\sigma_{g(A)}\sigma_{g(B)}}, \tag{2.20}$$

where $g(A)$ and $g(B)$ denote the rank converted variables A and B. Thus, Spearman correlation is simply the computation of Pearson correlation between two ranked input variables [Erl03, p.508]. Pearson and Spearman correlation can both be used to assess the agreement of the measured response vector Y (Y') with the predicted response vector $\hat{Y}$ ($\hat{Y}'$), on training (test) data.

Cross-validation for unbiased estimates

A commonly used methodology to obtain an unbiased estimate of the test error is cross-validation. In a k-fold cross-validation, the entire available data set is split into k equally sized, non-overlapping parts. A model is learned on the data composed of $k - 1$ parts and the k^{th} fraction of the data set is used to assess the test error. The special case that $k = n$ is known as leave-one-out-cross-validation. Using the quadratic loss function as above, we can compute the cross-validated test error $CV(\hat{f})$ of the predictor $\hat{f}$ as

$$CV(\hat{f}) = \frac{1}{k}\sum_{i=1}^{k}(Y^i - \hat{f}^i_X(X^i))^2, \tag{2.21}$$

where $\hat{f}_X^i$ denotes a predictor trained on all but the i^{th} fraction of X, X^i is the i^{th} fraction of the feature matrix X and Y^i is the i^{th} fraction of the response Y. Cross-validation can also used for parameter tuning. In this instance, it is called inner cross-validation. It partitions the training data into several parts and uses those to optimize, for instance, a hyper-parameter and returns its value leading to the minimal error achieved in the inner cross-validation. The final model is fitted using the identified hyper-parameter on the entire data set used for the inner cross-validation [H+06, p.241f].

A special form of cross-validation is Monte-Carlo cross-validation. In Monte-Carlo cross-validation, the data set is not equally split into k non-overlapping parts. The training and test set are randomly sampled from the entire data set k times. Therefore, the same data points can occur multiple times in the test and the training set [X+04].

Shrinkage methods for model selection

To achieve better model performance and to simplify model interpretation, it can be beneficial to select only a subset of all possible features at hand. This is achieved using model selection procedures [H+06, p.59]. In the context of this thesis, only shrinkage methods have been used and are illustrated in the following. In Section 3.3 of [H+06], further details are provided on related approaches, e.g. subset selection.

Ridge regression adds a weighted quadratic penalty concerning the regression coefficients to the objective function of the least squares minimization:

$$\hat{\beta}^{ridge} = argmin_\beta \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{m} x_{i,j}\beta_j)^2 + \lambda \sum_{j=1}^{m} \beta_j^2 \right\}, \tag{2.22}$$

with λ controlling the amount of shrinkage. The penalization of the regression coefficients especially solves issues with correlated variables in the input. For instance, in a not regularized linear regression problem two highly correlated variables could be cancelling out each other by achieving a strongly positive and negative coefficient with similar magnitude. Due to the size restriction of the ridge penalization, this can be avoided [H+06, p.63]. The problem shown in formula (2.22) can be as easily solved as the ordinary least squares function from (2.17) exists as a closed form solution:

$$\hat{\beta}^{ridge} = (X^T X + \lambda I)^{-1} X^T y, \tag{2.23}$$

where I is a $p \times p$ identity matrix [H+06, p.64]. Intuitively, ridge regression can been seen as a proportional shrinkage of all regression coefficients. Therefore, ridge regression does not produce sparse models, i.e. no feature is completely removed from the model. However, it might be assigned to an infinitesimally small regression coefficient [H+06, p.69ff].

Similar to ridge regression, lasso penalization adds an additive term to the penalization with regard to the magnitude of the regression coefficients:

$$\hat{\beta}^{lasso} = argmin_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{m} x_{i,j}\beta_j)^2 + \lambda \sum_{j=1}^{m} |\beta_j| \right\}. \tag{2.24}$$

As for ridge, the parameter λ acts as a weight for the penalty. There is no closed form solution for (2.24), however the lasso solution can be efficiently computed, e.g. using an adapted version of the Least Angle Regression (LAR) algorithm [E+04a] (c.f. Algorithm 1) as delineated in [H+06, p.74]. The iterative LAR algorithm can be seen as a modified version of the forward stepwise selection procedure where the addition of features to the active set and the value of their regression coefficient depend on the features correlation with the residual at each stage of the algorithm [H+06, p.58, 74].

Algorithm 1 Least Angle Regression [H+06, p.74]

Standardize the feature matrix X to have mean zero and unit norm.
$r \leftarrow Y - \bar{Y}$
$s \leftarrow min(n-1, m)$
$\mathcal{A} \leftarrow \{\}$
$\alpha = 10^{-4}$ {or user defined}
for all β_i **do**
  $\beta_i \leftarrow 0$
end for
$j \leftarrow argmax_j(cor(X_j, r))$
$\mathcal{A} \leftarrow \mathcal{A} \cup \{j\}$
for $i = 1$ **to** s **do**
  $r_k \leftarrow y - X_{\mathcal{A}}\beta_{\mathcal{A}}$
  $\delta \leftarrow (X_{\mathcal{A}}^T X_{\mathcal{A}})^{-1} X_{\mathcal{A}}^T r$
  while $\nexists\, l \in \{1..p\} \setminus \mathcal{A} : cor(X_{\mathcal{A}}, r) <= cor(X_l, r)$ **do**
    $\beta_{\mathcal{A}} \leftarrow \beta_{\mathcal{A}} + \alpha\delta$
    if $\exists j \in \mathcal{A} : \beta_j == 0$ **then**
      $\mathcal{A} \leftarrow \mathcal{A} \setminus \{j\}$
      $r_k \leftarrow y - X_{\mathcal{A}}\beta_{\mathcal{A}}$
      $\delta \leftarrow (X_{\mathcal{A}}^T X_{\mathcal{A}})^{-1} X_{\mathcal{A}}^T r$
    end if
  end while
end for
return β

The lasso has the attractive property to set regression coefficients exactly to zero, which is helpful for model interpretation if there are many features present in the data set [H+06, p.72]. However, the lasso is not robust under certain conditions, e.g. if there is multicollinearity in the data [SS18]. In such a case, lasso tends to

single out only of the correlated features randomly, while the ridge penalty would shrink all coefficients of those features together [HH05][H$^+$06, p.662].

To combine the benefits of both the lasso and the ridge penalty that is model sparsity and the retention of correlated yet predictive features in the model, Zou and Hastie proposed the elastic net penalty as a combination of both techniques [HH05]:

$$\hat{\beta}^{enet} = argmin_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{m} x_{i,j}\beta_j)^2 + \lambda \sum_{j=1}^{m} (\alpha\beta_j^2 + (1-\alpha)|\beta_j|) \right\}. \tag{2.25}$$

As before, the total shrinkage is regulated by λ and α determines the ratio between the ridge and the lasso penalty. Just as λ, also the value of α can be determined in cross-validation. By construction, *"elastic net selects variables like the lasso and shrinks together coefficients of correlated predictors like ridge"* [H$^+$06, p.73]. The latter characteristic is known as the grouping effect. This renders elastic net to be a favourable shrinkage technique when handling (epi)genomic data sets [H$^+$06, p.6620] [SS18]. As the elastic net optimization function can be rewritten into a lasso like structure, the already introduced LAR algorithm (c.f. algorithm 1) can be used in a slightly adapted version, delineated in [HH05].

2.2.2 Classification

In contrast to the quantitative response in regression, in classification we deal with qualitative, sometimes also referred to as discrete, response variables [H$^+$06, p.9]. There are several methods used in literature for classification, for instance, linear discriminant analysis (LDA) [H$^+$06, p.106] or logistic regression [H$^+$06, p.119]. The latter is explained here as it is used in this thesis.

Logistic regression

Logistic regression is a classification method using discriminant functions $\delta_k(x)$ for each class k. Sample x is classified according to the class achieving the highest value for $\delta_k(x)$. In logistic regression posterior probabilities $Pr(G = k|X = x)$ are used to model $\delta_k(x)$. The posterior probabilities for a model with K classes can be denoted by

$$Pr(G = k|X = x) = \frac{\exp(\beta_{k0} + \beta_k^T x)}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \beta_l^T x)}, \tag{2.26}$$

$$Pr(G = K|X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \beta_l^T x)}, \tag{2.27}$$

where the vector β holds the regression coefficients. In the general multi-class definition of logistic regression, the probability for class K is often use as a normalization factor in the logistic function, yielding:

$$\log\left(\frac{Pr(G = k|X = x)}{Pr(G = K|X = x)}\right) = \beta_{k0} + \beta_k^T x, \tag{2.28}$$

with $1 \leq k < K$. The sum of all posterior probabilities equals one. In a binary classification problem with two classes 1 and 2, we obtain a single linear function using the logit ($log[Pr(G = 1)/(1 - Pr(G = 1))]$) transformation [H+06, p.119]:

$$\log\left(\frac{Pr(G = 1|X = x)}{1 - Pr(G = 1|X = x)}\right) = \beta_{10} + \beta_1^T x. \tag{2.29}$$

Unlike the regression models presented in the previous section, logistic regression models are trained using maximum likelihood. The log-likelihood l for an entire data set with n samples, K classes and regression coefficients

$$\beta = \{\beta_{10}, \beta_1^T, ..., \beta_{(k-1),0}, \beta_{K-1}^T\}$$

is computed as [H+06, p.120]:

$$l(\theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} log(Pr(G = k|X = x_i; \beta)). \tag{2.30}$$

To exemplify this further, in a binary classification problem with the two classes 1 and 2 we define the probability terms $p_{i_1} = Pr(G = 1|X = x_i; \beta)$ and $p_{i_2} = 1 - Pr(G = 1|X = x_i; \beta)$, leading to [H+06, p.120]:

$$l(\beta) = \sum_{i=1}^{n} \{y_i \cdot log(p_{i_1}) + (1 - y_i) log(p_{i_2})\}, \tag{2.31}$$

$$l(\beta) = \sum_{i=1}^{n} \{y_i \cdot log(Pr(G = 1|X = x_i; \beta)) + (1 - y_i) log(1 - Pr(G = 1|X = x_i; \beta))\}, \tag{2.32}$$

where y_i encodes the class labels. It is $y_i = 1$ for class 1 and $y_i = 0$ for class 2 [H+06, p.120]. The likelihood function can be optimized by setting its derivative to zero. In matrix notation the first and second derivative are [H+06, p.121]:

$$\frac{\delta l(\beta)}{\delta \beta} = X^T(y - p), \tag{2.33}$$

$$\frac{\delta^2 l(\beta)}{\delta \beta \delta \beta^T} = -X^T W X, \tag{2.34}$$

with W being a $n \times n$ diagonal matrix with the i^{th} diagonal element given as $Pr(G = 1|X = x_i; \beta)/(1 - Pr(G = 1|X = x_i; \beta))$, y is a vector holding class labels and p refers to the probabilities computed for each class and sample. The optimization can be performed using Newton's method [Wal85] that finds the x satisfying $f(x) = 0$ in an iterative procedure. The general definition of Newton's method is

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}. \tag{2.35}$$

In case of logistic regression this gives us [H+06, p.121]:

$$\beta^{new} = \beta^{old} - \frac{\delta^2 l(\beta)}{\delta\beta\delta\beta^T}^{-1} \frac{\delta l(\beta)}{\delta\beta} \tag{2.36}$$

$$= \beta^{old} + (X^T W^{old} X)^{-1} X^T (y - p), \tag{2.37}$$

with W^{old} defined as W replacing β with β^{old}. The optimization can be started using $\beta = 0$. Note that the algorithm is not guaranteed to converge [H+06, p.121]. As for linear regression, also logistic regression can be combined with regularization such as lasso, ridge, or elastic net [F+10b]. As delineated in [H+06, p.121], Newton's method can be generalized for multi-class classification as well. For reasons of brevity this is omitted here.

Quality measures for classification problems

The performance of a classifier can be measured in several ways. Typical measures are accuracy (acc), which is especially used in binary classification problems, as well as precision (pre) and recall (rec) defined as [HM15]

$$acc = \frac{TP + TN}{TP + FP + TN + FN}, \tag{2.38}$$

$$pre = \frac{TP}{TP + FP}, \tag{2.39}$$

$$rec = \frac{TP}{TP + FN}. \tag{2.40}$$

Acc denotes the ratio of how many predictions have been made correctly related to all data points. Note that formula (2.29) is representing the binary classification problem. In this context true positives (TP) are samples correctly assigned to the positive class, true negatives (TN) are samples correctly assigned to the negative class, false positives (FP) are samples erroneously classified as positive and false negatives (FN) are samples mistakenly assigned to the negative class. In multi-class classification problems, *acc* can still be applied. In this case the nominator is the sum of all diagonal elements of the confusion matrix of the classification problem and the denominator represents the sum of all elements in the confusion matrix. The *acc* for random data can be obtained by $\frac{1}{\#classes}$, where $\#classes$ indicates the number of distinct class labels [HM15].

The formulas of *pre* and *rec* can be generally applied. Specifically, *pre* denotes how many predictions for a distinct class have been made correctly, while *rec* states how many of all samples belonging to a distinct class have been classified accurately. All introduced measures *acc*, *pre* and *rec* have a value range between $[0, 1]$, where 1 is the best value [HM15].

Also, the $F1$ measure, a combination of *pre* and *rec*, is frequently used. It is defined as the harmonic mean of *pre* and *rec* [ZZ09]:

$$F1 = 2\frac{pre \cdot rec}{pre + rec}. \tag{2.41}$$

Another way of combining performance measures are receiver operating characteristic curves (ROC-curve) as well as precision-recall curves. These are helpful if a model provides soft labels, e.g. a score indicative for the class assignments of data points [HM15, G$^+$15c]. In a ROC-curve, *rec* is plotted against the false positive rate (FPR) ($FPR = \frac{FP}{FP+TN}$). The area under the ROC-curve is a quantitative way of describing model performance. A value of 1.0 indicates that the curve has perfect shape, i.e. the model is highly sensitive while a value of 0.5 indicates that the model just performs as good as a random classifier.

It is known that the ROC measure is less suited for imbalanced data sets, as model performance may be judged overly optimistic. In such a setting, precision-recall curves should be used, as they are invariant to class imbalance. This is the case because TN are not considered in neither *pre* nor *rec*, which are contrasted in a precision-recall curve. As for ROC-curves, the area under the precision-recall curve is a common quantitative assessment of the curves shape and of model performance [G$^+$15c]. Note that there is no distinct value indicating randomness for the area under the precision-recall curve.

2.2.3 Principal Component Analysis (PCA)

PCA is a widely used method for the analysis of high-dimensional data sets. The method tries to unravel non-obvious patterns in the data by projecting it into lower dimensions, called principal components (PCs). The PCs are chosen such that they maximize the variance of the data points and such that the PCs are uncorrelated to each other. Thus, the first PC, PC1, can be seen as orienting the data points by the main source of variance across all features. Consequently, PC2 will be the projection with highest variance given PC1. In many applications, PCs other than PC1 and PC2 are not considered as the proportion of variance explained by those is often negligible [L$^+$17b].

In mathematical terms, a PCA finds a matrix W such that

$$T = XW, \tag{2.42}$$

where W holds the *eigenvectors* of X^TX. A PCA can be computed using the singular value decomposition (SVD) of the $n \times m$ data matrix X defined as [C$^+$09, p.65f]

$$X = UDV^T. \tag{2.43}$$

Here, U is $n \times n$ and V a $m \times m$ orthogonal matrix. Matrix D is an $n \times m$ diagonal matrix with the property $d_1 \geq d_2 \geq ... \geq d_p \geq 0$. The entries of D are called singular values of X. X is called singular if there exists at least one diagonal element d_i that equals zero. It holds that [C$^+$09, p.65f]

$$X^TX = VD^2V^T. \tag{2.44}$$

As it was shown that W is equivalent to V, the i^{th} principal component w_i can be computed according to [C$^+$09, p.65f]:

$$w_i = Xv_i = u_id_i, \tag{2.45}$$

where u_i is the i^{th} column of U and v_i is the i^{th} column of V, respectively.

The maximum number of PCs is given by the minimum number of samples or the number of features present in the data. Despite its popularity, PCA comes with several limitations. For instance, as PCA is not scale invariant, differences in scaling among different samples in a data set will make a difference, including outliers or samples analyzed with a different normalization or processing methods. Furthermore, PCA assumes that the data is linear and correlated sources of variance are difficult to resolve, because the PCs are uncorrelated by definition [L+17b]. An example for an alternative data visualization method unraveling non-linear associations in high-dimensions is the t-SNE approach [vdMH08].

2.2.4 Dynamic programming

Dynamic programming is a popular technique to solve optimization problems that is finding an optimal solution to a problem. Intuitively, the dynamic programming applies a divide-and-conquer methodology splitting the actual problems into smaller overlapping subproblems. Instead of solving overlapping subproblems repeatedly, their solution is stored and retrieved when necessary. Hence, the speed-up is obtained at the cost of using extra memory, an example for the time-memory trade-off [C+09, 363ff].

In the textbook, Introduction To Algorithms, a "recipe" to apply dynamic programming is provided [C+09, p.372f]:

1. *Characterize the structure of an optimal solution.*
2. *Recursively define the value of an optimal solution.*
3. *Compute the value of an optimal solution.*
4. *Construct an optimal solution from computed information.*

An optimization problem must have two properties to be solved with dynamic programming. First of all, it must exhibit the optimal substructure property. That means, any optimal solution of the problem contains optimal solutions of the subproblems [C+09, p.379f].

Secondly, it is required that the problem exhibits overlapping subproblems that is a recursive algorithm designed to solve the problem needs to solve the same subproblem multiple times [C+09, 384].

The number of subproblems directly effects the run time of the dynamic programming method, together with the number of possible choices the problem at hand allows for each subproblem [C+09, p.379f].

In Bioinformatics, dynamic programming is commonly used, e.g. to find the longest common subsequence (LCS) among two sequences [C+09, p.391ff], or in the Needleman-Wunsch algorithm for sequence alignment [NW70]. In the LCS problem the goal is to identify the longest common subsequence with length $c[i, j]$, where $c[i, j]$ is defined as the length of the longest common subsequence between the

prefixes $s_1[1,i]$ and $s_2[1,j]$. The problem can be formulated in a recursive fashion as

$$c[i,j] = \begin{cases} 0 & \text{if } i = 0 \text{ or } j = 0, \\ c[i-1,j-1]+1 & \text{if } i,j > 0 \text{ and } s_{1i} = s_{2j}, \\ max(c[i,j-1], c[i-1,j]) & \text{if } i,j > 0 \text{ and } s_{1i} \neq s_{2j}. \end{cases} \tag{2.46}$$

The first case of (2.46) represents the base case of the recursion that is at the beginning of the sequence, the common subsequence is obviously of length zero. The second case describes the equality of s_1 at position i and s_2 at position j, therefore one is added to the length of the previously longest subsequence at position $c[i-1,j-1]$. The third case applies if s_{1i} is not equal to s_{2j}. In this case, the length of the longest common substring is either contained in $c[i-1,j]$, $c[i,j-1]$ [C+09, p.393]. While matrix $\mathcal{C}$ containing all values of c provides information on the length of the LCS, the actual LCS can be retrieved from $\mathcal{C}$ either by backtracking or by storing the information on the LCS along with its length in each entry of $\mathcal{C}$.

2.2.5 Hidden Markov Models (HMMs)

Hidden Markov Models (HMMs) are used to solve the following problem: *"There is a sequence of discrete or continuous observations generated by an unknown process. Find a model that explains and characterizes the observed sequences"* [RJ86]. More formally, an HMM is a *stochastic process*. The stochastic processes is hidden and can only be observed through observations of a second variable. The most frequently used illustration for such a problem is the coin toss example: for each observation of head or tail, it needs to be decided whether or not a fair coin was used for the toss, an information not known to the observant [RJ86].

Mathematically, an HMM consists of several parts, introduced below using the nomenclature from [RJ86]:

- t = length of the sequence of observations,
- n = number of hidden states,
- m = number of observation symbols,
- $Q = \{q_1, q_2, ..., q_n\}$ the set of hidden states,
- $V = \{v_1, v_2, ..., v_m\}$ the set of possible observations,
- $A = \{a_{i,j}\}$, with $a_{i,j} = Pr(q_j, t+1 | q_i, t)$ the transition probability from q_i at observational index t to q_j at index $t+1$,
- $B = \{b_j(k)\}$, with $b_j(k) = Pr(v_k, t | q_j, t)$ the production probability for observation v_k at state q_j,
- $\Pi = \{\pi_i\}$, with $\pi_i = Pr(q_i, t = 1)$ the initial state distribution.

The sets A, B and Π are sufficient to define an HMM, compactly written as a triple $\lambda = (A, B, \Pi)$.

Going back to the coin-toss example, we can ask that given a sequence of observations O (heads and tails), what is the optimal sequence of hidden states $q_i \in Q$ (fair coin, unfair coin) that gave rise to O. If optimality is defined as singling out the path with the highest probability given λ, i.e. $Pr(O, I|\lambda)$, the Viterbi algorithm finds the optimal state sequence using dynamic programming [RJ86, For73].

Fitting the parameter of an HMM $\lambda = (A, B, \Pi)$ to a training data set to maximize $Pr(O|\lambda)$ requires to solve a maximum likelihood problem. This can be done in an iterative way using an expectation maximization method such as the the Baum-Welch algorithm [RJ86].

In this thesis, we also use a modified version of HMMs called Input Output Hidden Markov Model (IOHMM) [BF95]. In a standard HMM, the observation o_t at index t depends only on the hidden state q_t. However, in an IOHMM, there is an additional input layer directly influencing hidden state q_t, the observed output o_t and the transition probabilities between the hidden states q. Formally, this can be written as:

$$q_t = f(q_{t-1}, u_t), \tag{2.47}$$
$$o_t = g(q_t, u_t), \tag{2.48}$$

where u_t is the additional input specified at index t, f is a function returning the next hidden state and g returns the output o_t [BF95]. Thus, the observed output o_t might also be directly converted to the provided input u_t. Another fundamental difference to HMMs is that IOHMMs are trained in a supervised fashion. Due to brevity, we refer the reader to "An Input Output HMM Architecture" by Bengio and Frasconi for further reading [BF95].

2.2.6 Hypothesis testing

Hypothesis testing is a method used to *"testing a claim or hypothesis about a parameter in a population, using data measured in a sample. In this method, we test some hypothesis by determining the likelihood that a sample statistic could have been selected, if the hypothesis regarding the population parameter were true"* [Pri17].

To illustrate the concept, we look at an example. We formulate the claim that the average delay of any long distance train running in December 2018 in Germany is at least five minutes. To test whether this hypothesis is true, we obtain the publicly available information on train delays, covering only a small portion of all connections. We call these observed data points samples. This name arises from the notion that these observations are "sampled" from the entire population of observations.

The first step in performing the hypothesis test is to formulate two hypothesis, the null hypothesis (H_0) and the alternative hypothesis (H_1). The purpose of the test is to see whether the statement in the null hypothesis is true, in other words, we test H_0, because we actually believe that it does not hold. The alternative

hypothesis is the opposite of the null hypothesis, contradicting its meaning [Pri17]. For instance, in our train delay example we have:

- H_0: = The average delay of any long distance train running in December 2018 in Germany is at least five minutes,

- H_1: The average delay of any long distance train running in December 2018 in Germany is less than five minutes.

Next, we set a significance threshold α, which is typically 0.05, although stricter cut-offs are debated [Pri17, B$^+$17]. Subsequently, the actual test can be computed, providing information how likely the observed samples are if H_0 is true. There are several different so called test statistics available in literature. Which statistic should be used depends, among other things, on the hypothesis itself, on the distribution of the data at hand and on data abundance. An helpful guide how to choose the right test is provided, for instance, in [NH11]. Going back to the example, we like to test the mean of the delay and we do not know the variance of the entire population. Therefore, we can use the t-test in the computation, which is detailed below.

Given the value from the test statistic, we decide whether or not to reject the null hypothesis. The p-value *"is the probability under a specified statistical model that a statistical summary of the data (e.g., [the mean train delay]) would be equal to or more extreme than its observed value"* [WL16]. If the p-value is smaller or equal than the previously selected significance threshold α, we reject H_0, otherwise we accept it [Pri17]. Geometrically, the p-value resembles the area under the tail(s) of the test statistics distribution limited by the computed value of the test on the x-axis. Similarly, for the significance threshold α, the critical value of the test-statistic that is the value that just satisfies the significance threshold can be computed and also visualized as the area under the curve.

In the following, we define different test statistics used in this thesis, illustrate how to correct for multiple hypothesis testing, which is prevalent in computational biology.

t-test

The t-test can be used to test whether the mean of a population equals a distinct value μ_0, known as the one-sample t-test, or to assess whether two groups have the same mean (if their variance is identical), known as the two-sample t-test. Note that unlike the Gauss test, the t-test can be used if variance of the data is not known. However, the t-test assumes that the data follows a normal distribution. If this is not the case, or uncertainty exists about it, the Wilcoxon-Mann-Whitney test should be used [Zar10, p.130ff].

The one sample t-test T can be computed as

$$T = \sqrt{n}\frac{|\bar{X} - \mu_0|}{S}, \tag{2.49}$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{N} X_i, \tag{2.50}$$

$$S = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{n-1}}, \tag{2.51}$$

where X_i is the random variable representing the i^{th} sample, $\bar{X}$ is the sample mean and S is the sample standard deviation [Zar10, p.99ff].

In the two sample t-test, one distinguishes between paired and unpaired samples. Samples are paired, for instance, if the same patient has been tested twice, e.g. before or after a treatment. In the applications in this thesis, only the unpaired test with equal sample sizes has been used to test whether two groups exhibit the same mean. The test is defined as [Zar10, p.99ff]:

$$T = \frac{|\bar{X}_1 - \bar{X}_2|}{S\sqrt{\frac{2}{n}}}, \tag{2.52}$$

$$S = \sqrt{\frac{S_1^2 + S_2^2}{2}}, \tag{2.53}$$

$$S_1^2 = \frac{\sum_{i=1}^{N}(X_{1i} - \bar{X}_1)^2}{n-1}, \tag{2.54}$$

$$S_2^2 = \frac{\sum_{i=1}^{N}(X_{2i} - \bar{X}_2)^2}{n-1}. \tag{2.55}$$

To conclude the example from above, we use the one-sample t-test to test the H_0 hypothesis that "the average delay of any long distance train running in December 2018 in Germany is at least five minutes", using delay information obtained for $N = 100$ connections. Using the data points (data not shown), we obtain $\bar{X} = 3.1$ and $S = 1$. Thus, the test statistics evaluates to $T = 19$. The critical value t_0 for a t-test with 99 degrees of freedom and a one sided tail with a significance of 0.05 is $t_0 = 1.660391$. Because $t_0 < T$ we reject H_0.

Wilcoxon-Mann-Whitney test

The Wilcoxon-Mann-Whitney test is also known as the Mann-Whitney U test and as the Wilcoxon rank-sum test. It is used to test for independent samples whether two distributions are the same or not. More formally, given to random variables X, with elements x_i, and Y, with elements y_i as well as two cumulative distribution functions $f(X)$ and $g(Y)$, respectively, we test the H_0 hypothesis stating $f(X) = g(Y)$. In the original form of the test, the H_1 hypothesis states that X is stochastically smaller than Y, meaning that $\forall r \in \mathbb{R} : f(X > r) \leq g(Y > r)$ [MW47]. In

practice, a different definition is used. For two variables X and Y, with distribution functions $f(X)$ and $g(Y)$, we assume that $f(X)$ equals $g(Y)$ with respect to a shift a: $f(x) = g(Y - a)$. Thus, we define

- H_0: $a = 0$,
- H_1: $a \neq 0$.

The Mann-Whitney U test statistic U can be computed according to [WS10]:

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} S(x_i, y_j), \tag{2.56}$$

$$S(x, y) = \begin{cases} 1, & \text{if } y < x, \\ 0, & \text{else}, \end{cases} \tag{2.57}$$

where m is the number of measurements of X and n is the number of observations of Y, respectively. As this pairwise comparison can be computationally expensive for large data sets, an alternative computational strategy has been proposed [Zar10, p.162ff]:

$$U = min(U_x, U_y), \tag{2.58}$$

$$U_x = mn + \frac{m(m+1)}{2} - R_x, \tag{2.59}$$

$$U_y = mn + \frac{n(n+1)}{2} - R_y, \tag{2.60}$$

with R_x and R_y being the sum of the group specific ranks from a joint ranking of X and Y, respectively.

Correcting for multiple hypothesis testing

When many hypothesis test are performed, which is often the case in omics applications, the likelihood that we reject a null hypothesis simply by chance is increasing. Multiple testing correction adjusts for this [Nob09]. A p-value corrected for multiple hypothesis testing is also known as adjusted p-value, which can be computed, for instance, using the Bonferroni or Benjamini-Hochberg procedure. The former is a straight forward correction method that requires for a p-value to be significant that it is smaller than the adjusted significance threshold given by $\frac{\alpha}{n}$, where n is the total number of performed tests [Hay13]. The latter is based on controlling the *false-discovery rate (FDR)* rate [BH95].

2.2.7 Peak Calling

Peak calling throughout all projects described in this thesis was carried out with MACS2 [Z+08a] and/or JAMM [I+15]. Both methods are briefly characterized below.

MACS is short for Model-based Analysis for ChIP-Seq. The name itself points to the purpose of this tool, namely peak calling for ChIP-seq data [Z+08a]. However, MACS and the successor MACS2 have both been used in literature for DNaseI-seq and ATAC-seq peak calling as well [K+14b]. As there is no publication for the MACS2 algorithm, here we delineate the functionality of MACS, which is extended by MACS2.

In the first step, MACS removes all duplicated ChIP-seq reads that is reads occurring more often than expected from a random read distribution. The remaining reads can be modelled using a Poisson distribution.

Prior to the computation of the peaks, MACS shifts all ChIP-seq reads by $d/2$, where d is determined using the distance of the peak summits obtained from the bimodal ChIP-seq read enrichment patterns of 1000 high-quality peaks. These are identified using a sliding window approach with a window size that equals the sonication size and a significance test again a random distribution of reads. The shift is performed to ensure a better agreement with the center of biological activity, e.g. the binding site of the antibody-targeted protein of interest [Z+08a].

The shifted signal is used to find candidate peaks by shifting a window of size $2d$ across the genome. A region is selected as a candidate peak if it shows a significant enrichment according to the poisson distribution, using a p-value threshold of 10^{-5}. Overlapping candidate peaks are merged. The region with the highest read count is called center, or summit [Z+08a].

Finally, MACS scores each candidate peak by comparing it to either the input signal (the control experiment in ChIP-seq) or to a broader region around the peak region if no input is available. The parameter λ_{local} represents the background enrichment, it is determined as:

$$\lambda_{local} = \begin{cases} max(\lambda_{bg}, \lambda_{1kb}, \lambda_{5kb}, \lambda_{10kb}) & \text{if there is an Input,} \\ max(\lambda_{bg}, \lambda_{5kb}, \lambda_{10kb}) & \text{otherwise.} \end{cases} \tag{2.61}$$

Here, λ_{bg} is the poisson parameter estimated from the entire genome and λ_{xkb} the parameter fitted to $x$kb centered around the peak summit from either the input (if applicable), or the actual sample. If a peak is still significant using λ_{local} and passes a user defined p-value threshold, it is reported [Z+08a].

One of the major differences between MACS and MACS2 is that MACS2 uses $-\log_{10}$ converted p-values to generate a score for candidate peaks and to perform the final peak selection [Liu18].

In JAMM, a different methodology is pursued. First, JAMM identifies regions of the genome that show a global enrichment for the used signal. This is achieved by dividing the genome into bins (where the bin size is chosen specifically for each chromosome according to a cost function, or it is defined by the user). The enrichment test of a bin considers the read counts within the bin and a background bin as well as the signal-to-noise ratio within a bin compared to the entire chromosome. Neighbouring enriched bins are merged into enriched windows [I+15].

In the enriched windows, peaks are identified by fitting Gaussian mixture models with either two or three components, depending on the type of the called peak using

an expectation maximization approach. Compared to MACS2, peaks produced with JAMM are more pronounced and neighbouring peaks in peak dense regions are identified with high resolution and not merged into one peak [I+15].

2.2.8 An introduction to minimum description length

The minimum description length (MDL) principle is the basis for several methods to conduct inductive inference. While the purpose of any statistical inference method is to unravel hidden patterns in a data set, from a MDL perspective the presence of a pattern, or regularity in the data, can be interpreted as the ability to compress the data set [Gru07a, p.12, 22]. Thus MDL views *"learning as data compression"* [Gru07a, p.12]. The idea is that for any *"given set of hypotheses $\mathcal{H}$ and data set $\mathcal{D}$, we should try to find the hypothesis or combination of hypotheses in $\mathcal{H}$ that compresses $\mathcal{D}$ most."* [Gru07a, p.12].

In this section, we delineate some of the advantages of using MDL for model selection and provide essential nomenclature. This section is based on part one of the textbook The minimum description Length principle by Peter Grünwald, which provides a wide and detailed introduction to the topic.

As stated above, MDL is based on compression. The idea is that any regular pattern in the data allows for a shorter, more compact, yet unique description $\mathcal{D}'$ of the $\mathcal{D}$. MDL refers to $\mathcal{D}'$ also as an encoding, which can be used to fully reconstruct the original data $\mathcal{D}$. The encoding $\mathcal{D}'$ utilizes a description method, which converts an input sequence to a coding alphabet, which is typically a binary alphabet $\mathbb{B} = \{0, 1\}$. Using such a binary alphabet, any input sequence can be translated to a bit sequence. In MDL, each description method needs to satisfy the unique decodability property that is for any encoding $\mathcal{D}'$, there can be at most one $\mathcal{D}$ [Gru07a, p.6f].

Such a description method is the universal computer language (a language that can implement a universal Turing machine), as suggested by Ray Solomonoff. The idea is to find the shortest program in such a language (the actual choice of the computer language is not essential for large sequences $\mathcal{D}$) that returns the data $\mathcal{D}$ and terminates. The length of the shortest program that prints $\mathcal{D}$ is denoted by Kolmogorov complexity $L_{UL}(\mathcal{D})$. A small value of $L_{UL}(\mathcal{D})$ indicates that there is a large amount of regularity in $\mathcal{D}$ that is picked up by the model, which in turn means that the data can be represented in a very compact way. Unfortunately, Kolmogorov complexity can not be computed and the choice of language does matter for smaller sequences, which are more commonly occurring in practice. Therefore, compression using the universal computer language is termed idealized MDL. MDL approaches used in practice, known as practical MDL, use less general description languages that are often chosen depending on the data at hand [Gru07a, p.9ff].

A commonly used framework for practical MDL is the so called crude two-part MDL. It can be described as *"Let $\mathcal{H}_1$,$\mathcal{H}_2$, ... be a list of candidate models [..], each containing a set of point hypotheses [representing a distinct probability distribution]. The best point hypothesis $\mathcal{H} \in H$ to explain the data $\mathcal{D}$ is the one which minimizes the sum $L(\mathcal{H}) + L(\mathcal{D}|\mathcal{H})$, where $L(\mathcal{H})$ is the length, in bits, of the description of*

the hypothesis and $L(\mathcal{D}|\mathcal{H})$ is the length, in bits, of the description of the data when encoded with the help of the hypothesis." [Gru07a, p.14]. The outcome of this procedure suffers from uncertainty as the choice of the encoding of H can tremendously affect the results. Therefore codes that attempt to minimize, for instance, the worst-case total description length over all possible inputs, known as minimax codes, have been suggested. However, this requires that all possible solutions to a problem need to be explored [Gru07a, p.16].

In refined MDL, where codes are always designed following the minimax principle, *"we associate a code for encoding D not with a single $\mathcal{H} \in H$, but with the full model H. Thus, given model $\mathcal{H}$, we encode data not in two parts but we design a single one-part code with length $\bar{L}(\mathcal{D}|\mathcal{H})$. This code is designed such that whenever there is a member of H that fits the data well, in the sense that $L(\mathcal{D}|\mathcal{H})$ is small, then the codelength $\bar{L}(\mathcal{D}|\mathcal{H})$ will also be small"* [Gru07a, p.17].

Any approach following the MDL principle has several benefits that make its usage very appealing. For instance, MDL approaches, by construction, do not overfit and generalize well to unseen data. Further, they follow the spirit of *Occam's razor* that is they deliver a good trade-off between model complexity and goodness-of-fit, which simplifies model interpretation [Gru07a, p. XXV].

2.3 International efforts to characterize the (epi)genome of primary cells and cell types

In this thesis, we use publicly available datasets from ENCODE, Blueprint, Roadmap and the DEEP consortium.

The Encyclopedia of DNA Elements (ENCODE) project is funded by the NIH since 2007. ENCODE's mission is the characterization of regulatory elements in the human genome using a variety of experimental methods including RNA-seq, ChIP-seq of HMs and TFs and DNaseI-seq [D$^+$12b]. In January 2019, the ENCODE data portal (www.encodeproject.org) contained 9486 released samples, available for download as raw data files and as uniformly processed analysis files, e.g. peak calls or expression quantification. In addition to the data generation, also several computational methods have been developed in the scope of the ENCODE project, e.g. the widely used ChromHMM providing chromatin state segmentations [EK12].

The BLUEPRINT project, has been a large research project, involving 42 participants, funded by the European Union for 5 years until September 2016. Blueprint investigated epigenomic mechanisms of transcriptional regulation in diverse haematopoietic cell types covering both healthy and diseased cells. To do so, about 100 complete epigenomes have been generated from highly purified cells. The generated data allowed the discovery and validation of novel epigenetic markers for leukemia and Type 1 Diabetes [Stu18].

The NIH funded Roadmap Epigenomics Mapping Consortium started in 2008 with the characterization of DNA methylation, HMs, chromatin accessibility using DNaseI-seq and gene-expression in stem cells and tissue samples selected to *"represent the normal counterparts of tissues and organ systems frequently involved in human disease"* [B$^+$10b]. The final Roadmap data release contains 111 reference epigenomes, with each epigenome holding five histone marks (H3K4me3, H3K4me1, H3K27me3, H3K9me3 and H3K36me3) [K$^+$15].

Our group actively participated in the german epigenomics program (DEEP) that concluded in 2017. In a collaborative effort between several german research institutes and universities, DEEP provided epigenomic characterizations of cells relevant for chronic metabolic and inflammatory diseases. These include hepatocytes of human and mouse, adipocytes, fibroblasts, CD4+ T-cells, macrophages and monocytes. The estimated 70 epigenomes contain RNA-seq, chromatin accessibility using either DNaseI-seq, ATAC-seq, or NOME-seq data, HM-ChIP-seq and DNA methylation data [Con18].

The International Human Epigenomics Consortium (IHEC) is an international consortium functioning as an umbrella coordinating epigenomic data production, data quality and data standards all over the world. The goal of all IHEC members is to generate at least 1000 full epigenomes. ENCODE, Blueprint, Roadmap and DEEP contribute to IHEC. Hence, their epigenomic data is obtainable via the IHEC data portal [B$^+$16e].

3

Inferring key TFs from epigenetics and gene-expression data

A major focus of our work has been the elucidation of transcriptional regulation through TFs. Here, we first introduce the reader to mathematical ways of denoting the sequence preferences of TFs. Further, we discuss relevant models using statistical representations of TF binding to make genome wide Transcription Factor Binding Site (TFBS) predictions. We conclude the TF binding prediction part with the introduction of TEPIC, a method proposed by us to predict TF binding and at the same time generating TF scores on the gene level. Furthermore, we delineate various means to systemically analyze TF binding scores using machine learning approaches to improve our understanding of cell type specific transcriptional regulation.

This chapter summarizes our work published in four different articles [S+17a, K+17a, SS18, S+18b]. It also includes an extension of the Bachelor thesis by Fabian Kern [Ker16], which has been presented in a talk at the German Conference on Bioinformatics 2018 in Vienna.

3.1 Predicting TF binding *in silico*

This section delineates how to systematically describe the sequence preferences of TFs and how to use these representations in predictive models of TF binding.

3.1.1 Systematic description of the sequence preference of TFs

In Section 2.1.12, we describe several experimental ways to characterize TF binding preferences. Here, we illustrate how this information can be used to derive statistical models describing the binding behaviour of TFs.

Position Weight Matrices (PWMs)

No matter which experimental method has been used to assess TF binding *in vivo*, several statistical models require a multiple sequence alignment of all experimentally determined binding sites. For many years, researchers used the consensus sequence of this multiple sequence alignment X as a representation for a TF's binding preference [DM92]. To account for variability within the aligned sequences, which is

neglected by the consensus approach, a weighted matrix representation M has been proposed by Garry Stormo in 1982 [S$^+$82]. The idea behind the matrix representation is that the frequency of each nucleotide k in position i of the alignment is considered [Sto00]:

$$M_{k,j} = \sum_{i=1}^{n} I(X_{i,j} = k), \forall k \in \{A, C, G, T\}, \forall j \in \{1, ..m\}, \tag{3.1}$$

where $X_{i,j}$ refers to position j in the i^{th} sequence in the alignment, n is the number of all sequences in X, m is the length of X and I is the indicator function.

This formulation gives rise to the name Position specific Frequency Matrix (PFM) M, as the frequencies of each nucleotide are stored. Instead of the frequency based representation, PFMs are often transformed to Position specific Probability Matrices (PPMs) P, where an entry $P_{k,j}$ can be computed as

$$P_{k,j} = \frac{M_{k,j}}{\sum_{k \in \{A,C,G,T\}} M_{k,j}}, \forall j \in \{1, ..m\}. \tag{3.2}$$

To account for different background probabilities of the individual nucleotides b_k and to assess the information content of each position in the matrix, P can be transformed further to a so called Position Specific Scoring Matrix (PSSM), which is also known as Position Specific Weight Matrix (PWM) W [S$^+$86]:

$$W_{k,j} = P_{k,j} \cdot \log_2 \left(\frac{P_{k,j}}{b_k} \right). \tag{3.3}$$

To rank a DNA-sequence s composed of m base pairs with a TF binding motif of size m representing W, one computes the sum S of all respective scores in W as:

$$S = \sum_{i=1}^{m} W_{s_i,i}. \tag{3.4}$$

PWMs can be easily visualized in so called TF motifs [SS90], as exemplified in Figure 3.1. The height $h_{k,j}$ of nucleotide k at position j is computed according to

$$h_{k,j} = P_{k,j} \cdot R_j, \tag{3.5}$$

$$R_j = 2 - \sum_{k \in \{A,C,G,T\}} (-W_{k,j}) + e(n), \tag{3.6}$$

where $e(n)$ is a correction factor if there are only a few samples n, available, which can be computed as [S$^+$86]:

$$e(n) = \frac{3}{2 \cdot log(2) \cdot n}. \tag{3.7}$$

There are several open source databases such as Jaspar [K$^+$18c], Hocomoco [K$^+$18d], Uniprobe [H$^+$15b], or the commercial TRANSFAC service [M$^+$06] providing PWMs for hundreds of TFs and different species. For instance, in the latest release of our Tepic framework, we included PWMs for 30 species, including 561 non-redundant TF motifs for human, assembled from the aforementioned publicly available databases.

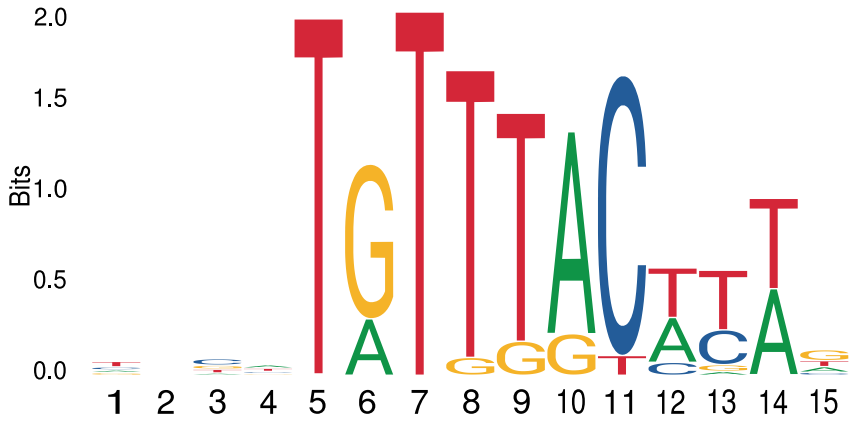


Figure 3.1: Binding motif for FOXA1, which has been obtained from the Jaspar database (MA0148.3) [K+18c]. The x-axis denotes the position in the motif, the information content at each position is shown on the y-axis. The size of a character is proportional to its information value.

Alternative representations for the binding preference of TFs

However, the usage of PWMs has several drawbacks, where the most important one is the simplifying assumption of independence between nucleotides at different positions. For instance, Eggeling *et al.* showed that there is substantial dependency between nucleotides in less conserved parts of the binding motif for CTCF [E+14a]. A straightforward way to address this issue is the usage of first-order Markov models, also known as Dinucleotide Weight Matrices (DWMs). These link the base occurring at position i to the one at position $i-1$ [Sid10]. However, Weihrauch *et al.* showed that DWM models do not offer a big advantage over simple PWM models in general, especially if the costs of a considerably larger parameter space are taken into account [W+13c]. The only minor improvement might be due to the still simplifying assumption of a first-order dependency between the nucleotides.

To model also more complex dependencies, several other approaches have been suggested, for instance SLIM models [KG15]. In SLIM models, *"the probability of a nucleotide at a certain position of a binding site may depend on any nucleotide observed at a preceding position"* [KG15]. Grau *et al.* fit SLIM models using sparse local inhomogeneous mixture models, allowing for weighted dependencies between nucleotides. Just like PWMs, SLIM models can also be efficiently visualized, including the modelled dependencies [KG15]. The advantages of SLIM models are stressed by the fact that one of the winning teams of the ENCODE DREAM challenge on *in vivo* TF binding site prediction used SLIM models as well [K+19]. Another measure, called BAMMS, was suggested by Siebert and Söeding. They proposed a Bayesian method to learn inhomogeneous Markov models to capture higher-order inter-dependencies without over fitting the model [SS16]. BAMMS can be seen as a generalization of the DWMs.

Although both SLIM models and BAMMS show better accuracy in pinpointing TF binding sites than simple PWMs, the latter are still dominantly used due to their simplicity and interpretability. The most important advantage of PWMs however is that they are available for many species and factors in the aforementioned databases. As there are no well curated databases for SLIM models nor BAMMS, these need to be learned *de-novo* for instance from ChIP-seq or PBM data, making it cumbersome and potentially, due to insufficient data, infeasible to apply these representations in every day usage.

3.1.2 Hit or no hit? Binary classification versus probabilistic modelling of TF binding

In Formula (3.4), we have already mentioned how a PWM W can be used to compute a score S for a distinct sequence s. In the field, there are two general approaches to assess from S whether the TF modelled by W would bind to s. In so called hit-based approaches, the statistical significance of the score S obtained for W applied to s is assessed compared against a background model, e.g. a zero-order null model with randomly generated sequences as used in FIMO [G+11]. If S is significant the site is reported, otherwise it will be discarded.

This "black and white" view of TF binding is very strict and does not reflect several aspects of biological reality that is the competition of TFs over a genomic location and the consideration of low-affinity binding sites [R+07]. Low affinity binding sites are sequences that do not fully match the actual binding site of a TF, but are still bound by the factor. It is a phenomenon that was shown to be highly relevant in biology [Tan06, C+15d].

Berg and von Hippel proposed a score to compute the estimated number of molecules bound to a distinct genomic location [BvH87], which was implemented by Roider *et al.* in a method called TRAP [R+07]. Compared to a hit-based classification, such affinity-based methods are well suited to rank different sequences in their likelihood for being bound by a TF [R+09]. Also, they have been successfully applied to analyse co-regulated genes [R+09], in assessing the effect of SNPs on TFBS [TC+11] and in the analysis of TF co-occurrence within DHS regions [vB15]. Nevertheless, hit-based approaches are still dominating the literature.

In the next section, we detail TRAP further, as it is widely used in this thesis. In Section 3.1.5, we provide the reader with an overview of other methods for TFBS prediction. Two of those are detailed in Sections 3.1.6 and 3.1.7. Later on, we compare our own TEPIC (Section 3.2) approach against those methods in several instances.

3.1.3 Transcription Factor Affinity Prediction (TRAP)

TRAP models the binding of TFs to any genomic region from a biophysical point of view, computing the number of bound molecules per genomic location s. To do so, the equilibrium between bound and unbound factors t at any site s can be

computed according to [R+07]:

$$p(s) = \frac{[t \cdot s]}{[s] + [t \cdot s]} \tag{3.8}$$

$$= \frac{K(s) \cdot [t]}{1 + K(s) \cdot [t]} \tag{3.9}$$

$$= \frac{K(s_0)e^{-\beta E(s)} \cdot [t]}{1 + K(s_0)e^{-\beta E(s)} \cdot [t]} \tag{3.10}$$

$$= \frac{R_0 \cdot e^{-\beta E(s)}}{1 + R_0 \cdot e^{-\beta E(s)}}. \tag{3.11}$$

Here, $K(s)$ is the equilibrium constant, specific for each site s. The concentration/activity of factor t and sequence s is denoted by squared brackets. The site with the highest affinity s_0, is assigned to the mismatch energy $E(s_0) = 0$ that is the site with the best match to a distinct factor t is used to calculate $K(s)$. Consequently the mismatch energy $E(s)$ for any site $s \neq s_0$ will be larger than 0. The parameter β is defined as $\frac{1}{\beta} = k_B \cdot T$, where k_B refers to the Boltzmann constant ($k_B = 1.3800649 \cdot 10^{-23} J/K$) and T denotes temperature. The parameter R_0 is defined as $R_0 = K(s_0) \cdot [t]$.

According to Berg and von Hippel [BvH87], who laid the foundations for the biophysical model used in TRAP, the mismatch energy $E(s)$ for a distinct factor t can be computed utilizing a PFM M^t for TF t according to [R+07]:

$$\beta E(\lambda, s, M^t) = \frac{1}{\lambda} \sum_{i=1}^{|M^t|} \sum_{k \in \{A,C,G,T\}} S_i^k \log \left(\frac{M_{i,max}^t}{M_{i,k}^t} b_k \right), \tag{3.12}$$

$$S_i^k = \begin{cases} 1, & \text{if } s_i = k, \\ 0, & \text{else,} \end{cases} \tag{3.13}$$

where $|M^t|$ denotes the number of positions of the frequency matrix M^t, $M_{i,max}^t$ represents the most frequent nucleotide at each position, $M_{i,k}^t$ is the entry in M^t for nucleotide k at position i and b_k denotes a background probability term.

Thus, with two parameters R_0 and λ, the expected number of TF molecules $\bar{N}$ bound to s can be computed as [R+07]:

$$\bar{N}_{M^t,s,\lambda,R_0} = \sum_{l=1}^{|s|-|M^t|} \frac{R_0 \cdot e^{-\beta E(\lambda, s_l, M^t)}}{1 + R_0 \cdot e^{-\beta E(\lambda, s_l, M^t)}}, \tag{3.14}$$

where $|s|$ denotes the length of sequence s. Thus $|s| - |M^t|$ denotes all possible positions for matrix M^t in s. On the basis of ChIP-Chip data for yeast, Roider *et al.* determined that a suitable value for λ is 0.7. For a fixed value of λ, the value for R_0 can be computed for each position frequency matrix M^t using its length m [R+07].

3.1.4 Differences between TF ChIP-seq data and predicted TFBS

In practice, TF ChIP-seq data is often used as a gold-standard to assess computational TFBS predictions. While it is definitely desirable to achieve the performance of the ChIP-seq experiments, we stress that some aspects of TF binding can not be adequately modelled. This is due to the conceptional design of both TF ChIP-seq and computational TFBS prediction methods as illustrated in Figure 3.2.

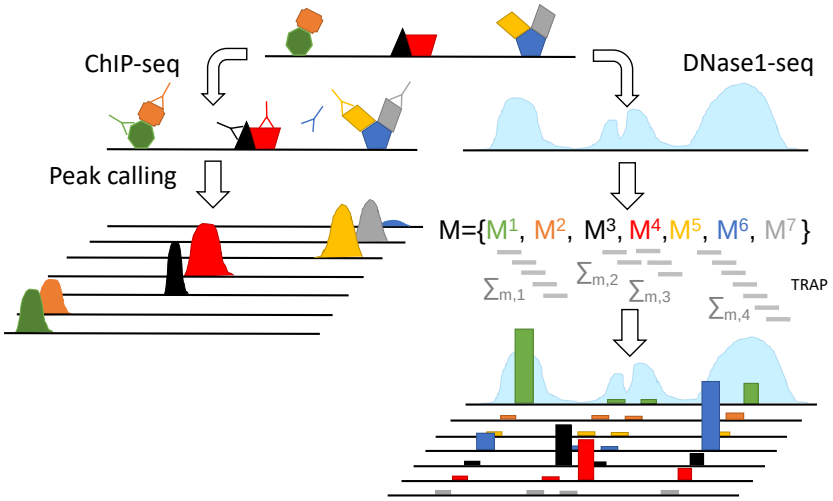


Figure 3.2: This figure depicts conceptual differences in detecting TFBS directly using ChIP-seq experiments or indirectly using computational predictions coupled with chromatin accessibility assays such as DNaseI-seq. Here, the result of TFBS predictions using TRAP for the indicated PFM M^t within DHSs are indicated by the colored bars. In contrast to ChIP-seq experiments, motif based prediction approaches are not able to model the activity of TFs in complexes, as TFs that bind indirectly that is they bind to other proteins instead of the DNA, can not be captured. This is illustrated by the orange TF, which can be pinpointed using ChIP-seq experiments. However, its position remains hidden using DNaseI-seq analysis. Another potential drawback of motif based predictions is shown for the TFBS predictions of M^1 in the fourth DHS. Although we do not see a ChIP-seq signal for the corresponding factor, we do observe a non-zero affinity, which is due to an confounding influence of the length of the DHS on the predicted motif scores. With an increasing size of the DHS, the random chance to observe a motif hit increases. Figure from Schmidt *et al.* [SS18].

Aside from the obvious advantage that ChIP-seq experiments generate a map of *in vivo* TF binding events, ChIP-seq experiments can also be used to screen TFs that bind indirectly that is instead of binding to the DNA, they bind to another TF via protein-protein interactions. Therefore, ChIP-seq experiments are well suited to study TF complexes as well. These can not be easily identified using computational

tools [N+16, WM16] that are considering only TF motif information.

However, this situation can be reversed, for instance, a TF that is bound to the DNA could be hidden from the antibody used for the ChIP-seq experiment, because it is fully blocked by other proteins. In this case, a motif based approach might find a motif for this TF, although no, or only a weak, ChIP-seq signal might be detected.

Furthermore, in case of ChIP-seq data, the decision whether a region is bound by a TF depends only the presence of a ChIP-seq peak. The intensity of a peak is not necessarily considered. In predictive approaches considering chromatin accessibility data, usually all possible sites within a candidate TFBS are used in computational prediction tools, e.g. in TRAP [R+07]. Therefore the length of a candidate region might influence TF scores as larger peaks could obtain a higher score by chance, if the size of the candidate regions is not taken into consideration.

3.1.5 Other computational approaches utilising PWMs

There are various approaches to predict TF binding *in silico*, which can be divided into three categories:

1. Methods that are purely based on the TF sequence specificity.
2. Site-centric methods that filter hits from (1) using epigenetics data.
3. Segmentation based methods utilizing epigenetics data to find candidate sites that are subsequently annotated using (1).

Predicting TFBS using only DNA sequence data

Jayaram *et al.* provide an extensive overview of the first class of methods that is purely sequence based approaches [J+16c], e.g. MATRIX-SCAN [T+08], CLOVER [F+04], the already mentioned FIMO [G+11], or POSSUMSEARCH [B+06b].

Both MATRIX-SCAN and CLOVER compute a log ratio score per sequence comparing the probability of a motif hit in a sequence s against a background model [T+08, F+04]. In addition to that, CLOVER determines a p-value to assess the scores significance using permutation experiments and also corrects for multiple testing [F+04]. The widely used method FIMO computes a log-likelihood ratio score for each distinct sequence position against a zero-order background model and computes a p-value per site using dynamic programming. By computing FDR, the p-values can be corrected for multiple hypothesis testing [G+11]. POSSUMSEARCH uses a suffix array constructed for the considered sequence to reduce search time. Via dynamic programming, motif specific thresholds based on a user defined p-value threshold can be computed [B+06b].

All of these methods can be installed locally and some are additionally available as a webserver, for instance FIMO [G+11]. Within the REGULATORTRAIL webserver [K+17a], we offer TFBS prediction using TRAP, which we have updated to allow for parallel execution within our TEPIC framework [S+18b].

While the methods mentioned so far utilise already known PWMs, recent deep-learning approaches, for instance DEEP-BIND [A+15b], attempt to infer the sequence specificity values of TFs *de novo* from large data sets. However, deep-learning methods have not been applied to many TFs and their practical usage for hands-on research is limited as they require a lot of data for a single TF and special hardware to be trained efficiently.

An overview of these purely sequence based methods is provided in Table 3.1 [S+18b].

Method	Availability	Motif scoring	Parallelized	Maintained
MATRIXSCAN [T+08]	Registration required	Hit-based	NA	NA
CLOVER [F+04]	Yes	Hit-based	No	No
FIMO [G+11]	Yes	Hit-based	No	No
POSSUMSEARCH [B+06b]	Yes	Hit-based	Yes	No
TRAP [R+07]	Yes	Affinity-based	No	No
TRAP(TEPIC) [S+17a]	Yes	Affinity-based	Yes	Yes
DEEP-BIND [A+15b]	Yes	NA	Yes	No

Table 3.1: Overview of purely sequence based TFBS prediction methods.

Methods utilizing epigenetics data to refine sequence based TFBS predictions

Applying TFBS prediction methods that consider only the sequence specificity of TFs have been shown to generate many false-positive hits compared to TF ChIP-seq experiments. By including epigenetics data into the TFBS predictions, the number of false-positive predictions can be greatly reduced because the search for TFBS is reduced to genomic sites of high regulatory activity, typically determined using chromatin accessibility assays, or HM ChIP-seq data [PR+11, G+16a]. The high agreement of *in vivo* TF binding as determined by TF ChIP-seq with DHSs is shown in Figure 3.3. Table 3.2 provides an overview of methods utilizing epigenetics data for TFBS prediction.

As mentioned before, in site-centric methods, genome wide TFBS predictions based on sequence matches with PWMs are classified, using epigenetics data, to be either truly bound or unbound. Many site-centric methods have been proposed, which can not all be listed here. One of the first and best known methods is CENTIPEDE [PR+11].

CENTIPEDE uses a hierarchical mixture model to predict bound TFBS incorporating chromatin accessibility data, histone modification ChIP-seq data, measurement of genomic conservation as well as the distance of a putative TFBS to the closest TSS [PR+11].

Another approach is taken by Cuellar-Partida *et al.* [CP+12], which we call FIMO-

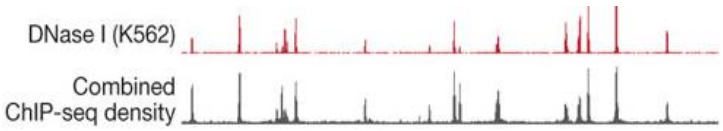


Figure 3.3: The figure illustrates the strong overlap of DHSs with TF ChIP-seq data, aggregated for 45 TF ChIP-seq experiments from ENCODE, obtained for the K562 cell line. Extracted from Figure 2 of Thurman *et al.* [T+12], obtained under the Creative Commons Attribution-Non-Commercial-Share Alike licence.

Prior. They compute an epigenetic prior using only DNaseI-seq signal and use this prior to reevaluate sequence based TFBS predictions (Section 3.1.6). Interestingly, although conceptually very simple, their approach was shown to perform on bar with the more involved Centipede approach [CP+12]. Another frequently used method in the field is PIQ (Section 3.1.7), which uses Bayesian inference to predict true TFBS [S+14c].

While the aforementioned methods are unsupervised, several supervised methods have been published, for instance MILLIPEDE [LH13] and BinDNase [KL15]. Both apply a binning strategy around candidate TFBS to learn characteristic DNaseI-seq profiles for truly bound TFBS. One of the top performing methods in the ENCODE-DREAM *in vivo* TFBS prediction challenge, called Catchitt also utilizes a supervised, site-centric strategy in a logistic-regression like fashion [K+19].

Method	Availability	Motif scoring	Parallelized	Maintained
Centipede [PR+11]	Yes	Hit-based	No	No
Fimo-Prior [CP+12]	Yes	Hit-based	No	No
PIQ [S+14c]	Yes	Hit-based	Only using gsub	No
Milipede [LH13]	Yes	Hit-based	No	No
BinDNase [KL15]	No	Hit-based	NA	No
DNase2TF [S+14d]	Yes	NA	No	No
Wellington [P+13c]	Yes	NA	Yes	Yes
HINT(BC) [G+16b, L+18a]	Yes	NA	Yes	Yes
TEPIC [S+17a]	Yes	Affinity-based	Yes	Yes
Catchitt [K+19]	Yes	Hit-based incl. Slim-models	Yes	Yes

Table 3.2: Overview of site-centric and segmentation based methods for TFBS prediction. Note that none of the footprinting methods have an integrated TFBS prediction module.

Many recently developed methods follow the segmentation-based strategy which is to first highlight a set of candidate regions, which are subsequently annotated with TFBS. Typically, segmentation based methods rely on TF footprints, which are, informally speaking, dips in DHSs, or also in ATAC-seq peaks. The footprints are most likely arising because the DNA is blocked at the occupied binding sites and is thus inaccessible for the DNaseI and the Tn5 enzyme. As a consequence, the DNaseI enzyme can not cut the DNA and the Tn5 enzyme can not insert the transposable element [K+18a]. Examples for the signal distribution of DNaseI and HMs at footprints are shown in Figure 3.4a. By considering only footprints in the search of TFBS, the prediction task is greatly simplified as the search space is considerably smaller and due to the assurance that sites are active, false positive predictions are reduced.

A variety of methods has been suggested to identify footprints, e.g. DNase2TF [S+14d], or Wellington, which uses a binomial test to identify footprints by comparing the read count within a footprint to the flanking region [P+13c]. One of the currently most sophisticated footprint callers is the HINT method [G+16b]. HINT is based on a HMM modelling DNaseI-seq and HM signal around footprints. The HMM architecture of HINT is depicted in Figure 3.4b. The already trained HMM can be obtained online and applied to unseen data. To reduce search space, HINT is applied to sites of signal enrichment only, e.g. DHSs. One of the advantages of HINT, is an integrated bias correction for DNaseI-seq cleavage bias [K+13b, G+16b]. This version is named HINT-BC. Recently, an updated version of HINT-BC was released that also accounts for an inherit sequence bias of ATAC-seq data [Mad15, L+18a]. None of the listed footprint calling methods have a built-in function to predict TFBS within the footprints.

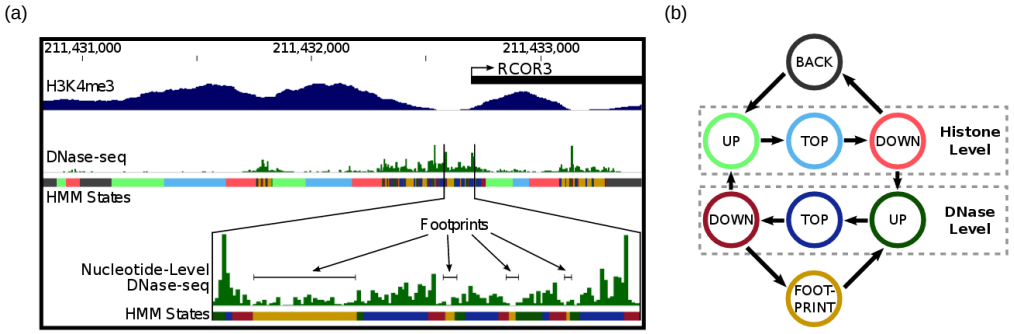


Figure 3.4: (a) Characteristic DNaseI and HM profile around TF-footprints: The H3K4me3 mark is flanking the NFR, which shows DHSs. Footprints are the sites with depleted DNaseI-seq signal within the DHS. (b) Architecture of the HMM used within HINT to identify TF-footprints. Figure following Figure 1 from Gusmao *et al.* [G+14d], obtained under Oxford University Press license 4511401158801.

Importantly, we note that some TFs are known to cause only weakly pronounced

footprints, because their residence time at the DNA is not sufficiently long to create a strong footprint signature [S+14d]. For these TFs, purely peak-based models, omitting the footprint calling step, might be a better choice. Furthermore, it was shown that pioneering TFs have the ability to bind to the genome even at heterochromatic sites. These are not straight forward to model using neither footprints nor peaks.

We have noticed that none of the segmentation based strategies have been combined with an affinity-based annotation of TFs, as offered by TRAP. Furthermore, the possibility of segmenting the genome based on peaks only, which are used for footprint detection anyways, has not been systematically analysed either. We followed up on these ideas and developed TEPIC, a framework for TFBS prediction using biophysical scores derived from TRAP, which can be easily applied to both peaks, derived from any chromatin accessibility assay, or to TF footprints [S+17a]. In addition to the TFBS prediction, TEPIC offers the computation of TF scores on the gene-level. Also, our framework includes several machine learning approaches using TFBS predictions to derive information on key regulators [S+18b]. Furthermore, TEPIC is included in the REGULATORTRAIL webserver [K+17a]. Before we introduce TEPIC in detail, we describe two related methods for TFBS prediction, which we used to benchmark TEPIC.

3.1.6 Epigenetic priors to compute TFBS predictions

To reduce the number of false positive TFBS predictions, Cuellar-Partida *et al.* proposed a combination of PWM derived motif hits with an epigenetic prior computed, for instance, from DNaseI-seq signal [CP+12]. They designed the prior to be a linear function whose value summed over all genomic positions equals the number of all TFs that are bound to any site in a tissue. The epigenomic signal is re-scaled to be within the range of $[0, 1]$, denoted by $g(y_i)$, where y_i is the original epigenetic signal at position i in the genome. Using $g(y_i)$, they define their prior function $f(y_i)$ as [CP+12]:

$$f(y_i) = \beta \frac{g(y_i)}{\sum_{j=1}^{n} g(y_j)}, \tag{3.15}$$

where n denotes the size of the genome and β is *"the total number of binding sites of all TFs"* [CP+12]. Thus, the final score $\hat{S}(s_i, y_i)$ indicating whether region s_i is

bound or not is given by:

$$S(s_i) = \sum_{j=1}^{m} W_{s_{i_j},j}, \tag{3.16}$$

$$P(y_i) = \log\left(\frac{f(y_i)}{1 - f(y_i)}\right), \tag{3.17}$$

$$\hat{S}(s_i, y_i) = S(s_i) + P(y_i), \tag{3.18}$$

$$\hat{S}(s_i, y_i) = \sum_{j=1}^{m} W_{s_{i_j},j} + \log\left(\frac{f(y_i)}{1 - f(y_i)}\right), \tag{3.19}$$

where $S(s_i)$ is a motif score based on a PFM W as defined above, m is the size of the considered motif, $P(y_i)$ is the ratio of the prior function testing whether s_i is bound or not according to the epigenetic signal y_i and the final score $\hat{S}(s_i, y_i)$ is the sum of $S(s_i)$ and $P(y_i)$.

3.1.7 Protein Interaction Quantification (PIQ)

The basic idea of PIQ is similar to that of FIMO-PRIOR. PWM based motif hits are reevaluated using an epigenetic prior, which in case of PIQ is explicitly obtained from DNaseI-seq data. In PIQ, reads are modeled from a Gaussian process, were the per-base read count μ_i is defined as [S$^+$14c]:

$$\mu_i = N(\mu_0, \Sigma), \tag{3.20}$$

$$\Sigma_{(i,j)} = Cov(\mu_i, \mu_j) = \sigma_0 k_{|i-j|}, \tag{3.21}$$

where $k_{|i-j|}$ defines both the degree and the smoothness of the signal across the genome. Details on parameter inference can be found in the Supplement of Sherwood *et al.* [S$^+$14c]. The core assumption taken in PIQ is that the DNaseI-seq profile around the center of a binding site y_M, where a factor is bound, would be different from a site where the factor is not bound. In other words, the assumption is that each TF generates a distinct, strand specific, DNaseI-seq pattern $(\hat{\mu}_i^+, \hat{\mu}_i^-)$ that is different from the unbound one (μ_i^+, μ_i^-). The strand specific adapted binding rate at position i can be computed according to [S$^+$14c]:

$$\hat{\mu}_i^+ = \mu_i^+ + \begin{cases} \beta_{i-j}^+ \text{ if } |y_m - j| \leq |M| & \text{and } I_M = 1, \\ 0 & \text{else.} \end{cases} \tag{3.22}$$

Here, $|M|$ denotes the length of the motif, y_M is the center of the motif, I_M is an indicator function evaluating to 1 if the factor is bound to y_M and to 0 otherwise. The parameter β^+ is the factor specific profile parameter. For the negative strand, the formula is computed analogously [S$^+$14c]. The strand specific scores are used in a logistic function that returns a probability p_j indicating whether a distinct genomic site j is bound by a factor. This probability is combined with two prior

functions f and g representing the PWM score and the overall DNaseI-seq counts for a distinct region j, respectively, to the final likelihood L_j by [S+14c]:

$$L_j = f_j + g_j + logit(p_j). \tag{3.23}$$

Intuitively, the three components of the PIQ score reflect the sequence specificity of TF binding (f_j), the general accessibility of the chromatin (g_j) as well as the factor specific DNase1-seq profile, modelled by $logit(p_j)$. Further details on how $logit(p_j)$ is computed are provided in the Supplement of Sherwood *et al.* [S+14c].

3.2 TEPIC for fast and accurate TFBS prediction

3.2.1 Usage of TEPIC for TFBS prediction

One of the fundamental features of our TEPIC framework is the ability to compute TFBS predictions. Within TEPIC we exploit the advantages of a biophysical scoring as provided by TRAP (Section 3.1.3) and integrate it into a segmentation based scoring method that takes a *bed*-file as input containing candidate binding sites obtained from either peak or footprint calling. Importantly, we do impose particular restrictions on the type of the assay used to find the candidate sites. Thus, the candidate regions could be, for instance, derived from either DNaseI-seq data, NOMe-seq data, or ChIP-seq of HMs. TEPIC annotates all sequences contained in the bed-file with TF affinities for each factor a TF motif is available for. Within the TEPIC repository (`www.github.com/Schulzlab/TEPIC`), we provide a maintained collection of TF motifs for various species obtained from Jaspar [K+18c], Hocomoco [K+18d] and the Kellis ENCODE Motif database [KK14], containing, for instance, sets for *homo sapiens*, *mus musculus* and *vertebrata* composed of 561, 380 and 690 TF motifs, respectively.

Using the command:

```
./TEPIC.sh -g genome.fa -b regions.bed -o Example -p Motifs.PSEM
```

TEPIC computes a matrix A with n rows and m columns, were n is the number of sites contained in `regions.bed`, corresponding genomic sequence are generated from `genome.fa` using BEDTOOLS [QH10] and m is the number of TF motifs contained in the entire motif collection $\mathcal{M}$ obtained from `Motifs.PSEM`. The motif set $\mathcal{M}$ does not contain typical PWMs as introduced before, instead it contains position specific energy matrices (PSEMs). The parameter λ and R_0 required for TRAP are already incorporated into the value of the PSEMs. Matrix A is computed using formula (3.14) as defined above:

$$A_{i,t} = \sum_{l=1}^{|s_i|-|M^t|} \frac{\mathcal{M}^t_{R_0} \cdot e^{-\beta_{\mathcal{M}^t} E(\lambda_{\mathcal{M}^t}, s_{i_l}, M^t)}}{1 + M^t_{R_0} \cdot e^{-\beta_{\mathcal{M}^t} E(\lambda_{\mathcal{M}^t}, s_{i_l}, M^t)}}, \forall i \in \{1, .., n\}, \forall t \in \{1, ..., m\}, \tag{3.24}$$

where s_i is the considered sequence, $|s_i|$ denotes the length of s_i in base pairs, s_{i_l} denotes the subsequence of s_i starting at index l, n is the number of sequences

to be scored, m is the number of considered PSEMs, M^t is the current TF motif, $|M^t|$ denotes the length of the motif, λ_M is the λ parameter chosen for all motifs in the motif set M and β_{M^t} is the energy parameter calculated for M^t. The sum iterates over all possible positions of M^t in s, denoted by l.

The summation of contributions from all possible binding sites within a candidate region as well as the energy based calculation of TF-binding scores ensures that also sites with a less likely binding motif still contribute to the overall score.

3.2.2 Validation of TFBS predictions using TF ChIP-seq data

To check the quality of the predictions computed in this manner, we have computed TFBS predictions using TEPIC within footprint calls from HINT-BC [G+16b], using FIMO-PRIOR and using PIQ for HepG2, GM12878, K562 and H1-hESC. Additionally, we obtained 33, 24, 19 and 22 preprocessed TF ChIP-seq datasets from ENCODE for those cell lines, respectively. ENCODE accession IDs and details on the execution of the software are provided in Section B.1.

The quality of the TFBS predictions is evaluated in terms of precision-recall AUC values (c.f. Section 2.2.2) computed using the PRROC package [G+15c]. This is a well suited strategy to evaluate the ranking of sites based on the affinity values and to deal with the imbalance of bound and unbound sites. In our evaluation, the cell type specific gold standard $\mathcal{G}_t$ for TF t is composed of all binding sites of t predicted with FIMO that overlap a ChIP-seq peak of t. The negative set $\mathcal{N}_t$ is holding all predicted sites not overlapping a ChIP-seq peak. Here, let c denote the score threshold used in the PR-AUC computation. Further, we define a site s_t as a TP if s_t overlaps a site $g \in \mathcal{G}_t$ and the score $a(s_t) > c$. Consequently, a FP is a site s_t not overlapping a $g \in \mathcal{G}_t$ and the score $a(s_t) > c$. A TN is a site $n \in \mathcal{N}_t$ that is either not overlapping any predicted site s_t, or $a(s_t) \leq c$. Lastly, a FN is a site $g \in \mathcal{G}_t$ not overlapping any of our predictions [S+18b].

As shown in Figure 3.5, our unsupervised TEPIC approach clearly outperforms the state of the art methods FIMO-PRIOR and PIQ. The superior performance is achieved by the combination of high-quality footprint calls from HINT-BC with the affinity-based scoring of TFs. Importantly, we outperform the TF specific DNaseI-seq models postulated in PIQ. However, for some TFs such as NRF1 and REST, FIMO-PRIOR outperforms both PIQ and TEPIC. It seems that for these TFs, an approach that considers footprints is not well suited. Due to their conceptual difference, We have not considered comparisons against supervised approaches like BINDNASE or CATCHITT.

3.2.3 Runtime analysis of TEPIC for TFBS predictions

During the course of TEPIC's development, we have published two versions of the framework, where the essential difference with respect to TFBS computation is a tremendous speed-up between TEPIC 1.0 and TEPIC 2.0 [S+17a, S+18b]. This speed-up could be achieved, for instance, by optimization of data processing steps, e.g. avoiding unnecessary annotations. By fare the most important difference has

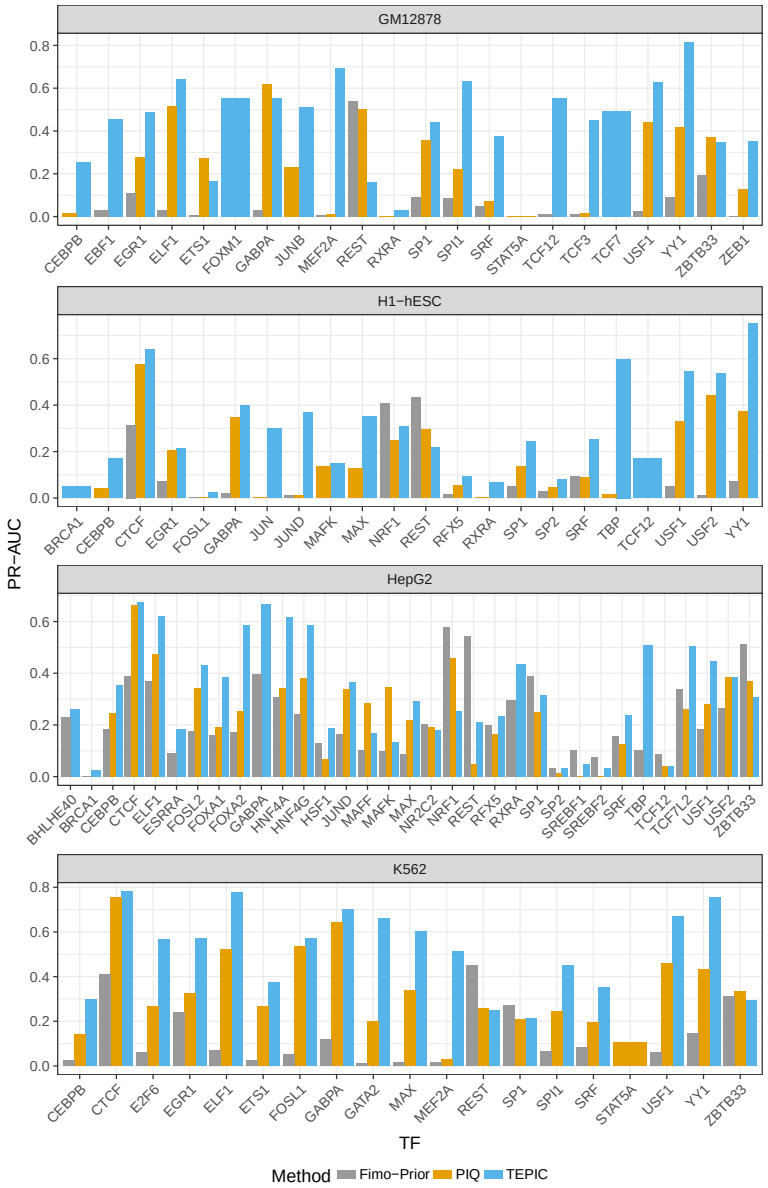


Figure 3.5: The y-axis shows the area under the precision-recall curve computed for several TFBS predictions obtained from various TFs on GM12878, H1-hESC, HepG2 and K562. The color code indicates which prediction method has been used. Overall, TEPIC combined with footprints from HINT, outperforms both Fimo-Prior and PIQ. Figure from Schmidt *et al.* [S+18b].

been the replacement of an R-implementation of Trap, by a C++ implementation previously used in the PASTAA webserver. However, the latter was not designed to be used in parallel, a functionality added by us. Note that the definition of the

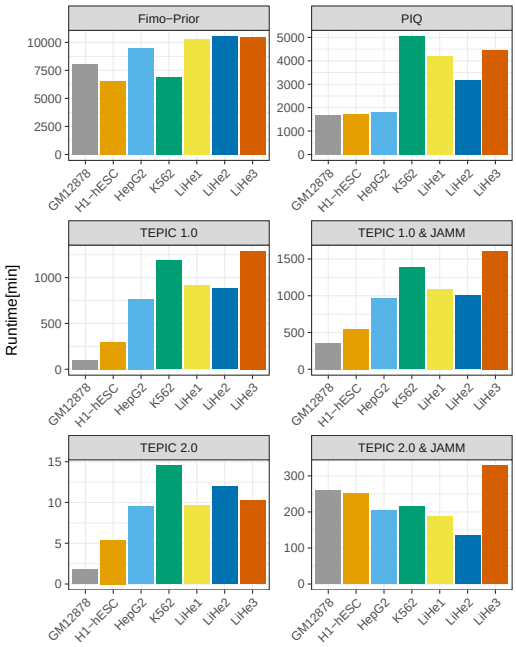


Figure 3.6: The y-axis shows the runtime in [min] for computing genome-wide TFBS predictions using FIMO-PRIOR, PIQ, TEPIC 1.0 [S+17a] and TEPIC 2.0 [S+18b]. As TEPIC requires the input of a candidate region file, we also include the runtime of this preprocessing step using the peak caller JAMM to provide a fair comparison to FIMO-PRIOR and PIQ. However, even including the peak calling step, both versions of TEPIC outperform the competitors, while TEPIC 2.0 again shows a significant speed-up compared to its predecessor. Figure from Schmidt *et al.* [S+17a, S+18b].

TF affinities was not changed in TEPIC 2.0. As shown in Figure 3.6, both version of TEPIC perform the TFBS prediction task much faster than both FIMO-PRIOR and PIQ, even if we add the time required for peak calling, for instance, using JAMM [I+15]. The runtime experiments have been conducted for seven samples, the cell lines mentioned above (HepG2, K562, GM12878, H1-hESCs) as well as three primary human hepatocytes samples from DEEP (LiHe1-3). Data IDs as well as details on processing of the DEEP data are provided in Section B.1. We assessed the runtime using the UNIX time utility (`/usr/bin/time`) on a compute server using an Intel Xeon CPU E7-8837 processor with 1TB of main memory. Within TEPIC and JAMM, 16 cores were used for the computation of TFBS prediction for 458 TFs, using the original TF motif set from Schmidt *et al.* [S+17a]. As above, for details on the execution of the software we refer the reader to Section B.1.

Both, the ChIP-seq and the runtime validation illustrate that TEPIC is capable of computing accurate and genome-wide TFBS predictions in very short time. In

the next section, we illustrate how these predictions can be aggregated to the gene-level, a feature of TEPIC that is not supported as an easy-to-use built-in function by any other tool listed in Table 3.2.

3.3 Aggregating genome-wide TFBS to the gene level with TEPIC

In this section, we explain how both genome wide TFBS predictions and TF ChIP-seq peaks obtained for one sample can be summarized onto the gene-level.

3.3.1 Common strategies to aggregate TFBS predictions to the gene-level

There are two main strategies how TFBS predictions obtained in peaks or footprints p can be linked to genes. The classical approach is to consider the distance $d_{p,g}$ of a predicted site p to each gene g. Next, p is assigned to the gene with the smallest value of $d_{p,g}$ [G$^+$15b]. By definition, in this approach a site p can only be linked to one distinct gene g, which especially in gene-dense regions might be an issue. Alternatively, in window based approaches, a gene is associated with regulatory regions p that are located within a defined genomic region, typically centered around a gene's TSS. In this thesis, windows are always centered at the most 5′ TSS of a gene. As the search windows around genes can be overlapping, a distinct site p can be assigned to more than one gene g [O$^+$09, M$^+$12b]. Figure 3.7 illustrates the two different concepts.

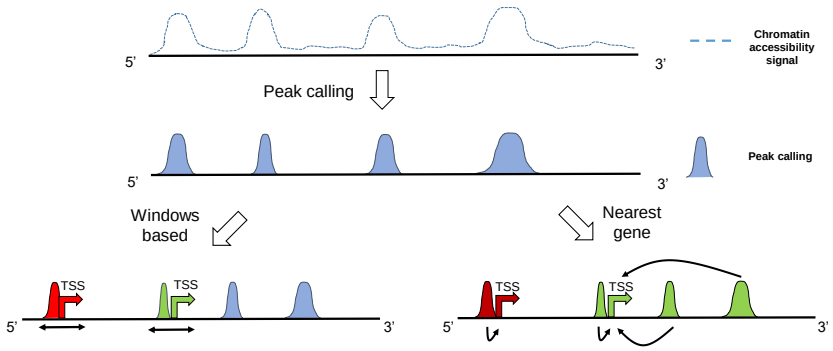


Figure 3.7: Here, the workflow of assigning TFBS to genes is illustrated exemplary for DHS identified using peak calling. In window based approaches, the peaks are assigned to a target gene using a window based linkage that is DHS sites are overlapped with a window centered at the TSS of a gene and overlapping peaks are matched. In nearest gene approaches, candidate regulatory sites are linked to a gene following the nearest gene approach that is a purely distance based association of each peak to its closest target gene in genomic space.

3.3.2 Computation of TF-gene scores in TEPIC

TEPIC's TF-gene score computation can be readily applied to TFBS predictions, but it can also be used to aggregate TF ChIP-seq data, as described in the following two sections. In Table 3.3, we provide an overview of all used variations of TF-gene scores.

Integration of predicted TFBS

Within TEPIC, we have tested both window based and nearest gene assignments of regulatory regions to genes. Let $\mathcal{P}_{g,w}$ denote the set of all regions assigned by linkage paradigm w to gene g.

Ouyang *et al.* suggested an exponential decay formulation for the aggregation of TF ChIP-seq data [O+09]. The assumption of the exponential decay is that the regulatory influence of a region on a gene declines with increasing distance to the gene. Although this is an oversimplification and neglects the occurrence of chromatin looping and long range interactions, it was shown to be a good approximation reducing noise added to the TF-gene scores [O+09, M+12b]. Therefore, we use the exponential decay formulation in TEPIC as well.

As mentioned in Section 3.2.1, TEPIC computes a score matrix $A_{p,t}$, denoting the TF affinity of TF t to site p by summing up the contribution of all possible TFBS in p.

In the original TF-gene score annotation, termed *Epi-Decay*($\mathcal{E}$), TF-gene scores $a^E_{g,t}$ are computed as (Eq. 3.23) [S+17a]

$$a^E_{g,t} = \sum_{p \in \mathcal{P}_{g,w}} a_{p,t} e^{-\frac{d_{p,g}}{d_0}}, \tag{3.25}$$

where d_0 is a constant set to 5000 [O+09].

The *Epi-Decay-Scaled* ($\mathcal{ES}$) annotation directly integrates the epigenetic signal s_p of region p into the TF-gene score $a^{ES}_{g,t}$ (Eq. 3.24)

$$a^{ES}_{g,t} = \sum_{p \in \mathcal{P}_{g,w}} a_{p,t} s_p e^{-\frac{d_{p,g}}{d_0}}. \tag{3.26}$$

In Schmidt *et al.* [SS18], we suggested *"normalized TF-gene scores"* [SS18] $\bar{a}^E_{g,t}$ (Eq. 3.25) that not only account for a bias caused by the length of region $|p|$ but also for a bias introduced by the number of potential TFBSs within p, given as $|p| - |m| + 1$. Furthermore, in addition to the TF affinities, we proposed three region-based features per gene: the number of considered regions c^E_g (Eq. 3.26), the length of those regions l^E_g (Eq. 3.27) and the combined epigenetic signal across

all considered regions f_g^E (Eq. 3.28):

$$\bar{a}_{g,t}^E = \sum_{p \in \mathcal{P}_{g,w}} \frac{a_{p,t}}{|p| - |m| + 1} e^{-\frac{d_{p,g}}{d_0}}, \tag{3.27}$$

$$c_g^E = \sum_{p \in \mathcal{P}_{g,w}} e^{-\frac{d_{p,g}}{d_0}}, \tag{3.28}$$

$$l_g^E = \sum_{p \in \mathcal{P}_{g,w}} |p| e^{-\frac{d_{p,g}}{d_0}}, \tag{3.29}$$

$$f_g^E = \sum_{p \in \mathcal{P}_{g,w}} s_p e^{-\frac{d_{p,g}}{d_0}}. \tag{3.30}$$

We name the set $\mathcal{EN} = \{\bar{a}^E, c^E, l^E\}$ *Epi-Decay-Normalized* and refer to $\mathcal{ESN} = \{\bar{a}^E, c^E, l^E, f^E\}$ as *Epi-Decay-Scaled-Normalized*, to $\mathcal{EPF} = \{c^E, l^E\}$ as *Epi-peak-features*, and to $\mathcal{EPFS} = \{c^E, l^E, f^E\}$ as *Epi-peak-features and signal*.

As mentioned above, to our knowledge, no other TFBS prediction tool provides this aggregation of predictions to the gene-level as a build in function.

Aggregation of TF ChIP-seq data

In a similar fashion TF-gene scores $a_{g,t}^C$ for gene g and TF t are computed for TF ChIP-seq data:

First, we get for $a_{g,t}^C$ (Eq. 3.29) as in Ouyang *et al.* [O+09]:

$$a_{g,t}^C = \sum_{p \in \mathcal{P}_{g,w}} c_{p,t} e^{-\frac{d_{p,g}}{d_0}}, \tag{3.31}$$

where we sum all ChIP-seq scores $c_{p,t}$ for TF t, weighted by their distance to the TSS $d_{p,g}$. The scores $c_{p,t}$ are defined as the $-log$ of the p-value computed for peak p by ENCODEs uniform peak processing pipeline. As above and proposed by Ouyang *et al.* [O+09], the parameter d_0 is set to 5000. We call this score design ChIP-seq TF-features ($\mathcal{C}$).

Also for ChIP-seq data, we suggested *"normalized TF-gene scores"* [SS18] $\bar{a}_{g,t}^C$ (Eq. 3.30). We defined $\bar{a}_{g,t}^C$ as the fraction of $a_{g,t}^C$ and the total number of ChIP-seq peaks c_g^C (Eq. 3.31). Additionally, we considered c_g^C and the total peak length l_g^C (Eq. 3.32), as extra features:

$$\bar{a}_{g,t}^C = \frac{\sum_{p \in \mathcal{P}_{g,w}} c_{p,t} e^{-\frac{d_{p,g}}{d_0}}}{c_g^C}, \tag{3.32}$$

$$c_g^C = \sum_{t \in \mathcal{T}} \sum_{p \in \mathcal{P}_{g,w}} I(c_{p,t}) e^{-\frac{d_{p,g}}{d_0}}, \tag{3.33}$$

$$l_g^C = \sum_{t \in \mathcal{T}} \sum_{p \in \mathcal{P}_{g,w}} I(c_{p,t}) |p| e^{-\frac{d_{p,g}}{d_0}}. \tag{3.34}$$

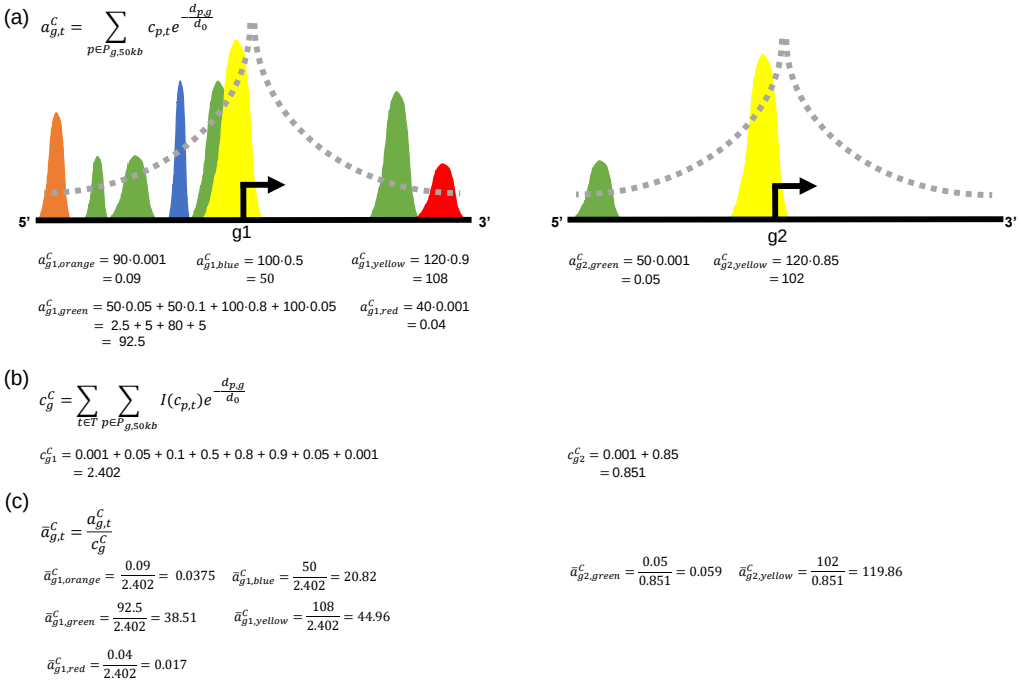


Figure 3.8: (a) The computation of the TF-gene scores $a^C_{g,t}$ from ChIP-seq data is shown for two genes $g1$ and $g2$ using several chipped TFs. In (b), we show how the normalization factor c^C_g is computed. Part (c) illustrates how the normalized TF-gene scores $\bar{a}^C_{g,t}$ are computed. As one can see, the scores for $g2$ are increasing, as there are not many peaks located in the vicinity of that gene. Simultaneously all scores of $g1$ are shrinked as this gene is residing in a region that is heavily bound by TFs. Figure from Schmidt *et al.* [SS18].

Here, the set comprised of all chipped TFs is denoted by $\mathcal{T}$, the length of peak p is represented by $|p|$ and I is the indicator function returning one if there is a ChIP-seq peak for TF t among all peaks in $\mathcal{P}_{g,w}$ for gene g and window w and zero otherwise, i.e. the function checks whether $c_{p,t}$ is non-zero. We point out that the additional features c^C_g and l^C_g are also weighted by the distance of the respective regions. Normalized scores are denoted by ChIP-seq TF-features normalized ($\mathcal{CN}$). An example for the impact of the normalization is provided in Figure 3.8.

The combination of features c^C_g and l^C_g is called *ChIP-seq peak-features* ($\mathcal{CPF}$). As pointed out in Schmidt *et al.*, c^C_g and l^C_g (Eq. 3.31, 3.32) *"capture the regulatory activity in the vicinity of a gene measured with ChIP-seq experiments. Thus, [these scores] can be seen as an aggregated view for the activity of transcriptional regulation"* [SS18].

Table 3.3: Overview of different TF-gene score variations used in this thesis [SS18].

	Abbreviation	Equation	Included features
Epi Decay	$\mathcal{E}$	(3.23)	a^E
Epi Decay-Scaled	$\mathcal{ES}$	(3.24)	a^{ES}
Epi Decay normalized	$\mathcal{EN}$	(3.25,3.26,3.27)	$\bar{a}^E, c^E, l^E$
Epi Decay-Scaled normalized	$\mathcal{ESN}$	(3.25,3.26,3.27,3.28)	$\bar{a}^E, c^E, l^E, f^E$
Epi peak-features	$\mathcal{EPF}$	(3.26,3.27)	c^E, l^E
Epi peak-features and signal	$\mathcal{EPFS}$	(3.26,3.27,3.28)	c^E, l^E, f^E
ChIP-seq TF features	$\mathcal{C}$	(3.29)	a^C
ChIP-seq TF features normalized	$\mathcal{CN}$	(3.30)	$\bar{a}^C$
ChIP-seq peak features	$\mathcal{CPF}$	(3.31,3.32)	c^C, l^C

3.4 Gene-expression modelling using TF-gene scores

Utilizing TF-gene scores as described in Section 3.3.2, we proposed to build a linear regression model predicting gene-expression, with the aim of inferring which TFs have a regulatory role in the analysed tissue or cell type. This is known as *per-sample learning*, because the models attempt to find features that predict gene-expression well over all genes within one sample. In instances where only a few samples are available, *per-gene learning*, which aims at identifying these associations for a distinct gene, can not be applied and only per-sample models can be used.

In the literature, several approaches can be found that suggest interpretable gene-expression models based on a variety of different features [O$^+$09, C$^+$11b, N$^+$12, M$^+$12b, W$^+$13a, OB14, B$^+$15a, LCHG15, S$^+$16b]. Aside from predicting gene-expression, the purpose of these models is to identify and to interpret those features that can be meaningfully associated with gene-expression. Novel insights on the overall importance of TFs both within [O$^+$09, S$^+$17a] and between samples [O$^+$09, C$^+$12a, D$^+$16e] can be obtained by methods that are utilizing either TF ChIP-seq or predicted TF binding data.

Due to the large amount of epigenetics data produced in consortia like ENCODE [D$^+$12b], Roadmap [K$^+$15] and Blueprint [A$^+$12], *in silico* models of transcriptional regulation have gained popularity in the community. For instance Ouyang *et al.* predicted gene-expression in mouse embryonic stem cells (mESC) from TF ChIP-seq data and used it to model differential expression between mESCs and embryoid bodies [O$^+$09].

Within this section, we present the machine learning model used by us, illustrate several analyses we have conducted to understand our model's performance, detail dependencies on various model characteristics and illustrate applications of the models to different primary cell types in the scope of the DEEP project. An overview of the used data sets and details on how those have been processed are provided in Section B.1.

3.4.1 Statistical model

Throughout this chapter, linear regression models using elastic net regularization, as implemented in the glmnet R-package [FHT10] are used to predict gene-expression. As features, we considered TF-gene scores derived either from Tepic's predicted TFBS or from TF ChIP-seq data. As explained in Section 2.2.1, models regularized with elastic net are sparse and thus interpretable. Furthermore, the grouping effect preserves correlated features. In the problem sets at hand features are correlated frequently, for instance, due to cooperation and co-occurrence of TFs. These desirable characteristics of the elastic net are achieved by a combination of the ridge and the lasso penalty terms:

$$\hat{\beta} = \arg\min_{\beta} ||y - X\beta||^2 + \lambda[\alpha||\beta||^2 + (1-\alpha)||\beta||]. \tag{3.35}$$

Here, β is the feature coefficient vector, $\hat{\beta}$ are the estimated coefficients, X is the TF-feature matrix, y the response vector holding gene-expression estimates and the parameter λ controls the total amount of regularization. The feature matrix X holds the features explained in Section 3.3.2. For example, using $\mathcal{C}$-scores the rows of X refer to genes and the columns contain TF-gene scores based on ChIP-seq data. Consequently, the entry $X_{g,t}$ reflects the TF-gene score $a^{C}_{g,t}$ for gene g and TF t. In Section B.1.9, we exemplify the composition of the feature matrix depending on the TF-gene scores listed in Table 3.3.

Both X and y are log-transformed considering a pseudo-count of 1 and are subsequently centered (subtraction of the column means from the column values) and scaled (the centered columns of X are divided by their standard deviations). The parameter α controls the trade-off between the ridge and lasso penalty terms. It is optimized in a grid search considering the interval [0.0,1.0] with a step-size of 0.01.

As described in Schmidt *et al.* we assessed the quality of the model using a ten-fold Monte-Carlo cross-validation procedure considering a randomly sampled hold-out test data set that is comprised of 20% of the complete data, while the remaining 80% of the data are used for training. The *cv.glmnet* procedure is used to fit the parameter λ in a six-fold inner cross-validation procedure. The λ is selected according to the minimum cross validated error, which is computed as the average mean squared error (MSE) on the inner folds *(lambda.min)*. The entire learning procedure is visualized in Figure 3.9. The selected λ and the entire training data set is used to compute the final regression coefficients. The total number of non-zero regression coefficients is denoted with $||\beta||_0^{model}$.

Conclusions on which TFs are relevant regulators can be drawn from the coefficients $\hat{\beta}$ computed by the model. Their sign and magnitude can be seen as an indicator for the explanatory power of TFs for gene-expression, averaged over all considered genes within the analysed sample. Before we start interpreting the model coefficients (Section 3.4.4), we examine the performance of the models and characterise factors influencing model behaviour (Section 3.4.2).

This section is a slightly adapted version of Section 2.4 from Schmidt *et al.* [SS18].

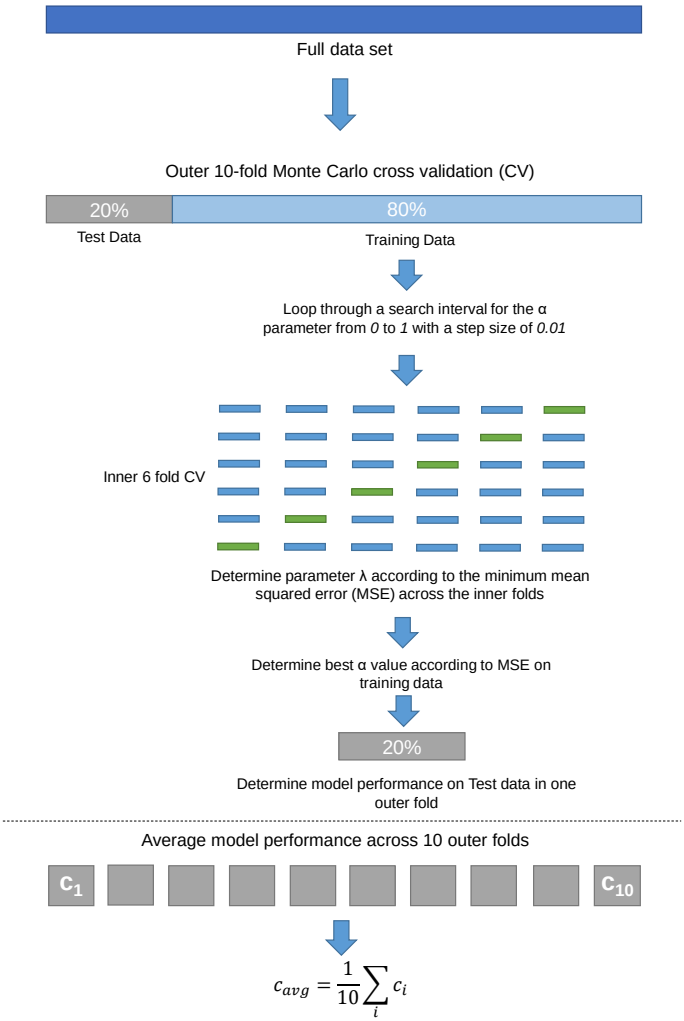


Figure 3.9: Schematic overview of the learning paradigm used to derive cell type specific transcriptional regulators from TF-gene scores. We assess the performance of a linear model using elastic net regularization in a 10-fold Monte-Carlo cross-validation procedure. Model parameters are learned in a 6-fold inner cross validation procedure, while the parameter α is optimized in a grid search in [0.0,1.0] with a step-size of 0.01. Figure from Schmidt *et al.* [SS18].

3.4.2 Model performance and evaluation

In the following, we describe how variations in the generation of the TF-gene scores influence model performance. Having models with a good performance is essential to draw reliable conclusions on which of the modeled factors are important for transcriptional regulation. Details on the data used in this section are provided in

Section B.1. Throughout all figures in this section, model performance is assessed in terms of Pearson correlation, if not stated otherwise. Except for the next paragraph, all results of this section are published in Schmidt *et al.* [S+17a].

Nearest gene compared to window based approaches

Before assessing the performance of models that utilize the actual TF binding information, we investigate the performance of models based only on chromatin accessibility and ChIP-seq activity information that is models $\mathcal{EPF}$ and $\mathcal{CPF}$, respectively. We have learned $\mathcal{EPF}$ and $\mathcal{CPF}$ models for five different cell lines (GM12878, HeLa, HUVEC, IMR90 and K562) utilising both the nearest gene and the window based linkage, the latter with two different window sizes: 3kb and 50kb.

As shown in Figure 3.10a, the window based models almost always outperform their nearest gene counterpart across all cell lines and two different sizes of w, namely 3kb and 50kb. Using ChIP-seq data, we observe that model performance is generally boosted as opposed to DNaseI-seq data and that window based models constantly outperform nearest gene models (Figure 3.10b). In Figure 3.10c the mean squared error (MSE) for 9000 randomly selected individual genes from the $\mathcal{EPF}$ model for HeLa cells are shown (Here a 50kb window is used). Note that the HeLa models shows on average a similar performance with either association strategy. Therefore, it is a nice example to illustrate the gene-specific advantages of one of the linkage paradigms. For example the MSE of *RPL7A(ENSG00000148303)* is nearly twice as high using the nearest gene than the window based annotation. As shown in Figure 3.11a there seems to be a bidirectional promoter for *RPL7A* and *MED22*. The model suggests that this can not be adequately covered by the nearest gene approach. A different scenario is depicted in Figure 3.11b for the gene *HINT1(ENSG00000169567)*. This gene is located in a gene sparse region surrounded by several DHS peaks which seem to add large portions of noise in the nearest gene approach. In contrast to that, for the gene *APOA2(ENSG00000131096)*, the nearest gene approach leads to a better performance as it neglects, in contrast to the window based model, several DHS sites that seem to be associated with *TOMM40L* (Figure 3.11c). Each of these genes, *RPL7A, HINT1* and *APOA2* is highlighted in Figure 3.10c.

Overall, these results suggest that neither the window based, nor the nearest gene annotation generalise well across all genes, although the window based approach performs slightly better on average. Therefore, for the remaining analysis, we stick to the window based approach.

In Chapter 8, we present a method that overcomes the limitations of both window and nearest-gene based assignments by learning regulatory regions *de novo* from large-cohorts. However, for now, we stick to the case where only few samples are available and only per-sample models can be considered.

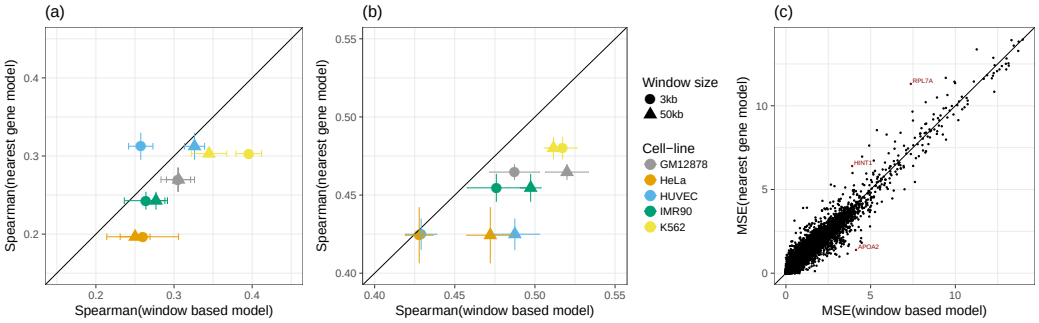


Figure 3.10: (a) Performance of $\mathcal{EPF}$ models comparing the nearest gene against the window based approach. (b) Performance of $\mathcal{CPF}$ models using either the nearest gene compared to the window based annotation. (c) Mean squared error shown for individual genes from the HeLa cell line comparing $\mathcal{EPF}$ models using the window based and nearest gene annotation. Joint work Fabian Kern in the scope of the extension of his bachelor thesis, presented at GCB 2018.

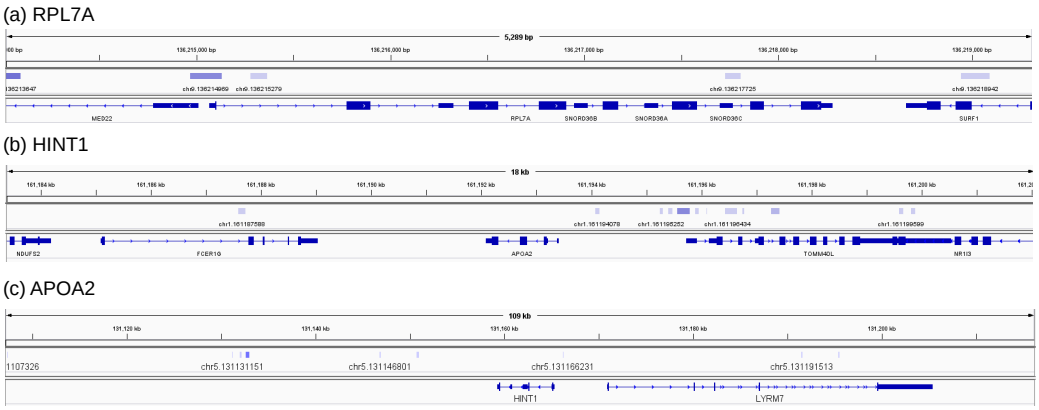


Figure 3.11: IGV browser tracks for (a) *RPL7A*, (b) *HINT1* and (c) *APOA2* illustrating the benefits of the linkage approaches over each other, depending on genomic context. Joint work Fabian Kern in the scope of the extension of his bachelor thesis, presented at GCB 2018.

Performance of predicted TFBS in window based models

We have tested the $\mathcal{E}$ and $\mathcal{ES}$ models using two different window sizes, a 3kb and a 50kb window, along with DNaseI-seq data from ENCODE obtained for GM12878, H1-hESC, K562, DEEP DNaseI-seq data for HepG2 and three primary human hepatocyte samples (LiHe1-3) as well as NOMe-seq data from DEEP for six CD4+ T-cell samples (T1-6). See Section B.1 for details on the data and on data preprocessing.

As depicted in Figure 3.12, including the signal intensity within candidate TFBS into the TF-gene scores improves model performance. On average, we observe that

the $\mathcal{ES}$ models perform better than $\mathcal{E}$ models and that 50kb models outperform 3kb models. Furthermore, combining the exponential decay in the 50kb window with scaling of the TF-gene scores using the chromatin accessibility signal outperforms all other tested setups.

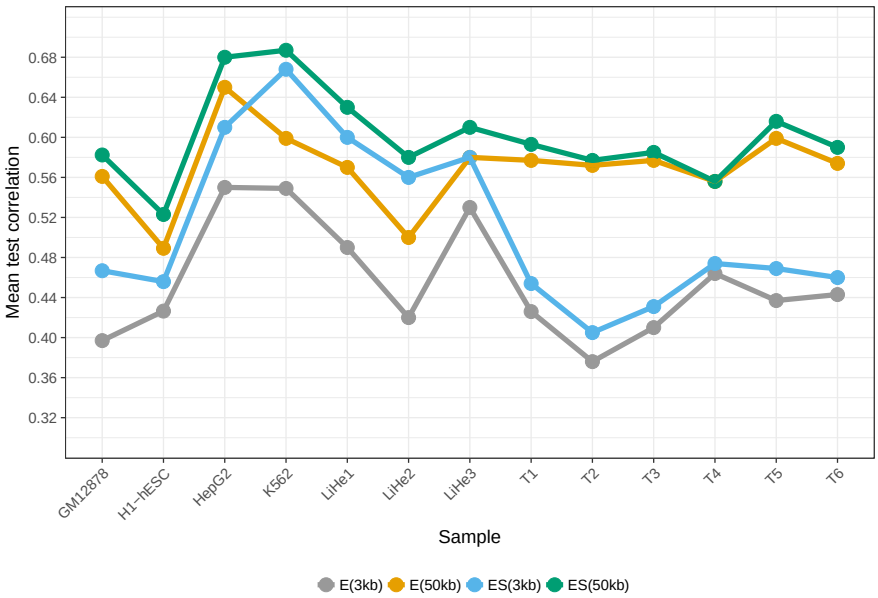


Figure 3.12: Mean test correlation for all linear regression models using predicted TFBS score aggregated using the $\mathcal{E}$ and $\mathcal{ES}$ models with two different windows. $\mathcal{ES}$ models outperform $\mathcal{E}$ models and the 50kb window leads to better results than the 3kb window. On average, the CD4+ T-cell samples (T1-T6) screened with NOMe-seq tend to perform worse than the remaining samples, which were screened with DNaseI-seq, especially considering the 3kb window. Figure from Schmidt *et al.* [S+17a].

This observation not only points out that incorporating distal TF binding events is important to model gene regulation accurately, it also suggests that the quantitative information on how accessible a TFBS is in a pooled sample provides insights on the regulatory activity of this site. We observe that the positive effect of scaling is stronger for DNaseI-seq than for NOMe-seq samples. This difference might be due to the inherent differences of the assays.

According to in-house quality control, the DNaseI-seq experiment for LiHe2 is of low quality, which might explain the poor performance of this sample in our gene-expression models. The poor quality is also reflected by a rather low number of DHSs detected in LiHe2 compared to, for instance, LiHe3 (1,166,618 versus 1,800,917 DHS called by JAMM).

Peak numbers influence model quality

We analysed the relation between the number of considered accessible sites and the performance of the linear regression models using TEPIC's TF-gene scores (Figure 3.13) by constructing twelve different peak sets using HepG2 DNaseI-seq data. The sets are including DHS sites ranked by their JAMM peak score. We considered $10,000$, $50,000$, $100,000$, $200,000$, ... , $900,000$ and all filtered peaks that is $1,023,463$.

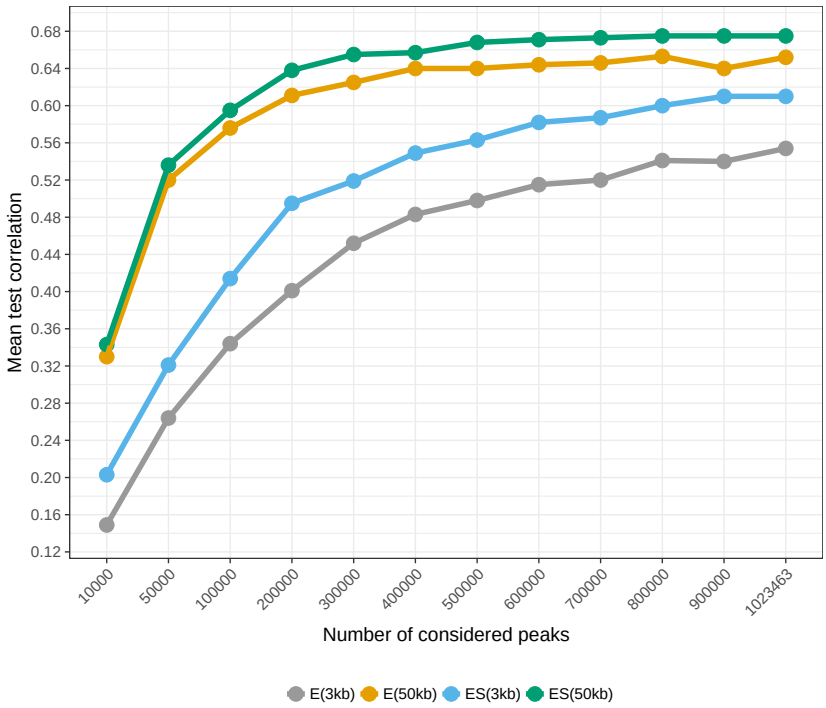


Figure 3.13: Throughout all annotations setups, model performance increases with increasing peak numbers. For the 50kb models, the performance remains unchanged for $\geq 500,000$ peaks, whereas it keeps rising for 3kb models until all peaks are included in the model. Figure from Schmidt *et al.* [S+17a].

We observe that the performance of the 50kb setups ($\mathcal{E}$ and $\mathcal{ES}$) remains roughly constant for peak numbers $\geq 500,000$, while the performance of the 3kb setups rises steadily until all peaks are included. This may be considered as support for the hypotheses that more distal regulatory events captured by the 50kb window are vital to model gene regulation. Furthermore, we notice that the difference between the setups pertaining to the same window size with and without the incorporation of the open-chromatin signal, respectively, rises with increasing peak numbers. This suggests that it is truly important to prioritize certain peak regions using the chromatin accessibility signal.

Footprints harbour essential transcription factor binding sites

So far, most *segmentation-based* methods identify TF binding sites by predicting footprints [G+16b]. There, we compared a footprint-based segmentation to a peak-based segmentation using DHSs. The peak-based segmentation has the advantage that it does not require specifically designed footprint-calling methods and that the actual footprint calling can be omitted after the already performed peak calling procedure.

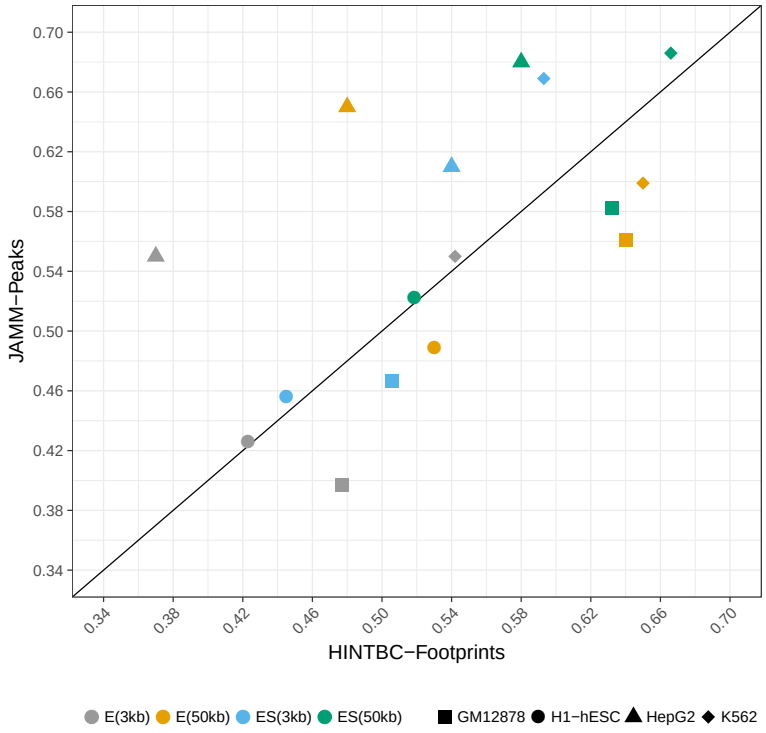


Figure 3.14: This scatter plot contrast the performance of linear models predicting gene-expression using TF-gene scores that are based on either HINT-BC footprints or JAMM-Peaks. The shape of the points indicates which cell line was used for the experiment, the color indicates the used annotation. There is no clear trend, which method performs best in this scenario, although one can see that the performance depends on the considered sample, e.g. HepG2 performs better with peaks, whereas, GM12878 constantly achieves better performance with footprints. Figure from Schmidt *et al.* [S+17a].

To conduct the comparison, we considered 452, 281 footprints in HepG2, 738, 707 footprints in K562, 598, 500 footprints in GM12878 and 1, 023, 559 footprints in H1-hESC identified with an accurate footprinting method for DNaseI-seq data, HINT-BC [G+16b]. The footprint calls were provided to us by Eduardo Gusmao,

the developer of HINT-BC. We used Tepic to compute TF-gene scores considering the regions around each footprint. As the footprints are often < 10bp in length, we computed TF-gene scores in small windows, either 24bp or 50bp that are centered in the middle of the footprints to ensure that scores for all TFs can be computed. We found that the results for both window sizes are very similar. Therefore, we present only the results for the slightly better 50bp window in the thesis. See Schmidt *et al.* [S$^+$17a] for the 24bp result.

Figure 3.14 illustrates the comparison between Tepic applied to footprints and DHSs. The DHS-based approach outperforms the footprints in HepG2 and K562. Furthermore, the DHSs have a slight performance advantage over footprints in H1-hESC, whereas in GM12878 the footprint based approach outperforms DHSs.

As before with DHSs, we see that including the chromatin accessibility signal, comparing *ES* against *E* scores, is beneficial for footprint based TFBS predictions as well. Although the peak based models do achieve a better performance on average, it is remarkable that the rather small footprint regions seem to capture most of the important binding sites. Using only 22.98%, 25.33%, 36.02% and 91.2% of base pairs in footprint regions compared to peaks in HepG2, K562, H1-hESC and GM12878 , respectively, illustrates that indeed most of the important regulatory TFBSs overlap the footprint calls.

The choice of peak callers tremendously influences model performance

During the development of Tepic, we compared two different peak callers JAMM and MACS2 [I$^+$15, Z$^+$08a]. While the latter one is the main peak caller used within the DEEP consortium, JAMM has been specifically designed to account for the narrow peaks present in DNaseI-seq data [I$^+$15], a design aspect that has not been considered specifically in MACS2. Section B.1 provides details on the command calls and parameters used for peak calling. For this analysis, we have used the DEEP DNaseI-seq samples for the previously introduced samples HepG2, LiHe1, LiHe2 and LiHe3.

As depicted in Figure 3.15 JAMM peaks lead to a better correlation between predicted and measured gene-expression than MACS2 peaks. This observation might be explained by the difference between the overall number of called peaks that is JAMM computes far more DHS sites, than MACS2. For instance, on HepG2 JAMM calls $1,023,463$ DHS sites, while MACS2 calls only $65,497$ peaks. As we found that the overall number of peaks also influences the learning (Figure 3.13), this might be an explanation for the poor performance of MACS2 peaks.

Interestingly, our results show that scaling the TF-gene scores using the DNaseI-seq signal within MACS2 peaks does not necessarily improve the learning result as compared to JAMM peaks. For example, in LiHe3 the correlation drops form 0.3 to 0.26 comparing the *E* and the *ES* setup with a 3kb window. A possible explanation for this behaviour is the already mentioned better resolution of JAMM for DNaseI-seq data, resulting in sharper peaks compared to MACS2: the mean width of peak in HepG2 called with JAMM is 96bp while it is 534bp for MACS2 peaks. Presumably, the chromatin accessibility information in the more narrow

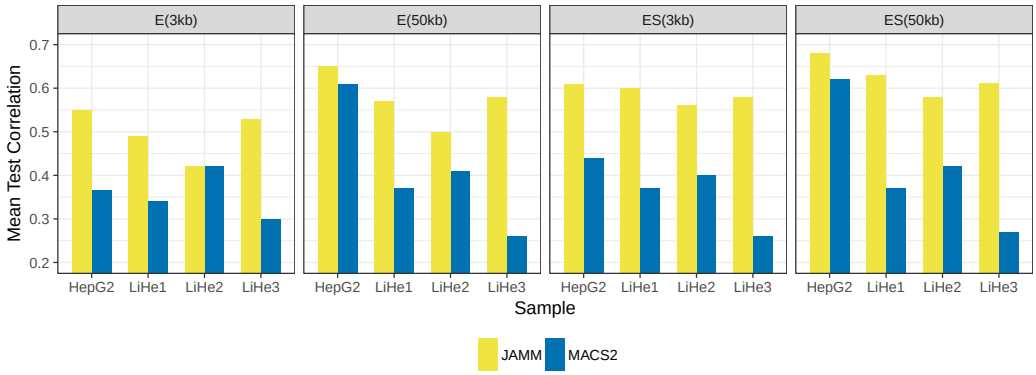


Figure 3.15: Mean test correlation of linear models for HepG2 and LiHe1-3 using DHS sites produced with JAMM or MACS2. Throughout all samples and annotation versions, JAMM peaks lead to a better performance of the regression models. Figure from Schmidt *et al.* [S+17a].

regions is less noisy and thus provides a clearer link between gene regulation and expression.

Choice of PWMs influences model performance

In the original Tepic publication, we have considered a TF motif set comprising 439 motifs from Jaspar (release 2014) [K+18c] and Uniprobe [H+15b]. We have compared this set against the complete set of human mono-nucleotide profiles from Hocomoco, version 10, containing 641 TF motifs [K+18d]. We compared the performance of TF-gene scores computed for both sets in gene-expression learning using the samples HepG2, LiHe1, LiHe2 and LiHe3. As shown in Figure 3.16, the Hocomoco motifs perform worse than the set obtained from Jaspar and Uniprobe. Therefore, for all further application scenarios within Schmidt *et al.* [S+17a], we used the Jaspar and Uniprobe set.

This analysis highlights that the quality of TF motifs is essential to infer well performing models. As a consequence, the motifs also directly influence the interpretability of the model coefficients.

Hit-based vs Affinity-based TFBS predictions

As explained in Section 3.2, Tepic uses a biophysical measure to quantify the binding of TFs. We have compared this affinity-based measure to traditional hit-based annotations in two different ways:

Firstly, we have replaced Trap with Fimo [G+11] to annotate DHS sites with TFBS, used the Fimo log-likelihood ratio scores (c.f. Section 3.1.2) in the TF-gene score computation. Therefore, we do not log-transform the Fimo TF scores in the elastic net model, comparable to McLeay *et al.* [M+12b], as this would imply taking the log two times. Secondly, we have compared Tepic against a state-of-the-art

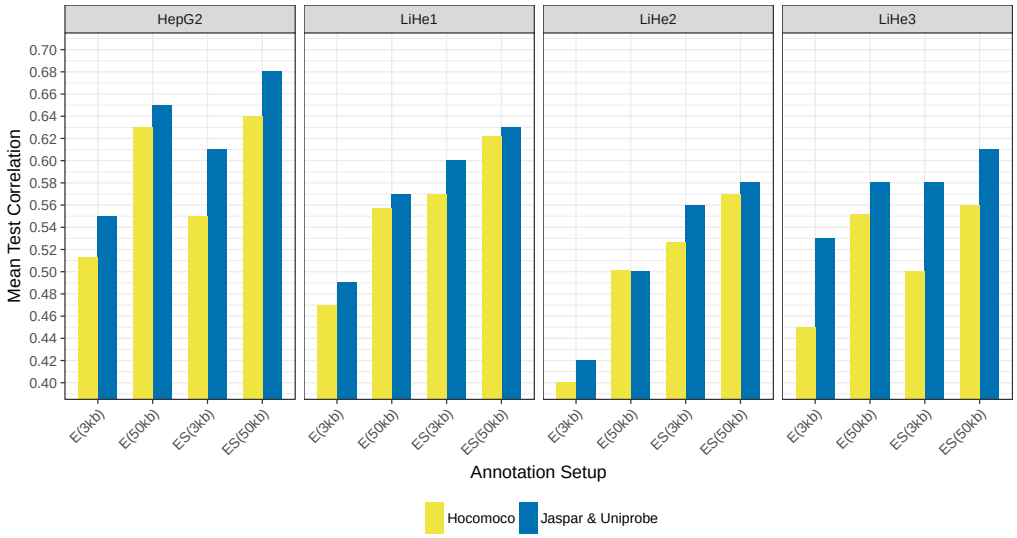


Figure 3.16: Mean test correlation of linear models for HepG2 and LiHe1-3 using TF motifs from Jaspar & Uniprobe compared to models using Hocomoco TF motifs. For all samples and annotation versions, the Hocomoco motifs perform worse than the combined Jaspar & Uniprobe set. Figure from Schmidt *et al.* [S+17a].

TF binding prediction method by Cuellar-Partida *et al.* that extends FIMO with an epigenetic prior [CP+12]. Just as in Gusmao *et al.* [G+14d], we refer to this method as FIMO-PRIOR, c.f. Section 3.1.6.

To run FIMO, we have used its default parameters, except for the parameter `max-stored-scores`, which we set to 200,000 instead of its default 100,000. This ensures that more binding sites can be retrieved by the tool. We applied FIMO to the samples HepG2, K562, GM12878, H1-hESC, LiHe1, LiHe2 and LiHe3.

Similar to the TF ChIP-seq comparison shown in Section 3.2.2, the results provided in Figure 3.17a illustrate that the incorporation of low-affinity binding sites using TRAP outperforms the traditional hit-based TFBS predictions methods in the gene-expression models. This underlines the importance of including low-affinity binding events into consideration. Similarly, in Figure 3.17b and c, we contrast the performance of TEPIC TF-gene scores using the *ES* models with 3kb and 50kb windows against FIMO-PRIOR and PIQ, respectively. Overall, we see that TEPIC performs favourably compared to both approaches.

Importantly, our results indicate that the performance of FIMO-PRIOR on K562, on H1-hESC and on GM12878 decreased if 50kb windows are considered compared to the 3kb window. This observation might be related to how the epigenetic signal is used in FIMO-PRIOR. As stated before, FIMO-PRIOR is a site-centric approach considering all binding sites in the 50kb window. Although the open-chromatin signal is used for reweighing the TFBS predictions, it may be that still too many false positive hits are considered in the final TF-gene score computation.

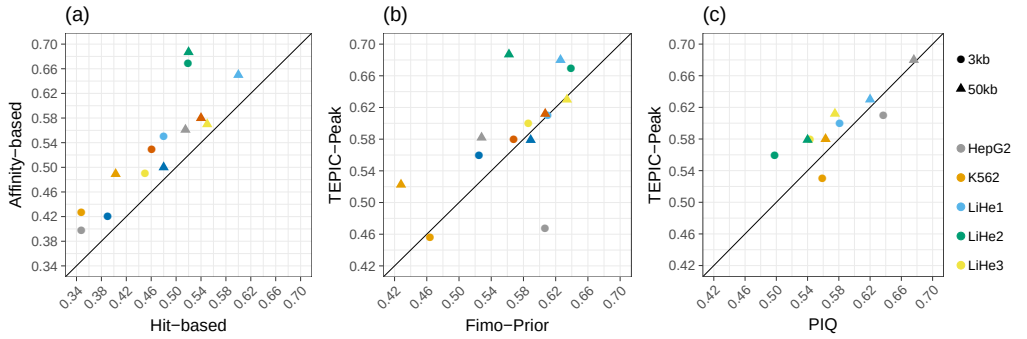


Figure 3.17: (a) Scatter plot comparing the performance of linear models using TF-gene scores composed of hit-based TFBS predictions computed with FIMO compared to the affinity-based scores used in TEPIC. The affinity-based scoring clearly outperforms the hit-based scores. Comparisons of FIMO-PRIOR against TEPIC and against PIQ are shown in (b) and (c), respectively. Overall, TEPIC scores achieve a better correlation between predicted and measured gene-expression than the ones from FIMO-PRIOR and PIQ. Figure from Schmidt *et al.* [S+17a].

Another important aspect is the runtime of the considered methods. As indicated in Figure 3.6, TEPIC runs much faster than both FIMO-PRIOR and PIQ, which renders FIMO-PRIOR and PIQ to be a bad choice if many samples need to be annotated.

Histone marks are well suited to pinpoint TFBS as well

Not only DHSs, but also Histone Marks (HMs) have been successfully used in predicting TFBSs [CP+12, B+15a, PR+11, G+16b]. Using preprocessed ENCODE ChIP-seq data of the active chromatin marks H3K4me3 and H3K27ac obtained for HepG2, K562, GM12878 and H1-hESC we show that HMs can also be used in TEPIC to pinpoint candidate TFBS. We have applied TEPIC separately to sites enriched for the active promoter mark H3K4me3 as well as on the active enhancer mark H3K27ac [H+09, C+10b]. Figure 3.18 holds model performance for these models and as it can be seen, both HMs lead to good performance in gene-expression learning.

Similar to the DNaseI-seq data, we note that using a larger window improves the learning results and that incorporating the abundance of the ChIP-seq peaks into the TF-gene scores improves model performance further in most cases. Besides, regions enriched for H3K4me3 lead to better prediction performance than regions enriched in H3K27ac across all samples. This could be related to the strong association of H3K4me3 to active promoters, whereas H3K27ac is rather related to potentially very distal enhancer regions [H+09, C+10b]. Particularly, this might explain the reduced performance of H3K27ac peaks in the 3kb windows compared

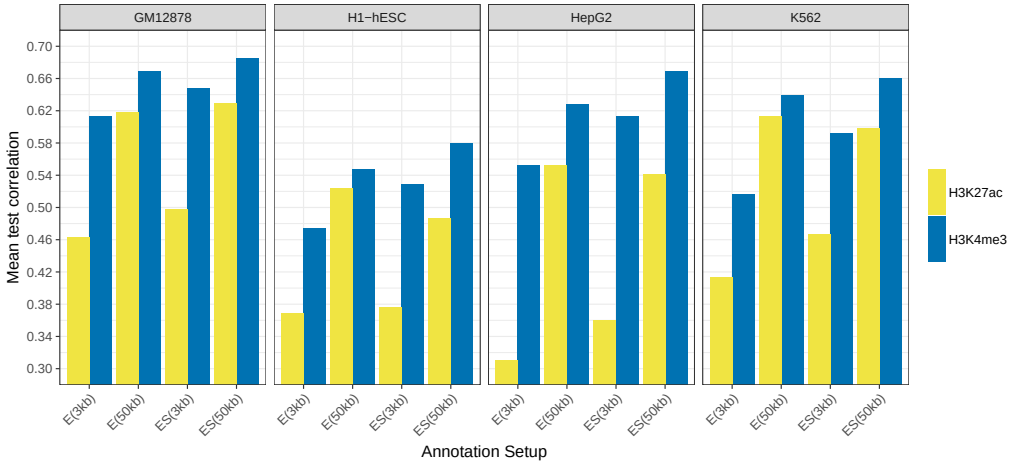


Figure 3.18: Performance of gene-expression models using sites enriched for either H3K4me3 or H3K27ac instead of DNaseI-seq or NOMe-seq peaks. We observe that H3K4me3 leads to better model performance than H3K27ac on average. Interestingly, the difference between both approaches is more prominent with 3kb and 50kb models. Figure from Schmidt *et al.* [S+17a].

to H3K4me3, as the 3kb window is likely to capture mostly TF binding events at the core promoter of a gene, whereas the effect of enhancers can be captured better in the bigger 50kb window.

Predicted TF-gene scores through TEPIC vs TF ChIP-seq

Another way of assessing the quality of both TEPIC's TF-gene scores and of the linear models is to compare them to models based on TF ChIP-seq data, as proposed by Ouyang *et al.* [O+09]. In Figure 3.19, the learning results for HepG2, K562, GM12878 and H1-hESC are shown. To state the relation between the different TFBS prediction methods, the figure holds the best correlation achieved by applying FIMO within DHSs (labelled as Hit-based), using FIMO-PRIOR, using TEPIC in footprints and in DHSs as well as using TF ChIP-seq data.

In HepG2 and K562 cells, TEPIC applied to DHSs outperforms all other TFBS prediction approaches, including FIMO-PRIOR as used in McLeay *et al.*[M+12b] and achieves correlation values that are close to what is obtained by using TF ChIP-seq data. In GM12878 and H1-hESC, TEPIC applied to footprints, outperforms the competitive prediction methods and also achieves an acceptable prediction performance.

These results are yet another indicator for the good quality of our TFBS predictions. The performance difference between the ChIP-seq and the best performing TFBS approach is not surprising, as all TFBS prediction approaches are still missing many binding events of TFs and all mechanisms of TF binding are not yet fully

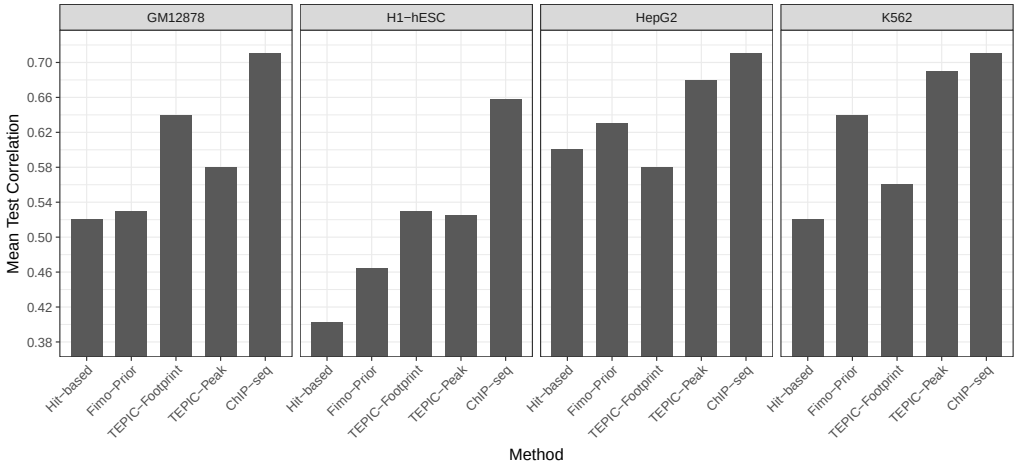


Figure 3.19: The bar plots indicated the mean test correlation for linear models using either a hit-based scoring, FIMO-PRIOR, TEPIC with footprints, TEPIC with peaks, or TF ChIP-seq data. The ChIP-seq based models outperform the ones utilizing predicted TFBS. However on HepG2 and K562, predicted TFBS through TEPIC on peaks achieve almost the same correlation, underlying the quality of our models. Figure from Schmidt *et al.* [S$^+$17a].

understood [K$^+$19]. Furthermore, the position of TF-complexes in the genome, which can be described by TF ChIP-seq data can not be easily predicted using computational models, as there is no direct interaction of all proteins with the DNA and thus there is no detectable binding motif.

3.4.3 Robustness of TF-gene scores derived from ChIP-seq and predicted TFBS in gene-expression models

As stated above, the main purpose of the gene-expression models is to deduce insights on important regulatory factors. Therefore, the models should be robust and lead to reproducible results. Especially correlation between different features was reported to effect model interpretability and might lead to wrong conclusions about the biological questions at hand [B$^+$18a].

Indeed, it has been shown that chromatin accessibility data [M$^+$12b], HM abundance and TF-binding data [B$^+$15a] are similarly predictive for gene-expression, arguing for the presence of shared information between the biological signatures, even across different biological experiments. For TF ChIP-seq data, redundancy has been reported between individual TFs [R$^+$11, Y$^+$13, R$^+$15a, D$^+$16a].

Partially this redundancy might be related to the wealth of *"known biases influencing various chromatin profiling experiments, e.g. the so called expression bias of ChIP-seq data [P$^+$13a], ChIP-seq antibody quality, PCR amplification biases, sequencing depth, or outlier samples. Those biases have been investigated in detail*

and methods have been suggested to account for them [K+11, D+12a, Y+14, R+15a, G+16b, W+17]." [SS18].

We realized that none of these approaches investigated data from a gene-centric perspective. Therefore these methods do not account for biases introduced through data integration. In Schmidt *et al.* [SS18], we reviewed confounders in modeling TF-gene scores from both TF ChIP-seq as well as chromatin accessibility data and studied their effect on gene-expression prediction and on the interpretation of the models.

The material presented in this section is published in the manuscript `On the problem of confounders in modelling gene-expression` [SS18]. Results dealing with the performance of linear models are postulated in terms of Spearman correlation. In the Supplement of our manuscript, also Pearson correlation and MSE are provided. As the conclusions are invariant for all performance measures, we do not show all three, but stick to Spearman correlation for brevity.

Row-wise permutation of the feature matrix

To test whether the input data for the gene-expression models contains a systematic bias, we permuted the original feature matrix X_o per gene that is per row and obtained a randomized matrix X_r. This was suggested before in Bessiere *et al.* [B+18a]. Confounding factors that affect all TF-gene scores for one gene are retained by a per-gene permutation. However, TF specific confounders are removed. The randomized matrix X_r is used as input for the linear regression model throughout this chapter, whenever we refer to permuted input data or permuted features. A graphical illustration of the permutation is shown in Figure B.1.

Multicollinearity in TF ChIP-seq data

Several studied showed TF ChIP-seq data is well suited to predict gene-expression using *in silico* models [O+09, R+15a]. However, Bessiere *et al.* observed that per-gene permuted TF-gene scores derived form TF ChIP-seq data have almost the same predictive power than the original data [B+18a]. However, Bessiere *et al.* did not try to elucidate the reason of that behaviour.

In Schmidt *et al.* [SS18], we tried to reproduce their findings by performing a similar experiment and trained linear regression models using elastic net regularization to predict gene-expression in four different cell lines. Specifically, we obtained ENCODE TF ChIP-seq data for K562, HepG2, GM12878 and H1-hESC cells (see Section B.1 for ENCODE accession IDs). We found that the performance of models based on randomized input is significantly worse compared to the original data (Figure 3.20a). However, the absolute value of the performance measure is not indicating that the randomized models are indeed based on an erroneous data set. Therefore, we hypothesized that *"the presence of any TF ChIP-seq peak in the vicinity of a gene is predictive for gene-expression"* [SS18]. This hypothesis is backed up by Yan *et al.* [Y+13]. They showed that a majority of TFs tend to bind in dense clusters throughout the genome. This observation suggests that the TF-gene

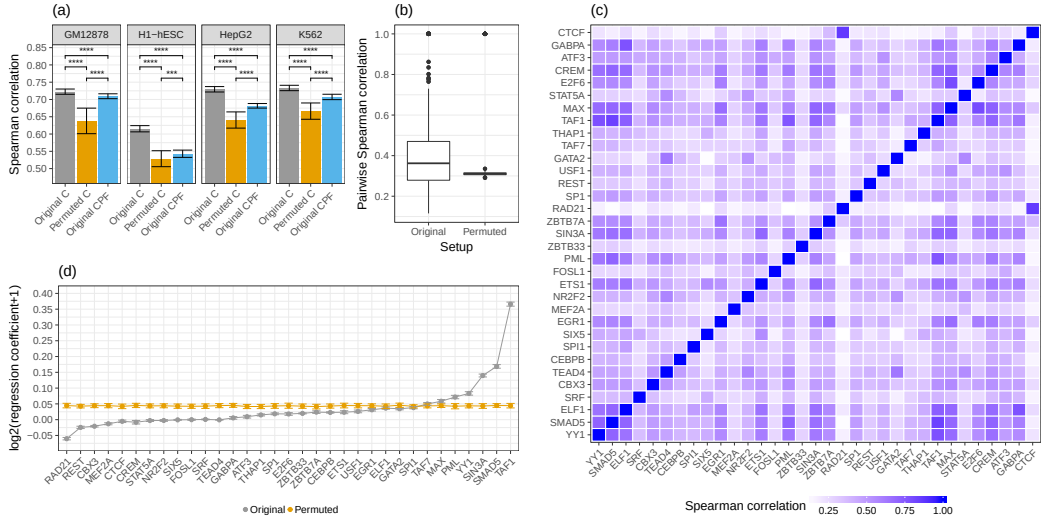


Figure 3.20: (a) The box plots indicate the Spearman correlation achieved by linear regression models using TF-gene scores derived from TF ChIP-seq data for four different cell lines using three different feature setups: $\mathcal{C}$ scores, permuted $\mathcal{C}$ scores and $\mathcal{CPF}$ scores. (b) The distribution of pairwise Spearman correlations of TF-gene scores for 33 TFs for K562 is shown for original and permuted $\mathcal{C}$ scores. (c) A detailed inspection of these pairwise correlations is provided for the original $\mathcal{C}$ scores in a heatmap. (d) The regression coefficients for $\mathcal{C}$ scores and permuted $\mathcal{C}$ scores are visualized here. Statistical significance in (a) is computed with a Wilcoxon test, where **** refers to a significance level of 10^{-4}. Figure from Schmidt *et al.* [SS18].

score vector for an actively regulated gene is not sparse, but is composed of many non-zero values, which might cause the scores to be exchangeable without a strong reduction in model performance. To test whether this hypothesis is plausible, we used the $\mathcal{CPF}$ scoring, considering only peak count and peak length per gene. We found that $\mathcal{CPF}$ models perform worse than the original $\mathcal{C}$ models, but at the same time show better performance than the permuted $\mathcal{C}$ models, thereby supporting our hypothesis (Figure 3.20a).

To better understand this finding, we determined the pairwise Spearman correlation between all TF ChIP-seq scores for 33 TFs in K562 cells, shown in Figure 3.20b. As indicated, the median correlation between original scores (0.362) is only marginally higher than the correlation on the randomized version (0.311). This explains why the permuted $\mathcal{C}$ scores and also the $\mathcal{CPF}$ scores achieve good model performance compared to original $\mathcal{C}$ scores. The high pairwise correlation of the TFs renders a large portion of the input features to be exchangeable. However, permuting the original input matrix causes the correlation between a few originally highly correlated TF pairs to be diminished. To learn whether these high values point out to biologically meaningful links between the factors, we inspected all pairwise

correlations in detail using a heatmap, depicted in Figure 3.20c. We could verify using the literature that some highly correlated TFs are indeed known interaction partners. For instance, CTCF is known to interact with RAD21 [G$^+$14b] (Spearman correlation: 0.862). Furthermore, GABPA and ELF1 are both belonging to the ETS TF-family [Sha01] (Spearman correlation: 0.776).

With respect to the non-robustness of gene-expression models based on TF ChIP-seq data in randomization experiments, Bessiere *et al.* raised concerns that such model might be misinterpreted easily [B$^+$18a]. To address this concern, we inspected the regression coefficients obtained by the linear models in detail. We observe that regression coefficients determined for original $\mathcal{C}$ scores are spread over a wide range of values with a standard deviation of 0.056. Unlike that, regression coefficients determined for permuted $\mathcal{C}$ scores have a similar value across all factors with a standard deviation of 0.0053 (Figure 3.20d). As the regression coefficients are stably selected in this fashion and are thus not interpretable at all in case of the randomized data, a wrong interpretation is unlikely. Despite the high pairwise correlation of the original $\mathcal{C}$ scores, only regression coefficients deduced from them can be meaningfully interpreted. For instance, TAF1, an indispensable TF to initiate transcription [B$^+$14a], has the highest regression coefficient.

Characterisation of the correlation

As stated above, also the $\mathcal{CPF}$ scores (Eq.3.31, 3.32) achieve a reasonable performance (Figure 3.20a). This puts forward that *"aggregating ChIP-seq data across several TFs resembles a measure of regulatory activity, which itself is highly predictive for gene-expression."* [SS18]. We associated this hypothesis to studies by Ramachandran *et al.* [R$^+$15a]. They trained gene-expression prediction models using only single TF ChIP-seq experiments as input and compared those to models based only on DNaseI-seq data. They found that only a handful of factors, e.g. TAF1 or POL2, are highly predictive for gene-expression. Furthermore, they suggest that chromatin accessibility data can substitute ChIP-seq data for all other TFs.

We followed up on this idea by computing the overlap of ChIP-seq peaks to DHSs in HepG2, K562, GM12878, as well as in H1-hESCs considering two cases. First, we considered all ChIP-seq peaks throughout the genome and secondly only ChIP-seq peaks located in a 50kb window around the 5′TSS of all protein coding genes. As shown in Figure 3.21, 71% of all genome-wide ChIP-peaks overlap a DHSs and even 81% of all ChIP-seq peaks around the 5′TSS of protein coding genes are overlapping a DHSs. These results indicate that the pure presence of a peak can be seen as an equivalent to the presence of a DHS site, arguing for the interchangeability of TF ChIP-seq data as well as its usage in an aggregated fashion.

Attempting to adjust for the correlation

By accounting for the number of ChIP-seq peaks around a genes 5′TSS, we tried to improve the robustness of ChIP-seq derived TF-gene scores against permutation.

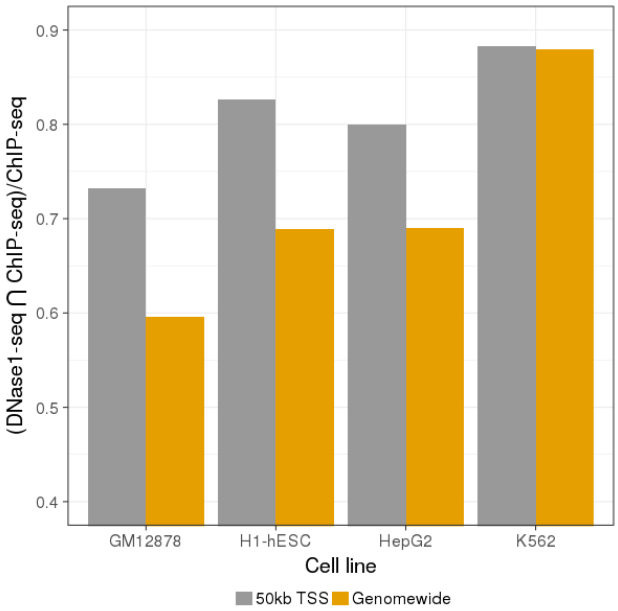


Figure 3.21: The fraction of ChIP-seq peaks that overlap a DNase peak is shown for all genome-wide ChIP-seq peaks and for all ChIP-seq peaks in a 50kb window around the 5′TSS of protein coding genes. Figure from Schmidt *et al.* [SS18].

This is implemented in the $\mathcal{CN}$ (Eq. 3.30) scoring. The observation that the c^C feature, which represents the number of peaks, has a strong positive regression coefficient in $\mathcal{CPF}$ models (Figure 3.22) motivates the novel score as it implies that this quantity itself covers a large portion of the information contained in TF ChIP-seq data. The value of c^C is high if there are many TF ChIP-seq peaks within the considered window and these peaks are close to the 5′TSS of the considered gene (Figure 3.8). Thus, normalizing by c^C leads to a general depletion of TF-gene scores if many ChIP-seq peaks are present around a gene and simultaneously increases TF-gene scores if there are only a few peaks located in the considered window. Intuitively, $\mathcal{CN}$ scores strengthen individual peaks and weaken the importance of peaks occurring in dense clusters. As depicted in Figure 3.23a using $\mathcal{CN}$ scores performs as expected and results in a significant reduction of model performance on permuted input compared to permuted $\mathcal{C}$ scores. Therefore, we believe that $\mathcal{CN}$ scores are more robust against permutations than $\mathcal{C}$ scores, because essential information is lost through the permutation. To our surprise, we found that $\mathcal{CN}$ scores also caused a significant performance increase on original data for three out of four cell lines (Figure 3.23a).

According to a Wilcoxon test, the normalization implemented in the $\mathcal{CN}$ scores reduced the pairwise correlation between TF-gene scores significantly for both original and permuted data (Figure 3.23b). From an hands-on perspective, these results imply that model performance and the pairwise correlation among TF-gene scores

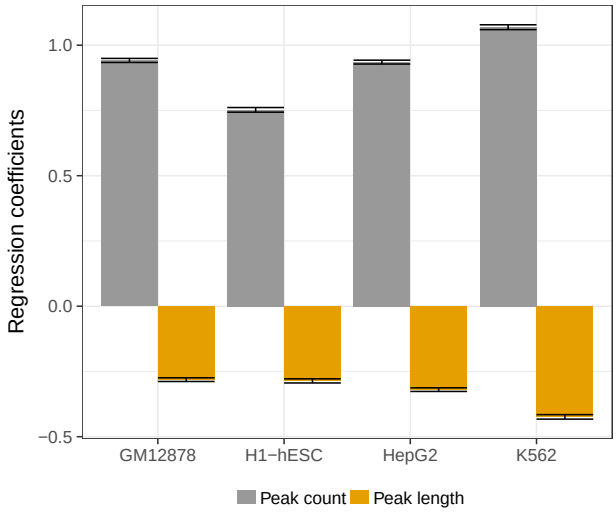


Figure 3.22: The regression coefficients of Peak count c^C and Peak length l^C in models using only peak features derived from TF ChIP-seq data of GM12878, H1-hESC, HepG2 and K562 is shown. Figure from Schmidt *et al.* [SS18].

computed for $\mathcal{CN}$ scores are more suitable than those computed for $\mathcal{C}$ scores to spot errors occurring during data handling or processing.

As above, we closely investigated the values of the pairwise correlations. We found that the normalization introduced a negative correlation between several TFs (Figure 3.23c), for instance, between TAF1 and CTCF (-0.282). The negative correlation between these two factors has been reported previously in the literature [K$^+$07]. With the original $\mathcal{C}$ scores, this TF-pair achieved a correlation of (0.181). The difference suggests that the $\mathcal{CN}$ scores seem to improve the modeling of interactions among TFs. As a consequence of the altered correlation between the TF features, the regression coefficients for some TFs are altered as well (Figure 3.23d). Only using $\mathcal{CN}$ scores, several TFs that are known to be repressors such as E2F6 [G$^+$04], REST [B$^+$06c] and EGR1 [A$^+$08] also obtain a negative regression coefficient.

Open chromatin characteristics impact predicted TFBS and TF-gene scores

Although ChIP-seq experiments deliver genome-wide insights into *in vivo* TF-binding, it is infeasible to obtain ChIP-seq data for all TFs in all tissues. Therefore, predicting TFBS using chromatin accessibility data became a standard procedure in the field. Hence, we additionally examined confounders affecting TF-gene scores based on predicted TFBS.

We trained linear regression models with elastic net regularization to predict gene-expression for seven distinct DNaseI-seq samples using Tepic TF-gene scores computed according to the $\mathcal{E}$ setup (Eq. 3.23). As reported before [B$^+$18a], we also

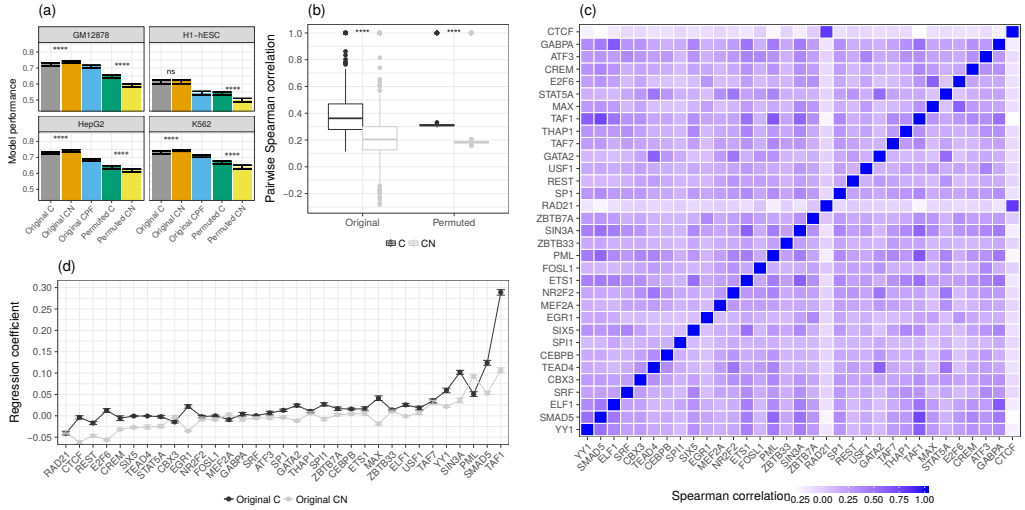


Figure 3.23: (a) The model performance of linear regression models predicting gene-expression based on original ($\mathcal{C}$) scores is compared to the normalized ($\mathcal{CN}$) scores. (b) The boxplots indicate the pairwise Spearman correlation of TF-gene scores based on TF ChIP-seq data computed for 33 TF ChIP-seq assays in K562 for $\mathcal{C}$ and $\mathcal{CN}$. (c) The heatmap shows the pairwise correlation between 33 TFs for $\mathcal{CN}$ scores. (d) The regression coefficients inferred for $\mathcal{C}$ and $\mathcal{CN}$ scores are contrasted. Statistical significance in (a) and (b) is computed with a Wilcoxon test, where **** refers to a significance level of 10^{-4}. Figure from Schmidt *et al.* [SS18].

find that model performance drops marginally on randomized input (Figure 3.24a) and thus renders performance to be inadequate to judge model reliability. We contrasted the performance of a model considering only peak count and peak length per gene as features ($\mathcal{EPF}$) (Eq. 3.26, 3.27) against a model using the full feature matrix ($\mathcal{E}$) to test whether chromatin-accessibility data itself might be a confounder that is inherently contained in TF-gene scores. As it could have been expected from the ChIP-seq experiments, also $\mathcal{EPF}$ models considering DHS based features show good performance (Figure 3.24b). Similar observations were made for the $\mathcal{ES}$ setup (Eq. 3.24). These are omitted here for brevity and are provided in Schmidt *et al.* [SS18]. As noted by others [M$^+$12b, R$^+$15a], this shows that chromatin-accessibility itself is predictive for gene-expression. Furthermore, it also supports the idea that TF-gene scores might be linked to chromatin specific features such as peak count and peak length.

To follow up on that hypothesis, we calculated the pairwise correlation between all TF-gene scores across all genes within each DNaseI-seq sample. As shown in Figure 3.24c, several TFs are highly correlated, especially TFs with a similar binding motif such as HEY1 and CLOCK, or TEAD1, TEAD3 and TEAD4. Correlation that is due to similar sequence preferences between TFs would be lost in a per-gene

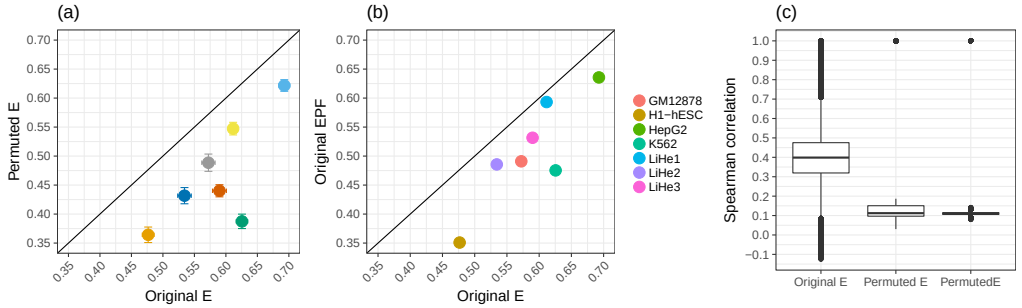


Figure 3.24: (a) Spearman correlation values of linear regression models predicting gene-expression using TEPIC scores ($\mathcal{E}$) as input compared against permuted ($\mathcal{E}$) scores. (b) $\mathcal{E}$ scores are compared against a $\mathcal{EPF}$ scores considering peak length and peak count as features. (c) Boxplots showing the pairwise Spearman correlation between TF-gene scores, for both original and permuted data across all DNaseI-seq samples using the $\mathcal{E}$ setup. Figure from Schmidt *et al.* [SS18].

randomization. However, correlation that is caused by confounders affecting each gene should not be removed by a per-gene randomization. Therefore, the remaining correlation on permuted data, which is shown in Figure 3.24c, is probably due to confounding variables representing chromatin context. As shown in Figure 3.25a peak length, peak count and peak signal are indeed highly correlated to TF affinities. An example is shown in Figure 3.25b and c. It illustrates the correlation between TF-gene scores of HOXA3 and peak length l^E (0.9568) as well as peak count c^E (0.6786), respectively.

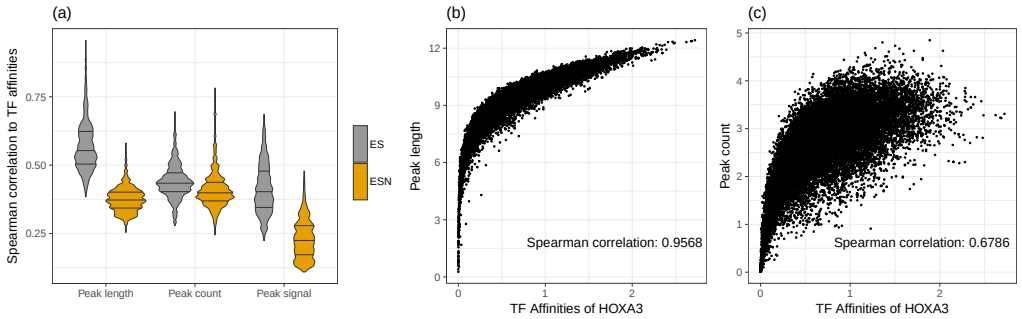


Figure 3.25: (a) The pairwise correlation of $\mathcal{ES}$ and $\mathcal{ESN}$ scores against peak length l^E, peak count c^E and peak signal f^E is shown. In (b) the correlation of TF affinities for HOXA3 using the $\mathcal{E}$ setup are plotted against peak length l^E, whereas in (c) they are plotted against peak count c^E. Figure from Schmidt *et al.* [SS18].

Adjusting for the confounding effects

In the computation of TF affinities following the $\mathcal{E}$ and $\mathcal{ES}$ setup, the contributions of all possible TFBS within a DHS are considered in the final score. Therefore, the length of the DHSs is indirectly incorporated into TF-gene scores. We attempt to account for this by normalizing TF affinities per DHS in $\mathcal{EN}$ scores (Eq. 3.25). Per DHS, we divide the TF affinities by the number of possible binding sites $|p|-|m|+1$, where $|p|$ is the length of the region p and $|m|$ is the length of the current binding motif. We apply the same normalization to the $\mathcal{ES}$ setup leading to $\mathcal{ESN}$ scores. Additionally, we consider the DNaseI-seq signal as an extra feature, instead of multiplying it with the TF affinities as performed in the $\mathcal{ES}$ setup.

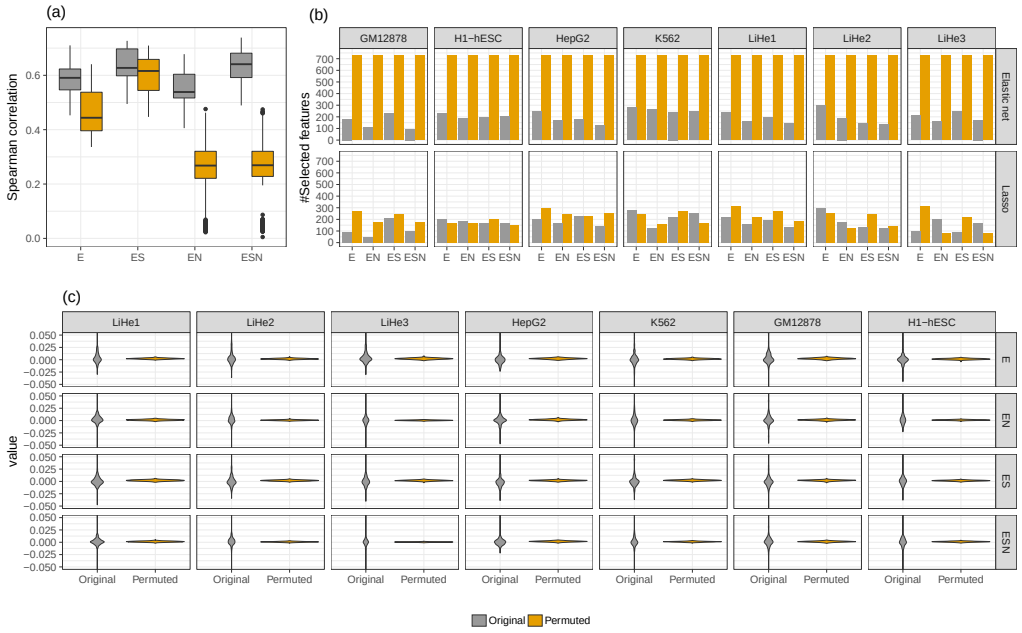


Figure 3.26: (a) Illustration of the performance of gene-expression models based on four different annotation setups ($\mathcal{E}$, $\mathcal{ES}$, $\mathcal{EN}$, $\mathcal{ESN}$) for original and permuted data. (b) Here the number of selected features is shown for all annotation variants in original and permuted data using elastic net or lasso regularization. In (c), the range of regression coefficients per sample inferred from permuted and not permuted data for TF-gene scores computed using the $\mathcal{E}$, $\mathcal{EN}$, $\mathcal{ES}$ and $\mathcal{ESN}$ setup and elastic net regularization are depicted. Figure from Schmidt *et al.* [SS18].

Using $\mathcal{EN}$ and $\mathcal{ESN}$ TF-gene scores as input to the linear models shows that the performance compared to $\mathcal{E}$ and $\mathcal{ES}$ scores on original, not permuted data, changed only marginally. As expected, we observe a significant drop in model performance for permuted data, achieving a median Spearman correlation of 0.268 and 0.269, respectively (3.26a). This observation is invariant to the used regularization method,

we observed it for both elastic net and lasso regularization. The drop in model performance is linked to a reduction of the correlation between TF-gene scores and chromatin-accessibility features using $\mathcal{EN}$ and $\mathcal{ESN}$ scores. For instance, as outlined in Figure 3.25a, the correlation between TF-gene scores for HOXA3 in LiHe1 and peak length could be decreased from 0.9568 to 0.5808.

Impact of the regularization on judging model performance

In addition, we looked at the number of non-zero features derived on permuted and not permuted data observing the general trend $||\beta||_0^D \geq ||\beta||_0^{DS} \geq ||\beta||_0^{DN} \geq ||\beta||_0^{DSN}$ (Figure 3.26b). We made the striking observation that elastic net constantly selects all features in each annotation setup on permuted data, while lasso selects only a few representative ones. Due to the grouping effect, the elastic net considers all predictors and assigns them similar regression coefficients if the predictors are part of a group of highly correlated features [HH05]. Using permuted $\mathcal{C}$, $\mathcal{E}$ and $\mathcal{ES}$ scores, all contained features form one group of correlated predictors with similar pairwise correlations. Therefore, elastic net selects all of them and assigns them similar regression coefficients that are close to zero (Figure 3.26c).

Model evaluation using a gold-standard set of gene regulation in primary human hepatocytes

In order to get an impression on the quality and reliability of the TF predictions depending on the different TF-gene score systems, we compared TFs selected by the linear models as tissue-specific regulators in three primary human hepatocyte samples from DEEP against a manually curated gold-standard ($\mathcal{GS}$) set, To avoid any biases existing in a literature curated $\mathcal{GS}$, we considered all TFs that are expressed by at least 5 Transcripts Per Million (TPM) in liver RNA-seq expression data obtained from the Human Protein Atlas [U$^+$15]. Additionally, the TFs need to be included in our TF motif collection. With respect to these two constraints, we obtained a gold-standard set comprising 200 TFs (c.f. Section B.1.12). To evaluate the predictions, we refer to the area under the precision-recall (AUPR) curve computed using the PRROC package [G$^+$15c]. A TP is defined as a TF retrieved by the model that is contained in the $\mathcal{GS}$, a FP is a TF that is inferred by the model but is not included in the $\mathcal{GS}$ and a FN is a TF that is listed in the $\mathcal{GS}$ but is not retrieved by the model. In PRROC, TFs are sorted according to their regression coefficients.

As listed in Table 3.4, the absolute number of non-zero features varies across the tested annotation versions and samples, while the AUPR is similar across all annotation setups. However, we note a drop in terms of non-zero features and AUPR for LiHe2 with the $\mathcal{ESN}$ annotation. As mentioned before, DEEP quality control suggested that the DNaseI-seq data for this sample might not be optimal, which could explain the difference to the other primary human hepatocyte replicates. Notably, there is a slight advantage for the original approaches $\mathcal{E}$ and $\mathcal{ES}$ in terms of AUPR.

Table 3.4: Number of selected features and AUPR scores computed in a gold standard comparison of regulatory TFs suggested by linear models using elastic net regularization for primary human hepatocytes.

	#selected features				AUPR			
	E	ES	EN	ESN	E	ES	EN	ESN
LiHe1	274	210	156	143	0.341	0.360	0.333	0.368
LiHe2	301	145	227	107	0.355	0.346	0.347	0.292
LiHe3	193	297	238	160	0.347	0.333	0.311	0.319

The differences in the number of non-zero features might be linked to the correlation between features existing in $\mathcal{E}$ and $\mathcal{ES}$ scores: Keeping in mind that elastic net attempts to find a balance between sparsity and the inclusion of correlated yet predictive features, the number of selected features might be higher in $\mathcal{E}$ and $\mathcal{ES}$ models compared to $\mathcal{EN}$ and $\mathcal{ESN}$ models. Importantly, this analysis did not clearly argue in favor of one of the tested approaches to compute TF-gene scores in terms of biological relevance.

Conclusions for interpreting and handling gene-expression models

Predictive models of gene-expression became prevalent in computational biology. While our analysis showed, similar to the earlier work by Bessiere *et al.* [B$^+$18a] that row-wise permutation of TF ChIP-seq data does not remove the entire signal, an erroneous interpretation of the models is unlikely. Due to the grouping effect of the elastic net, correlated features are down weighted and their values are shrinked towards zero, which, in case of permuted data, leads to small regressions coefficients for all factors. The widely used lasso regularization does not show this helpful characteristic. Therefore, it should be used carefully to avoid wrong conclusions.

The normalized scores for ChIP-seq ($\mathcal{CN}$) and predicted TF-gene scores ($\mathcal{EN}$, $\mathcal{ESN}$) improve model robustness and, in case of $\mathcal{CN}$ scores, even model performance. However, no scoring strategy could completely alleviate the correlation present in TF-gene scores. As illustrated in Figure 3.2, a complete removal of the correlation should not be expected as the correlation is partially due to biological and experimental reasons. For example, ChIP-seq data captures the signal of TFs forming complexes via protein-protein interactions, thereby yielding correlated scores. Furthermore, it is known that TFs tend to bind in clusters [Y$^+$13], ChIP-seq data is very sensitive to that and causes features to be correlated as well. Nevertheless, the correlation could also be of technical nature, for instance, due to similar binding motifs or open chromatin characteristics.

Although we investigated ways how to reduce this correlation, it is inherent to the data and the problem setting and thus, to some extend, unavoidable.

In Schmidt *et al.* [SS18] we stressed this point to make the community aware of the potential pitfalls arising from gene-expression modelling. We demonstrated

how the number of non-zero features and the magnitude of regression coefficients are indicators for model quality and can pinpoint researchers to potential flaws in feature design or data handling. Importantly, our work led to the conclusion that results obtained using approaches like the $\mathcal{C}$ scores [O$^+$09] or $\mathcal{E}$ scores [M$^+$12b] are not necessarily incorrect, but highlighted the complexity of prioritizing meaningful TFs due to confounders investigated here.

3.4.4 Linear models suggest expressed and known transcriptional regulators

Within this section, we delineate how we can use predicted TFBS within our linear models to learn about candidate tissue-specific regulators. The results presented in this section are based on $\mathcal{ES}$ models, published in Schmidt *et al.* [S$^+$17a].

Regression coefficients harbour cell type specificity

To learn about whether our linear models highlight tissue-specific regulators, a Principal Component Analysis (PCA) (Section 2.2.3) was conducted on the regression coefficients vectors of all 13 samples used in Schmidt *et al.* [S$^+$17a]. As depicted in Figure 3.27, the primary human hepatocyte samples (LiHe1-3) are distinctly placed away from the other samples, while according to PC1, HepG2, a human liver cancer cell line, is their next neighbour. The T-cell samples T1-T6 are positioned in the right half of the PCA plot. Their nearest neighbour is GM12878, which is a lymphoblastoid cell line. Because lymphoblasts can differentiate into T-cells, the position of GM12878 in the PCA plot is sensible as well. However, we note that PC1 might also capture an experimental difference between the T-cell and the remaining samples. Chromatin accessibility in all DEEP T-cell samples has been screened using NOME-seq, whereas chromatin accessibility in all other samples was investigated with DNaseI-seq.

Keeping the PCA analysis in mind, we performed a cross-sample comparison using our models. That means, a model has been trained for a distinct sample $s_i \in \mathcal{S}$, where $\mathcal{S}$ is the set of all 13 samples. Next the model learned for s_i is used to predict the expression in all other samples s_j, with $j \neq i$. As shown in Figure 3.28, this experiment argues for a tissue-specificity of our models as well. The dendrograms resemble a clear similarity/dissimilarity between related/unrelated cell types. Therefore, we concluded that it is worthwhile to investigate the feature coefficients in more detail to learn about tissue-specific regulators.

Primary human hepatocytes

To investigate the role of TFs in the primary human hepatocyte samples (LiHe1-3), we computed the overlap between the features with a non-zero regression coefficient using the $\mathcal{ES}$ annotation with a 50kb window, visualized in Figure 3.29a using a Venn diagram. We found that 65 (38.5%) TFs are commonly selected among the three replicates.

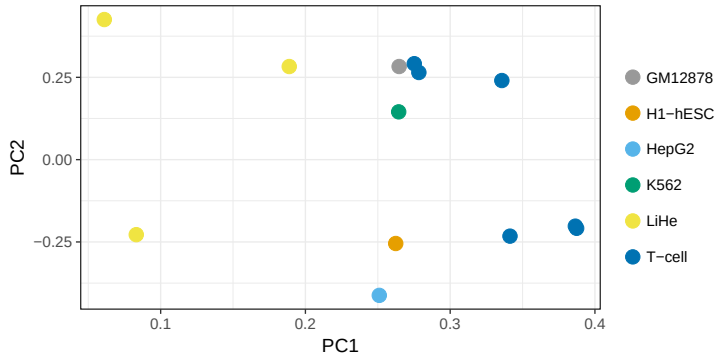


Figure 3.27: PCA of the regression coefficients learned from linear gene-expression models using TEPIC TF-gene scores for 13 datasets. Three primary human hepatocyte sample cluster separately at the left hand-side of PC1, while six T-Cell samples cluster on the right hand side. Cell lines are arranged in the middle of the PCA. Figure from Schmidt *et al.* [S+17a].

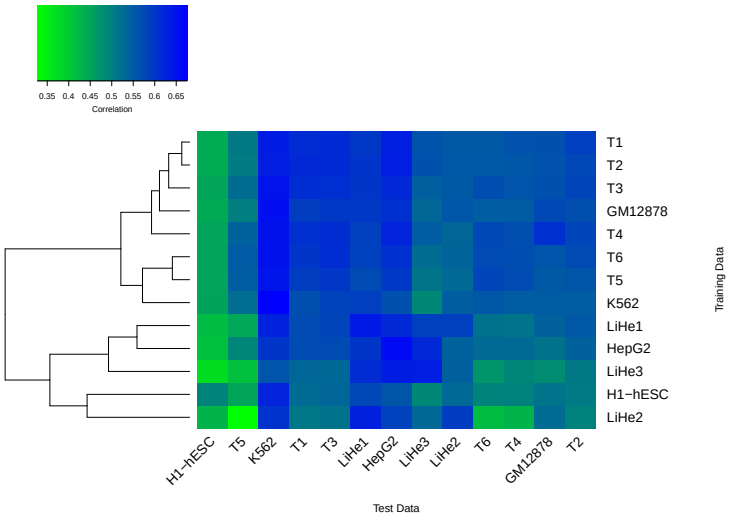


Figure 3.28: Clustered heatmap, using euclidean distance, showing the Pearson correlation values obtained by a between sample learning experiment. They y-axis shows the data used for training, the x-axis shows the data used for testing. As indicated by the dendrogram, the coefficients behave similarly across related samples. Figure from Schmidt *et al.* [S+17a].

In Figure 3.29b, we show the top 10 positive and top 10 negative features selected by our models ranked by the mean regression coefficients across the replicates.

By searching the literature we retrieved evidence for 52 of the 65 factors to be known to have a function in hepatocytes. All sources are listed in the Supplementary

Material of Schmidt *et al.* [S+17a]. Within the top 10 positive and negative features, we found for example, the hetero-dimer PPARG::RXRA. This complex is known to have a key role in hepatic transcription [Q+00]. Another TF is CEBPA, which is known to be involved in liver regeneration [Die98, C+03]. The TF GATA4 has been shown to be involved in liver induction [B+16c]. Similarly CTCF was found to have a role in imprinting liver [HdL13, G+12b] and NRF1 has been shown to possess a protective function against oxidative stress in liver [X+05].

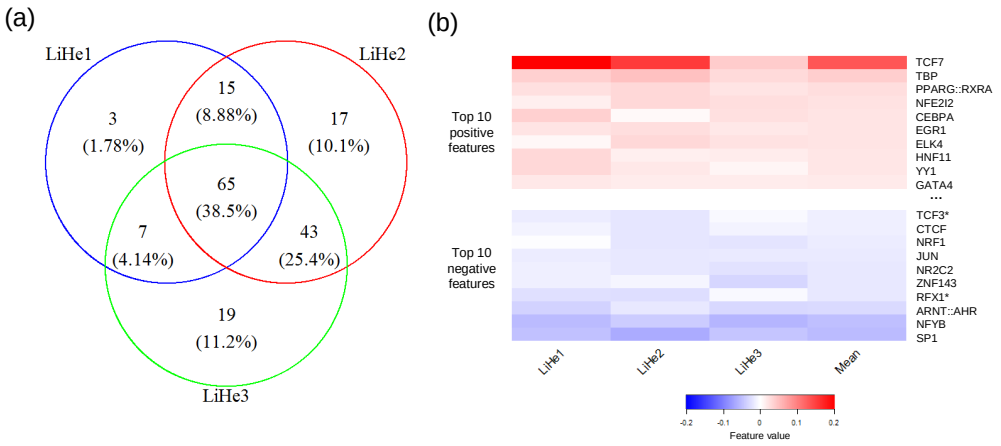


Figure 3.29: (a) Overlap of non-zero regression coefficients determined for the primary human hepatocyte models (LiHe1-3). The top 10 positive and negative features among the 65 shared ones ranked by the mean regression coefficient are shown in (b). TFs labelled with a * could not be related to hepatocytes by literature research. Figure from Schmidt *et al.* [S+17a].

CD4+ T-cells

As for primary human hepatocyte samples, we have computed the commonly selected features across all six CD4+ T-cell samples. In total, there are 53 (39%) TFs commonly selected. The overlap between the individual T-cell replicates is shown in Figure 3.30a. We suggest that those 53 TFs are potential key regulators within CD4+ T-cells. In literature, we found evidence that 42 out of the 53 suggested regulators are known to be related to the immune system, see the Supplementary Material of Schmidt *et al.* [S+17a] for the complete list. For instance, among the top 10 positive and negative coefficients (Figure 3.30b) we found the factor GMEB1, which was shown to inhibit T-cell apoptosis [Ko12]. Another TF with a positive coefficient is ETS1, which is known to be essential for T-cell development [E+04b]. In agreement with its negative regression coefficient, the TF ZBTB7B, is an established repressor in CD4+ T-cells [W+08].

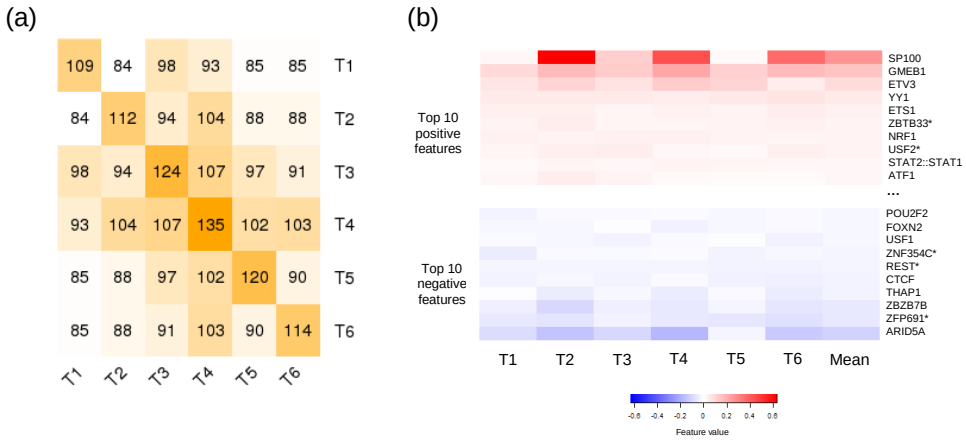


Figure 3.30: The pairwise overlap of non-zero regression coefficients learned for DEEP T-cell samples is shown in (a). In total 53 TFs are common between all six samples. The top 10 positive and negative coefficients are shown in (b), ranked by the mean regression coefficient. TFs labeleld with a * could not be linked to T-cells by literature search. Figure from Schmidt *et al.* [S+17a].

TF expression and its relation to model performance

Another sanity check for the selected features is to check how many of the selected TFs that is TFs with a non-zero regression coefficient, are expressed in the analyzed samples. We found that the mean expression level of selected TFs is higher than the mean expression level of the ones that have not been selected (Figure 3.31, Figure 3.32). Similarly, in Schmidt *et al.* [SS18], we have shown that the expression of TFs highlighted by models from permuted data (c.f. Section 3.4.3) is significantly lower than that of TFs inferred by original models.

These insights not only suggests that the model selects meaningful regulators, but they also point out that it might be possible to reduce the feature space by removing factors that are not expressed, without a loss in model performance. Therefore, we have repeated the gene-expression learning with an expression filtered set of TFs using a FPKM cut-off of 1.0 and additionally removed all TFs that could not be mapped to a ENSEMBLE gene ID. Figure 3.33 shows that this reduction of TFs does not significantly affect model performance. However, due to the reduced feature space, it can help to simplify model interpretation and speeds model fitting as well.

3.4.5 Integration of conformation capture data into TF-gene scores

The window based exponential decay formulation explained in Section 3.3.2 is not able to capture long Promoter-Enhancer-Interactions (PEIs), mediated e.g. by DNA-looping, as delineated in Section 2.1.13.

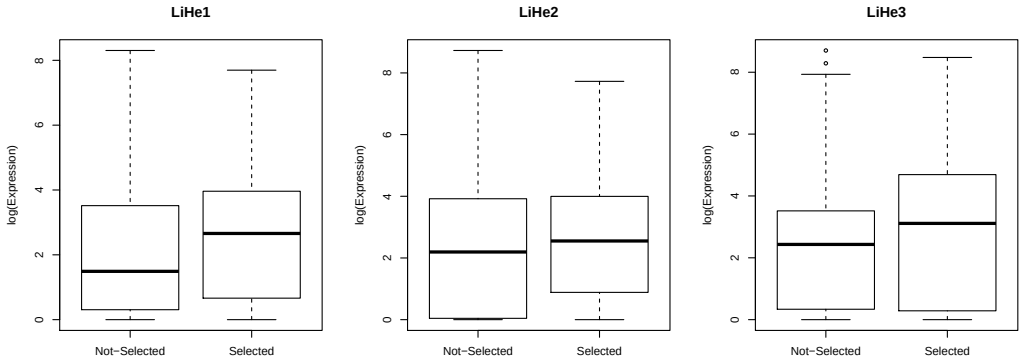


Figure 3.31: Expression [log2(FPKM+1)] of TFs with a non-zero regression coefficient (selected) vs not selected factors in the three primary human hepatocyte samples from DEEP. Figure from Schmidt *et al.* [S$^+$17a].

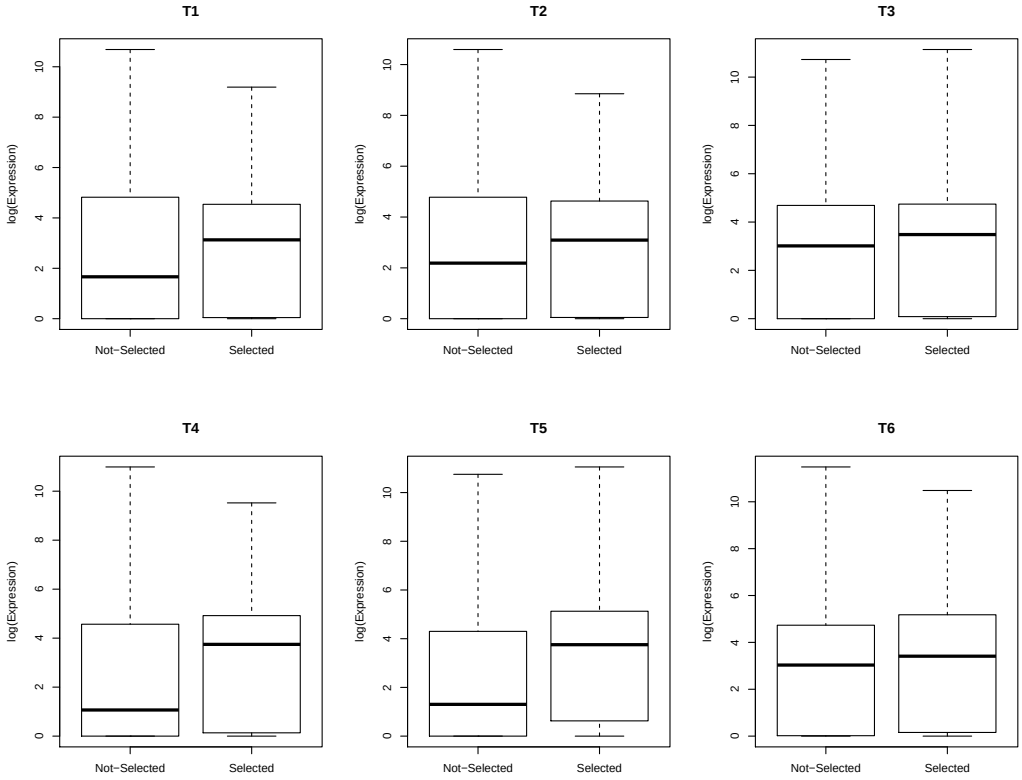


Figure 3.32: Expression [log2(FPKM+1)] of TFs with a non-zero regression coefficient (selected) vs not selected factors in the six T-cell samples from DEEP. Figure from Schmidt *et al.* [S$^+$17a].

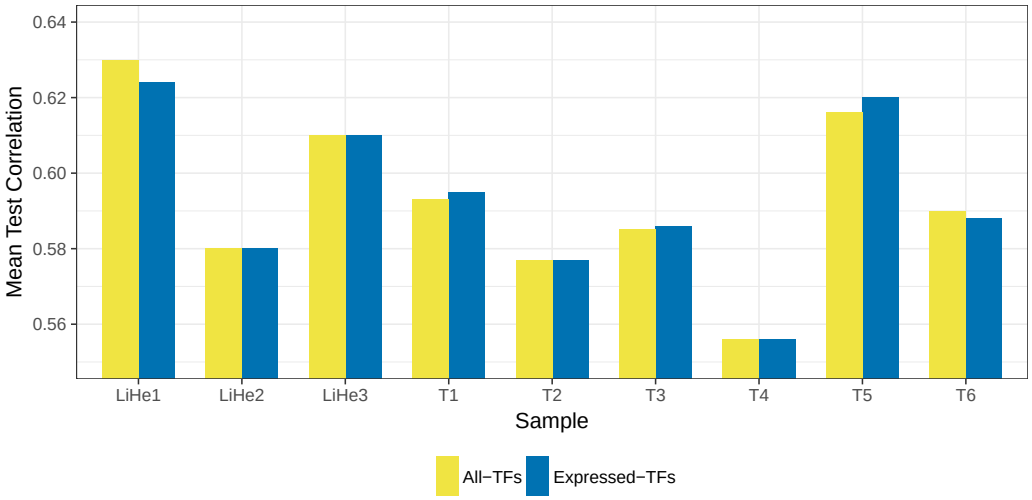


Figure 3.33: Mean test correlation of models if TFs that are not expressed by at least one FPKM are removed from the features space. As it can be seen in the figure, the reduction of the feature space does not negatively affect model performance. Figure from Schmidt *et al.* [S+17a].

We extended our *TEPIC* framework to account for regulatory elements derived from chromatin conformation-capture experiments, specifically form Hi-C data. Experimental details on Hi-C are delineated in Section 2.1.11.

To avoid additional complexity introduced by considering individual TF binding events and to test whether the inclusion of Hi-C data into our gene-expression models is beneficial at all, we decided to model gene-expression only from chromatin accessibility or aggregated ChIP-seq data in a first step, similar to the models $\mathcal{EPF}$ and $\mathcal{CPF}$. Furthermore, we include chromatin state segmentations from ChromHMM into our modelling, attempting to further refine the considered genomic regions. Details on data processing are provided in Section B.1. The material presented in this section is joint work with Fabian Kern extending his bachelor thesis [Ker16]. It was presented in an oral presentation at GCB 2018 in Vienna.

Model extension

We compared a window based annotation that incorporates Hi-C data against a local window based linkage, as introduced in Section 3.3.2. As before, candidate regulatory sites are derived from DHS as well as TF ChIP-seq peaks and additionally from a chromatin state segmentation computed with ChromHMM [EK12].

As before, we use the following notation: For a DHS site $d \in \mathcal{D}$, where $\mathcal{D}$ is the set of all DHS sites, we denote the length of d with $l(d)$ and DNaseI-seq signal in d with $s(d)$. For a distinct ChIP-seq peak $c_t \in \mathcal{C}_t$, where $\mathcal{C}_t$ refers to all ChIP-seq peaks for TF t, we denote its length with $l(c_t)$ and the peak score with $s(c_t)$. The set of all chipped TFs is denoted with $\mathcal{T}$.

We obtained chromatin state segmentations for several cell lines: GM12878, HeLa, HUVEC, IMR90 and K562 from ENCODE (see Section B.1 for details). Each segmentation contains 15 states generated with ChromHMM [EK12]. From those 15 states, we extract two promoter states, specifically `TssA` (1) and `TssAFlnk` (2), as well as three enhancer states, namely `EnhG`(6), `Enh`(7) and `EnhBiv`(12). Neighbouring segments of the same type are merged into one segment h, representing a distinct ChromHMM state. The set of all considered segments for gene g is denoted with $\mathcal{H}_g$. The length of a distinct segment $h \in \mathcal{H}_g$ is denoted with $l(h)$ and the DNaseI-seq/aggregated TF ChIP-seq signal within h is denoted with $s(H)$. Accordingly, we define three ChromHMM features pl_g^h, pc_g^h and ps_g^h as:

$$pl_g^h = \sum_{h \in \mathcal{H}_g} l(h) e^{-\frac{\text{dist}(h,g)}{d_0}}, \tag{3.36}$$

$$pc_g^h = \sum_{h \in \mathcal{H}_g} e^{-\frac{\text{dist}(h,g)}{d_0}}, \tag{3.37}$$

$$ps_g^h = \sum_{h \in \mathcal{H}_g} s(h) e^{-\frac{\text{dist}(h,g)}{d_0}}, \tag{3.38}$$

where pl_g^h is the length of the considered ChromHMM segments for g, pc_g^h are the distance weighted counts and ps_g^h is the distance weighted, aggregated epigenetic ChIP-seq or DNaseI-seq signal within the segments. Note that in the ChIP-seq case $s(h)$ aggregates the ChIP-seq signal across all available TFs, neglecting TF specificity. In this way, the ChIP-seq signal can be easily compared against the chromatin accessibility signal deduced from DNaseI-seq.

Furthermore, we define an intersection operation $\cap_{\mathcal{H}}$ between $\mathcal{D}_{g,w}$ or $\mathcal{C}_{g,w}$ and $\mathcal{H}_{g,w}$ such that only $d \in \mathcal{D}_{g,w}$ or $c \in \mathcal{C}_{g,w}$ are retained that overlap by at least one 1bp with any $h \in \mathcal{H}_{g,w}$. The variable w indicates the width of the window used for a window based aggregation of regulatory elements to the gene level. Formally that is:

$$\mathcal{X}_{g,w} \cap_{\mathcal{H}} \mathcal{H}_{g,w} = \{x | x \in X_{g,w} \ \wedge \ \exists h \in \mathcal{H}_{g,w} : x \cap m \neq \emptyset\}, \tag{3.39}$$

where $\mathcal{X}_{g,w} = \mathcal{D}_{g,w}$ or $\mathcal{X}_{g,w} = \mathcal{C}_{g,w}$ and $x \cap h$ indicates the overlap in genomic space of peak x and segment h.

We apply the $\cap_H$ intersection operation to $\mathcal{D}_{g,w}$ and $\mathcal{C}_{g,w}$, thereby removing regions that do not overlap with either an enhancer or promoter state from $\mathcal{H}_{g,w}$, obtaining the reduced sets $\mathcal{D}'_{g,w}$ and $\mathcal{C}'_{g,w}$, respectively:

$$\mathcal{D}'_{g,w} = \mathcal{D}_{g,w} \cap_{\mathcal{H}} \mathcal{H}_{g,w}, \tag{3.40}$$

$$\mathcal{C}'_{g,w} = \mathcal{C}_{g,w} \cap_{\mathcal{H}} \mathcal{H}_{g,w}. \tag{3.41}$$

Because the window based linkage of genes achieved a better model performance than the nearest-gene linkage in gene-expression experiments presented in Section 3.4.2, we have extended the window based linkage to include chromatin interactions derived from Hi-C data. Specifically, we used the loop files as provided

by the Lieberman-Aiden group [R+14b] for GM12878, HeLa, HUVEC, IMR90 and K562. The loop files contain, for each sample, the extracted Hi-C loops with a specific resolution. In case of the Hi-C datasets used in this work the loops are of 5kb, 10kb and 25kb resolution, respectively.

Recall from Section 2.1.11 that a loop is defined as a pair of genomic loci that are in arbitrary genomic distance from each other, but, at the same time, are in close spatial proximity. The Hi-C resolution defines the number of base pairs to which both genomic loci can be pin-pointed in the genome. Therefore, the better the resolution of the Hi-C experiment (numerically smaller), the shorter is the region around the true interaction site in each locus. Here, the Hi-C resolution called *All* refers to loops of an arbitrary resolution, a more conservative approach where we consider all available loops. For reasons of simplicity, less frequent inter-chromosomal loops, are excluded.

The loops are modelled by considering an additional window v inferred from contacts of a Hi-C experiment (equations 3.40-3.48). We link a Hi-C contact to g if one of its loop regions is located within a 50kb radius of g. All DHS sites p, ChIP-seq peaks c and ChromHMM regions h within v, denoted with $\mathcal{D}_{g,v}$, $\mathcal{C}_{g,v}$ and $\mathcal{H}_{g,v}$, respectively, are included in the score computation. The set $\mathcal{C}_{g,t,v}$ denotes all ChIP-seq peaks for TF t that are associated with gene g via the additional window v. Because the Hi-C experiment suggests a direct interaction of the potentially far away region v with gene g, we do not apply an exponential decay to peak signals of that region. However, we did test whether applying the exponential decay in the Hi-C regions would be benefital for model performance and found that it is indeed not the case (results not shown).

The updated Hi-C formulas for DNaseI-seq data are:

$$pl_g^d = \sum_{d \in \mathcal{D}_{g,w}} l(d) e^{-\frac{\text{dist}(d,g)}{d_0}} + \sum_{d \in \mathcal{D}_{g,v}} l(d), \tag{3.42}$$

$$pc_g^d = \sum_{d \in \mathcal{D}_{g,w}} e^{-\frac{\text{dist}(d,g)}{d_0}} + |\mathcal{D}_{g,v}|, \tag{3.43}$$

$$ps_g^d = \sum_{d \in \mathcal{D}_{g,w}} s(d) e^{-\frac{\text{dist}(d,g)}{d_0}} + \sum_{d \in \mathcal{D}_{g,v}} s(d), \tag{3.44}$$

for ChIP-seq data:

$$pl_g^c = \sum_{t \in \mathcal{T}} \left(\sum_{c_t \in \mathcal{C}_{g,w}} l(c_t) e^{-\frac{\text{dist}(c_t,g)}{d_0}} + \sum_{c_t \in \mathcal{C}_{g,v}} l(c_t) \right), \tag{3.45}$$

$$pc_g^c = \sum_{t \in \mathcal{T}} \left(\sum_{c_t \in \mathcal{C}_{g,w}} e^{-\frac{\text{dist}(c_t,g)}{d_0}} + |\mathcal{C}_{g,t,v}| \right), \tag{3.46}$$

$$ps_g^c = \sum_{t \in \mathcal{T}} \left(\sum_{c_t \in \mathcal{C}_{g,w}} s(c_t) e^{-\frac{\text{dist}(c_t,g)}{d_0}} + \sum_{c_t \in \mathcal{C}_{g,v}} s(c_t) \right), \tag{3.47}$$

and for ChromHMM promoter/enhancer segments we get:

$$pl_g^h = \sum_{h \in \mathcal{H}_{g,w}} l(h) e^{-\frac{\text{dist}(h,g)}{d_0}} + \sum_{h \in \mathcal{H}_{g,v}} l(h), \tag{3.48}$$

$$pc_g^h = \sum_{h \in \mathcal{H}_{g,w}} e^{-\frac{\text{dist}(h,g)}{d_0}} + |\mathcal{H}_{g,v}|, \tag{3.49}$$

$$ps_g^h = \sum_{h \in \mathcal{H}_{g,w}} s(h) e^{-\frac{\text{dist}(h,g)}{d_0}} + \sum_{h \in \mathcal{H}_{g,v}} s(h). \tag{3.50}$$

Finally, for the window based and nearest gene annotation, we intersect $\mathcal{D}_{g,v}$ and $\mathcal{C}_{g,v}$ with $\mathcal{H}_{g,v}$ and obtain $\mathcal{D}'_{g,v}$ and $\mathcal{C}'_{g,v}$ to reduce the number of regions associated with g from the distal region v. The considered annotation versions are explained in Figure 3.34.

Association of Hi-C loops to accessible chromatin

Before learning any models using Hi-C data, we tried to get a better understanding of the characteristics of Hi-C data and its relation to chromatin accessibility in general. To this end, we assessed the overlap between DHS and Hi-C loops. As shown in Figure 3.35a, the fraction of Hi-C loops overlapping with at least one DHS increases with a decreasing Hi-C resolution.

The tremendous differences between the various resolutions suggest that the choice of the used Hi-C resolution will likely effect any downstream analysis relying on DHS sites. Taken into account each Hi-C loop across all resolutions, at least 80% of the identified Hi-C loops overlap with at least one DHS site in four out of five cell lines.

These observations trigger the question how many DHSs that can be associated with the loop are already occupied by factors such as CTCF or Cohesin that are required for mediation and maintenance of the chromatin interactions. Assuming that these factors spatially fully occupy the accessible chromatin, it might be likely that regulatory factors interacting with the mediator complex and the transcriptional initiation machinery need to bind to other regions and thus such, we term them structural DHSs, might be a source of noise in our models.

The effect of the Hi-C resolution on the number of genes that are linked to a chromatin loop is depicted in Figure 3.35b-c. Generally, we observe that the number of genes associated with a loop reduces with a more precise, i.e. numerically smaller, Hi-C resolution. The search window used to link a Hi-C loop to a gene also influences the number of mapped genes. As expected, with an increasing window size, the number of genes that are linked to a loop is rising. Simultaneously the slope of the increase depends on the utilised Hi-C resolution. For example, as shown in Figure 3.35b-c, the increase in the number of genes is only marginal for the best resolution (5kb), while it is more than three times as strong for the lowest one (25kb).

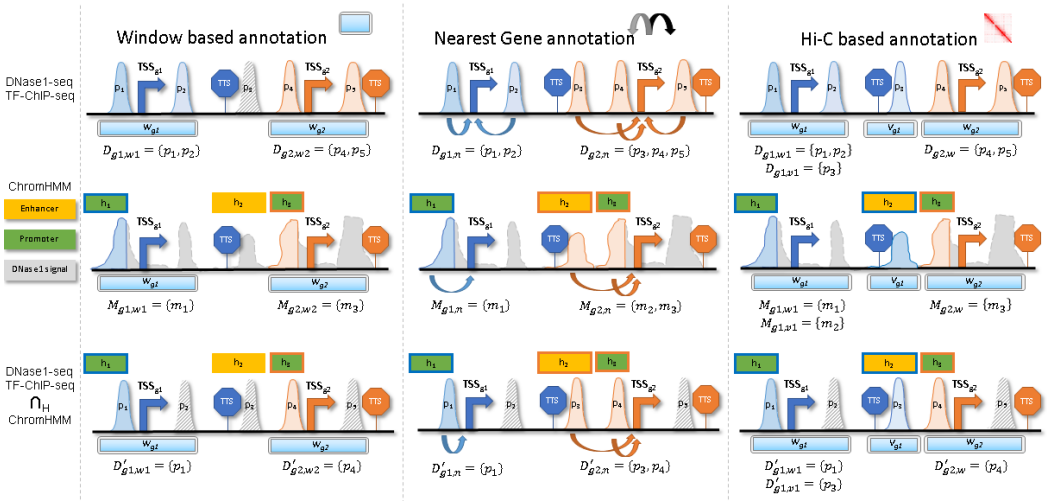


Figure 3.34: Figure 1 illustrates several ways how regulatory regions can be linked to genes. In a window based annotation, DNaseI-seq or TF ChIP-seq peaks are linked to a gene if they are located within a window w centred at the 5' transcription start site (TSS) of a gene of interest. Using Hi-C, a second window v covering the looped region is considered in addition to the TSS window. We computed three variants per linkage strategy. In the first row, we depict the linkage using only DNaseI-seq or TF ChIP-seq peaks and aggregating their signal across all associated peaks. The second row shows the aggregation of the signal of an epigenetic signature in promoters and enhancers identified by ChromHMM and in the third row, we consider the intersection of peak regions with ChromHMM promoter and enhancer segments. The figure illustrates these setups for two genes $g1$ and $g2$. The color code of peaks and the border color of segments indicate to which gene a peak or segment is assigned. Peaks with a striped filling are not assigned to any gene. To improve clarity, we show for the DNaseI-seq case the peak/segment sets. Joint work with Fabian Kern, presented at GCB 2018.

Performance of models including Hi-C loops

We trained linear models of gene-expression on the Hi-C gene set using window based models of DNaseI-seq ($\mathcal{D}_{g,w}$, Figure 3.36a) and ChIP-seq ($\mathcal{C}_{g,w}$, Figure 3.36b) data with two different window sizes, 3kb and 50kb. Those models are compared to Hi-C models incorporating DHS/ChIP-seq peaks in the Hi-C window v both with and without an intersection using promoter/enhancer regions from ChromHMM. The results shown are based on Hi-C loops from all resolutions. They did not change if a subset of resolutions has been used (data not shown).

In case of DNaseI-seq (Figure 3.36a), we do not find a clear trend to argue that a certain setup constantly performs best. However, we can see that the Hi-C models

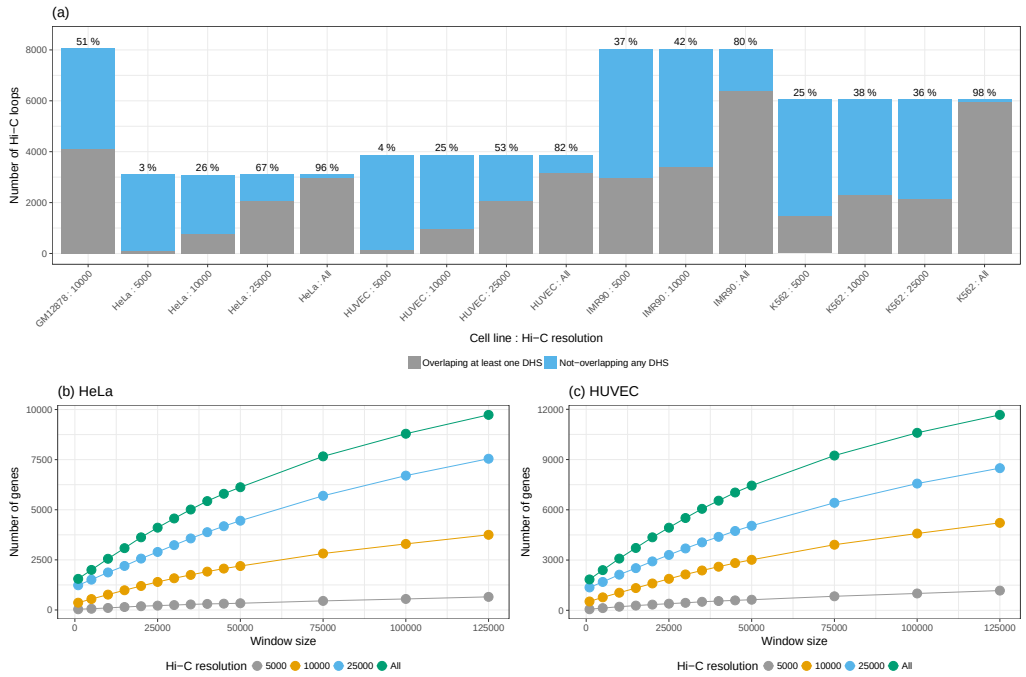


Figure 3.35: a) The total number of Hi-C loops (aggregated across all available Hi-C resolutions) per cell line is shown and the percentage of how many loops intersect with at least one DNaseI-seq peak is indicated. Except for GM12878, a large fraction of Hi-C loops overlap at least one DHS site. b)-c) indicate the number of genes with a least one Hi-C loop for varying search window sizes and Hi-C resolutions. With an improving resolution of the Hi-C experiments, i.e. it's value gets numerically smaller, the number of overlapping genes seems to be reducing. With an increasing search window size around the TSS of genes, the number of overlapping genes is monotonically increasing although the increase tends to be less pronounced for high resolution experiments. Joint work with Fabian Kern, presented at GCB 2018.

never perform better than the conceptually simpler peak-based models. Also, we note that the ChromHMM intersection improves the performance significantly in only one cell line.

In the ChIP-seq case (Figure 3.36b), we observe a generally increased model performance compared to the DNaseI-seq data. Further the simple 50kb window models tend to outperform the remaining annotation versions, including the Hi-C models. In contrast to the DNaseI-seq case, the ChromHMM intersection does significantly improve model performance in four out of five samples. At the same time four of these models perform similar to the 3kb window models neglecting Hi-C data.

Overall, including long-range PEI deduced from several Hi-C experiments by

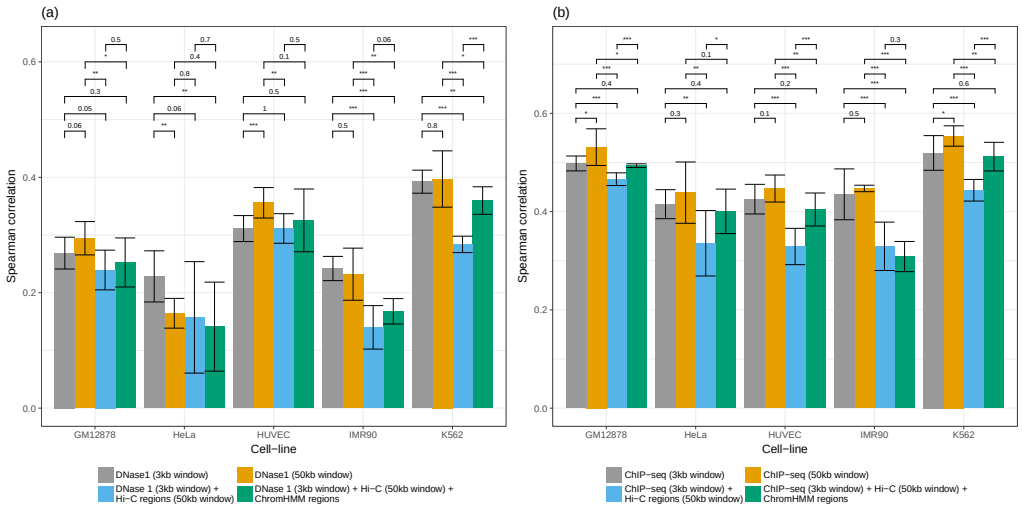


Figure 3.36: In a) the performance of gene-expression models, assessed via Spearman correlation, based on DNaseI-seq data is depicted using four different annotation setups: [1] DHSs in a 3kb window, [2] DHSs in a 50kb window, [3] DHSs in a 3kb window combined with DHS sites in Hi-C loops searched in a 50kb window, [4] same as [3] considering chromatin state segmentation from ChromHMM. Although no clear trend in terms of model performance is identifiable, it appears that purely DNaseI-seq based models perform slightly better than models including Hi-C data. Figure b) is analogous to a) but using TF ChIP-seq data instead of DNaseI-seq data. Here, best model performance could be achieved by considering the ChIP-seq signal within a 50kb window around the TSS of genes, neglecting the Hi-C data. However, it can be seen that combining the Hi-C data with ChromHMM segmentations improves model performance. A p-value ≤ 0.001 is indicated by ***, a p-value ≤ 0.01 is indicated with ** and a p-value ≤ 0.05 is indicated with *. Joint work with Fabian Kern, presented at GCB 2018.

Aiden *et al.* [R+14b] turned out not to be beneficial for modelling gene-expression as performed in this study.

Our models indicate that reducing the number of considered peaks from the Hi-C loop windows is beneficial for model performance, suggesting that the focus on fewer regions in the distal loop window is eliminating irrelevant signal such as pure noise or the aforementioned structural DNaseI-seq peaks that are relevant for loop formation but not for the actual expression regulation. Furthermore, it might also be possible that not all chromatin contacts are directly linked to transcriptional regulation and gene-expression, as suggested by Ray *et al.* [R+19].

Refining the considered genomic space via chromatin state segmentations

To gain a better understanding why the performance of the Hi-C models could be improved by a stricter selection of potential regulatory regions, we had a closer look at the performance of simple window based models using the reduced sets of DNaseI-seq or TF ChIP-seq peaks, respectively ($\mathcal{D}'_{g,w}$ and $\mathcal{C}'_{g,w}$). Additionally, we considered models based on DNaseI-seq, or TF ChIP-seq signal within the selected ChromHMM segments ($\mathcal{E}_{g,w}$).

In Figure 3.37a, the results for the models trained on DNaseI-seq data using a 50kb window are shown.

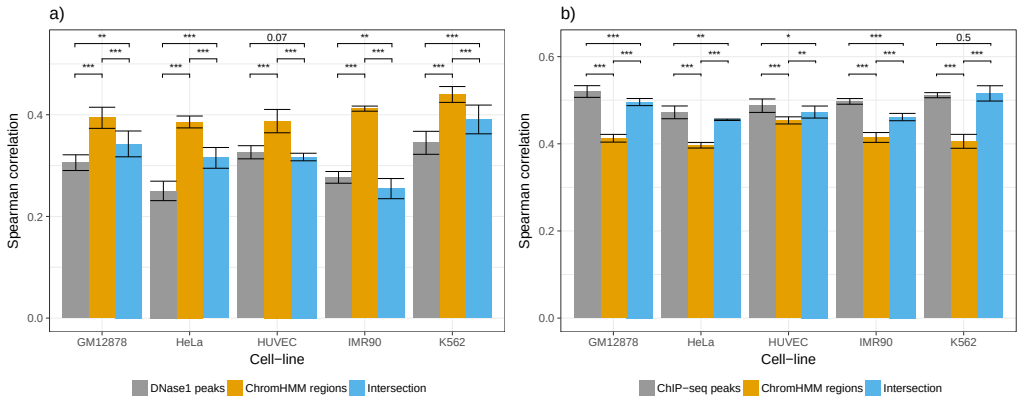


Figure 3.37: a) Spearman correlation of expression models based on DNaseI-seq data is shown for five cell lines and three annotation versions using a 50kb window: [1] DNaseI-seq peaks only, [2] ChromHMM regions and [3] the intersection of both. Considering only the DNaseI-seq signal within ChromHMM segments outperforms the other two approaches, while the intersection of DHS-sites with the ChromHMM segments outperforms the purely DHS based features. b) This sub figure is analogous to (a) with the difference that DNaseI-seq data is replaced with ChIP-seq data. Here, considering the ChIP-seq signal within the ChromHMM segments performs worse than the alternative annotation versions. A p-value ≤ 0.001 is indicated by ***, a p-value ≤ 0.01 is indicated with ** and a p-value ≤ 0.05 is indicated with *. Joint work with Fabian Kern, presented at GCB 2018.

Interestingly, models based exclusively on promoter/enhancer segments ($\mathcal{H}_{g,w}$) perform significantly better than models solely relying on DHS sites ($\mathcal{D}_{g,w}$). The intersection of regulatory segments with DHS sites ($\mathcal{D}'_{g,w}$) significantly improved over the $\mathcal{D}_{g,w}$ models in three out of five cases. However, the intersected feature space still leads to significantly worse performance than the state based models ($\mathcal{H}g, w$).

We have characterised the peaks removed by the intersection in terms of peak quality and functional annotation using ChromHMM. As exemplified for GM12878

in Figure 3.38a, the mean q-value of removed DHS peaks in the $\mathcal{D}'_{g,w}$ case is lower than that of peaks retained after the intersection suggesting that more pronounced and reliable peaks are maintained. The same observation can be made for the other cell lines (data not shown).

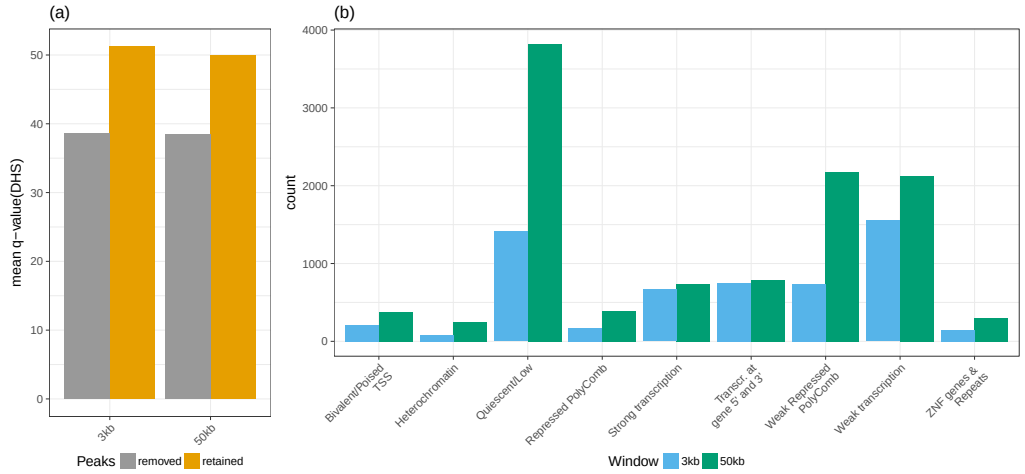


Figure 3.38: a) The mean of JAMM q-values is shown for DHS sites in GM12878 that are removed/retained by an intersection with promoter and enhancer segments derived from ChromHMM using two different window sizes. Overall, the score of the retained peaks is higher than that of the removed peaks. In b), we show the ChromHMM states for those peaks being removed from consideration. We observe differences in the count distribution between the 3kb and the 50kb windows. By increasing the window from 3kb to 50kb, the amount of peaks falling into *Quiescent/Low*, *Weak Repressed/Polycomp* and *Weak transcription* segments is increasing considerably, suggesting that many DHS sites that would be considered by the enlarged window might not be relevant for the regulatory activity. Joint work with Fabian Kern, presented at GCB 2018.

A large portion of removed peaks are linked to *Quiescent/Low, Weak Repressed Polycomb* and *Weak transcription* ChromHMM states (Figure 3.38b), which also indicates that the removed peaks are not relevant for transcriptional regulation.

In the ChIP-seq case, depicted in Figure 3.37b, we find a different behaviour. Here, models based only on the enhancer segments ($\mathcal{M}_{g,w}$) perform worse than the other two models ($\mathcal{C}_{g,w}$, $\mathcal{C}'_{g,w}$). Note that the ChIP-seq peaks intersected with the promoter/enhancer regions ($\mathcal{C}'_{g,w}$) perform significantly worse than the purely peak based models ($\mathcal{C}_{g,w}$) in four of five cases. However, we do observe, as for the DNaseI-seq data that the mean score of the removed ChIP-seq peaks ($\mathcal{C}'_{g,w}$) is lower than that of the retained peaks. Evaluating the ChromHMM segments linked to the removed peaks again shows that many peaks are linked to *Quiescent/Low, Weak Repressed Polycomb* and *Weak transcription* states. We conclude that currently

available Hi-C resolutions lead to many false positive associations, preventing an adequate modelling of long range PEIs, in this per sample based gene-expression prediction problem. Furthermore, our analysis also puts forward the hypothesis that only very few genes are involved in long-distal regulatory interactions. It is likely that other methods, e.g. ChIA-PET or capture Hi-C [D+14, F+09] (Section 2.1.11), which can enrich the sequencing libraries for distinct regions such as promoters of interest, lead to more precise contact maps. Leveraging these more fine-grained technologies for gene-expression modeling seems to be an opportunity to improve the prediction performance and to increase our understanding of the underlying regulatory processes.

3.5 INVOKE - A pipeline for integrative analysis of TFBS prediction and gene-expression data

3.5.1 Motivation

To provide users with a comfortable way of using both TEPIC and computing linear models from TF-gene scores, we include the INVOKE pipeline in the TEPIC repository. In general terms, INVOKE is an integrated analysis pipeline of epigenetics data, e.g. open-chromatin data (DNaseI-seq, ATAC-seq, NOMe-seq) and gene-expression data to suggest key transcriptional regulators in the analysed sample.

The INVOKE analysis is split up into two main steps:

1. Computing TF-gene scores on the basis of epigenetic data using TEPIC.
2. Learning a linear regression model to predict gene-expression from TF-gene scores computed in (1).

As illustrated above for T-cells and primary human hepatocytes, we use TF-gene scores as features in a linear regression setup to predict gene-expression. In such a *per sample* approach, we stick to the simplifying assumption that all genes are regulated similarly. Features with a high regression coefficient can be suggested to be key regulators in the analysed sample, as they seem to affect the expression of a large portion of the genes under consideration. However, the results of this method should be seen as suggestions for possible regulators and not as the absolute truth.

3.5.2 Implementation details

The INVOKE pipeline can be easily executed using a single bash script, controlling TFBS annotation through TEPIC and learning the linear model in R. Model parameters can be easily adjusted in a configuration file. An executable example is provided in the TEPIC repository.

We offer three different regularization techniques: lasso, ridge and elastic net.

As described in Section 2.2.1, lasso regularization leads to sparse models and can be optimized quickly. But, lasso cannot properly deal with correlated features, e.g. instead of distributing the coefficients among them, only one feature is selected.

Also, lasso solutions are not stable and therefore should be interpreted with caution. Nevertheless, lasso regularization is good to get a first impression of model performance. The disadvantage of ridge regression is that it cannot produce sparse models (many coefficients being exactly 0), which may hinder interpretability. Elastic net regularization on the other hand, resolves the correlation between features by distributing the feature weights among them and simultaneously leads to sparse and stable models.

The data matrix X, containing TF-gene scores and the response vector y, containing gene-expression values, are log-transformed, with a pseudo-count of 1, centered and scaled. Regression coefficients are computed in an inner cross validation, the α parameter of elastic net regularization is optimized with a default step size of 0.1.

We offer two ways to use our learning pipeline:

1. Learn a model for feature interpretation without computing performance measures: In order to provide a time efficient way of obtaining an interpretable model and to prevent a potential loss of information by considering only a portion of the full data set for model training, the regression coefficients are determined on the entire data set.

2. Learn a model for feature interpretation and compute model performance: Nested cross-validation is used to learn the models and to assess their performance. Per default, 20% of the data are used as test data and 80% are used as training data. Model performance is assessed in an outer cross validation. We report the mean Pearson correlation, the mean Spearman correlation and the MSE over the outer folds as measures of model performance. Additionally, a model is learned on the entire data set as described in (1) for interpretation of the coefficients.

All parameters mentioned in this section can be changed by the user.

3.5.3 Required input

In addition to the input required for the computation of TF-gene scores in TEPIC that is a reference genome file, a TF motif file, a set of candidate regions as well as a genome annotation file, gene-expression data must be provided to run Invoke. The gene-expression data is a tab delimited file that should be structured such that Column 1 contains the gene identifiers and Column 2 holds expression values.

3.5.4 Output and hints for interpretation

The user is always provided with the following files:

- a list of regression coefficients computed on the entire data set and
- a bar plot showing the regression coefficients with an absolute value > 0.025.

The larger a regression coefficient, the stronger is the inferred effect of the corresponding TF on gene-expression. Positive coefficients suggest an activating influence of TFs, negative TFs suggest an inhibiting effect.

If model performance was assessed, the following is available in addition:

- a summary on model performance containing the aforementioned measures (Pearson correlation, Spearman correlation, MSE),
- a list of regression coefficients determined in the outer cross validation,
- a heatmap visualizing the regression coefficients determined in the outer cross validation for at most the top 10 positive and negative features, sorted according to their median value.
- an image showing a box plot for Pearson and Spearman correlation, respectively and
- scatter plots showing the predicted vs the measured gene-expression for each outer cross validation fold.

The heatmap can be easily used to judge model performance, as it shows the regression coefficients of all outer-cross validation runs and thus indicates whether the coefficients are stable. The box plots provide further insights into model performance and stability across the outer folds of the cross validation.

3.6 Regulator Trail

3.6.1 Purpose of RegulatorTrail

To provide an even more user-friendly option to use Tepic and the Invoke pipeline, we included both approaches in the RegulatorTrail webserver [K+17a]. RegulatorTrail is a web service for both identification and prioritisation of key transcriptional regulators using different methods. The webserver provides an extensive documentation as well as tool-tips to guide users while performing the analyses. Example data is available to exemplify all potential applications. RegulatorTrail allows four general use cases [K+17a], which are briefly sketched in the following Section.

3.6.2 Supported use cases

To RegulatorTrail, we have contributed the code, examples and documentation for use-cases three and four.

Over-representation analysis

The first use case requires a user to upload a list of differentially expressed genes. Using RegulatorTrails rich collection of known regulatory-target interactions

(RTI) transcriptional regulators, whose set of target genes have a significant overlap with the uploaded gene list are identified. To do so, three statistical tests are offered: a binomial test, a hyper-geometric test and the Fisher's exact test. P-value adjustment methods are available as well. In the end, the user is provided with a list of regulators sorted according to the adjusted p-values.

Regulator effect analysis

In the second use-case, gene-expression data for two groups of interest, e.g. disease and control, should be provided. Gene-expression differences are identified either via simple statistical measure like the z-score or the fold-change, or via more sophisticated methods like DESeq2 [AH10].

In a second step, RegulatorTrail utilizes user defined lists of up and down regulated genes in approaches that utilize expression correlation between regulators and targets to prioritize the regulatory factors. To do so, RegulatorTrail offers several approaches, including Reggae developed by Kehl *et al.* [K$^+$18b]. Besides, the sorted regulator lists information on whether the regulators have an activating or repressing effect on the selected genes is provided.

Annotation of TFBS using TEPIC

This scenario refers to the annotation of candidate regulatory sites using Tepic to compute TF-gene scores. To improve the runtime of the annotation, only candidate TFBS in the vicinity of genes are considered (according to a user defined window). The resulting affinities are aggregated into TF-gene scores, which can be used, for instance, in enrichment analysis or used as input for a linear model, as described in the next Section.

INVOKE analysis to determine tissue specific regulatory factors

This use-case is the webserver implementation of the Invoke analysis. First, TF-gene scores are computed as in use-case 3. Next a linear regression model using either lasso, ridge, or elastic net regularization is fitted to predict gene-expression in a sample of interest. The gene-expression file needs to be uploaded by the user.

The webserver provides a user-friendly way to set the parameters of the method and to interpret the determined regression coefficients.

3.7 Contributions of all researchers involved in the projects described here

The work in this Chapter is based on contributions from several people: Florian Schmidt (Saarland University), Nina Gasparoni (Saarland University), Gilles Gasparoni (Saarland University), Kathrin Gianmoena (IfADo), Cristina Cadenas (IfADo), Julia K Polansky (German Rheumatism Research Centre, currently at Charité University Medicine), Peter Ebert (Max Planck Institute for Informatics,

currently at Saarland University), Karl Nordström (Saarland University), Matthias Barann (Chris
tian-Albrechts-University), Anupam Sinha (Christian-Albrechts-University), Sebastian Fröhler (Max-Delbrück Center for Molecular Medicine), Jieyi Xiong (Max-Delbrück Center for Molecular Medicine), Azim Dehghani Amirabad (Saarland University), Fatemeh Behjati Ardakani (Saarland University), Barbara Hutter (DKFZ Heidelberg), Gideon Zipprich (DKFZ Heidelberg), Bärbel Felder (DKFZ Heidelberg), Jürgen Eils (DKFZ Heidelberg), Benedikt Brors (DKFZ Heidelberg), Wei Chen (Max-Delbrück Center for Molecular Medicine), Jan G. Hengstler (IfADo), Alf Hamann (German Rheumatism Research Centre), Thomas Lengauer (Max Planck Insitute for Informatics), Philip Rosenstiel (Christian-Albrechts-University), Jörn Walter (Saarland University), Marcel H. Schulz (Saarland University and Göthe University Frankfurt), Fabian Kern (Saarland University), Lara Schneider (Saarland University), Tim Kehl (Saarland University) and Hans-Peter Lenhof (Saarland University).

Specifically, DEEP primary hepatocyte samples (LiHe1, LiHe2, LiHe3) and HepG2 data have been generated by Kathrin Gianmoena, Cristia Cadenas and Jan G Hengstler. CD4+ T-Cells from DEEP (T1-T6) were obtained by Julia K Polansky and Alf Hamann. RNA-seq of all DEEP samples was carried out by Sebastian Fröhler and Wei Chen. The data was processed by Matthias Barann, Anupam Sinha and Philip Rosenstiel. DNaseI-seq and NOMe-seq experiments for DEEP data were performed by Gilles Gasparoni, Nina Gasparoni and Jörn Walter. Data management and inital processing such as alignment was carried out by Bärbel Felder, Barabara Hutter, Gideon Zipprich, Benedikt Brors and Jüergen Eils. Aligned DNaseI-seq data was processed by Peter Ebert (MACS2) and Florian Schmidt (JAMM). Aligned NOME-seq reads were analyzed by Karl Nordström and Gilles Gasparoni.

All ENCODE DNaseI-seq an RNA-seq data has been downloaded from the ENCODE data portal and has been curated by Florian Schmidt.

Florian Schmidt developed the TEPIC approach including all scoring schemes listed in Table 3.3. All experiments and analysis shown in Figures 3.2, 3.5-3.9 and 3.11-3.33 were performed by Florian Schmidt. He was advised by Marcel H Schulz.

Fabian Kern performed the comparison of window based to nearest gene approaches shown in Figure 3.10 and developed the Hi-C extension for Tepic (Figures 3.34). He was advised by Florian Schmidt, who suggested the project and the analyses shown in Figures 3.35-3.38 and Marcel H Schulz.

Fatemeh Behjati and Azim Dehghani Amirabad contributed to the first version of the INVOKE pipeline by contributing a script for merging several feature matrices and an R-script for linear regression, respectively. With the ongoing development of Tepic, both scripts were replaced by improved versions written by Florian Schmidt, which are now included in the Tepic framework.

The RegulatorTrail webserver has been developed at the chair of Hans-Peter Lenhof. In detail Florian Schmidt provided the code for TEPIC and INVOKE modules, while Tim Kehl programmed the web-interface and the remaining func-

tionality of RegulatorTrail. Lara Schneider supported both and assisted in testing the webserver and contributed to the tool-tips and the documentation.

4

Identification of regulators linked to differential gene-expression

This chapter delineates our contributions to DEEP sub-project SP5-1 *Epigenetics of inflammatory T-cells-features functions and implications for the clinic*, published in Durek *et al.* [D$^+$16e].

4.1 Research questions of the project

4.1.1 Problem setting in SP5-1

The goal of this sub-project was to characterize the impact of epigenetic modifications on the differentiation of memory T-cells. Particularly, a better understanding of the directions of differentiation among various memory T-cell sub-types should have been obtained. Also, essential regulatory signatures involved in regulating the differentiation should have been determined.

To this end, two replicates each of naive CD45RA+ CD4+ T-cells from blood (TN), central memory cells (TCM) and effector memory cells (TEM) were subject to genome-wide DNA methylation, histone modification, DNA accessibility (using NOMe-seq) and gene-expression profiling.

Details on the data used within this chapter as well as on data processing are provided in Section B.2.

4.1.2 Our contributions

We made two main contributions to this work. We devised a similarity score used to argue for a linear differentiation from TN to TCM to TEM cells, c.f. Figure 2c in Durek *et al.* [D$^+$16e]. For reasons of brevity, this is not further discussed within this thesis.

Furthermore, we designed an integrative approach that suggests TFs exhibiting a differential binding behaviour between cell types that is predictive for observed gene-expression changes. Using knockout experiments in mice, the TF FOXP1, which was suggested by our method, was validated as an essential regulatory TF. Within the remainder of this chapter, we describe this approach, termed Dynamite, in more detail.

4.2 The DYNAMITE pipeline

4.2.1 Overview

DYNAMITE infers TFs exhibiting a differential binding behavior between two cell types or conditions that results in differential gene-expression. As input, DYNAMITE requires both chromatin accessibility and gene-expression data for both groups of interest, as depicted in Figure 4.1. For gene-expression data, we use the information

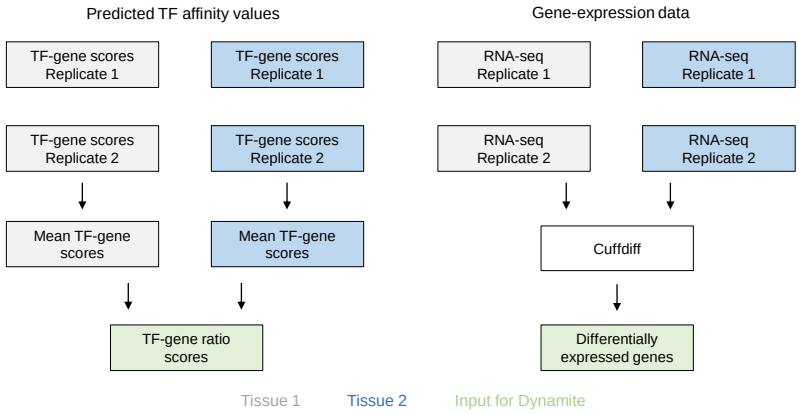


Figure 4.1: Here, the overall score computation for the feature matrix used within DYNAMITE and for the response, i.e. the differential gene-expression, is illustrated.

of all available replicates to determine differentially expressed genes, e.g. using CUFFDIFF [T+10]. Genes that are differentially expressed between two conditions are labelled with 1 if they are up-regulated in group 1 compared to group 2 and with 0 otherwise. Genes that are not differentially expressed are not considered within a DYNAMITE analysis. The computation of TF-gene ratio scores is detailed in the next Section.

4.2.2 A differential TF-gene score

Differential TF-gene scores should reflect differences in the binding behaviour of TFs between two cell types or two conditions. We compute TF-gene scores $a^E_{g,t,s}$ using TEPIC in candidate TFBS deduced from chromatin accessibility data in each available replicate s separately. Within the two considered groups of samples $\mathcal{U}_1$ and $\mathcal{U}_2$, we compute the mean TF-gene scores $a^E_{g,t,U1}$ and $a^E_{g,t,U2}$ for each gene g

and TF t, according to:

$$a^E_{g,t,\mathcal{U}_1} = \frac{1}{|\mathcal{U}_1|} \sum_{s \in \mathcal{U}_1} a^E_{g,t,s}, \tag{4.1}$$

$$a^E_{g,t,\mathcal{U}_2} = \frac{1}{|\mathcal{U}_2|} \sum_{s \in \mathcal{U}_2} a^E_{g,t,s}, \tag{4.2}$$

where $|\mathcal{U}_1|$ is the number of samples of group one and $|\mathcal{U}_2|$ denotes the number of samples of group two.

The mean values $a^E_{g,t,\mathcal{U}_1}$ and $a^E_{g,t,\mathcal{U}_2}$ are converted into a ratio score $a^E_{g,t,r}$ denoting the changes in TF binding:

$$a^E_{g,t,r} = \frac{a^E_{g,t,\mathcal{U}_1} + 1}{a^E_{g,t,\mathcal{U}_2} + 1}. \tag{4.3}$$

To ensure that the score can be computed also if $a^E_{g,t,\mathcal{U}_2} = 0$, we add a pseudo-count of 1 to both the nominator and the denominator. The ratio score $a^E_{g,t,r}$ is > 1 if the computed TF-gene score for TF t in samples $s \in \mathcal{U}_1$ is higher than in samples $s \in \mathcal{U}_2$. Analogously, $a^E_{g,t,r}$ is < 1 if the computed TF-gene scores for TF t in samples $s \in \mathcal{U}_1$ is smaller than in samples $s \in \mathcal{U}_2$. The ratio score evaluates to 1 if the binding behaviour is identical for TF t. Note that this mostly happens in case that a TF does not bind to gene g. An example is provided in Figure 4.2.

We refer to the TF-gene ratio matrix composed of all $a^E_{g,t,r}$ as X_r, where the rows are genes and the columns are TFs.

4.2.3 Logistic regression to classify genes as up- or down-regulated

We use X_r as input for a logistic regression classifier, as explained in Section 2.2.2. The classifier predicts gene-expression labels that is whether a gene is up- or down-regulated between two groups of samples, using the TF-gene ratio scores.

We are using the same model fitting and evaluation procedure as with the linear models in Chapter 3, c.f. Section 3.4.1 with the only difference that we are using accuracy as a performance measure and not the MSE. Thus, the optimization function is:

$$\min_{\beta_0,\beta} \left\{ \sum_{i=1}^{N} [y_i \log(p(x_i;\beta)) + (1-y_i)\log(1-p(x_i;\beta))] - \sum_{j=1}^{p} (\alpha\beta_j^2 + (1-\alpha)|\beta_j|) \right\}, \tag{4.4}$$

where N is the number of considered genes, p is the number of features and β is the regression coefficient vector as before. The parameter α determines the weights between lasso and ridge regularization, as introduced in Chapter 3. The optimisation problem is solved using the GLMNET R-package [FHT10].

As explained in Chapter 3 in context of linear gene-expression prediction models, the advantage of the elastic net regularization is that the inferred models are sparse yet correlated features are preserved in the model.

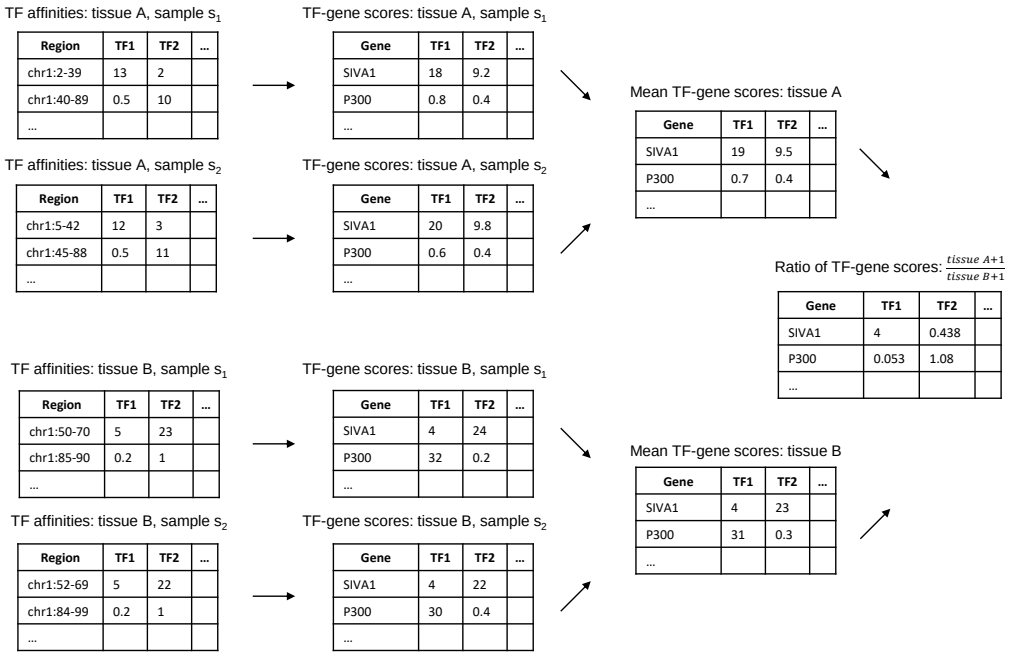


Figure 4.2: Example of the differential TF-gene score computation between two tissues with two replicates each. We consider two different tissues A, B with two samples each $A = \{s_1, s_2\}$ and $B = \{s_1, s_2\}$. First, we compute TF affinities for all candidate TFBS. Next, these are aggregated to TF-gene scores. From these scores, mean TF-gene scores (Eq. 4.1 and 4.2) are computed. Finally, the TF-gene ratio score is calculated (Eq. 4.3).

4.2.4 Availability and Usability of DYNAMITE

Similar to the INVOKE, DYNAMITE is integrated into our TEPIC framework. We provide a bash script that automatically computes TF-gene scores for provided candidate regions, calculates the TF-gene ratio scores and fits the logistic regression model. Parameters can be easily adjusted via a configuration file and an illustrated example is provided in the repository as well.

4.2.5 Required input

To run *DYNAMITE*, a user most provide candidate regions of TFBS for two groups $\mathcal{U}_1$ and $\mathcal{U}_2$, e.g. control and diseased. These can be derived, for example, from open chromatin experiments such as DNaseI-seq, or NOMe-seq. It is important for the performance of the model that the candidate regions reflect the characteristics of chromatin organization in the analysed tissues with high accuracy. In addition, a list of differentially expressed genes in terms of $\log_2$ fold changes of the expression are required.

4.2.6 Output and model interpretation

Model performance is reported both in a text file and visually in a bar plot using mean test and training accuracy as well as the F1 measure. Additionally, we report confusion matrices for all outer cross validation folds. A heatmap shows the regression coefficients of the selected features in the outer cross validation folds, providing insights on model stability. Besides, a bar-plot is generated showing the regression coefficients of all features selected in the final model. A positive coefficient is used by the model to predict genes as up-regulated, a negative coefficient is related to genes that are predicted as down-regulated.

To further simplify model interpretation, we provide an additional plotting script within the TEPIC repository that can be applied to the output files of a DYNAMITE run. As shown in Figure 4.3, density and scatter plots as well as a one column heatmap showing the regression coefficient of a distinct feature are generated to help elucidating why a particular feature was selected by the model. The density plots show the distribution of the mean feature values for both considered groups using the full data set as well as the 0.9 quantile, which removes extreme values. The scatter plots show the values for a distinct gene, colored according to the genes $\log_2$ fold change.

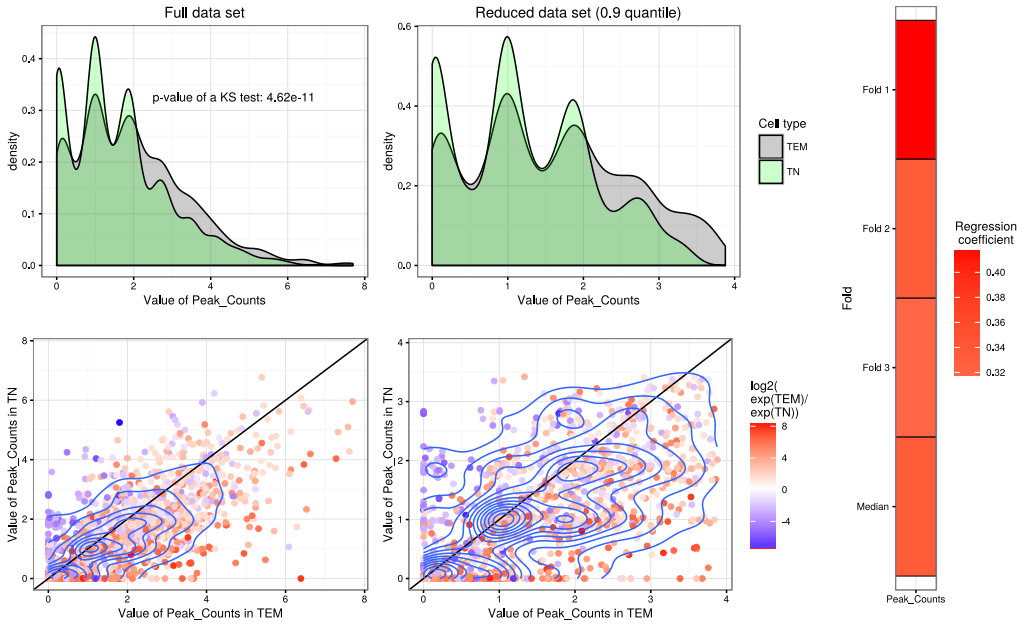


Figure 4.3: Example for an automatically created feature analysis figure generated on the example data provided in the TEPIC repository. The density plots show the distribution of feature values, the scatter plot relates them to the observed expression changes. The miniature heatmap shows the regression coefficients determined during the outer cross validation.

4.3 Application of DYNAMITE to CD4+ T-cell differentiation

We apply DYNAMITE to suggest key regulators involved in CD4+ T-cell differentiation. Within DEEP, 2 replicates each of TN, TCM and TEM cells were profiled. Here, we use both NOMe-seq and RNA-seq data, see Section B.2 for details on the data. In total, we consider three comparisons:

- TN versus TCM (1223),
- TCM versus TEM (614),
- TN versus TEM (2259),

where the numbers in brackets indicate the number of differentially expressed genes according to a q-value threshold of 0.01.

We use our initial set of TF motifs, comprising 450TFs [S$^+$17a] to compute TF-gene scores for all 6 replicates. As sketched in the previous chapter, we compute TF-gene ratio scores between the different cell types and use those as input for a logistic regression classifier to predict differential gene-expression. The logistic regression model is trained with the same scheme used for INVOKE: We use a 10-fold outer Monte-Carlo cross validation procedure to assess model performance and 6-fold inner cross validation for parameter learning. The step-size to optimise α is again set to 0.01.

4.3.1 Prediction results

The bar-plots in Figure 4.4 show the mean test-accuracy across the outer folds. Note that a random model would deliver an accuracy of 0.5. In our application all models are performing sufficiently well to be subject to feature interpretation.

At the right hand-side of Figure 4.5, we show which features obtained a median regression coefficient across the 10 outer-folds in the individual comparisons (marked by a red box). This filtering is performed to focus on features that are consistently selected on different compositions of the training data, thereby improving model performance. The heatmap on the left hand-side illustrates of Figure 4.5 how the expression of these TFs varies among the different replicates of the T-cell cell types, in relation to the mean expression level across all T-cell replicates. Note that the continuous expression information is not used directly in DYNAMITE and is shown here only for illustration purposes.

Interestingly, our model suggests several TFs exhibiting distinct expression profiles during T-Cell development. For instance, the TF genes *RUNX2*, *BCL6*, *FOS*, *ETV6*, *REL*, *BATF::JUN* and *JUN::FOS* are becoming more expressed during the suggested differentiation trajectory from TN to TCM to TEM cells. In contrast to that, *LEF1*, *KLF7*, *FOXP1* and *SREBF1* seem to be down-regulated during that differentiation. Also, several selected TF genes seem to be under epigenetic control, for instance, *AHR*, *FLI1*, *FOXP1* and *RUNX3*, as they overlap differentially methylated regions arising between the different T-cell sub-types [D$^+$16e].

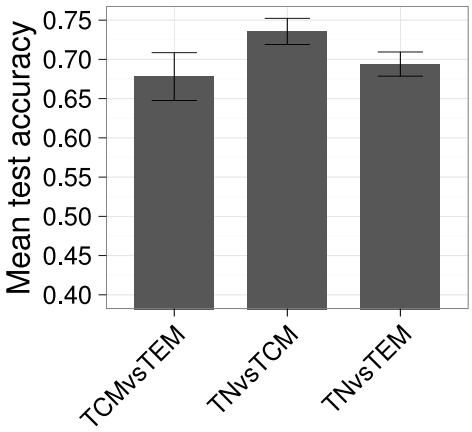


Figure 4.4: Mean test accuracy across a 10-fold Monte-Carlo cross validation of logistic regression models. The best classification performance can be achieved between TN and TCM. All models perform better than random models, which would achieve an accuracy of only 0.5. After Figure S4a from Durek *et al.* [D$^+$16e] (open access).

4.3.2 Experimental validation

The TF FOXP1 was of particular interest in this study as the *FOXP1* gene is gaining DNA methylation during differentiation (TN to TCM to TEM), implying that the methylation is exerting a silencing effect on *FOXP1* expression. Further, our DYNAMITE analysis highlighted FOXP1 as being an essential regulator, related to expression differences between TN and TEM. Our predictions are backed up further by an iREGULON analysis that identified TFBS of FOXP1 to be enriched in genes forming T-cell sub-type specific clusters [D$^+$16e]. Therefore, the role of FOXP1 was elucidated experimentally using a knockout-experiment in mice, which is detailed in Durek *et al.* [D$^+$16e].

CD4+ T-cells were isolated from spleens of both FOXP1 expressing and knock-out mice. The results of sorting these extracted cells according to the markers CD44 and CD62L are show in Figure 4.6. In the knockout mice, shown at the right hand side of the figure, the number of TN cells is diminished, while the number of TCM and TEM cells is increased. This suggests that FOXP1 is a TF that keeps T-cells in a naive state, a so called naive keeper. Importantly, the experiment also validates our DYNAMITE prediction for FOXP1.

4.4 Related approaches

In 2012, Chen *et al.* have shown that there is a substantial correlation between differential TF-binding derived from ENCODE TF ChIP-seq experiments and differential gene-expression between cell lines [C$^+$12b]. Specifically, they have used the $\log_2$ ratio of the ChIP-seq signals of 22 TFs available for both K562 and GM12878

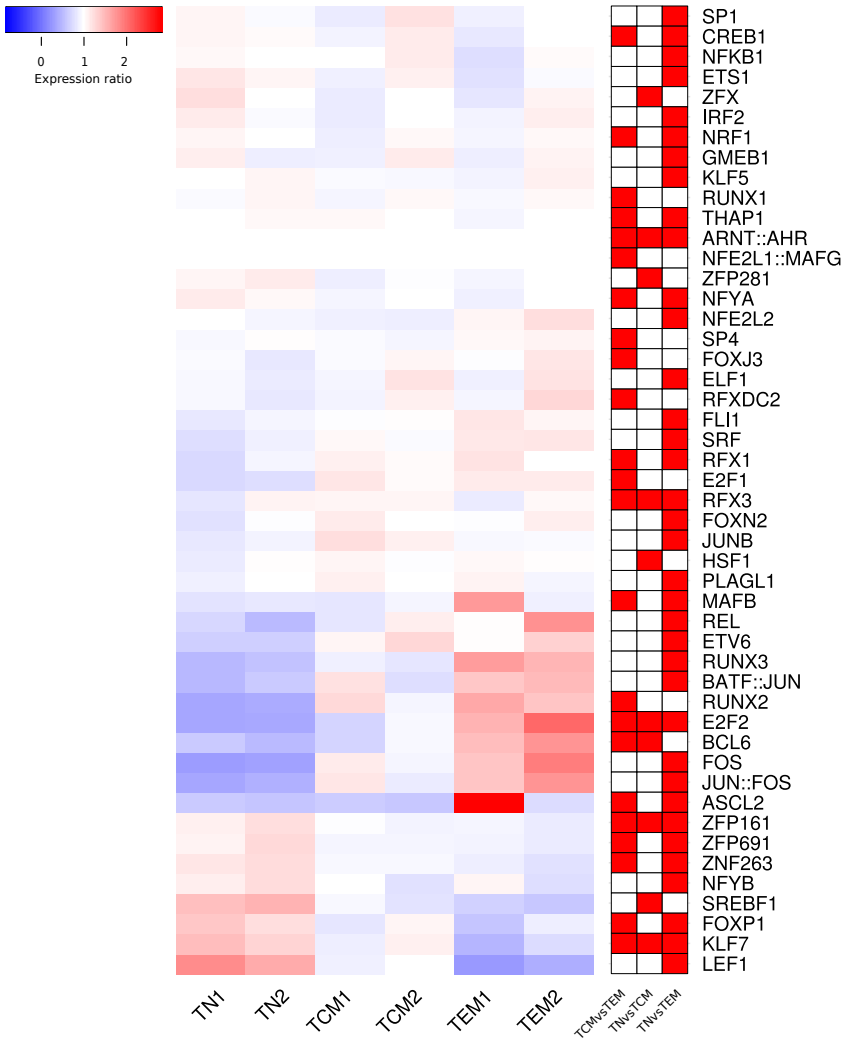


Figure 4.5: TF with a non-zero median regression coefficient across the outer cross validation for the individual comparisons are indicated at the right hand site by red boxes. The heatmap indicates the variation of gene-expression of these factors related to the mean expression of the factors across all six considered T-cell replicates. Extended version of Figure 4b from Durek *et al.* [D$^+$16e] (open access).

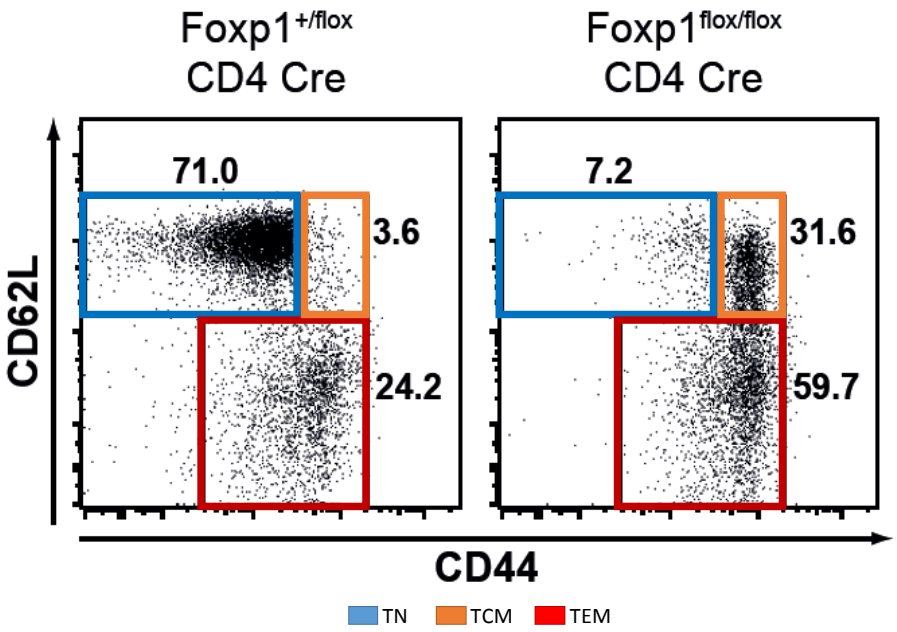


Figure 4.6: Results of a FOXP1 knock-out experiment in mice. The left rectangle shows sorted CD4+ T-cells extracted from spleen of mice expressing FOXP1, whereas the right rectangle represents the knock-out mice. The colored areas indicate the cell type of the extracted T-cells, according to the markers CD44 and CD62L. After Figure 5a from Durek *et al.* [D+16e] (open access).

cells as features in four different machine learning approaches, including random forests and support vector regression, to predict the $\log_2$ ratio of differentially expressed genes. The ChIP-seq data has been aggregated in 2kb windows centered around genes TSSs.

Their work can be seen as a proof of concept of the idea to link differential gene-expression to differential TF-binding, which in their case has been determined *in vivo*.

Within DYNAMITE, we extended their idea to account for predicted TFBS. Another difference to this earlier work is that we are treating the problem as a classification and not as a regression problem. We choose for this formulation of the problem as it is an easy way to reduce the influence of noise in the gene-expression data on the model.

Another fairly related work has been presented by Gonzales *et al.* [G+15b] in 2015. They obtain DNaseI-seq data from Roadmap for six cell types: human embryonic stem cells, hematopoietic stem and progenitor cells, monocytes, B-cells, T-cells and NK-cells. All identified DHSs are treated as candidate regulatory elements and linked to the closest gene in genomic space. Next, DHSs are annotated with

TFBS and those are used as features for a ridge regression model predicting gene-expression changes within HSPC, B-cells, T-cells, Monocytes and NK-cells.

This approach is different from DYNAMITE, because it does not consider differential binding of TFs, between conditions, but rather a combination of binding events in fully differentiated cells. Therefore, the explicit description of differential binding formulated in DYNAMITE is not made by Gonzales *et al.*. Furthermore, the ridge regression model is less suited for model interpretation than the elastic net regularization used in DYNAMITE, because the ridge solution will not be sparse.

4.5 Contributions of all researchers involved in the described project

The research presented in this chapter is part of a major collaborative project in the scope of DEEP to which the following people contributed [D$^+$16e]: Pawel Durek (German Rheumatism Research Centre), Karl Nordström (Saarland University), Gilles Gasparoni (Saarland University), Abdulrahman Salhab (Saarland University), Christopher Kressler (German Rheumatism Research Centre), Melanie de Almeida (German Rheumatism Research Centre), Kevin Bassler (University of Bonn), Thomas Ulas (University of Bonn), Florian Schmidt (Saarland University), Jieyi Xiong (Max-Delbrück Center for Molecular Medicine), Petar Glazar (Max-Delbrück Center for Molecular Medicine), Filippos Klironomos (Max-Delbrück Center for Molecular Medicine), Anupam Sinha (Christian-Albrechts-University), Sarah Kinkley (Max Plank Institute for Molecular Genetics), Xinyi Yang (Max Plank Institute for Molecular Genetics), Laura Arrigoni (Max Planck Institute of Immunobiology and Epigenetics), Azim Dehghani Amirabad (Saarland University), Fatemeh Behjati Ardakani (Saarland University), Lars Feuerbach (DKFZ Heidelberg), Oliver Gorka (Technical University Munich), Peter Ebert (Max Planck Institute for Informatics, currently at Saarland University), Fabian Müller (Max Planck Institute for Informatics, currently at Department of Genetics, Stanford University School of Medicine), Na Li (Max Plank Institute for Molecular Genetics), Stefan Frischbutter (German Rheumatism Research Centre), Stephan Schlickeiser (Charité University Medicine), Carla Cendon (German Rheumatism Research Centre), Sebastian Fröhler (Max-Delbrück Center for Molecular Medicine), Bärbel Felder (DKFZ Heidelberg), Nina Gasparoni (Saarland University), Charles D Imbusch (DKFZ Heidelberg), Barbara Hutter (DKFZ Heidelberg), Gideon Zipprich (DKFZ Heidelberg), Yvonne Tauchmann (Charité University Medicine), Simon Reinke (Berlin-Brandenburg Center for Regenerative Therapies), Georgi Wassilew (Charité University Medicine), Ute Hoffmann (German Rheumatism Research Centre), Andreas S Richter (German Rheumatism Research Centre), 0lina Sieverling (DKFZ Heidelberg), Hyun-Dong Chang (German Rheumatism Research Centre), Uta Syrbe (Charité University Medicine), Ulrich Kalus (Charité University Medicine), Jürgen Eils (DKFZ Heidelberg), Benedikt Brors (DKFZ Heidelberg), Thomas Manke (Max Planck Institute of Immunobiology and Epigenetics), Jürgen Ruland (Technical University Munich), Thomas Lengauer (Max Planck Institute for Informat-

ics), Nikolaus Rajewsky (Max-Delbrück Center for Molecular Medicine), Wei Chen (Max-Delbrück Center for Molecular Medicine), Jun Dong (German Rheumatism Research Centre), Birgit Sawitzki (Charité University Medicine), Ho-Ryun Chung (Max Plank Institute for Molecular Genetics), Philip Rosenstiel (Christian-Albrechts-University), Marcel H Schulz (Saarland University, Göthe University Frankfurt), Joachim L Schultze (University of Bonn), Andreas Radbruch (German Rheumatism Research Centre), Jörn Walter (Saarland University), Alf Hamann (German Rheumatism Research Centre) and Julia K Polansky (German Rheumatism Research Centre, currently at Charité University Medicine).

The overall project was designed by Julia K Polansky, Jörn Walter, Alf Hamann supported by Nina Gasparoni. Samples were prepared by Stefan Frischbutter, Stefan Schlickeiser, Ulrich Kalus, Carla Cendon, Yvonne Tauchmann, Georgi Wassilew, Simon Reinke, Uta Syrbe, Birgit Sawitzki, Jun Dong, Hyun-Dong Chang, Alf Hamann and Julia K Polansky. NOMe-seq data considered within our DYNAMITE approach has been generated by Giles Gasparoni, Nina Gasparoni and Jörn Walter. It has been post-processed by Karl Nordström and Giles Gasparoni. RNA-seq data has been generated by Sebastian Fröhler, Wei Chen and Charles D Imbusch. It has been computationally processed by Anupam Sinha. The data was managed by Bärbel Felder, Gideon Zipprich, Karl Nordström, Peter Ebert, Charles D Imbusch, Barbara Hutter, Benedikt Brors and Jüergen Eils. FOXP1 knockout experiments shown in Figure 4.6 were performed by Oliver Gorka and Jürgen Ruland.

Florian Schmidt designed the DYNAMITE approach as sketched in Figures 4.1 and 4.2. The classifier is based on a script for linear regression by Azim Dehghani Amirabad. However, the script has been heavily rewritten for speedup, clarity and stability. Also, an automated generation of Figures was added by Florian Schmidt. To further simplify usage and interpretation of DYNAMITE, Florian Schmidt integrated the classifier into the TEPIC framework as a fully automated pipeline. Also, together with Marcel H Schulz, he devised a strategy for improved interpretation of the selected features (Figure 4.3). Further, Florian Schmidt evaluated the models and generated the results shown in Figure 4.4 and 4.5.

Besides, Florian Schmidt suggested the cosine similarity as a measure to argue for cell type (dis)similarity in context of the linearity of the differentiation path of considered T-cell sub types. Fatemeh Behjati Ardakani came up with the idea of boosting to generate a statistical significance for this result (data not shown in the thesis).

Further details on author contributions covering parts of the manuscript that are not essential for the work presented in this chapter can be found in the Cell Immunity article [D$^+$16e].

5

EPIC-DREM - Identification of key regulators from time-series data

In this chapter, we illustrate our contributions to the article `"Temporal enhancer profiling of parallel lineages identifies AHR and GLIS1 as regulators of mesenchymal multipotency"` [GSo18], which was joint work with Deborah Gérard and Lasse Sinkkonen from the University of Luxembourg. The manuscript is used as a scaffold for this chapter.

5.1 Project description

5.1.1 Motivation and research objectives

The main research question in Gérard *et al.* [GSo18] was the elucidation of shared regulatory factors involved in the differentiation of multipotent bone marrow stromal progenitor cells towards either osteoblasts or bone marrow adipocytes. These are so called mesenchymal cell types.

A better understanding of the differentiation path of these cell types is of great (bio)medical interest for several reasons. Due to the common progenitor cells of osteoblasts and bone marrow adipocytes, there is a reciprocal balance in the number of fully differentiated cells between the two cell types. This dependency might explain why a reinforced commitment of progenitors to differentiate to adipocytes was associated with the inhibition of, for instance, bone healing [A$^+$17a]. This observation has been made related to obesity and high age. The accurate and complete differentiation of osteoblasts is not only important in recovering bone fractures and osteoporosis. Additionally, hormones produced by osteoblasts carry out important cellular functions, for instance, in the metabolism [L$^+$07, RC14]. Furthermore, bone marrow adipocytes have been shown to be an essential source of hormones responsible for metabolic well being such as the sensitivity to insulin concentration [C$^+$14b]. These examples illustrate how important the balance between osteoblast and bone marrow adipocyte differentiation is. In the following, we refer to bone marrow adipocytes simply as adipocytes.

5.1.2 Generated data and used methods

To identify shared as well as unique regulators important for the (de)differentiation of osteoblasts and bone marrow adipoctyes, time-series epigenomic (H3K27ac,

H3K36me3, H3K4me3) and transcriptomic profiling using RNA-seq was performed by our collaborators. In the profiling experiments, we investigated six distinct time points spread over 15 days during the differentiation processes of multipotent bone marrow stromal cell line cells (ST2) towards both osteoblasts as well as adipocytes.

Using TF-footprints called with the software HINT-BC[G+16b] on H3K27ac data, we computed TFBS predictions and devised a new approach to use TF-gene scores, computed at various time points, as input for the Dynamic Regulatory Events Miner (DREM). The purpose of DREM is to identify essential regulatory elements by associating regulatory information to temporal transcriptomic data (c.f. Section 5.2). We refer to our novel approach as EPIC-DREM. It is detailed in Section 5.3. To simplify the interpretation of EPIC-DREM predictions, we devised a network visualization strategy to indicate the interplay between predicted factors as well as the regulatory impact of individual TFs on their target genes.

Additionally, our coauthors identified super-enhancers (SEs) from H3K27ac data. SEs are large genomics intervals with enhancer function. Differences in the activity of super-enhancers over time have also been used to highlight putative regulatory proteins. The identified SEs are merged across all analysed time points. By considering the activity of these merged SEs across different time points, dynamic activity profiles of SE for both differentiation processes are generated. These profiles were used to further prioritize potential transcriptional regulators which we identified using EPIC-DREM. Computational details on SE identification are provided in Section B.3.3.

5.1.3 Results

Both the super-enhancer based identification of regulators as well as the EPIC-DREM analysis have highlighted FOXN1, AHR and GLIS1 as potential regulatory factors. While AHR has been reported earlier to inhibit both osteoblast [W+19a] and adipocyte [Ao98] differentiation, GLIS1 and FOXN1 have not been associated with these differentiation trajectories before.

In the scope of Gérard *et al.*, the regulatory function of AHR1 and GLIS1 was experimentally validated using both over-expression and silencing of the corresponding genes. The experimental findings are further described in Section 5.6. A more detailed analysis of FOXN1 is subject to future work.

5.2 Dynamic Regulatory Events Miner (DREM)

DREM is a tool designed to combine time-series expression data with protein binding data, to *"infer an annotated global temporal map"* [E+07, S+12b]. The global temporal map has two key purposes. Firstly, it denotes the major transcriptional regulatory events resulting in the measured expression patterns. Secondly, the map points the user to the regulatory factors linked to these regulatory events.

Within DREM, regulatory events are modelled as bifurcation events. A bifurcation event refers to a split point of gene-expression development that is a group of

genes that shares similar expression levels until some time point t_s where they split and follow different paths. Exactly at t_s, the bifurcation event happens. For several organisms, e.g. *escherichia coli*, a multi-layer hierarchical model of gene regulation has been proposed. The bifurcation events can be seen as a representation of this structure [B$^+$05a, E$^+$07]. A bifurcation event is not limited to be a binary split, also splits into more branches are possible.

Overall, DREM performs the following tasks:

- Identify bifurcation events in time-series transcriptomic data.
- Link TFs to bifurcation events.
- Assign genes to paths in a temporal map, which denotes a gene's expression development over time.
- Present this information in a global temporal map, exemplified in Figure 5.1a.

To fulfill these tasks, DREM uses an IOHMM, an extension of classical HMMs, introduced in Section 2.2.5. In the original version of DREM, either ChIP-chip experiments or motif data are used as the extra input. The output used in the models is time-series expression data. The hidden states are linked to the individual time points. Gaussian distributions model gene-expression for genes linked to the time point. An important change to this was made with DREM2.0 [E$^+$07], enabling the use of dynamic, time point-specific, regulatory information in the input.

By constraining the transitions between the hidden states, a tree structure is enforced that allows the bifurcation events to be modeled. DREM investigates many feasible tree structures. It selects the best one in a cross-validation procedure to optimize parameters. Specifically, for each tree structure, a logistic regression classifier (Section 2.2.2) is trained for each hidden state. The classifier is used to map the provided input to the individual hidden states and maps it to its transition probabilities (Figure 5.1b).

The global dynamic map is computed from the transitions that are inferred between the hidden states. Finally, for the best model, each gene is assigned to a distinct path within the global map according to its time-series expression information as well as the regulatory information used in the input. Subsequently, DREM computes association scores for TFs at bifurcation events utilizing a hypergeometric enrichment test [B$^+$05a].

Throughout this thesis, whenever we refer to DREM, we refer to DREM2.0, specifically to version 2.0.3. As mentioned above, DREM2.0 allows to use varying regulatory input features at each time point. Moreover, the included set of regulatory features in DREM2.0 has been extended compared to the original version and DREM2.0 accepts continuous regulatory scores as input. Also, the expression of TFs can be incorporated into the model.

In the past, DREM became an established tool to analyse regulatory networks from time-series transcriptomic data in several species [C$^+$12c, S$^+$16c].

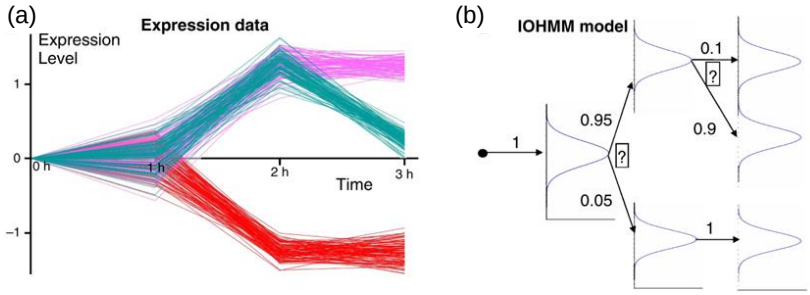


Figure 5.1: (a) Illustration of three temporal paths of gene-expression changes. Different colors encode different groups of genes. Until time 1h, all genes follow one trajectory until they split. While the expression of the genes labeled in red decreases until it stays constant at time 2h, the genes marked in blue and pink gain expression until time 2h and then are split again. Here, the expression of pink genes remains constant while the expression of blue genes drops again. DREM attempts to pinpoint such split-events. (b) Illustration of the logistic regression classifiers (marked with a "?") learned at the split points during the fitting procedure of the IOHMM model. Adapted from Figure 1 of Ernst *et al.* [B+05a] obtained under the Creative Commons Attribution License.

5.3 EPIC-DREM

5.3.1 Workflow of EPIC-DREM

In EPIC-DREM we replace the static set of regulatory interactions, which are included in DREM by predicted, time point-specific, regulatory scores. To do so, we suggested a novel approach to compute time point-specific TFBS predictions in TF footprints. To compute a p-value for the significance of TF affinity values for each TF at each time point, a randomization strategy, which accounts for GC-content and footprint length, is utilized. Time point-specific TFBS predictions computed in this way can be used as input for DREM to infer a temporal map of gene-expression development and regulation (Figure 5.2).

5.3.2 Method description

DREM2.0 supports continuous scores for TF-gene relationships. However, we have encountered not only issues in the computation of results using the continuous scores (both in terms of time and memory constraints) but also in the interpretation of the resulting models, simply because the feature space of the considered regulatory factors is very large.

Therefore, we developed a new TEPIC module that allows to compute a TF specific affinity cut-off to derive a binary measure of TF binding from TF affinities.

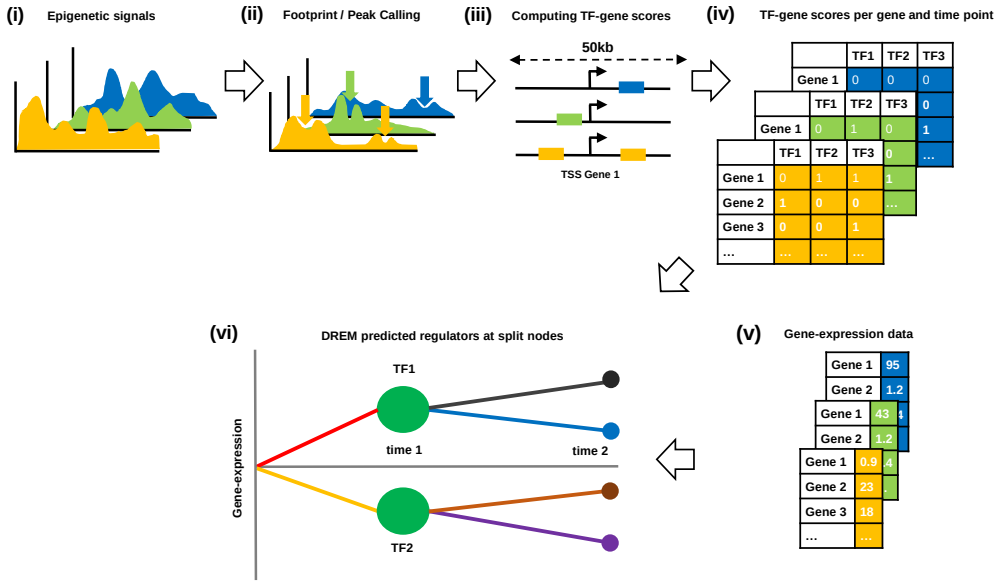


Figure 5.2: (i) From temporal epigenomic data, e.g. chromatin accessibility data, putative TFBS are computed (ii) through either peak or footprint calling. (iii) Using a novel TEPIC sub-module, TF- and time point-specific affinity thresholds are calculated to obtain binary TF affinities per candidate TFBS and time point. (iv) Using TEPIC, the binary TF affinities are aggregated into binary, time point-specific TF-gene scores. (v) The binary time point-specific TF-gene scores together are used as input. DREM generates a temporal map of gene-expression development and identifies related regulators. (vi) In the example, DREM highlights two gene sets, visualized by the red and yellow lines, respectively. They represent two distinct expression patterns, which are associated with the TFs 1 and 2. At the next time point, DREM predicts the formation of four sub-groups. The functionality to compute TFBS predictions as input to EPIC-DREM is part of TEPIC 2.0 [S+18b]. Figure based on Figure 2a from Gérard *et al.* [GSo18] (open access).

This is achieved by comparing TF affinities computed on the actual candidate sites to TF affinities computed on a set of random sequences that are designed to mirror both the GC content and length distribution of the true candidate binding sites.

Formally, let $a_{r,t}$ denote the affinity for TF t computed in the random region $r \in \mathcal{R}$, where $\mathcal{R}$ is the set of all random regions and $|r|$ is the genomic length of r. Analogously, let $a_{o,t}$ denote the affinity for TF t computed in the actual region $o \in \mathcal{O}$, where $\mathcal{O}$ is the set of all true candidate TFBS and $|o|$ is the genomic length of o. Similar to the normalized TF-gene scores introduced in Chapter 3, we compute length normalized TF affinities $a'_{r,t}$ and $a'_{o,t}$ for TF t in each region $r \in \mathcal{R}$ and

$o \in \mathcal{O}$, respectively:

$$a'_{r,t} = \frac{a_{r,t}}{|r|}, \forall r \in \mathcal{R}, \tag{5.1}$$

$$a'_{o,t} = \frac{a_{o,t}}{|o|}, \forall o \in \mathcal{O}. \tag{5.2}$$

Using the distribution of $a'_{r,t}$ across all $r \in \mathcal{R}$, we calculate a threshold z_t for TF affinities according to a p-value cut-off of c by considering the $1-c$ quantile of $a'_{r,t}$ ($\forall r \in \mathcal{R}$) to determine the value of z_t. With respect to z_t we can calculate a binary affinity value $b_{o,t}$ indicating whether TF t binds to region o or not:

$$b_{o,t} = \begin{cases} 1, a'_{o,t} > z_t, \\ 0, else. \end{cases} \tag{5.3}$$

The binary TF affinity scores $b_{o,t}$ are used to compute a binary TF-gene link $a_{g,t}$ for gene g and TF t:

$$a_{g,t} = \begin{cases} 1, \exists\ o \in \mathcal{O}_{g,w} : b_{ot} = 1, \\ 0, else. \end{cases} \tag{5.4}$$

In this application, $\mathcal{O}_{g,w}$ refers to all footprint regions that occur within a window of size w, which is centered at the TSS of gene g. Together with time-series gene-expression data, binary TF-gene links can be used as input to DREM.

5.3.3 Validation of TF-specific affinity cut-offs using ChIP-seq data

In order to ensure that the affinity thresholding introduced in the previous section leads to a good distinction between bound and unbound TFBSs, we compared the predicted to experimentally determined TFBSs. In this comparison, ENCODE TF ChIP-seq data for K562, HepG2 and GM12878 as well as H3K27ac data is used (see Section B.3 for details).

Candidate TFBS were determined by footprint calling with HINT-BC [G+16b]. TF affinities were computed both in the footprints as well as in the set of random sequences that reflect the characteristics of the footprint regions in terms of length and CG content. The script to compute the random sequences was provided by Peter Ebert (MPI for Informatics, Saarland University).

To see how the binarization relates to *in vivo* TF-binding data, we computed TF affinity thresholds for several p-values: 0.01, 0.025, 0.05, 0.075, 0.1, 0.2, 0.3, 0.4 and 0.5. Affinities that are smaller than the selected threshold z_t are set to zero, the remaining ones are assigned to one. We assess the accuracy of the discretization using a peak-centric measure in terms of precision and recall. The peak-centric performance measure has been suggested before in Cuellar-Partida *et al.* [CP+12] and has also been used in Schmidt *et al.* [S+18b]. Here, *"the positive set of the gold standard is comprised of all ChIP-seq peaks that contain a motif predicted by* FIMO *[G+11], the negative set contains all remaining ChIP-seq peaks. A prediction*

is counting as a true positive (TP) if it overlaps the positive set, it counts as a false positive (FP) it if overlaps the negative set. The number of false negatives (FN) is the number of all entries in the positive set that are not overlapped by any prediction" [GSo18]. We compute precision and recall as introduced in Section 2.2.2.

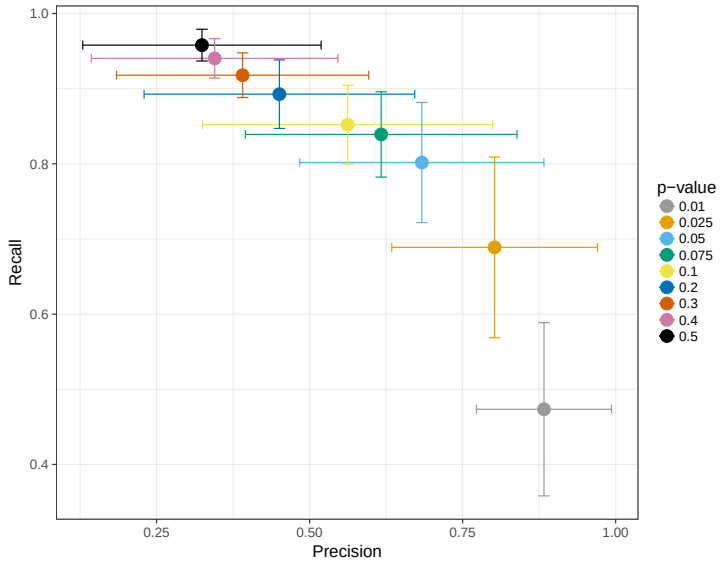


Figure 5.3: This figure shows mean precision and recall computed in a TF ChIP-seq comparison across 36 TFs in HepG2, 18 TFs in K562 and 24 TFs in GM12878 for several p-values to assess the performance of the TF affinity binarization. Figure following Supplementary Figure 2a from Gérard *et al.* [GSo18] (open access).

As indicated in Figure 5.3, choosing a stricter p-value threshold improves precision at the cost of lowering recall. Because 0.05 appears to achieve an acceptable trade-off between precision and recall, we used this p-value throughout our work in Gérard *et al.* [GSo18].

5.3.4 Considered related methods

We are comparing Epic-Drem against three related approaches: One approach is to use the static regulatory information included in Drem2.0 instead of predicted point-specific TFBS predictions.

Another strategy, which can be seen as a baseline is to use predicted TFBS within the promoter region of all genes. To do so, we computed TF affinities for all genes within 2kb windows, which are centered at the 5′ TSS of the genes and apply the same binarization approach as explained above for Epic-Drem. We term this approach Drem-Trap. It represents a class of methods that is solely based on sequence and annotation and does not consider dynamic changes in the chromatin.

One more sanity test of Epic-Drem is a permutation experiment of the actual Epic-Drem input matrix. Specifically, we permuted the columns 1 (TFs), 2 (target genes) and 4 (time points) of the TF-gene score input matrix. By this permutation, the number of TFs, the number of target genes and the number of time points listed in the matrix are not changed. However, the biological signal of the associations is lost. We refer to this approach as Random.

5.3.5 Aggregation of DREM enrichment scores at split nodes

Drem utilizes a hypergeometric distribution to compute the so called split-score. It measures the association of a TF to genes in a distinct path A at split S. The lower the value of the split score, the stronger is the association between the TF and the genes. The split score for TF t is computed according to

$$\sum_{i=c_A}^{min(c_S,n_A)} \frac{\binom{c_S}{i}\binom{n_S-c_S}{n_A-i}}{\binom{n_S}{n_A}}. \tag{5.5}$$

Here c_A is the total number of genes regulated through TF t in path A, c_S is the number of genes in split S regulated by TF t, n_S is the total number of genes in path A entering split S and n_A is the total number of genes in path A leaving S.

In our application, we perform many tests per split, at most as many as we have TFs in Tepic's TF motif database. Therefore, the split-scores are corrected for multiple testing using Bonferroni correction (Section 2.2.6).

The Bonferroni adjusted Drem split-scores are $-log_2$ transformed and visualized in violin plots to allow a easy comparison between enrichment scores for various inputs. In the remainder of the manuscript, we refer to these scores as Drem split score. The higher the value of the Drem split score the stronger is the association of the TFs to its target genes across the split points.

While the Drem split scores provide insights on the importance of regulators at a distinct bifurcation event, we also want to find a ranking of all TFs that belong to bifurcation events at a distinct time point. To do so, we aggregate individual p-values following Fisher's method:

$$X_{t,n} = -2\sum_{j=1}^{s} \log(p_{t,j,n}). \tag{5.6}$$

Here, $X_{t,n}$ refers to the combined score for TF t at time point n, j refers to the current split, s is the number of splits at time point n and $p_{t,j,n}$ refers to the split score for TF t in split j at time point n. Sorting $X_{t,n}$ according to time point n provides a ranking of the most influential TFs per time point.

5.3.6 Generation of TF-TF interaction networks

To improve the interpretability of Epic-Drem results, we generated TF-TF interaction networks at individual split points in the global temporal map. At a split of interest, we obtain the top 25 regulators $t \in \mathcal{T}$, ranked by the Drem split score.

For each regulator $t \in \mathcal{T}$, we determine its target genes $\mathcal{G}_t$ considering any gene g to be a target of TF t, if and only if $a_{g,t} = 1$. To maintain the networks readable and to avoid them to be visually overcrowded, only interactions among the top regulators are shown. For a directed edge from t to g in the TF-TF network to be present, it is required that $g \in \mathcal{T}$. For each TF t, we only show the top 10 interactions, ranked by the numerical affinity values across all $a'_{g,t}$. Furthermore, the diameter of each node in the network is scaled according to the total number of target genes $|\mathcal{G}_t|$ of a TF $t \in \mathcal{T}$. We considered a set of several intervals: $\{[0, 2000], [2001, 4000], [4001, 6000], [6001, 8000], [8001, 10000], [10000, \infty[\}$. Thereby, the importance of the regulators can be judged also aside from the top TFs for the depicted split. Expression changes of TFs compared to time point zero of the time-series data are color coded: orange indicates down-regulation, blue indicates up-regulation. Networks visualization is performed with GRAPHVIZ and the NEATO layout algorithm [GKN05].

5.4 A data driven approach to investigate mesenchymal multipotency

5.4.1 Experimental setup and preliminary analysis

Our collaborators performed epigenomic and transcriptomic profiling for the differentiation of mouse multipotent bone marrow stromal ST2 cell line cells to both adipocytes and osteoblasts. RNA-seq and ChIP-seq experiments screening for H3K4me3, H3K27ac and H3K36me3 were performed at six different time points during the differentiation which lasted for 15 days: day 0, 1, 3, 5, 9 and 15. Three replicates were generated at each time point. The successful differentiation of ST2 cells to adipocytes and osteoblasts was verified via microscopic inspection with regard to cellular morphology as well as using osteoblast and adipocyte specific marker genes [GSo18]. A PCA performed on the obtained RNA-seq data illustrates that two distinct lineages are formed (Figure 5.4).

In total, our collaborators identified 5156 significantly differentially expressed genes using DESEQ2 [L$^+$14c] across all time points in adipocytes and 2072 significantly differentially expressed genes in osteoblasts. A gene counts as differentially expressed if it passes an FDR threshold of 0.05 and has a $\log_2$ fold change of at least 1. Of all these genes 1401 are affected in both lineages, all be it often in lineage specific ways, i.e. they are up-regulated in one lineage, but down-regulated in the other one. We used EPIC-DREM as a data driven approach to identify regulators linked to these differential gene-expression patterns. Experimental details and an overview of the data generated in the scope of this project is listed in Section B.3.

5.4.2 Application of EPIC-DREM

Time point-specific regulatory input

As input for EPIC-DREM, we used TF footprints in the H3K27ac data called with HINT-BC [G$^+$16b]) (version 0.9.9). Next, we computed binary TF-gene scores (c.f.

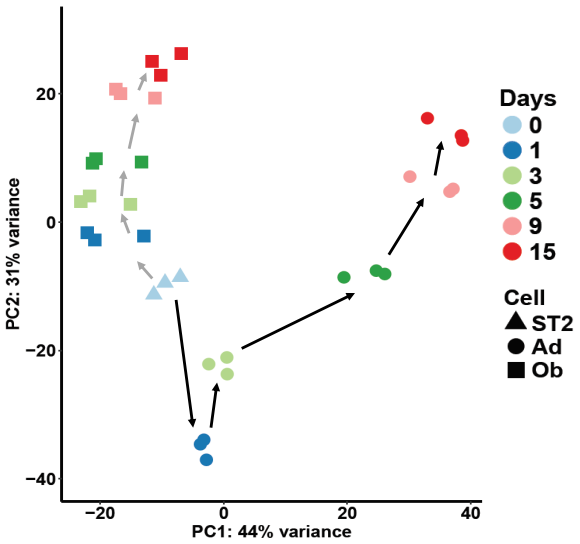


Figure 5.4: PCA on the time-series gene-expression data. Different cell types are indicated by the shape of the points, their color indicates different time points. Replicates of identical cell types and time points cluster together describing the differentiation of the cells in PCA space. Analysis performed by D Gérard. Figure following Figure 1b from Gérard *et al.* [GSo18] (open access).

Section 5.3.2) exploiting the H3K27ac footprints as candidate TFBS. We applied a p-value threshold of 0.05 to determine binary TF affinity values and considered a window of size 50kb centered at the 5′ TSS of genes to aggregate the affinities to TF-gene scores. In addition to all raw data (Section B.3), the TF-gene score matrices are available online at datadryad.org using doi:10.5061/dryad.r32t3.

Here, we explicitly decided to use H3K27ac footprints as these would rather be linked to enhancers than the H3K4me3 footprints, which are more closely linked to active promoters. As cell type specific regulation is typically mediated by enhancers, we focused our analysis on these regions. In total, we obtained TF-gene scores for 687 TF motifs specific for *Mus musculus* included in the Tepic repository using the mouse genome version mm10 (GRCm38).

Output of EPIC-DREM

Applying Epic-Drem to our time-series data sets resulted in two temporal maps for adipocyte and osteoblast differentiation, respectively. They are shown in Figure 5.5a and 5.5b. As explained in Section 5.2, Drem clusters co-expressed genes across different time points, infers bifurcation events indicating a branching of gene-expression development and links transcriptional regulators to these split points. Because the total gene regulatory networks are consisting of several thousands nodes and edges, they can not be visualized completely. Therefore, we show only a refined set of TF-TF networks illustrating the top 25 TFs associated with distinct paths in the temporal maps. We refer the reader to the Supplement of Gérard *et al.* [GSo18] for a complete list of all regulators at each path. The Epic-Drem results are also available online at `datadryad.org` under doi:10.5061/dryad.r32t3.

The diameter of nodes in the TF-TF networks illustrates that TFs involved in regulatory events right after the differentiation initiation are associated with the highest number of predicted target genes. An extreme case is the TF HES1, with more than 10,000 predicted targets. Interestingly, HES1 is a known regulator for both adipogenesis and osteoblastogenesis [Far06, H$^+$04].

A detailed investigation of the shown cell type specific TF-TF networks revealed that many of the found top TFs are established activators such as KLF5, CEBPA and TGIF2 as well as known repressors of adipogenesis, for instance, HES1, NR4A3 and FOXC1 [Far06, H$^+$08, O$^+$14, C$^+$08]. In case of osteoblastogenesis, HES1, TEAD2 and BHLHE40 are examples for known transcriptional regulators [H$^+$04, H$^+$14a, I$^+$06].

Suggested shared regulators between adipo- and osteogenesis

The major objective of studying two parallel lineages was the identification of common regulatory factors that could mediate the (de)differentiation of both adipocytes and osteoblasts. Therefore, we merged the Drem split scores obtained from all bifurcation points per time point and derived separate lists for the top 20 TFs that have the strongest associations at each time point in both lineages (Figure 5.6).

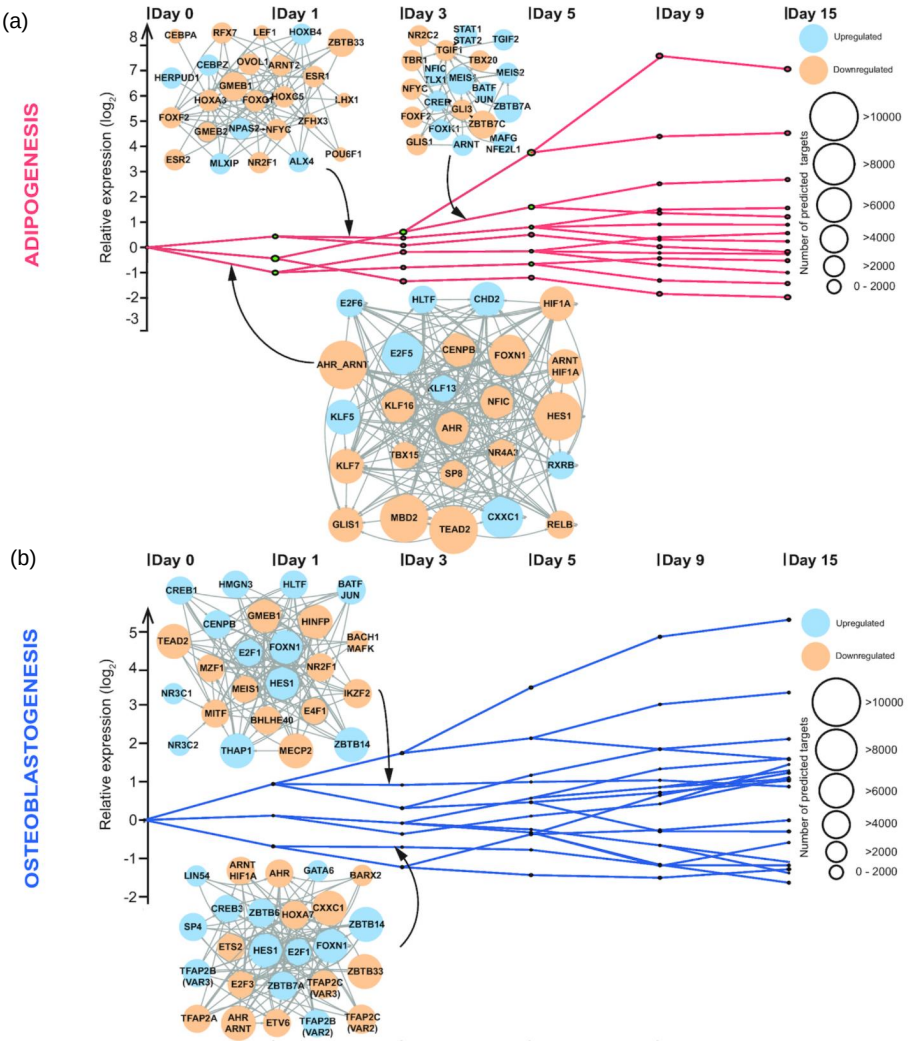


Figure 5.5: Temporal maps of gene-expression development derived by EPIC-DREM for the differentiation of adipocytes and osteoblasts, shown in (a) and (b), respectively. At split points, DREM maps TFs, according to their time point-specific binding scores to gene-expression changes occurring over time. Thousands of regulatory interactions are derived for each unique path in the map, enabling us to compute TF-TF networks that represent the interplay of the regulatory factors. Here, TF-TF networks for the top 25 TFs are shown for two selected paths in osteoblast differentiation as well as for three selected paths in adipocyte differentiation. Whether the expression of a TF is up- or down-regulated compared to its expression in ST2 cells is indicated by a color code. Blue represents up-regulation whereas orange represents down-regulation. Figure based on Figure 3ab from Gérard *et al.* [GSo18] (open access).

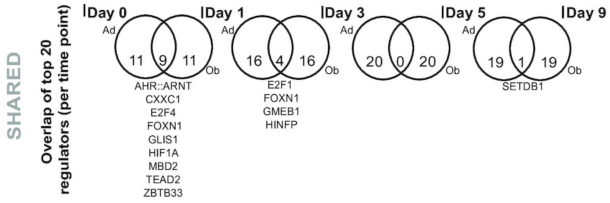


Figure 5.6: For both adipocytes and osteoblasts rankings of the top 20 TFs per time point are assembled by aggregating all predictions from the individual splits. The overlap between lineage-specific top TFs at each time point as well as the shared top TFs are shown in the Venn diagrams. Figure based on Figure 3c from Gérard *et al.* [GSo18] (open access).

We overlapped the top 20 TFs at each time point to unravel how many and which TFs are shared among the top ones. As expected, we found the highest number of shared regulatory factors in the beginning of the differentiation, where 9 out of 20 TFs are present in both osteoblasts and adipocytes (Figure 5.6). It is likely that TFs that are common in both lineages at the early stage are essential for maintaining and orchestrating the multipotent state of the ST2 cells. Therefore, these TFs might act as repressors of the differentiation process.

In fact, AHR::ARNT, E2F4, GLIS1, HIF1A and TEAD2 have been identified to be be involved in gene regulation within different stem cell types [G^+17a, S^+11a, M^+11, L^+17a, L^+17d, F^+10a, T^+11]. Another interesting factors is FOXN1, which is a shared TF at two time points and is also highly connected in TF-TF networks. However, the factor seems to carry out opposing roles in the two lineages, as it is highly expressed in osteoblasts, but down-regulated in adipocytes [GSo18].

Quality assessment of Epic-Drem predictions

To ensure that Epic-Drem's predictions are reliable, we compare Epic-Drem to alternative methods, as described in Section 5.3.4. Specifically, we use ChIP-seq datasets of TFBS as provided in DREM2.0 as input. These are not time point-specific. Also, we use Drem-Trap, where TFBS predictions are computed in 2kb windows centered at the 5′TSSs of all genes, neglecting any epigenomic data. Further, we use a randomized version of the actual Epic-Drem feature matrix.

A first indication for the reliability of Epic-Drem is that the Drem split scores, which are explained in Section 5.3.5, are overall higher for Epic-Drem as compared to the other tested methods (Figure 5.7). This suggests that TFs prioritized by Epic-Drem have a superior explanatory power for gene-expression dynamics than those inferred by the other methods.

To test the reliability of Epic-Drem in yet another way, we combined the top 15 TFs with the highest Drem split scores from each split at day 0 into lists of potential master regulators suggested for both lineages and each prediction approach

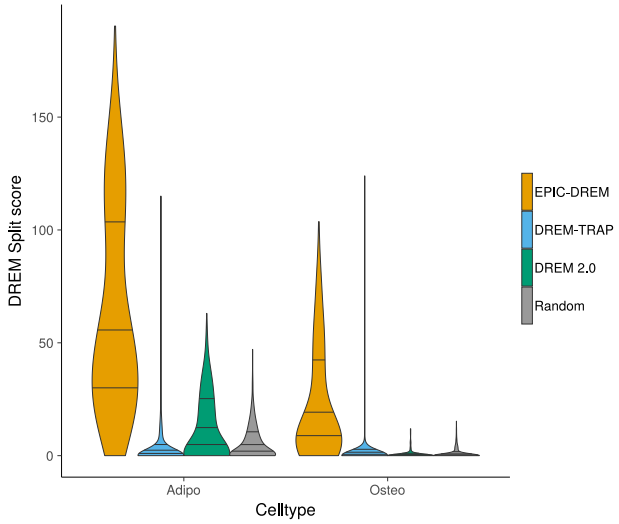


Figure 5.7: The distributions of the computed DREM split scores per method and lineage are depicted. EPIC-DREM typically obtains the highest scores at most split points, indicating a superior explanatory power of gene-expression differences through EPIC-DREM. Figure based on Figure 2c from Gérard *et al.* [GSo18] (open access).

(see Section 5.3.5 for details on the aggregation). We examined the literature for evidence indicating that the predicted TFs in these lists have been previously report to be involved in adipogenesis or osteoblastogenesis. The complete list showing all references can be found in the Supplement of Gérard *et al.* [GSo18].

As shown in Figure 5.8, only $20\% - 30\%$ of the identified factors using the RANDOM scores are mentioned in the literature to be related to osteogenesis or adipogenesis. This percentage increases to $53\% - 59\%$ using DREM-TRAP or DREM2.0. EPIC-DREM yields an even better precision: 92% of TFs linked to adipogenesis and 74% of TFs linked to osteoblastogenesis are known in the literature [GSo18].

Both evaluation approaches indicate that considering the epigenomic landscape at different time points indeed improves the prediction results that can be obtained with DREM. This also strengthens the findings of our application of EPIC-DREM in analyzing the differentiation of adipocytes and osteoblasts.

5.5 Mapping of super-enhancers to their targets suggests regulatory factors as well

To complement our analysis with EPIC-DREM, our collaborators generated super-enhancer (SE) profiles using the generated H3K27ac acetylation data. Methodological details on SE identification are provided in Section B.3. They hypothesized that key transcriptional regulators of the studied differentiation processes would also be under strong regulatory control, potentially by temporal changes in the activity of

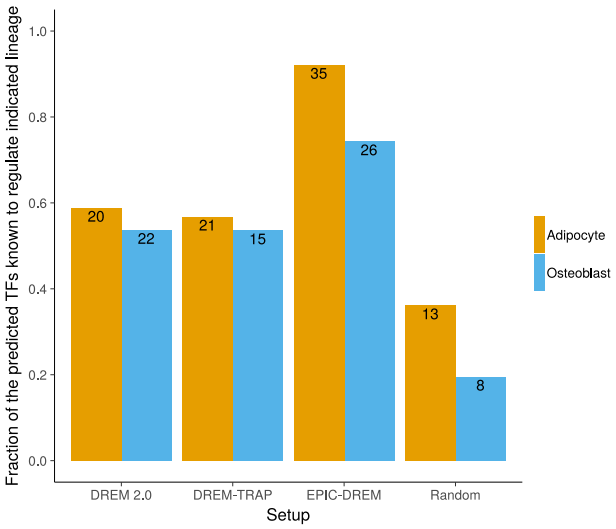


Figure 5.8: Benchmarking of DREM2.0, EPIC-DREM, DREM-TRAP and the randomized EPIC-DREM input matrix using regulators of adipo- or osteoblastogenesis, respectively, identified across all time points of the differentiation. We consider the top 15 TFs sorted by their DREM score across all bifurcation events in the initial split node at day 0. For all identified TFs, we conduct a literature search to check for evidence suggesting a function of the TFs in osteoblastogenesis, adipogenesis or in maintaining multipotency. The fraction of the suggested TFs that are backed up by literature is shown for each method and lineage separately. Further, the total number of identified top TFs is provided at the top of the bars. Figure following Figure 2b from Gérard *et al.* [GSo18] (open access).

SEs. In their analysis, a SE must be a region of at least 10kb that is enriched for H3K27ac signal.

To quantify changes in SE signal across time, overlapping SEs are merged to form one region capturing the entire genomic space covered by the SEs at different time points. To describe changes in the regulatory activity of SEs, H3K27ac ChIP-seq read counts are normalized to day 0. A SE is declared to be dynamic, if it shows a $\log_2$ fold change ≥ 1 in at least one time point. Applying this criterion resulted in a list of 120 and 79 dynamic SEs for adipocytes and osteoblasts, respectively [GSo18]. These SEs are assigned to putative target genes by calculating the Pearson correlation between the H3K37ac signal within the SEs across replicates and time points to the gene-expression of all genes located within 500kb up- and downstream of the individual SEs. As regulatory interactions are typically limited to topological associated domains, this is a reasonable distance threshold. Following this strategy 151 genes were identified.

Importantly, the SE analysis identified the genes *Ahr*, *Glis1* and *Hoxa10* to be

under SE control in both adipo- and osteoblastogenesis. Interestingly, *Ahr* was linked to four distinct SEs, more than any other regulatory factor in the SE analysis. Especially AHR and GLIS1 are promising candidates as they are suggested to be top regulators by Epic-Drem, occurring in the TF-TF networks at the early stages of differentiation, with GLIS1 occurring also within the top 25 regulators at day 3 of adipogenesis [GSo18].

The SE analysis together with the transcriptomics data suggested that *Ahr* is strictly repressed during adipogenesis from day 1 on wards, while the repression in osteoblastogenesis is more linear during the differentiation. Similar observations are made for *Glis1* although the expression pattern of *Glis1* in adipocytes shows an induction at day 5 [GSo18]. These findings indicate that both AHR and GLIS1 act as gatekeepers of mesenchymal multipotency, which might explain their repression during differentiation. Due to the strong evidence for a biological role of AHR and GLIS1 in both adipo- and osteoblastogenesis, our collaborators decided to validated our predictions for AHR and GLIS1 in the lab using over-expression as well as knock-down experiments.

5.6 Experimental validation of candidate regulators

5.6.1 Over-expression experiments of *Ahr* and *Glis1*

Experimental setup

To test whether omitting the down-regulation of *Ahr* or *Glis1* upon differentiation prevents successful differentiation, our collaborators generated stable doxycyclin inducible ST2 cell lines that are able to over-express *Ahr* or *Glis1* (ST2-TetOn-AHR and ST2-TetOn-GLIS1 cells, see Section B.3 for details).

Inducible CopGFP cells that is cells expressing GFP upon doxycyclin treatment, were generated (ST2-TetOn-GFP). CopGFP expression was confirmed using fluorescence microscopy, after a 24h treatment of the generated ST2-TetOn-GFP cell lines with doxycycline (Dox+). Thereby the inducibility of the cell lines was verified, too.

Upon the positive control experiment, all generated cell lines were differentiated towards either adipocytes or osteoblasts in two different conditions: Either in presence (Dox+) or in absence (Dox-) of doxycycline. To study the impact of the over-expression of the factors, RNA and protein samples were extracted at day 5 and day 9 of differentiation.

The successful induction of *Ahr* and *Glis1* expression is shown using RT-qPCR and, in case of AHR, also with western blotting. Western blotting is not performed for GLIS1 due to the absence of a specific antibody. We note that for both TFs slightly elevated expression levels were also detected in the Dox- conditions, which argues for a baseline activity of the Tet-On 3G promoter in the absence of doxycycline as well. However, the induction in Dox- is several fold less compared to Dox+ conditions [GSo18].

Impact on adipocytes

The consequences of the over-expression of AHR and GLIS1 on adipocyte differentiation are assessed in two ways. Firstly, our collaborators have performed *Oil Red O staining* of lipid accumulation on day 5 and day 9 of adipogenesis.

A faint red staining in ST2-TetOn-GFP control cells at day 5 shows a minor accumulation of lipids. However, such staining is neither observed for ST2-TetOn-AHR nor ST2-TetOn-GLIS1 at day 5. At day 9, the control cells accumulated many lipids, resulting in a strong red staining in both Dox+ and Dox- setting. In contrast to that, AHR expressing cells show no lipid accumulation at all in Dox+ setting and only accumulation of lipids in the Dox- setting (Figure 5.9a).

For GLIS1 expressing cells, the behavior after day 5 could not be screened by Oil Red O staining because cells expressing GLIS1 are loosing their adherence with progressing differentiation [GSo18].

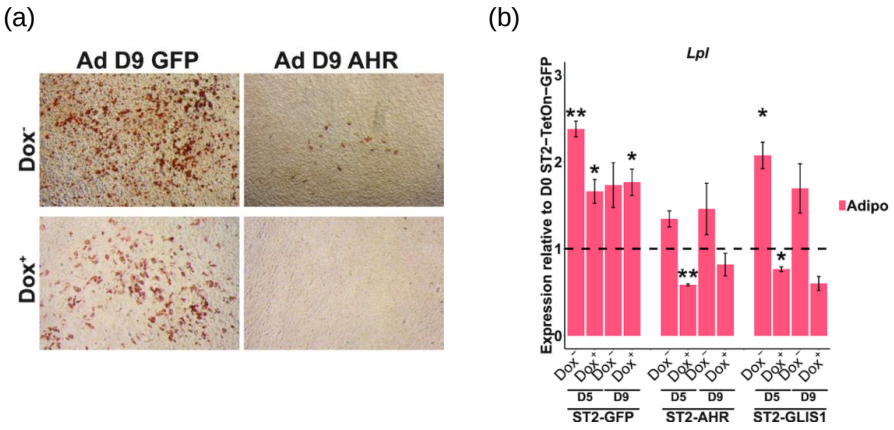


Figure 5.9: (a) Oil Red O staining at day 9 of adipocyte differentiation. In the AHR over-expressing Dox+ sample, no staining can be observed. (b) Statistical significance for RT-qPCR measurements of *Lpl* expression is assessed by a comparison to the *Lpl* expression in the undifferentiated ST2-TetOn-GFP cells by a one sample t-test. $* = p < 0.05, ** = p < 0.01$ and $*** = p < 0.001$. Data points represent the mean of 3 independent stable cell lines. Figure following Figure 6de from Gérard *et al.* [GSo18] (open access).

Secondly, RT-qPCR of the known adipocyte marker gene Lpl was conducted. The RT-qPCR results, depicted in Figure 5.9b support the Oil Red O staining: While in ST2-TetOn-GFP cells Lpl is up-regulated by day 5 and remains constantly high till day 9, independent of doxycycline. Upon *Ahr* over-expression in Dox+ cells, no additional induction in ST2-TetOn-Ahr cells was observed. The same is observed for *Glis1* over-expressing cells. Taken together these results indicate an unsuccessful differentiation towards adipocytes upon the over-expression of either *Ahr* or *Glis1* [GSo18].

Impact on osteoblasts

The impact of *Ahr* over-expression on osteoblastogenesis can not be clearly assessed, because doxycycline caused an increase in the expression of the marker gene Sp7, suggesting that doxycycline itself is involved in osteoblastogenesis, which has also been reported in the literature [W+99, G+17c]. Nevertheless, *Glis1* over-expression

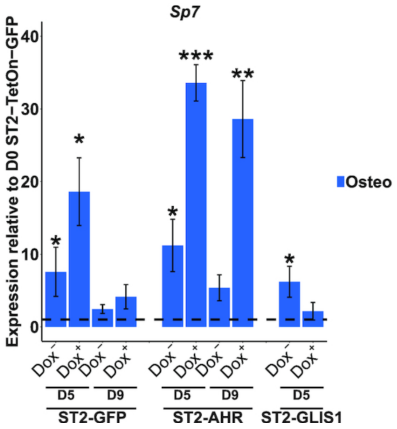


Figure 5.10: Statistical significance for RT-qPCR measurements of *Sp7* is assessed by a comparison to its expression in undifferentiated ST2-TetOn-GFP cells by a one sample t-test. $* = p < 0.05, ** = p < 0.01$ and $*** = p < 0.001$. Data points are the mean of 3 samples from independent stable cell lines. Figure based on Figure 6f from Gérard *et al.* [GSo18].

in ST2-TetOn-GLIS1 cells leads to a complete loss of Sp7 in day 5 in Dox+ condition, suggesting that *Glis1* expression prevents osteoblast differentiation (Figure 5.10). Because, *Glis1* over-expressing cells undergo increased rates of cell death especially in osteoblastogenesis, data for day 9 could not be collected [GSo18].

5.6.2 Silencing

Our collaborators have performed a knock-down of endogenous *Ahr* and *Glis1* in ST2. To confirm a successful differentiation, the expression of lineage-specific marker genes (Adipocytes: Cebpa, Pparg, Lpl; Osteoblasts: Runx2, SP7, Bglap) was tested (Figure 5.11a). For *Ahr* a knock-down of 50% is shown for mRNA as well as protein, while the mRNA reduction of *Glis1* is around 30%. Due to the lack of a GLIS1 specific antibody, the reduction of the protein can not be measured. Nevertheless, GLIS1 knock-down leads to a significant induction of the marker genes *Cebpa*, *Lpl* and *Bglap* expression, while AHR knock-down affects the *Lpl* marker gene, consistent with our results from the over-expression experiments. These findings indicate that both AHR and GLIS1 are involved in maintaining the multipotent state of ST2 cells [GSo18].

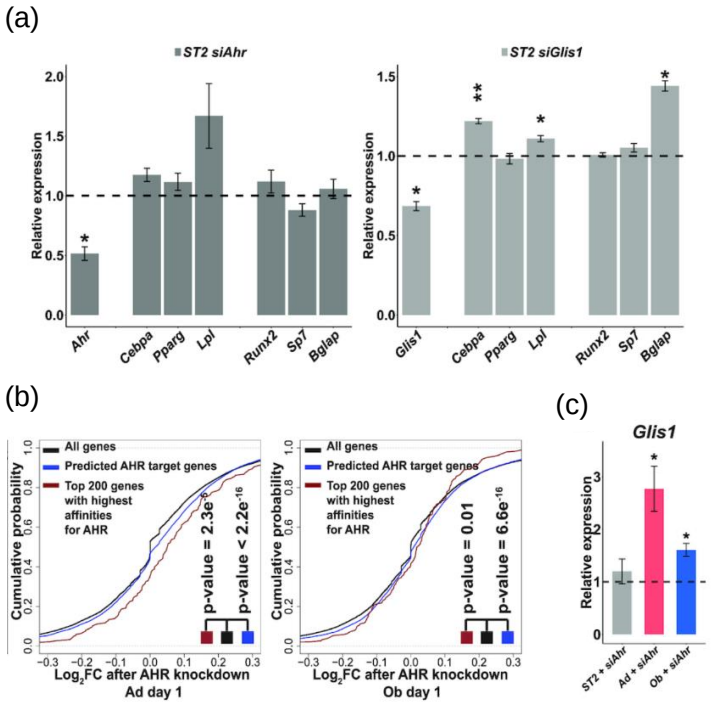


Figure 5.11: Validation silencing: (a) The y-axis shows the relative expression of marker genes for adipocyte and osteoblast differentiation upon *Ahr* or *Glis1* knock-down in ST2 cells. Statistical significance for RT-qPCR measurements of knock-down cells against cells transfected with siControls was determined by a two-tailed Student's t-test: $* = p < 0.05, ** = p < 0.01$ and $*** = p < 0.001$. Data points represent the mean of 3 independent samples from stable cell lines. (b) The cumulative distribution of the log2 fold change upon the depletion of AHR is shown separately for all expressed genes, all AHR target genes predicted by EPIC-DREM as well as for the top 200 predicted AHR target genes ranked by AHR affinities. A Kolmogorov-Smirnov test was used to assess the significance of the fold changes. (c) The relative expression of *Glis1* mRNA upon *Ahr* knock-down is shown for ST2 cells, cells differentiating to adipocytes and cells differentiating to osteoblasts. Figure following Figures 7a, 7d and 7e from Gérard *et al.* [GSo18] (open access).

Because the *Ahr* knock-down is more robust than the *Glis1* knock-down, we continue our analysis with *Ahr* knock-down only. To assess whether the endogenous activity of AHR is related to keep ST2 cells in their multipotent state, *Ahr* was knocked-down in ST2 cells. Those have been differentiated towards both osteoblasts and adipocytes. Gene-expression was measured two days after the knock-down using RNA-seq.

We used the *Ahr* knock-down information to further validate Epic-Drem's predictions. That is, we tested whether the depletion of AHR had an impact on the targets of AHR predicted by Epic-Drem predicted AHR targets at the corresponding time point were indeed affected by depletion of AHR [GSo18].

Figure 5.11b illustrates that the predicted AHR targets are significantly more affected by the AHR knock-down than all genes on average for both adipogenesis and osteoblastogenesis. This holds especially for the genes with the highest TF affinity scores for AHR at day 1 of adipogenesis, as those genes are clearly up-regulated by the knock-down. This is indicated by a shift to the right in the cumulative distribution plot. These results stress the reliability of Epic-Drem's predictions and further strengthen the role of AHR as a repressor. Furthermore, the TF-TF networks shown in Figure 5.5 states that GLIS1 is regulated by AHR at day 0 during early adipogenesis. Via RT-qPCR, we test this prediction and show a significant change in Glis1 expression upon AHR depletion in adipocytes as well as in osteoblasts (Figure 5.11c), which further supports the predictions of Epic-Drem [GSo18].

5.6.3 Conclusions made for mesenchymal differentiation

Taken together, the experimental results support a role of AHR and GLIS1 as likely *"guardians of mesenchymal multipotency"* [GSo18]. Importantly the wet-lab experiments validate not only the general prediction of Epic-Drem that both AHR and GLIS1 act as important regulatory factors, but also showed that even detailed aspects of our predictions are correct, e.g. the target genes of AHR.

5.7 Interactive visualization of dynamic regulatory networks (iDREM)

During the revision of our manuscript, a successor of the original Drem method called iDrem [D$^+$18] has been published. Their new method attempts to integrate more datasets then the original Drem approach such as time-series *proteomics* and epigenomics data. Another interesting feature is the possibility to superimpose single-cell RNA-seq data to the temporal maps generated from bulk data.

A major difference between iDrem and Epic-Drem is how the additional input information is incorporated into the model. In iDrem proteomics data can be used to model the activity of TFs that is TF activity acts as a prior for the static regulatory information included in Drem and can thus prioritize highly active TFs and at the same time neglect inactive TFs. Epigenomics data such as DNA methylation or HM ChIP-seq data is incorporated into the model in the same way.

In contrast to that, in Epic-Drem we consider time point-specific regulatory predictions, which is a conceptually different approach as no ChIP-seq data is used and only regulatory sites that are most likely active are considered at all. An advantage of Epic-Drem compared to iDrem is that more regulatory factors can

be included in the model and that more species can be considered, because the build-in set of regulators is no limiting factor.

Overall, the simultaneous development of Epic-Drem and iDrem irrevocably indicates the need for novel approaches to integrate both time-series transcriptomics and epigenomics data to gain better insights on transcriptional regulation.

5.8 Contributions of all researchers involved in the described project

The following people contributed to the work presented in this Chapter: Deborah Gérard (University of Luxembourg), Florian Schmidt (Saarland University), Aurélien Ginolhac (University of Luxembourg), Martine Schmitz (University of Luxembourg), Rasi Halder (University of Luxembourg), Peter Ebert (MPI Inf, currently at Saarland Univeristy), Marcel H Schulz (Saarland Univeristy, currently at Göthe University Frankfurt), Thomas Sauter (University of Luxembourg) and Lasse Sinkkonen (University of Luxembourg).

The overall project was designed by Deborah Gérard, Thomas Sauter and Lasse Sinkkonen. Deborah Gérard and Aurélien Ginolhac performed the RNA-seq and ChIP-seq analysis. Martine Schmitz and Deborah Gérard performed western blotting. The knock-down experiments were performed by Rasi Halder. All other wet-lab experiments as well as bioinformatics data preprocessing such as alignment, peak and footprint calling as well as initial analysis was performed by Deborah Gérard (Figures 5.4, 5.6, 5.9-5.11). Besides, Deborah Gérard performed the SE identification and analysis advised by Marcel H Schulz and Lasse Sinkkonen.

The Epic-Drem approach, as depicted in Figure 5.2, was developed jointly by Marcel H Schulz and Florian Schmidt. Peter Ebert contributed a script to generate random DNA sequences given a template set. The regions provided by that tool are used in an extension of TEPIC written by Florian Schmidt to generate binary TF binding scores (Figure 5.3). Florian Schmidt computed TEPIC predictions for all considered samples and time points, generated TF feature matrices and used them to run Drem. A flexible python script to generate regulatory networks has been written by Florian Schmidt as well (Figure 5.5). Besides, Florian Schmidt performed all benchmark experiments of the Epic-Drem approach (Figures 5.7 and 5.8).

The main manuscript was written by Lasse Sinkkonen, supported by Deborah Gérard, Florian Schmidt and Marcel Schulz.

6

Same same but different - Diversity of chromatin accessibility assays

The work presented here is based on the article by Nordström *et al.* `"Unique and assay specific features of NOMe-seq, ATAC-seq and DNaseI-seq data"` published in the journal Nucleic Acids Research [N+19].

6.1 Motivation and research objectives

In Chapters 3 and 4, we have seen how chromatin accessibility data can be used in various applications to assess the regulatory activity of genomic regions and to predict TFBS. In Section 2.1.10, we introduced three different assays to identify nucleosome depleted regions (NDR) in a cell: DNaseI-seq, NOMe-seq and ATAC-seq. The first two, DNaseI-seq and NOMe-seq, have been used in Chapters 3 and 4 already. Due to the utmost importance of chromatin accessibility assays in understanding gene regulation and functional annotations, we perform a systematic comparison of DNaseI-seq, NOMe-seq and ATAC-seq by generating chromatin accessibility profiles with all methods for identical samples and applying joint bioinformatics analysis.

In light of upcoming single-cell applications, especially of ATAC-seq [C+18a], it is important to characterise potential biases of the assays to avoid a wrong interpretation of the noisy single-cell readout. In an earlier work, Song *et al.* have shown that there are certain regions in the genome exclusively identified by DNaseI-seq and *FAIRE-seq*, respectively [S+11b], arguing for the presence of similar specificity's as well with the more frequently used ATAC-seq and NOMe-seq approaches.

In the remainder of this Chapter, we present our contributions to Nordström *et al.* [N+19], as well as other essential aspects of that paper required for its general understanding.

6.2 Generated data and experimental setup

To directly compare DNaseI-seq, NOMe-seq and ATAC-seq chromatin accessibility profiles, we generated novel data for the HepG2 cell line. All chromatin accessibility assays were performed using an identical stock of cells and under the same cultivation conditions to reduce technical confounding variables in the lab of Jörn Walter at Saarland University. The data was generated according to established IHEC and BLUEPRINT protocols. Because HepG2 is a major cell line used in

ENCODE, several external datasets, for instance, TF ChIP-seq data, are available for further validation. Details on the experimental and computation processing are provided in Section B.4.

6.3 Results

We evaluated the agreement between the three assays both on signal and peak level. The signal based annotation allows an unbiased genome-wide view on the data, while the peak level focuses on truly accessible sites. The three libraries were sequenced at sufficient sequencing depth that is 10x genome wide coverage of GpCs with NOMe-seq, 180 million reads with DNaseI-seq and 60 million reads with ATAC-seq, allowing a reliable quantification of enriched sites.

6.3.1 Signal level

To obtain a peak caller independent, genome-wide impression on the agreement between the different assays, we suggested a comparison based on the actual signal of the assays. To this end, Karl Nordström computed the genome wide GCH methylation levels of the NOMe-seq data as well as FPKM values obtained for DNaseI-seq and ATAC-seq reflecting the number of $5'$ read ends across the genome. Recall that H represents A, C and T. Next, the correlation between the raw signal distributions of the different assays aggregated in 500bp bins was assessed. As shown in Figure 6.1, we observe a Spearman correlation of 0.41 between DNaseI-seq and ATAC-seq data. However, the agreement between DNaseI-seq or ATAC-seq to NOMe-seq is far less with a Spearman correlation value of only 0.25 and 0.21, respectively.

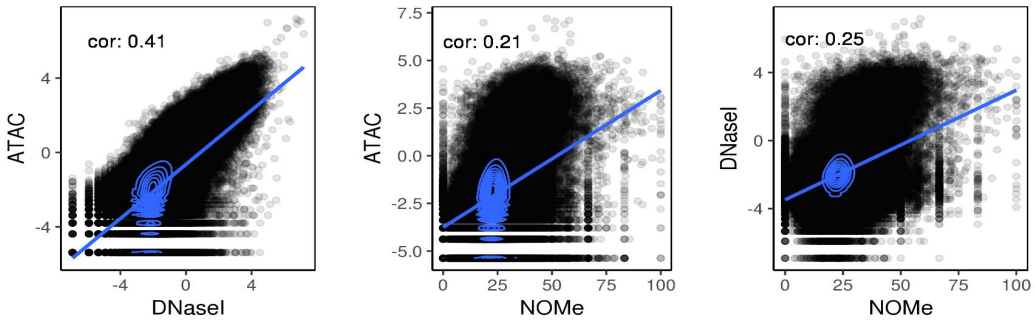


Figure 6.1: Pairwise correlation of the binned genome-wide signal between DNaseI-seq, ATAC-seq and NOMe-seq. Figure based on Supplementary Figure 1 from Nordström *et al.* [N+19].

This genome-wide comparison already illustrates that there seems to be a high variability in the signal between the different methods. The observation that ATAC-seq and DNaseI-seq are more similar to each other than they are to NOMe-seq makes intuitive sense as the ATAC-seq and DNaseI-seq are both enrichment based

methods and thus more similar in terms of the experimental design to each other than they are to NOMe-seq. Furthermore, it has been reported in literature before that ATAC-seq and DNaseI-seq do produce fairly similar chromatin accessibility maps on the peak-level, although (dis)similarities have not been investigated in detail [B+13d].

6.3.2 Peak level

Typically, researches use chromatin accessibility data to identify accessible sites, also known as nucleosome free regions (NFR) or nucleosome depleted regions (NDR), as these can be used for functional annotations and interpretations, c.f. Chapter 3-5. Often, researchers refer to such regions simply as peaks. In this study, peaks in DNaseI-seq and ATAC-seq data were called using MACS2 [Liu18]. For NOMe-seq, we determined the position of NDRs with a HMM termed gNOMeHMM. This method was developed in the lab of Jörn Walter at Saarland University together with Nico Pfeiffer and Marcel H Schulz from MPI for Informatics and Saarland University, respectively. It provides a robust genome wide NDR annotation. Details on gNOMeHMM and on the usage of Macs2 are provided in Section B.4. Figure 6.2 provides an overview of the peak overlap between the assays.

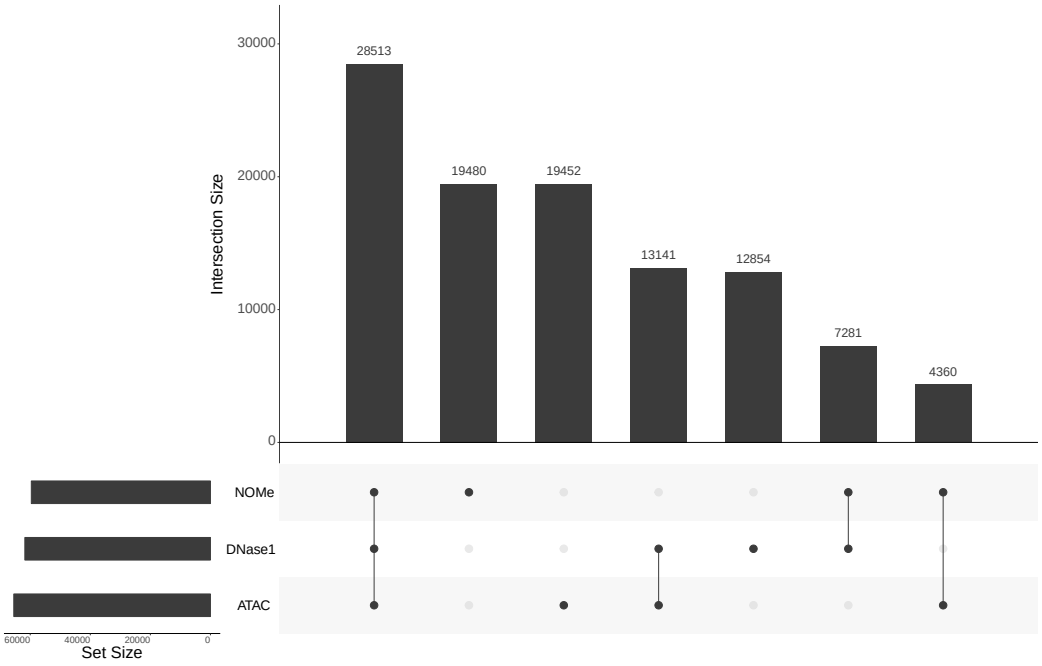


Figure 6.2: Overlap of peak-calls for DNaseI-seq, ATAC-seq and NOMe-seq.

Interestingly, the overall number of detected accessible sites is very similar across the three assays: There are 65,683 NDRs for NOMe-seq, 62,365 for DNaseI-seq and 67,675 for ATAC-seq.

Overall, there are 105,081 NDRs detected by at least one of the three assays. In

total, we find 24% of NDRs to be shared by all methods and 27% to be supported by two methods. However, 49% of the peaks are identified by only one distinct assay. Specifically, 19,480 unique NDRs are identified using NOMe-seq, 12,854 with DNaseI-seq and 19,452 with ATAC-seq.

As could be expected from the signal based comparison, most shared open regions are found between ATAC-seq and DNaseI-seq data. We note that in addition to the already mentioned experimental commonalities, e.g. the enzymatic reaction, also the shared peak calling methodology might have contributed to this result as well.

NDRs commonly reported by all three assays are frequently longer than unique ones and show a pronounced above-average signal in all assays (Figure 6.3a). Accessible sights uniquely identified with NOMe-seq are an exception to this observation. In contrast to the other two assays, the unique NDRs show a slightly stronger signal as opposed to the commonly retrieved ones. In general, assay-unique NDRs show a rather strong signal in the assay the are found with. The signal intensity for those region falls or even vanishes completely in the remaining assays (Figure 6.3b).

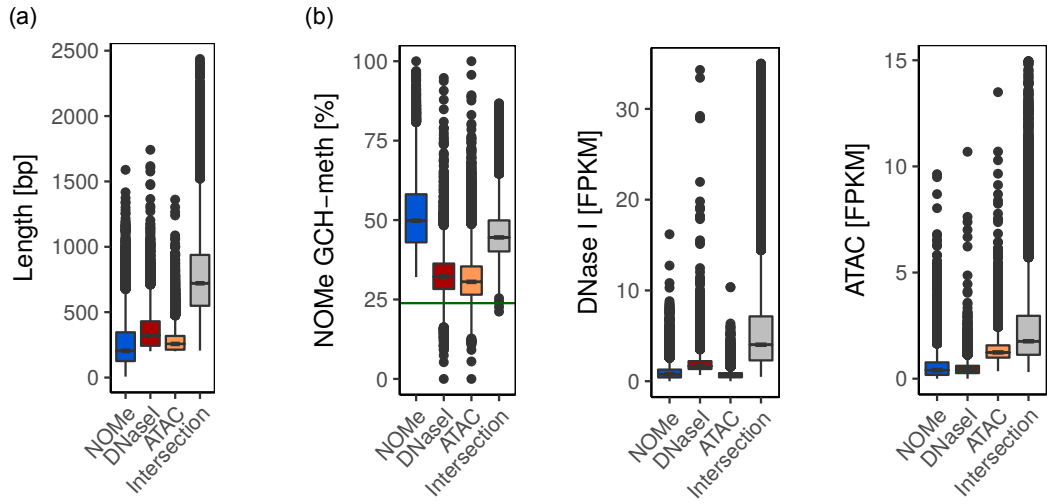


Figure 6.3: (a) Length of unique and common peaks [bp] (b) Signal intensity in terms of NOMe GCH-methylation, DNaseI and ATAC read counts. The green line indicates the average GCH methylation. Figure following Figure 3b-e from Nordström *et al.* [N+19].

6.4 Targeted deep amplicon sequencing of unique NDRs

To assess the correctness of NDRs detected by NOMe-seq, DNaseI-seq and ATAC-seq, 17 NDRs exhibiting distinct recognition patterns across the different assays were selected for validation using targeted deep amplicon bisulfite sequencing followed by a NOMe treatment.

In accessible regions reliably identified by several methods our collaborator's measured a median GCH-methylation level of 38%-70% , while NDRs not called by any

of the assays exhibit a lower median GCH methylation of 9%-26% (Figure 6.4a). An example control locus is HSPA5(up) shown in Figure 6.4b, while MLH1, for instance, is a site that is reliably detected by all assays.

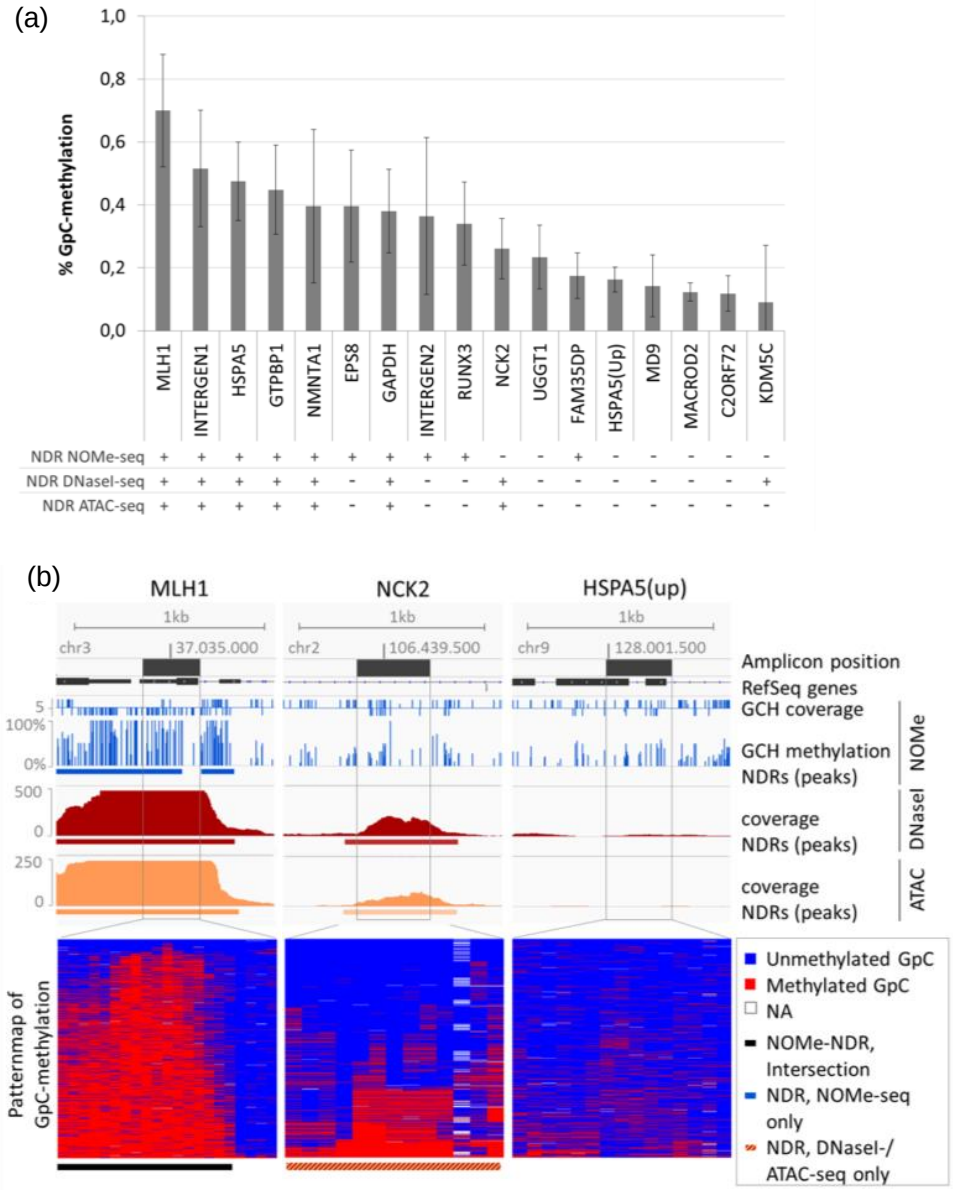


Figure 6.4: (a) GCH methylation signal of the targeted amplicon sequencing (b) ATAC, DNaseI and NOMe signal at three regions: NDR at MLH1, found by all sites, NDR at NICK2, which is not called by the actual NOMe data and a closed control region HSPA5(up). Figure based on Supplementary Figure S9 from Nordström *et al.* [N$^+$19]. Data generated in the lab of Jörn Walter.

Importantly, the deep amplicon sequencing enables us to describe interesting patterns of GCH-methylation in regions not identified by GNOMEHMM on the actual NOMe-seq data, but only by the other assays. One such example is the DNaseI-seq and ATAC-seq specific open region NCK2 (Figure 6.4b). There, we observe an obvious GCH-methylation pattern in a fraction of the sequences suggesting that a proportion of the cells fulfills the criteria of a NOMe-specific NDR. This example not only clarifies that some accessible regions might be neglected by GNOMEHMM due to low GCH-methylation levels but also underlines that unique DNaseI-seq and ATAC-seq NDRs can be truly accessible regions. Another reason why the GNOMEHMM peak calling might miss an open site is a low GCH-density in the respective genomic loci resulting in an insufficient methylation signal.

6.5 Clustering of NDRs is linked to functional associations

To unravel commonalities in the signal patterns among different accessible sites, we computed the average NOMe-signal, ATAC-seq signal and DNaseI-seq signal in 10bp bins spanning a 2kb window centered on NDR summits. NDR summits are determined using NOMe-seq peaks first, following DNaseI-seq, following ATAC-seq only peaks. Using k-means clustering, the resulting signal matrix is clustered into 15 clusters. The obtained clustering is shown in Figure 6.5a.

These clusters were overlapped with a CHROMHMM segmentation for HepG2, computed by Peter Ebert on DEEP HepG2 data [K$^+$15]. Also, an enrichment analysis using TF-ChIP-seq data was carried out with LOLA [SB16]. The Locus Overlap Analysis (LOLA) is an enrichment analysis for sets of genomic regions. The query regions are compared against a curated database of regulatory elements and functional regions obtained, for instance, from ENCODE. The results of the cluster specific LOLA analysis are shown in Figure 6.5b.

According to the CHROMHMM segmentation, regions identified by any accessibility method are typically linked to active transcription start sites (Figure 6.5b, Clusters C3, C9, C11, C12, C13, C15). An enrichment analysis of assay-unique NDRs reveals that NOMe and ATAC unique regions are enriched for CTCF, Rad21 and SMC3 binding sites (Clusters C7), which are interacting in the Cohesin protein complex [G$^+$14b]. Unique DNaseI-seq regions are enriched for binding sites of FOXA1, FOXA2 and HNF4G (Cluster 1).

6.6 Unique accessible regions contribute information to gene-expression prediction models

To gain a better understanding on the regulatory relevance of the various open chromatin regions identified by the different assays and their relation to gene-expression, we trained linear regression models to predict gene-expression, as introduced in Chapter 3.

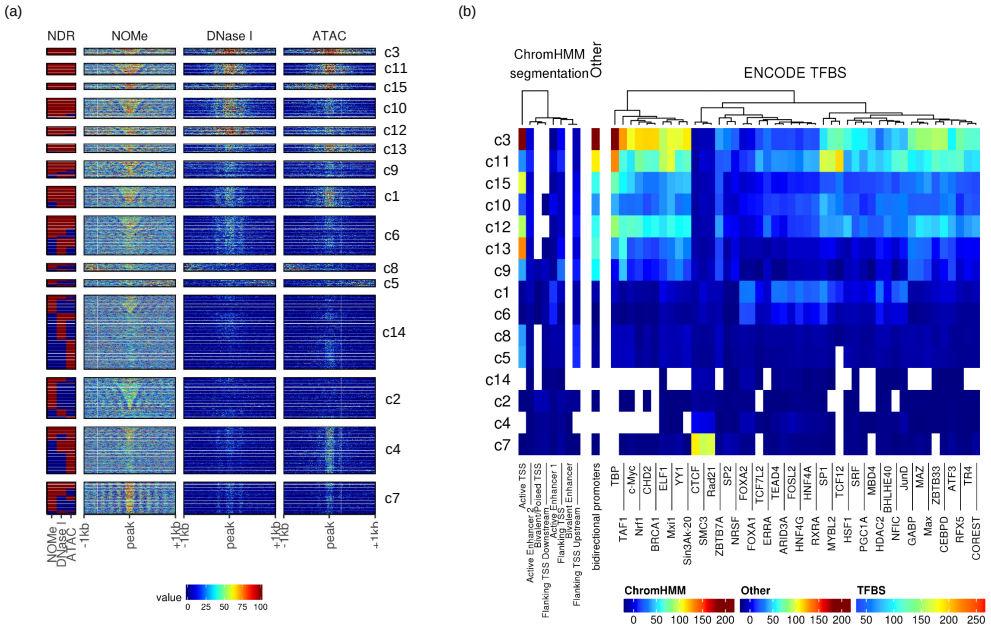


Figure 6.5: (a) Clustering of NDRs according to the epigenomic signal (b) Heatmap showing the enrichment tests result of the clusters from (a) compared to ChromHMM and ENCODE TFBSs. Figure following Figure 4 from Nordström *et al.* [N+19].

6.6.1 Feature definition

In this study, we computed TF-gene scores using TEPIC [S+17a] in the ATAC-seq $\mathcal{A}$, DNaseI-seq $\mathcal{D}$ and NOMe-seq $\mathcal{N}$ NDR sets. In addition, we included the intersection $\mathcal{I}$ of the three sets, as well as their union $\mathcal{U}$:

$$\mathcal{I} = \mathcal{A} \cap \mathcal{D} \cap \mathcal{N}, \tag{6.1}$$

$$\mathcal{U} = \mathcal{A} \cup \mathcal{D} \cup \mathcal{N}. \tag{6.2}$$

Moreover, we considered NDR sets extending $\mathcal{A}$, $\mathcal{D}$ and $\mathcal{N}$ to match $|\mathcal{U}|$ by sampling regions, not overlapping any of the real NDRs ($\mathcal{A}_R$, $\mathcal{D}_R$, $\mathcal{N}_R$). In the end, eight different NDR sets $\mathcal{P} = \{\mathcal{A}, \mathcal{A}_R, \mathcal{D}, \mathcal{D}_R, \mathcal{N}, \mathcal{N}_R, \mathcal{I}, \mathcal{U}\}$ are considered for model training. Here, $\mathcal{P}_j$, refers to the j^{th} set in $\mathcal{P}$.

For each NDR $p \in \mathcal{P}_j$, TF affinities $a_{p,t}$ for TF t are computed using *TRAP* ([R+07]) for a set of 726 TF motifs obtained from the TEPIC 2.0 repository ([S+18b]).

TF affinities $a_{p,t}$ are combined to normalized TF-gene scores $\bar{a}_{g,t}$ for gene g and TF t following the $\mathcal{EN}$ scoring as suggested in Section 3.3.2, Eq. 3.25 – 3.27, using a 3kb window as before.

6.6.2 Linear regression

For each set of normalized TF-gene scores, we learned a linear regression model using elastic net regularization as introduced in Section 3.4.1.

6.6.3 The union of all NDRs achieves the best model performance

The performance in terms of Spearman correlation of all linear models is shown in Figure 6.6.

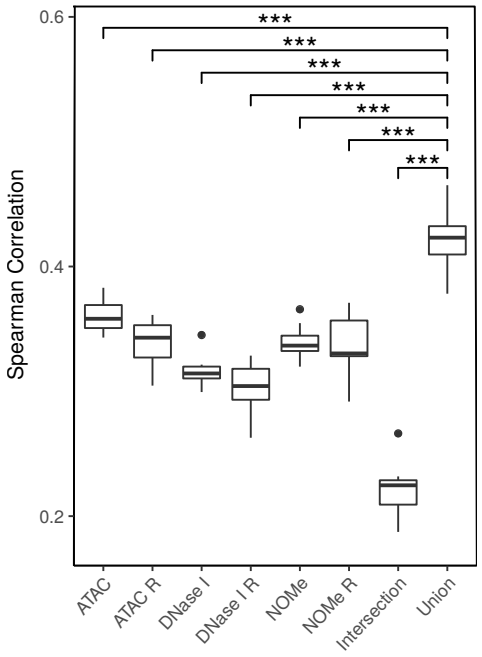


Figure 6.6: Spearman correlation of the linear models predicting gene-expression. Figure following Figure 5b from Nordström *et al.* [N+19].

Evaluating the linear models shows that the worst performance is obtained using the intersection of all three assays $\mathcal{I}$. This is surprising as it indicates that, although these regions have a strong signal, they do miss out many important regulatory sites. NDRs specific models, based exclusively on either the $\mathcal{A}, \mathcal{D}$ or $\mathcal{N}$ set perform comparable to each other with a marginally better performance of ATAC-seq over NOMe-seq and DNaseI-seq.

TF-gene scores based on the union of all three NDR sets $\mathcal{U}$ lead to the best model performance, unmatched by no other NDR set. To ensure that the improved performance is not only due to the larger number of NDRs, we extend each assay specific NDR set with randomly generated peaks to be as large as the union set: $\mathcal{A}_R$, $\mathcal{D}_R$, $\mathcal{N}_R$. Models based on these NDR sets models perform constantly worse than their not extended counterparts $\mathcal{A}$, $\mathcal{D}$ and $\mathcal{N}$.

Overall, we conclude that the individual assays are unable to describe a distinct part of the chromatin landscape. Therefore, the union of assay specific NDRs allows

to model the regulatory landscape of gene-expression with greater accuracy. This conclusion is backed up by the enrichment results presented in Section 6.5, because they illustrate that certain TFs are enriched in NDRs that are exclusively detected by a distinct chromatin accessibility method.

6.7 Shape, sequence and methylation characteristics at the active sites of the participating enzymes

While the previous analyses have shown that the assay specific NDRs are functionally relevant, we attempted to better understand the reason behind why certain regions are uniquely identified by a distinct method. Therefore, we investigated a potential bias of the enzymes used in the various reactions with respect to DNA shape, DNA sequence and DNA methylation.

Specifically, we profiled these three molecular signatures on the 5′-cut-sites/GpC sites retrieved from the aligned reads using BEDTools [QH10]. As pointed out to us by Ivan Costa, ATAC-seq sequences need to be shifted by 4bp upstream, to account for the 3D structure of the Tn5 transposase dimer [B$^+$13d]. In case of NOMe-seq, reads are sampled with respect to the GpC methylation.

6.7.1 DNA shape

The DNAshapeR R-package [C$^+$16a] was used to obtain estimates for the minor groove width (MGW), Roll, propeller twist (ProT) and helix twist (HelT), as introduced in Section 2.1.1. To predict DNA shape characteristics, we randomly selected 2 million sequences per assay, constructed as described above. This reduction is necessary due to memory limitation of the DNAshapeR R-package. All spatial features were predicted in a 31bp window centered at the enzyme active site.

For NOMe-seq, we find a striking signal for an increased Helix Twist (HelT) and Propeller Twist (ProT) at the M.CviPI enzyme recognition site 5'GpC3'. Also, we observe an increased Minor Groove Width (MGW) flanking the GC site. For DNaseI-seq, we predict the MGW to be enlarged around the cut site, as reported before by Lazarovici *et al.* [L$^+$13c], coupled with a slightly increased base roll. In contrast to the monomers, M.CviPI and DNaseI, the Tn5 transposase is a dimer. Consequently, we observe bidirectional changes in MGW, ProT and Roll oscillating around the Tn5 insertion site. The DNA shape predictions are visualized in Figure 6.7. They were reproduced in various other samples, too (Figure B.5).

6.7.2 DNA sequence

Using 59,850,858 DNaseI-seq sequences, 30,108,148 ATAC-seq sequences and 126,202,679 NOMe-seq sequences, we generated sequence logos using the ggseqlogo R-package [Wag17]. Although the bias motifs reported in literature are relatively short that is 6bp for DNaseI-seq [K$^+$13c] and 20bp for ATAC-seq [B$^+$13d],

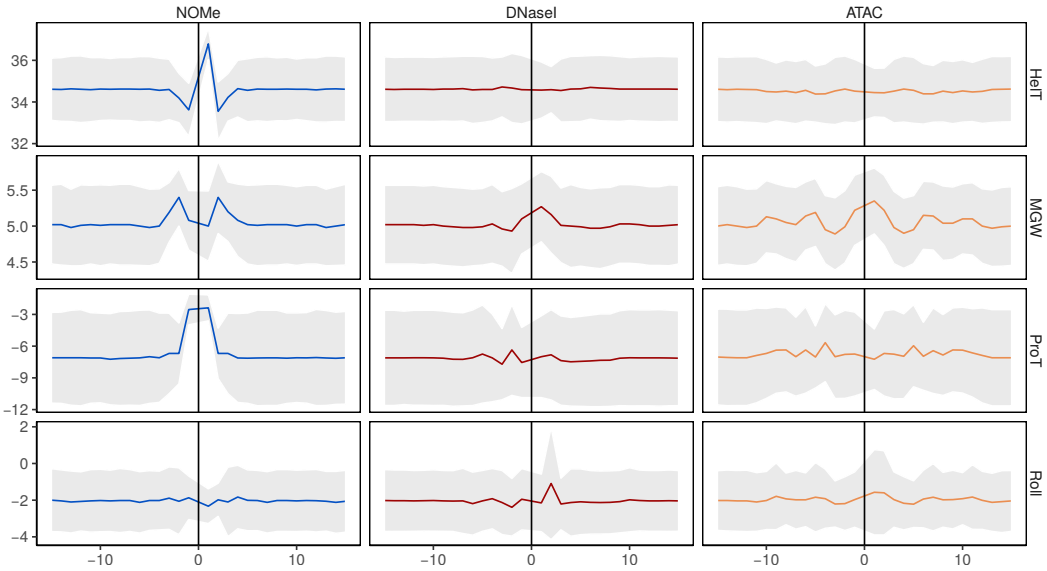


Figure 6.7: Predictions for HelixTwist (HelT), Minor Groove Width (MGW), Proppeller Twist (ProT) and Roll on 2 million randomly sampled 5′ sites of the distinct assays. Figure based on Figure 2b from Nordström *et al.* [N+19].

we used a 31bp window centered on the enzyme activity sites of each assay to harmonize the sequence logos with the remaining figures.

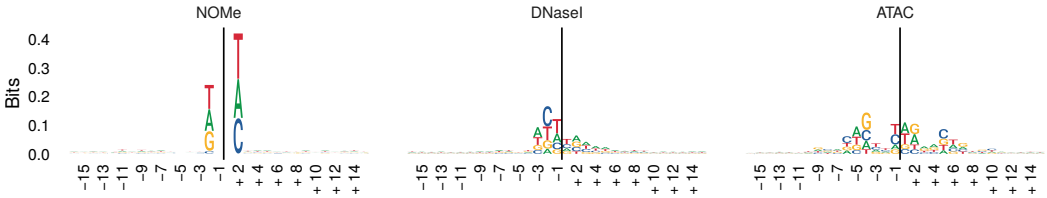


Figure 6.8: The detected sequence bias is visualized using sequence logos. For NOMe-seq, the artificial GCG bias is obvious, while for DNaseI-seq and ATAC-seq, a slightly more complex bias can be observed. Figure based on Figure 2a from Nordström *et al.* [N+19].

In concordance to earlier work [B+13d, K+13c] we find a distinct sequence preference at the active sites of the DNaseI-seq and the Tn5 enzymes, as shown in Figure 6.8. An additional analysis by Karl Nordström reveals the existence of these profiles also in various other samples (Figures B.2, B.3, B.4). For M.CviPI, only a minor sequence preferences at the flanking −2 and +2 position is found, depicted in Figure 6.8. However, this is only due to the *in silico* exclusion of ambiguous GCG sites, which are overlapping with endogenous CpG methylation and are therefore inconclusive.

Taken together, the sequence based analysis already indicates that each assay

has strong sequence and structural preferences that impact the genome wide signal distribution.

6.7.3 DNA methylation

We obtained CpG methylation data for the same set of sequences used to generate the sequence-motifs reported above. On the basis of that data, the average DNA methylation, weighted to coverage, was calculated for each relative position.

As reported by Lazarovici *et al.*, the DNaseI-seq enzyme has a slight but pronounced preference for an increased CpG methylation around its cut sites [L$^+$13c], while neither the M.CviPI nor the modified Tn5 show a position-specific relation to CpG methylation. Nevertheless, we do notice that the overall amount of 5mC around DNaseI-seq and modified Tn5 cutting sites is obviously lower compared to M.CviPI active sites. Our results on DNA methylation are shown in Figure 6.9. As for the sequence motifs and shape predictions, the DNA methylation profiles were verified in various other samples as well (Figure B.5).

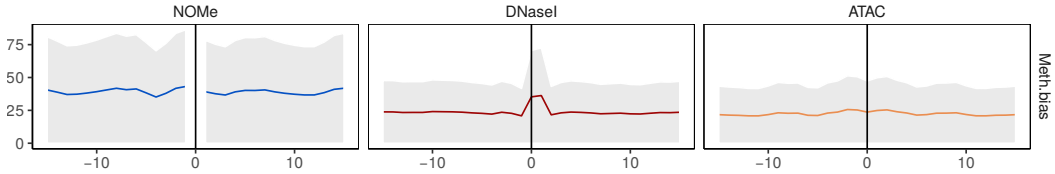


Figure 6.9: DNA methylation (CpG) characteristics around the active sites of the enzymes. Figure following Figure 2b from Nordström *et al.* [N$^+$19].

6.8 A logistic regression classifier to classify assay specific NDRs

As shown in the previous section, there are distinct DNA methylation, sequence and shape characteristics detectable for assay-specific open chromatin regions. To systematically assess whether those can be used to categorize assay-specific NDRs, we trained a multi-class logistic regression classifier considering various sequence based features. Specifically, we computed:

- A, T, C and G content,
- CG content,
- CpG and GpC count,
- average CpG methylation,
- NOMe-seq coverage,
- and counts for all 1024 5-mers.

Two features were purposefully excluded from the model, as they might have overshadowed other relevant features: GpC methylation for its undoubtedly strong relation to NOMe-seq NDRs and the length of NDRs, which would have likely been a resemblance of peak caller artefacts [K+14c]. In total, we obtained the aforementioned features for 12,415 unique DNaseI-seq, 19,323 unique ATAC-seq and 19,453 unique NOMe-seq NDRs.

On the basis of these features, we trained a multi-class logistic regression classifier with elastic net regularization, as described in Section 3.4.1. As we are dealing with a classification problem, model performance is assessed in terms of accuracy (ACC) on a balanced hold-out test data set, represented in a $3x3$ confusion matrix C:

$$ACC = \frac{C_{1,1} + C_{2,2} + C_{3,3}}{\sum_{i,j} C_{i,j}} \tag{6.3}$$

We remind the reader that, because we perform a 3-class classification, a random classifier would obtain an accuracy of about 0.33.

The accuracy of the classifier including the counts of 5-mers in the regions, is 0.63, computed from the confusion matrix shown in Figure 6.10a. Compared to a model neglecting the k-mer information, which obtains an accuracy of only 0.55, the importance of both sequence composition and indirectly of DNA shape becomes obvious.

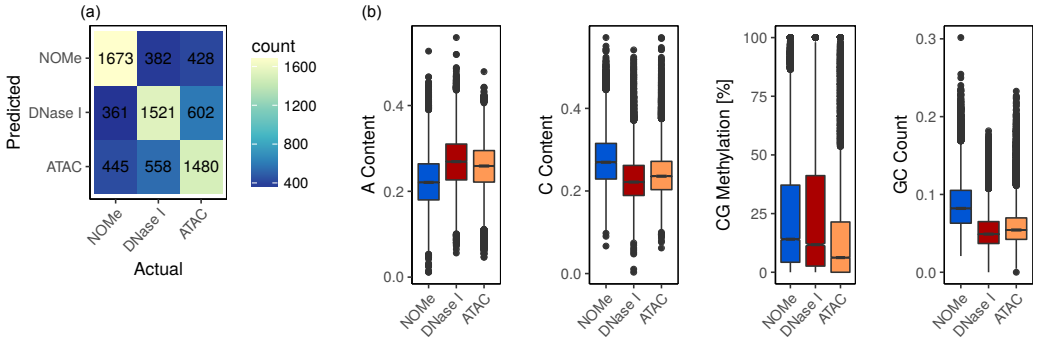


Figure 6.10: a) Confusion matrix of a 3-class classifier separating NDRs uniquely identified with just one assay from each other b) Feature values exemplified for A content, C content, CG methylation and GC Count. Figure based on Figure 2a from Nordström *et al.* [N+19].

As shown in Figure 6.10b, the regression coefficients suggest that A-content, C-content and GC count separate unique NOMe-seq regions from ATAC-seq and DNaseI-seq, while CG methylation seems to distinguish unique ATAC-seq NDRs from both NOMe-seq and DNaseI-seq regions. An overview of the top assay specific features is provided in Table 6.1.

As indicated in Table 6.1, some 5-mers are helpful in separating the classes as well. For example, for classifying NOMe-seq unique NDRs the 5-mer `CGCGC` is depleted, representing the aforementioned GCG effect, while 5-mers enriched in DNaseI-seq

Table 6.1: The table contains the top 15 enriched (enr.) and depleted (dep.) features to distinguish assay unique NDRs. The features are ranked by the weights determined by the logistic regression classifier.

DNaseI enr.	DNaseI dep.	ATAC enr.	ATAC dep.	NOMe enr.	NOMe dep.
G content	GACTC	A content	GpC count	GpC count	CAGTG
CATCG	AGGGC	T content	CG methylation	CG content	CTCTG
C content	CCTTC	C content	CCCGG	CG methylation	GCATG
TCCTG	CGGTC	G content	CCGGG	AAATT	CAGGC
AGCCA	AACT	CpG count	AGACT	GTCTT	CACTG
TTGCG	AGAGG	NOMe coverage	CACGG	ATGGT	CAGAG
TTGCA	GAAGC	CGCCG	TTCGA	AAGAC	AAAGA
TATCG	GCCAC	CGGCG	GGCCT	GTCTC	CCTTG
AATCG	GCGAT	CG content	ACCGA	GAGAC	CAAGG
CCCGC	GTGGC	GACTC	GTTTT	AATTT	CGCGC
GTCAA	ATCGT	GAGTC	GCCGA	AAGAT	T content
TTCAA	T content	TGACT	CAAAC	AGTGA	A content
TGCCA	A content	TGAGT	AGGCC	CCACT	CpG count
GTTTA	GpC count	ACTCA	TTAGG	GGTGG	C content
GACCT	CG content	TGCAG	AAAAC	TTTTC	G content

unique NDRs resemble the observed DNaseI-seq sequence bias motif. In general, the classifier is better suited to separate NOMe-seq from ATAC-seq and DNaseI-seq and the most miss-classifications arise between the ATAC-seq and DNaseI-seq unique regions. (see Figure 6.10a).

6.9 General conclusions

Although many of the identified accessible regions are called by all three or at least two assays, we observe that there are numerous more specific NDRs, which are suggested by one of the assays only. Using targeted deep amplicon sequencing, our collaborators have shown that assay unique NDRs can be assumed to be genuine accessible sites.

Via gene-expression modelling, we illustrate that the assay specific NDRs do contain essential biological signals, required for accurate gene-expression predictions. Furthermore, using a logistic regression classifier, we attempt to find genomic features that could help us to comprehend why certain regions are preferred by a distinct assay. Additionally, we look at chromatin shape predictions as well as at the sequence biases at the cut sites/active sites of the DNaseI, the Tn5 and the methyltransferase M.CviPI, respectively.

Overall, our comparative study suggests that all three assays enable researchers to identify the most pronounced accessible regions, which can be frequently associated with molecular functions, e.g. TF binding. However, we also see that assay specific NDRs are indispensable in gene-expression modelling. Our findings lead to the conclusion that single assays to characterise chromatin accessibility are less comprehensive than expected and seem to lead to a biased and incomplete picture of a cells chromatin landscape, all be it, following the recommended protocols.

6.10 Contributions of all researchers involved in the described project

Karl Nordström (Saarland University), Florian Schmidt (Saarland University), Nina Gasparoni (Saarland University), Abdulrahman Salhab (Saarland University), Gilles Gasparoni (Saarland University), Kathrin Kattler (Saarland University), Fabian Müller(MPI Inf, currently at Department of Genetics, Stanford University School of Medicine), Peter Ebert (MPI Inf, currently at Saarland University), Ivan G. Costa (RWTH Aaachen), DEEP consortium, Nico Pfeifer (MPI Inf, currently at University of Tübingen), Thomas Lengauer (MPI Inf), Jörn Walter (Saarland University) and Marcel H Schulz (Saarland University, currently at Göthe University Frankfurt).

The project has been suggested by Jörn Walter and Marcel H Schulz in the scope of the DEEP project. All chromatin accessibility data was generated in the lab of Jörn Walter by Nina Gasparoni, Gilles Gasparoni and Kathrin Kattler. The generated raw data was processed by Karl Nordström, Peter Ebert and Nina Gasparoni.

Florian Schmidt suggested the signal based comparison of the assays shown in Figure 6.1. The actual experiment was performed by Karl Nordström. Figure 6.2 is based on MACS2 peak calls generated by Karl Nordström using a pipeline provided by Peter Ebert in the scope of the DEEP project. Basic characteristics of these regions have been investigated by Karl Nordström (Figure 6.3). Nina Gasparoni, Gilles Gasparoni and Karl Nordström performed the amplicon analysis summarized in Figure 6.4. Further, Karl Nordström performed the clustering analysis shown in Figure 6.5. Florian Schmidt performed the gene-expression modelling and the required data processing, summarized in Figure 6.6.

The analysis of DNA shape, DNA sequence and DNA methylation (Figures 6.7-6.9) at unique sites has been suggested by Florian Schmidt. For DNA methylation, this analysis was performed by Karl Nordström, for DNA sequence and DNA shape it was performed by Florian Schmidt on HepG2. Karl Nordström applied Florian Schmidt's code later on an extended set of samples.

Using a logistic regression classifier to classify assay specific sites was suggested by Marcel H Schulz. Florian Schmidt implemented the classifier, generated the features and interpreted the results. He generated the content of Figure 6.10 and of Table 6.1

The main manuscript was written by Karl Nordström and Jörn Walter. Florian Schmidt provided the method and results description for the gene-expression modelling, DNA shape, sequence and (in parts) methylation analysis as well as the explanation of the multi-class classifier.

All researchers involved in the project as well as other researches from the DEEP consortium regularly discussed the preliminary results of the project and brain stormed on potential further analyses.

7

Objective assessment of batch effect adjustment methods

7.1 Motivation and research objective

In Chapters 3, 4 and 5 we introduced several approaches to combine epigenomics with transcriptomics data to identify key regulatory factors in per-sample approaches. With the advancement of consortia like IHEC, GTEx, or The Cancer Genome Atlas (TCGA) and the wealth of data that has been produced, even per-gene approaches become feasible. However, the existence of batch effects in the biological data sets makes it hard for researches to perform integrative analyses of large scale data sets. Batch effects can arise from different labs carrying out the experiment, different environmental conditions, different experimental protocols and so on. It was shown that neglecting batch effects can result in inaccurate conclusions [L$^+$10b] and the community accords that batch effects should be dealt with in any computational data processing problem [G$^+$17b]. Therefore, several Batch Effect Adjustment (BEA) approaches have been suggested by the community (Section 7.3).

Choosing an appropriate BEA method and deciding whether the adjusted data is improved compared to the original data set is not straight forward. However, an accurate adjustment of the data is essential to ensure the reliability of all downstream analyses. It is especially important to ensure that while adjusting for batch effect(s), no or only very little biological variation is removed from the data. A standard approach to assess BEA quality is the visual examination of the data in reduced dimensions, which can be achieved using PCA or t-SNE visualizations of the data before and after BEA. But, we and others believe that the visual check is rather subjective and non-interpretable, particularly in instances where the batch is not associated with the highest variance occurring in the data [R$^+$13]. Till now, an objective measure to determine the quality of batch effect adjustment that is applicable to heterogeneous data sets with only few, biologically highly diverse samples, has not been proposed.

Here, we present an innovative, novel method to judge the performance of BEA approaches. We use the Cell Ontology (Section 7.5) to generate a gold-standard of sample similarity; that is we compute a distance matrix holding pairwise similarity scores for all samples within a data set. This enables us to model the relationships between samples even in instances with low replicate numbers and diverse, hetero-

geneous data sets. Comparing the expected similarity matrix against a similarity score obtained for the original as well as the batch effect adjusted data, we can quantify the quality of the BEA method. Our score allows us to obtain not only a global picture of BEA performance, but also allows for a fine grained, sample and tissue specific evaluation of the quality of BEA.

In Section 7.2, we describe related scores to judge the performance of BEA methods and outline the limitations of these scoring strategies. Furthermore, in Section 7.3, we briefly sketch three commonly used tools and algorithms to adjust data for batch effects.

In the following, the term group variable, refers to a variable indicating a biological group of samples, e.g. the cell- or tissue-type. The term batch variable denotes a known confounder, for instance the consortia a sample originates from. This chapter is based on our article `"An ontology-based method for assessing batch effect adjustment approaches in heterogeneous data sets"` [S+18a].

7.2 Established methods to evaluate batch effect adjustment methods

As mentioned above, the most common way to assesses BEA performance is a visual examination of the data in a lower dimensional space, or instance using PCA or t-SNE [vdMH08].

A more quantitative approach would be to compute the overlap of samples originating from divers data sets before and after BEA considering their distance in a high-dimensional space. This can be done, for instance, by calculating the ratio of samples belonging to the same study and those belonging to another one for the k-nearest neighbors of each sample. This quantity is established in the field as the mixture score [L+13b]. Instead of the origin of the study, also other labels, e.g. cell type, can be used.

Furthermore, a correlation analysis of replicates can inform about BEA performance. Assuming that BEA is removing confounders and preserving the actual signal, the correlation between replicates is assumed to rise through BEA. Frequently, few or no replicates are available, for instance, in DEEP or in the other IHEC consortia, making this type of analysis challenging.

Another alternative to systematically evaluate BEA performance would be to train a classifier to predict group variables, e.g. cell type or tissue labels. A successful application of BEA methods is expected to result in a improved performance of the classifier on unseen test data when the model was trained on BEA corrected data in contrast to the original data set. The downside of this approach is that it is not suited be an indicator of BEA performance, because the low replicate numbers are not sufficient such that the data at hand can be equally split into training and test sets [L+10c].

Yet another approach is to asses the skewness of a gene's expression distribution across studies, for example by considering the cumulative density function of gene-expression values [L+13b]. However, in a setting with samples of high

biological variability, the gene-expression profiles are expected to have a high biological variability already, which renders this approach less suitable as well. The same line of arguments excludes the usage of differential gene-expression analysis [LS07, GBS12b].

A promising approach if batch variables are known is principal variance component analysis [L+10b, C+11a], which first identifies Principal Components (PCs) and subsequently uses those in variance component analysis to determine the impact of the known batch variables. A successful BEA should lower the contribution of the known batch variables. Unfortunately this type of analysis is unfeasible in most settings as for many data sets, only a very limited number of batch effects are described in the meta data.

Recently, probabilistic principal component and covariate analysis (PPCCA) has been suggested. This approach aims at circumventing some of the aforementioned problems by including covariates into a PCA and calculating significance tests for each PC individually to assess whether it is linked to a batch effect [N+17]. To perform PPCCA covariates have to be known. Therefore, PPCCA can not be utilized to quantify whether the BEA reduced the impact of the batch effects on the data at hand or not.

As delineated in this section, most BEA assessment methods are not suitable if sample groups are highly diverse and not sample numbers vary a lot between batches. This is exactly the scenario we are facing in many recent data sets, as generated by IHEC, GTEx, or TCGA.

7.3 Batch effect adjustment methods

Within this section, we briefly describe three commonly used BEA methods developed for bulk RNA-seq data sets: ComBat [J+07], Surrogate variable analysis (SVA) [LS07] and Removing unwanted variation ((RUV) [J+16a].

Here, we used the ComBat and SVA implementation in the sva R-package (version 3.24.4) as well as the RUV R implementation RUVNormalize (version 1.12.0).

7.3.1 Combat

A default approach to adjust for batch effects is to convert data from various batches such that it will exhibit a similar mean and variance for each gene. This approach is known as the Location and scale (L/S) adjustment. Mean and variance can be adjusted using both linear and non-linear transformations.

ComBat is a widely used method implementing the L/S strategy via an empirical Bayes method, pooling information across genes with similar expression profiles. Specifically, the underlying gene-expression model in ComBat for the expression $Y_{i,j,g}$ of gene g in sample j of batch i is:

$$Y_{i,j,g} = \alpha_g + X\beta_g + \gamma_{i,g} + \delta_{i,g}\epsilon_{i,j,g}. \tag{7.1}$$

Here α_g is the overall gene-expression, X is the experiment design matrix, β_g is the vector of regression coefficients corresponding to X, $\gamma_{i,g}$ and $\delta_{i,g}$ are the additive and multiplicative batch effects for batch i influencing g and ϵ is the error term following a normal distribution [J+07].

Using per-gene standardized expression data $Z_{i,j,g}$, the parameters $\gamma_{i,g}$ and $\delta_{i,g}$ are estimated with an empirical Bayes approach. See Johnson *et al.* [J+07] for details on parameter inference. The BEA gene-expression $Y^*_{i,j,g}$ can be computed according to:

$$Y^*_{i,j,g} = \frac{\hat{\sigma}_g}{\delta_{i,g}}(Z_{i,j,g} - \hat{\gamma}_{i,g} + \hat{\alpha}_g + X\hat{\beta}_g), \tag{7.2}$$

where $\hat{\alpha}_g$, $\hat{\beta}_g$ and $\hat{\sigma}_g$ are the model parameters estimated using an ordinary least-squares approach [J+07].

Since ComBat has been developed to correct for batch effects occurring in microarray data sets, it is well suited for small batch sizes with low biological variability. ComBat can be utilized either with or without informing about group variables. All batch variables must be known to the model [J+07].

In the scope of this project, ComBat is used with information on present batches that is the source of the data and the tissue or cell type label of each sample.

7.3.2 Surrogate variable analysis (SVA)

Also matrix factorization can be used for BEA by identifying and removing batch-associated factors, which can be automatically identified and removed from the data using.

A well known algorithm implementing matrix factorization is SVA [LS07]. SVA approximates the number of latent variables to be removed and can be utilized with as well as without considering group variables in the model. The algorithm attempts to estimate surrogate variables searching for frequent patterns in gene-expression variance. The challenge lies in designing the surrogate variables such that their signal is not due to the primary variable (e.g. cell- or tissue-type).

The method is split into two parts. First, unmodeled factors are detected, which are subsequently used to construct surrogate variables h_k [LS07]. Unmodeled factors are detected by generating a residual matrix R that is computed from the actual gene-expression matrix removing the influence of the primary variable y_j (e.g. cell- or tissue-type) on gene-expression:

$$X_{i,j} = \mu_i + f_i(y_j) + \epsilon_{i,j}, \tag{7.3}$$

$$R_{i,j} = X_{i,j} - \mu_i + f_i(y_j), \tag{7.4}$$

where μ_i is the baseline expression of gene i, y_j is the primary variable of interest, $f_i(y_j)$ denotes the relationship between $x_{i,j}$ and y_i according to $E(x_{i,j}|y_i) - \mu_i$, X is the normalized gene-expression matrix where i indicates genes and j indicates samples, and $\epsilon_{i,j}$ is an error term. The variables μ_i and $f_i(y_j)$ can be determined

using a regression approach and allow the computation of matrix R. An essential idea of SVA to model the unwanted factors is to split the error term $\epsilon_{i,j}$ into

$$\epsilon_{i,j} = \sum_{l=1}^{L} \gamma_{l,i} g_{l,j} + \epsilon^*_{i,j}. \tag{7.5}$$

Here L refers to the number of all unmodeled factors, $g_{l,j}$ denotes the contribution of factor l to sample j, $\gamma_{l,i}$ denotes the gene-specific influence of factor l and $\epsilon^*_{i,j}$ is true gene-specific noise [LS07]. The purpose of SVA is to estimate the linear combination $\sum_{l=1}^{L} \gamma_{l,i} g_{l,j}$. Firstly, a singular value decomposition of the residual matrix R is computed:

$$R = UDV^T. \tag{7.6}$$

From D, the amount of variance explained can be computed for each eigengene k that is the k^{th} diagonal element of D according to:

$$T_k = \frac{d_k^2}{\sum_{l=1}^{n-df} d_l^2}, \tag{7.7}$$

where *df* refers to the degrees of freedom of R. Using permuted versions of R, a p-value for the significance of each T_k can be computed [LS07].

In the second phase of the SVA algorithm, we test for each significant eigengene e_k, where $e_k = (e_{k,1}, ..., e_{k,n})^T$ is the k^{th} column of V from Equation 7.5, whether it significantly influences the expression of each in X. The number $\hat{m}$ of genes that are associated with e_k is determined and a residual gene-expression matrix X_r holding the expression of those genes is generated. Next, another set of eigengenes, specifically the eigengenes e^r_j for matrix X_r are computed and we define the estimated surrogate variable $\hat{h}_k$ as

$$\hat{h}_k = e^r_{\underset{1 \leq j \leq n}{argmax}(cor(e_k, e^r_j))}. \tag{7.8}$$

Using the estimated surrogate variables, the adjusted gene-expression can be determined according to

$$X^*_{i,j} = \mu_i + f_i(y_i) + \epsilon_{i,j} + \sum_{k=1}^{K} \lambda_{k,i} \hat{h}_{k,j}, \tag{7.9}$$

where $\lambda_{k,i}$ are the eigenvectors corresponding to e_k weighing the contributions of each surrogate variable $\hat{h}_k$.

7.3.3 Removing unwanted variation (RUV)

RUV is an alternative method that utilizes several control genes to remove batch effects in the considered data sets [J+16a]. Typically, a set of so-called negative

control genes that is gene whose expression is invariant under the variable of interest, are used to adjust the batch effect for all other genes. In RUV, the underlying gene-expression model is

$$Y = X\beta + W\alpha + \epsilon, \tag{7.10}$$

where Y is the observed gene-expression, X is the matrix containing the factors of interest, W represents the unwanted factors, e.g. the confounders causing the batch effect, ϵ is the noise term following a normal distribution and β and α are the coefficients corresponding to X and W, respectively. An estimator for W based on a set of negative control genes is termed $\hat{W}_2$ [GBS12a]. Assume further that X is not known that is $X\beta = 0$ in Equation 7.9. Than, we can determine α according to

$$\underset{\alpha \in \mathbb{R}^{kxn}}{min} ||Y - \hat{W}_2\alpha||_F^2, \tag{7.11}$$

where k is the number of unwanted factors and $||x||_F^2$ is the *Frobeniusnorm*. With an estimate of α, the observed gene-expression data Y can be adjusted ($\hat{Y}$) in a straightforward way according to:

$$\hat{Y} = Y - \hat{W}_2\alpha. \tag{7.12}$$

For RUV, we exploit several housekeeping genes suggested by [NT09] as negative control genes.

7.4 Data used in this study

7.4.1 IHEC data

Here, 36 fastq files for ENCODE [D$^+$12b] and 112 fastq files for Roadmap [K$^+$15] RNA-Seq experiments have been obtained. Additionally, we acquired fastq files for 12 DEEP RNA-seq samples [Con18] as well as for 56 BLUEPRINT RNA-seq samples [A$^+$12]. Gene-expression is reported in *transcripts per million (TPM)*. The quantification was performed with Salmon (version 0.8.2) [P$^+$17a] utilizing a reference transcriptom downloaded from Gencode v26 (GRCh38.p10). Data accession numbers for ENCODE and Roadmap data, sample IDs for DEEP and Blueprint data, as well as tissue and cell type assignments, sample numbers per consortia and command lines used to run Salmon are provided in Section B.5.

7.4.2 GTEx and TCGA data

We used fastq files for 6,575 distinct RNA-seq data sets from the GTEx project [C$^+$15c] and 741 RNA-seq samples from TCGA [W$^+$13b], respectively. In our application, only TCGA control samples are considered while tumour samples are not taken into account. We focused on five different tissues that are contained in both GTEx and TCGA: colon, liver, kidney, prostate and thyroid. Altogether, this

leads to 1,062 GTEx samples as well as to 274 TCGA samples. In Section B.5, we provide a detailed overview of sample counts per tissue and consortia.

The processing of the GTEx and TCGA RNA-seq data was performed by Engin Cukuroglu at the Genome Institut of Singapore. The RNA-Seq data was mapped against the human reference genome version hg19 using TopHat2 (version 2.0.12) [K+13a] with the Ensembl gene annotation v75. Mapped reads have been counted with the R package GENOMIC ALIGNMENTS [L+13a] and the parameter setting `mode=Union` and `inter.feature=FALSE`. Only primary read alignments have been retained. Data normalization has been performed using DESEQ2 [L+14c].

7.5 Cell Ontology

The Cell Ontology(CL) [B+05b, D+16b] provides a curated vocabulary of mostly vertebrate cell types. With the last major update of the Ontology in 2016, the CL contained 2,200 classes of cell types, with a focus on cell types occurring *in vivo* [D+16b]. The cell types are linked to each other in a directed acyclic graph structure, yielding a hierarchical structure with the root class Cell, denoted with the identifier `CL:0000000`. A heart cell, for instance, has the ID `CL:1000147`, whereas a cell of the large intestine has the ID `CL:1000320`. The CL can be freely accessed online, e.g. via the ONTOBEE webserver [O+17].

7.6 An ontology score to assess sample similarities

Figure 7.1 provides a schematic overview of the computation of our ontology score. In Section 7.6.1, we detail how the Cell Ontology is used to obtain a measure of an expected similarity between samples, Section 7.6.2 describes how a similarity measure is computed on the gene-expression data and the ontology score combining both measures is explained in Section 7.6.3. R-Code to use our method is freely available online at `https://github.com/SchulzLab/OntologyEval`.

7.6.1 Calculating expected sample similarities from the *Cell Ontology*

For all ontology terms contained in the CL [B+05b], we calculate the pairwise similarity $sim(t_i, t_j)$ between terms t_i and t_j using both jaccard coefficients (sim_{jac}) and cosine similarity (sim_{cos}) [P+09]. The function $A(t_i)$ is returning the set of ancestors for a term t_i in the CL, considering only subclass relationships. Here, $A(t_i)$ is defined such that $t_i \in A(t_i)$. Examples are shown in Figure 7.2a and b.

To compute the cosine similarity, a vector representation v_t for term t is required. The vector has |CL| entries, where |CL| is the total number of terms contained in the *Cell Ontology*, with a one to one mapping between terms of the CL and entries in v_t . At every index that corresponds to an entry in $A(t)$, we set v_t to one and to zero otherwise, see Figure 7.2a,b for an example. The jaccard similarity and cosine

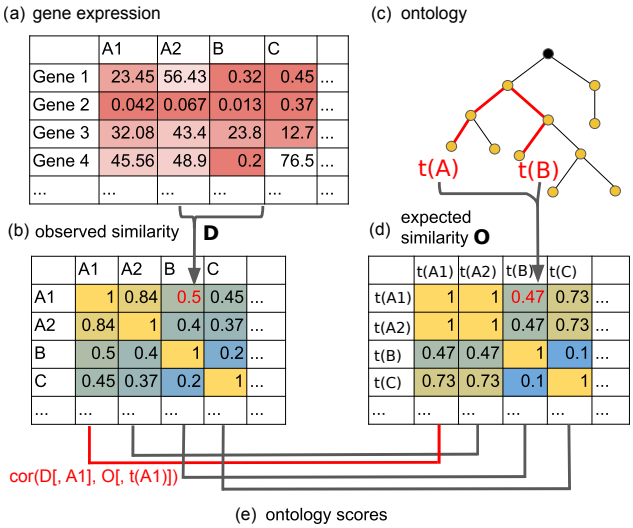


Figure 7.1: (a) Using a gene-expression data matrix, a similarity matrix D holding pairwise similarities of the measured gene-expression data is computed. (c) Using an ontology the lengths of the paths between all ontology terms associated with the samples in (c) are computed. These are used to computed matrix O of expected pairwise sample similarities (d). By computing the correlation between the observed sample similarities from (b) and the expected similarities from (d) the ontology score can be computed (e). Figure from Schmidt *et al.* [S+18a].

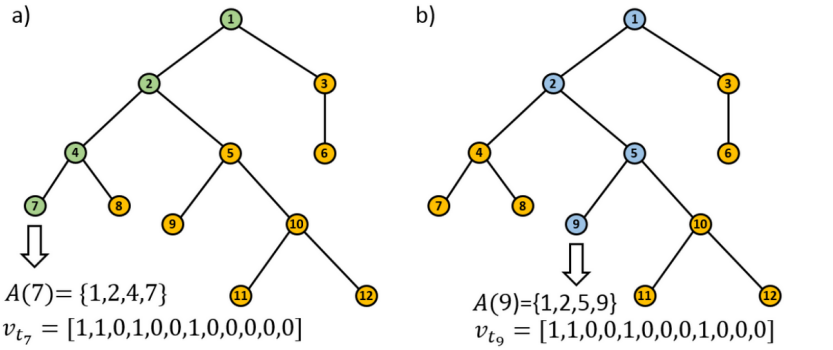
similarity between two terms t_i and t_j are defined as

$$\text{sim}_{\text{cos}}(t_i, t_j) = \frac{v_{t_i} \cdot v_{t_j}}{\sqrt{v_{t_i}^2} \cdot \sqrt{v_{t_j}^2}}, \tag{7.13}$$

$$\text{sim}_{\text{jac}}(t_i, t_j) = \frac{|A(t_i) \cap A(t_j)|}{|A(t_i) \cup A(t_j)|}. \tag{7.14}$$

Illustrative examples for the computation of these two measures are provided in Figure 7.2c and d.

In our application, we have manually mapped all samples from TCGA, GTEx and IHEC to CL terms. With these mappings, we generated matrices holding expected pairwise sample similarities $sim(s_k, s_l)$ for any combination of samples s_k and s_l (within one consortia) according to the similarity measures sim and CL terms t_{s_k} and t_{s_l}. Calculating $sim(s_k, s_l) = sim(t_{s_k}, t_{s_l})$ leads to two symmetric similarity matrices: O_{jac} and O_{cos}. These can be transformed to distance matrices according to $dist(s_k, s_l) = 1 - sim(t_{s_k}, t_{s_l})$.



c)

$$sim_{jac}(t_7, t_9) = \frac{|A(t_7) \cap A(t_9)|}{|A(t_7) \cup A(t_9)|} = \frac{|\{1,2,4,7\} \cap \{1,2,5,9\}|}{|\{1,2,4,7\} \cup \{1,2,5,9\}|} = \frac{|\{1,2\}|}{|\{1,2,4,5,7,9\}|} = \frac{1}{3}$$

d)

$$sim_{cos}(t_7, t_9) = \frac{v_{t_7} \cdot v_{t_9}}{\sqrt{v_{t_7}^2} \cdot \sqrt{v_{t_9}^2}} = \frac{\sum_{i=1}(v_{t_7 i} \cdot v_{t_9 i})}{\sqrt{\sum_{i=1}(v_{t_7 i})^2} \cdot \sqrt{\sum_{i=1}(v_{t_9 i})^2}} = \frac{2}{2 * 2} = \frac{1}{2}$$

Figure 7.2: We show the value of $A(t_7)$ and $A(t_9)$, as well as the vector representations v_{t_7} and v_{t_9} required to compute the cosine similarity in (a) and (b), respectively. An example for the computation of the jaccard similarity is shown in (c) and an example for the cosine similarity is shown in (d). Figure from the Supplement of Schmidt *et al.* [S+18a].

7.6.2 Using PCA to obtain a sample similarity matrix with respect to gene-expression data

To compute a matrix of observed sample similarities, we first perform a dimensionality reduction of the original gene-expression matrix using PCA. By definition, most of the gene-expression variability between the samples is captured by the first PCs. Therefore, these can be used to distinguish different samples from one another. This approach allows us to focus on the difference between samples. The differences might be harder to detected in normal data space due to overshadowing by housekeeping or faintly expressed genes. Using PCs1 – 4 (explaining 86% to 95% of the measured gene-expression variance), we computed a sample similarity matrix D using Spearman correlation, as shown in Figure 7.1b.

7.6.3 Contrasting expected distances with expression-based distances to obtain a quality score

By construction, matrix O with expected similarities from the ontology and matrix D with the observed similarities from the gene-expression data are n x n matrices, where n is the number of considered samples. Not only the dimensionality but also the order of samples in the observed similarities $D[, k]$ matches that in the

expected similarities in $O[,k]$ for each sample k. This enables us to assess the agreement between the expected similarities O and the inferred similarities D via global score that considers the similarity of both matrices, for instance, the inner product. Although such a score provides an overall judgment of BEA performance, sample-specific information would be lost. Therefore, we compute a vector u of *ontology scores* with a simple index based comparison, as depicted in Figure 7.1e, according to

$$u_k = cor(D[,k], O[,k]) \tag{7.15}$$

for each sample k and *cor* as either Spearman S_{sp} or Pearson S_p correlation. All figures shown in this chapter are based on O_{cos}, S_{sp}. We refer the reader to the Supplementary Material of Schmidt *et al.* for results on other score combinations [S$^+$18a]. These are omitted as changes to the scoring did not effect any of the drawn conclusions.

The per sample scoring permits us to investigate batch effects at different levels of granularity. For instance, we can look at individual samples, at groups of distinct samples, or at group variables, e.g. cell types.

An important point is that there is no threshold to determine which value the ontology score should reach to be satisfactory. Therefore, our ontology score should be used to perform comparisons on a relative level only.

7.7 Results

Before we apply our ontology score to real data, we used simulations to determine the effects of using the CL for computing ontology scores u. Secondly, to assess the robustness of the ontology scores u, we added Gaussian noise to GTEx data and analysed the impact of that on our scoring. Thirdly, we introduced an artificial batch effect on GTEx data and adjusted for it using ComBat to see if the ontology score adequately reflects the data manipulation events.

As explained in Section 7.7.4, we applied the ontology score to TCGA, GTEx and IHEC data to present the interpretability of our score when it is applied to heterogeneous data sets.

7.7.1 The ontology score leverages information captured in the Cell Ontology

To characterize the robustness of our ontology score, we conducted randomization experiments using GTEx data for five different tissues (see Section 7.4.2). Specifically, we generated 100 sets in which each sample was assigned to a random CL term from all available ones. For each set, we recomputed both the similarity and the ontology score.

As shown in Figure 7.3, the mean score obtained for the randomized ontology is significantly lower than that of the original ontology matrix.

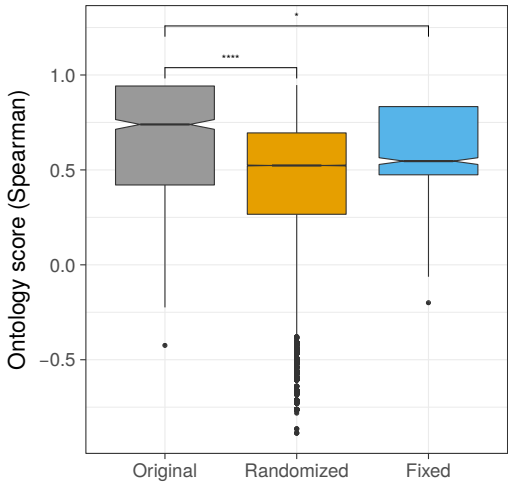


Figure 7.3: The y-axis shows the ontology score, in terms of Spearman correlation for GTEx normal samples. To compute the score, we used the original ontology matrix, a fixed similarity matrix, as well as 100 randomly generated ontology matrices. We see that the real ontology matrix results in scores that are significantly higher than those obtained for both alternative similarity measures (Wilcoxon-Mann-Whitney test: $*** = p < 0.0001$, $* = p < 0.05$). Figure from Schmidt *et al.* [S$^+$18a].

Another sanity check is to use a fixed similarity measure that assigns samples from the same tissue a score of 1.0 and 0.25 otherwise. This test reveals whether leveraging between sample type similarities leads to a significantly higher score than considering only the absolute identity between group variables as a measure. As shown in Figure 7.3, the between sample similarity score achieves a better ontology score than the fixed similarities. Overall, these experiments suggest that the ontology score does inform about sample similarities we can expect to observe in gene-expression data. They emphasize that a precise mapping of samples to CL terms will influence our novel score. In other words, the more accurate the used ontology is and the more precisely the sample mapping reflects the underlying biology, the more accurate the ontology score can be.

Here, we mapped samples to their ontology terms manually. We acknowledge that this is an error-prone procedure and have experienced that it can be challenging to come up with suitable mapping. Therefore, we encouraged our IHEC partners to include expert-curated associations of ontology terms to samples in the sample meta data.

7.7.2 The ontology score is sensitive to noise in the data

As we have seen that the ontology score is sensitive to the used ontology, we next tried to see how it depends on the quality of gene-expression data. To this end, we

have conducted two simulation studies to learn about the effect of artificial noise added to the expression data at hand.

Firstly, we added Gaussian noise $N(\mu = 10, \sigma = 1)$ to all genes across all tissues affecting $0\% - 50\%$ of all GTEx samples. As shown in Figure 7.4a, the score decreases when the fraction of samples that have been exposed to noise increases.

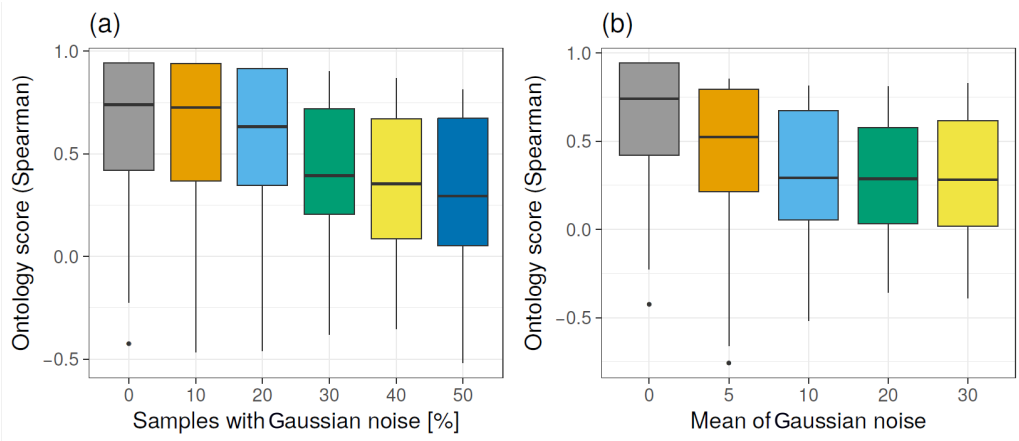


Figure 7.4: (a) The plots shows how the ontology scores reacts on Gaussian noise $N(\mu = 10, \sigma = 1)$ which is added to a subset of GTEx samples. (b) Here, the impact of a varying mean (μ) of Gaussian noise added to 50% of all samples on the ontology score is shown. With increasing mean, the ontology score drops. However, the score stays almost steady after $\mu > 10$. Figure from Schmidt *et al.* [S+18a].

Secondly, we added Gaussian noise with constant variance but with a varying mean ranging from 0% to 30% to 50% of all samples (Figure 7.4b). The ontology score drops rapidly with increasing noise intensity but stays almost constant after $\mu \geq 10$.

In summary, these results indicate that our novel scoring strategy is susceptible for increasing levels of distortion in data associated either to the number of affected samples or to the level of the considered noise.

7.7.3 The ontology score describes the performance of BEA

We extended the simulation study introduced before by trying to adjust for the introduced noise using ComBat. By design, ComBat should be well suited to adjust for the artificial linear shift we have introduced. As shown in a PCA analysis, the original GTEx data clusters in a tissue-specific manner (Figure 7.5a,d).

As expected, adding Gaussian noise to 50% of the samples dissolves the tissue-specific clustering of the original data and results in the formation of two large separate clusters (Figure 7.5b,e). Upon adjusting the data with ComBat the tissue specific clustering is mostly restored (Figure 7.5c,f). Encouragingly, the ontology score reflects the behaviour observed in the PCA. As shown in Figure 7.6, the score

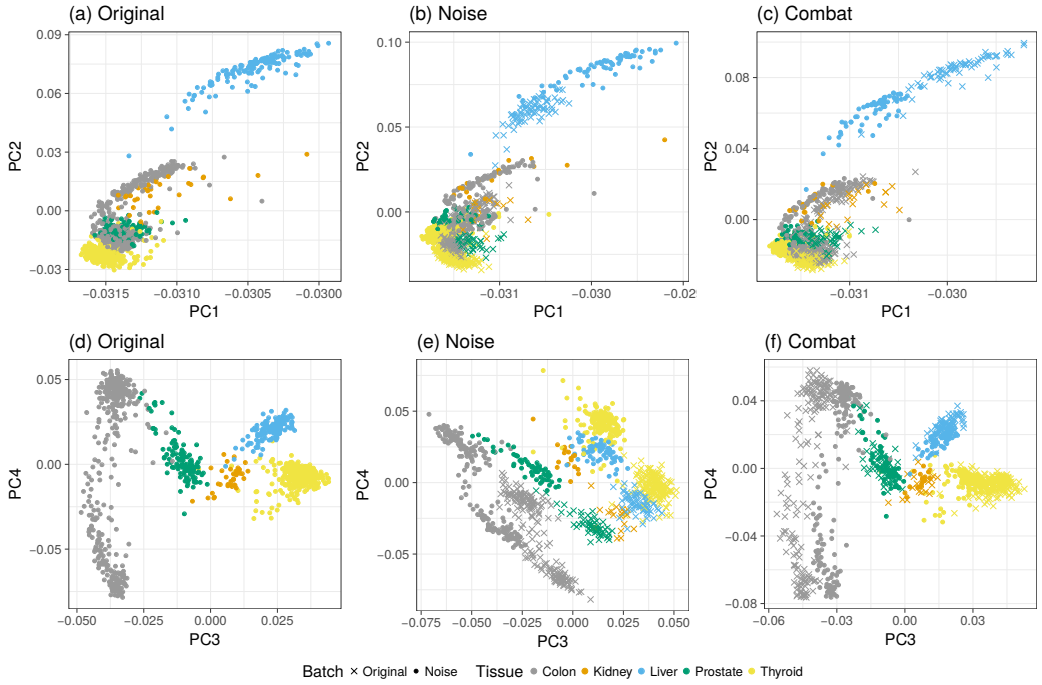


Figure 7.5: In (a-c), the 1*st* and 2*nd* PC, in (d-f) the 3*rd* and 4*th* PC of a PCA of original GTEX data is shown. The addition of Gaussian noise caused a linear shift in the data, which can be seen in (b) and (e). We observe that the addition of the noise leads to the formation of two distinct batches per tissue and the tissue specific clustering in PC3 and PC4, shown in (d) vanished almost completely. The results of the PCA after BEA adjustment using ComBat are shown in (c) and (f). The shift observed in (e) does not exist in (f) anymore. Overall, the PCA of the adjusted data is resembling that of the original data shown in (d). Nevertheless, the liver samples were not adjusted properly, as indicated in (c). There, a new shift, has been introduced by the BEA resulting in a different overall clustering as compared to the original data shown in (a). Figure from Schmidt *et al.* [S+18a].

is decreasing when noise is added and it is almost restored to its original value upon BEA via ComBat. However, we note that the score can not be restored for liver samples. There, PC1 and PC2 indicate that the artificial noise could not be adequately removed (Figure 7.5a,c). Hence, the ontology score does not improve either.

This example illustrates the practicality of our novel approach in a controlled setup. Surprisingly, we find it already challenging to assess the BEA only by a visual inspection of the PCA. Even in this simplistic scenario, the ontology score offers an interpretable and effective alternative measure to circumvent the subjective visual assessment of the PCA.

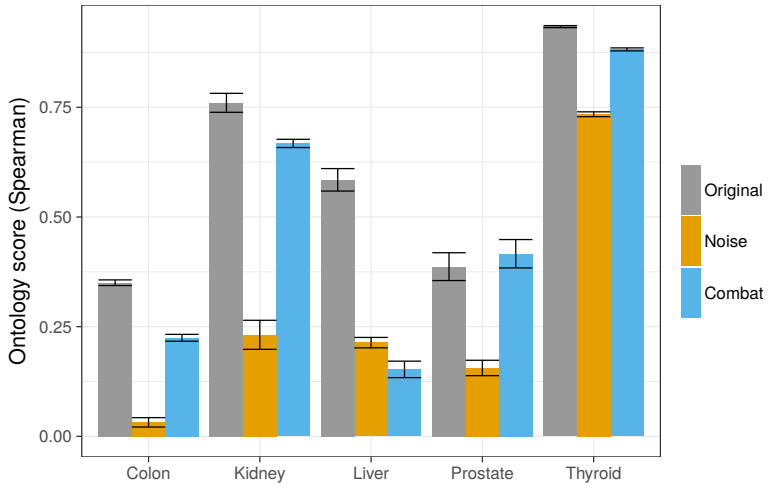


Figure 7.6: Here, the ontology score is depicted for GTEx normal tissues in three instances. Firstly, for the original GTEx data as obtained from the consortium. Secondly, for data that was subject to Gaussian noise $N(\mu = 10, \sigma = 1)$, which has been added to 50% of all samples and thirdly, for the instance that this artificial noise has been removed with ComBat. Figure from Schmidt *et al.* [S+18a].

7.7.4 Application to heterogeneous data sets

Sections 7.7.1 to 7.7.3 supported the reliability of our scoring scheme. Hence, we applied our method to two real use-cases: (1) we considered data from TCGA and GTEx together and (2) data from the epigenomic consortia Blueprint, DEEP, ENCODE and Roadmap.

We corrected for known batches using the BEA methods ComBat, RUV and SVA, explained above in Section 7.3.

In Figure 7.7a the behaviour of the ontology score in the first use case is documented separately for TCGA and GTEx data. Recall that we consider five different tissue-types in this comparison: colon, kidney, liver, prostate and thyroid. According to our score, the RUV method does not lead to an improvement of data quality. In contrast to that, SVA appears to properly adjust GTEx data, but fails to account for batches in TCGA data. Interestingly, the ontology scores for both data sets do show little improvements if ComBat was used for BEA.

In Figure 7.7b, we utilize the ability of our score to analyse the effect of BEA from a more fine-grained perspective. Specifically, we show the ontology score separately for each consortia and tissue-type. This is helpful to identify samples and batches which really benefit from BEA. Devoid of this detailed representation of the scores, only little insights can be gained (Figure 7.7a).

The detailed view on the ontology scores shows that SVA lead to negative scores for prostate samples (Figure 7.7b) and highlights that SVA was not able to improve

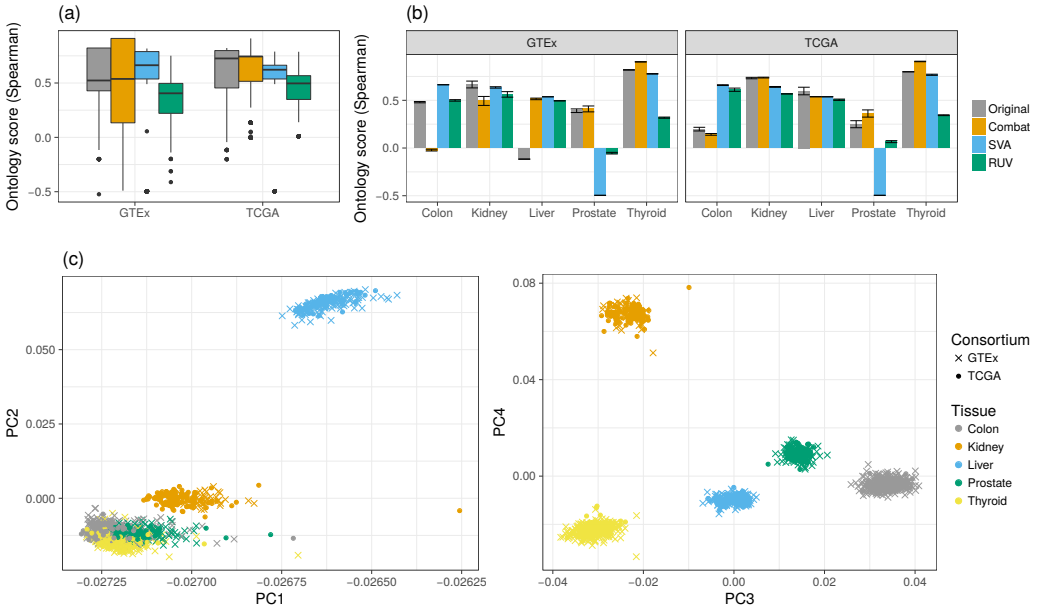


Figure 7.7: a) The boxplots show ontology scores calculated before and after BEA for GTEx and TCGA data belonging to five different tissues. (b) Same as (a) but with a tissue separation of the ontology scores. (c) PCA analysis of the *SVA* adjusted data showing PC1 vs PC2 and PC3 vs PC4. Although the clustering in PC3 and PC4 suggests an adequate adjustment of the data, one can observe in PC1 and PC2 that prostate samples overlap with colon and thyroid samples. Figure from Schmidt *et al.* [S+18a].

the quality of TCGA liver data, while it did improve GTEx liver data. These insights could not easily be obtained from a visual inspection of the PCA plots obtained for the various BEA approaches and the two data sets.

Importantly, our score suggests potential issues not easily deducible from a PCA. For instance, Figure 7.7c shows the PCA of SVA adjusted data. It seems that SVA results in a desirable separation of samples according to tissue-types. However, our score suggests that the cluster with the prostate samples is not behaving as expected in relation to the remaining tissue-types, suggesting that SVA might have caused an artifact.

In fact a detailed analysis of the pairwise correlations of prostate samples with the remaining ones confirmed this (Fig. 7.8). According to our ontology score, none of the BEA approaches seems to be able to improve the gene-expression data in this use-case. Secondly, we use our ontology score in a challenging scenario with 65 different tissues and cell types obtained from four consortia within IHEC. In this data set, only few replicates are available per tissue and cell type.

A PCA analysis depicted in Figure 7.9 reveals that the data clusters in a highly consortia specific manner, preventing any integrative analysis effort. As also, the

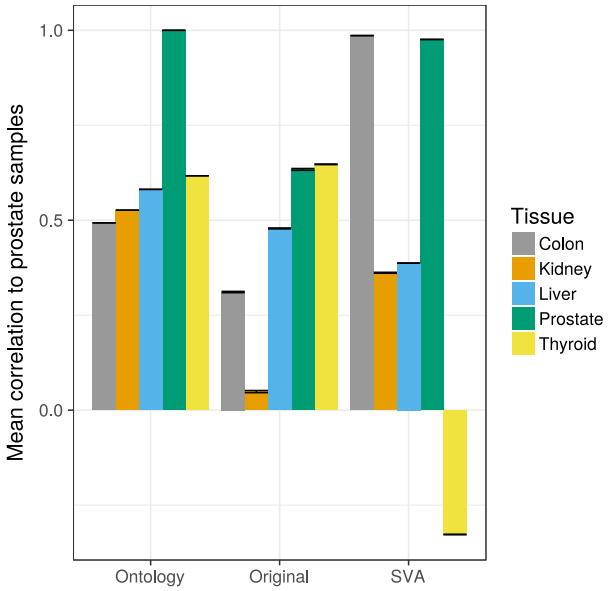


Figure 7.8: Pairwise correlation between the gene-expression of all samples to prostate samples in terms of Spearman correlation. In the SVA corrected data, the prostate samples are negatively correlated to thyroid and obtain a high correlation to the colon samples.

overlap of tissues and cell types across the consortia is small, BEA is challenging (Section B.5). Integrating such diverse data sets is a major challenge for the IHEC consortium. This holds not only for RNA-seq, but also for other data types, rendering this a highly important application scenario.

As shown in Figure 7.10a, the RUV algorithm appears to achieve favorable results on DEEP and Roadmap samples, whereas it is the worst method on Blueprint data. For the latter, ComBat shows the best score improvements. Surprisingly, according to the ontology score, there is no BEA approach that is able to adjust the ENCODE data.

An example of a cell type and tissue specific analysis is shown in Figure 7.10b. Here, RUV adequately adjusts gene-expression data of primary human hepatocytes across the consortia, while it does not work well on erythroblasts. This use case shows that on heterogeneous data sets such as the one produced by IHEC, the currently available sample numbers are not sufficient to utilize already established BEA methods without carefully evaluating the adjusted data. Our analysis raises severe concerns regarding the impact of BEA methods on the data, as in the worst case, the BEA might introduce an additional noise by eliminating true, biologically relevant, variation. Also in light of the large-numbers of single-cell data being produced nowadays, it is an important question how to integrate these data sets and how to quantify the quality of such data cohorts.

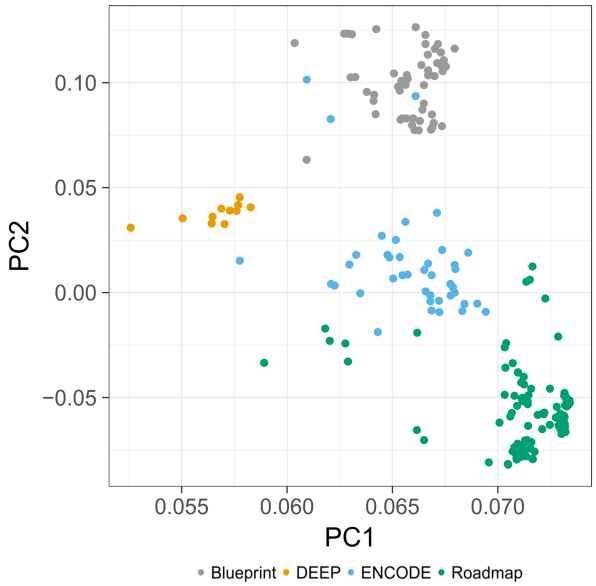


Figure 7.9: PCA analysis of the IHEC data set comprised of data from Blueprint, DEEP, ENCODE and Roadmap. With the exception of four samples, the RNA-seq data clusters in a highly consortia specific manner, suggesting the presence of a strong batch effect that hampers an integrative analysis of the data set.

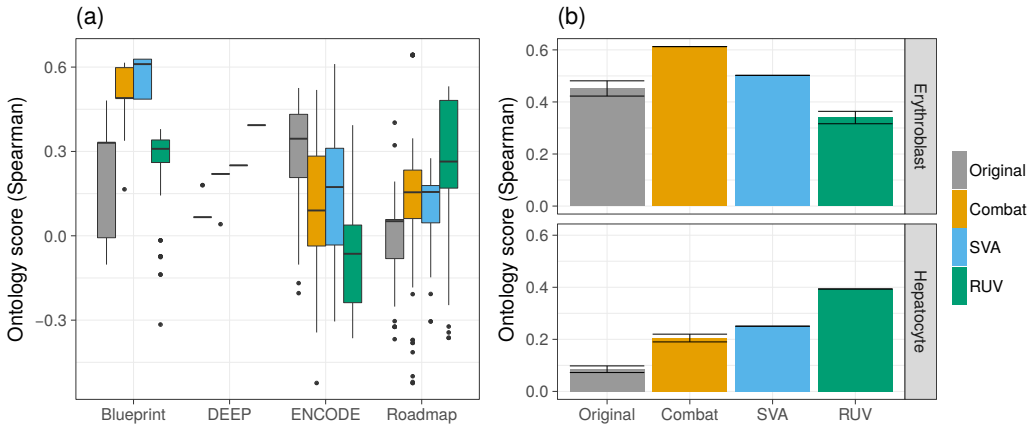


Figure 7.10: a) Box plots holding the ontology scores for BEA in a IHEC data set that is comprised of 65 different tissues and cell types. (b) Illustration for a tissue specific view of the ontology score across the consortia. Figure from Schmidt *et al.* [S+18a].

7.8 Conclusions and Impact of this work

The presented ontology score is, to our knowledge, the only approach to objectively and robustly assess the performance of BEA approaches on heterogeneous data. Moreover, our method is not only applicable to BEA but could also be applied to test data normalization approaches. Specifically, our method has two main advantages compared to previous approaches to assess the performance of BEA methods. Firstly, due to using an overall sample similarity score, no problems arise if only few samples of the same cell- or tissue-type are available. Secondly, our methods allows the comparison of several BEA methods with one another at various levels enabling users to identify strengths and limitations of the BEA approaches.

As already stated in the previous Section, our ontology score indicates that there is a need for novel BEA methods that are specifically designed to handle heterogeneous data sets, like the IHEC data set. Not being able to adequately integrate these data sets will severely hamper integrative analysis and prevent us from progressing in deciphering gene regulation in more detail. We do agree that designing such methods is complicated as unified cell lines or mock-up samples, processed by each consortia are not existing. These would have simplified the identification and adjustment of batches considerably. Also, in case of other data sets such as HM ChIP-seq, or DNaseI-seq, BEA approaches must be applied to the raw signal, which might turn out to be even more complicated than adjusting quantified gene-expression data. Our ontology score will be helpful to optimize and to develop such approaches.

7.9 Contributions of all researchers involved in the described project

The following people contributed to this project: Florian Schmidt (Saarland University), Markus List (MPI Inf, currently at Technical University Munich), Engin Cukuroglu (Genome Institute of Singapore), Sebastian Köehler (Berlin Institute of Health), Jonathan Göke (Genome Institute of Singapore) and Marcel H. Schulz (Saarland University, currently at Göthe University Frankfurt).

Florian Schmidt proposed the idea of a gold standard based comparison to judge batch effect adjustment approaches, Marcel H. Schulz suggested to use the Cell Ontology for that. Together with Markus List, Florian Schmidt devised the Cell ontology score in a brain storming session. The IHEC RNA-seq data was processed by Florian Schmidt. Engin Cukuroglu processed the GTEx and TCGA gene-expression data, provided by Jonathan Göke during a research visit of Florian Schmidt at the Genome Institute of Singapore. Sebastian Köhler generated the pairwise jaccard coefficients and cosine similarity scores for all Cell Ontology terms. Florian Schmidt generated code to apply Combat and RUV, Markus List provided code to use SVA. The Cell ontology score was implemented by Florian Schmidt, who also generated all Figures shown in this Chapter, except for Figure 7.1, which was designed by Markus List. All computational analyses have been performed by Florian Schmidt.

Markus List and Florian Schmidt actively discussed potential analyses and their interpretations. Furthermore, Jonathan Göke and Marcel H. Schulz advised Florian Schmidt. Markus List and Florian Schmidt contributed to the manuscript and are shared first authors of the respective publication in Bioinformatics [S$^+$18a]. Florian Schmidt presented a talk on this work at ECCB 2018 in Athens.

8

Suggesting regulatory sites on the gene-level

So far, in Chapters 3-5, we have looked at per-sample models of gene-expression. The increasing amount of available epigenetics data in IHEC triggered us to develop a novel approach to elucidate transcriptional regulation on the level of distinct genes, instead of averaging across all genes as done before. We call this method StitchIt, following the analogy that we stitch several data sets together to gain novel biological insights. This chapter is based on the manuscript `"Integrative analysis of epigenetics data identifies gene-specific regulatory elements"`, available at bioRxiv [S+19]. This work is accepted for presentation at the Great Lakes Bioinformatics Conference (GlBio) 2019 at the University of Madison at Wisconsin. In the course of this project, we have collaborated with Alexander Marx and Jilles Vreeken from the Exploratory Data Analysis group at the Cluster of Excellence for Multimodal Computing and Interaction at Saarland University.

8.1 Motivation and research objectives

As described in Chapter 1, the elucidation of transcriptional regulation is a major problem in computational biology. Especially regulatory elements (REMs) such as promoters (Section 2.1.4) and enhancers (Section 2.1.13), harbouring binding sites for TFs are essential to orchestrate cellular processes [V+09a].

Recall from Section 2.1.13 that according to the scanning model, REMs can influence a gene in close proximity. According to the looping model however, they can also influence genes that are several kb away - brought into spatial proximity by chromatin looping [BK98]. The identification of REMs throughout the genome has been addressed by international efforts such ENCODE and Roadmap. There, REMs have been identified using DNaseI-Hypersensitive Sites (DHS) [T+12] via distinct patterns of Histone Modifications (HMs), i.e. the co-occurrence of H3K27ac, H3K4me1 while H3K4me3 is absent [H+07], or from TF-ChIP-seq experiments of proteins such as EP300 [V+09b]. Typically such data sets are analyzed with peak calling algorithms. Although there is a plethora of peak callers available that were designed for ChIP-seq [T+16] and chromatin accessibility data [K+14c], they still have several limitations. For instance, the selection of the cut-off to determine peaks over background is non-trivial and also cell cycle stage [L+17c] or cell numbers [G+12a] can prevent us from accurately detecting all truly enriched regions.

Furthermore, the minimum level of enrichment to make a region biologically active is unclear [CZ10]. As illustrated in Figure 8.1, integrating peak calls across several diverse samples is not straightforward [LS15]. However, an integrated set of peaks is required if machine learning approaches should be utilized to associate a defined set of candidate REMs to potential target genes across many samples, which is required for per-gene learning. Therefore, we wanted to devise a method that does not rely on peak-calling to identify REMs.

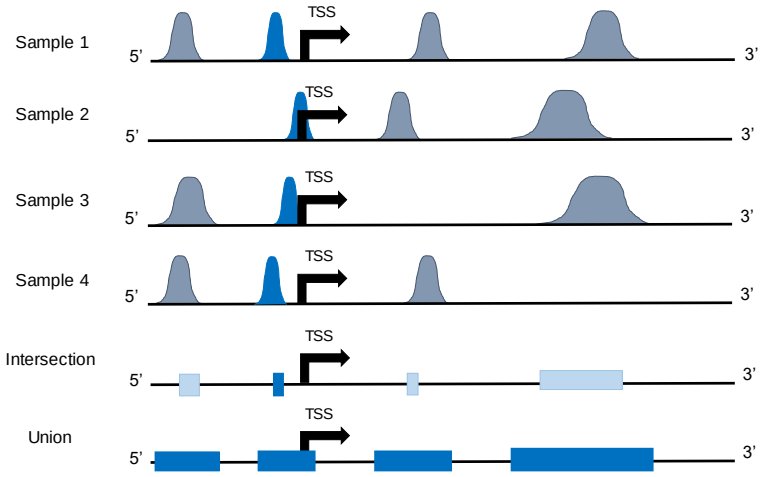


Figure 8.1: Illustration of two common ways to integrate DHSs across multiple samples. In the intersection case, only the accessible portion of the genome across all samples is available. In the union case, all overlapping DHSs are merged. The latter case preserves a large part of the accessible sites, but it looses some of the variance in the data. The intersection, on the other hand, might be too conservative. Figure from Schmidt *et al.* [S$^+$19].

However, identifying REMs throughout the genome does not suffice to learn about their function, as the REMs need to be associated with their target genes as well. In literature, especially in instances where only few replicates are available, putative REMs are often linked to their nearest gene according to genomic distance [G$^+$15b], or aggregated using window based approaches [S$^+$17a, M$^+$12b]. As explained in Section 2.1.13, especially enhancers and repressors do not regulate their nearest gene but may influence more distant genes. Yao *et al.* [Y$^+$15] described two general approaches to account for such long distance regulatory events: (1) methods based on physical interaction, i.e. capture Hi-C and (2) methods based on associating gene-expression to the activity of REMs, e.g. using DNaseI-seq [T$^+$12, H$^+$18b], or HM abundance [E$^+$11]. While methods based on physical interaction are laborious, time consuming and experimentally challenging, e.g. in terms of providing a sufficient resolution of long-range contacts [R$^+$14b], association based methods are predestined to use the plethora of available epigenetics data to link REMs to

their target genes. Several of these methods are detailed in Section 8.3. However, most of them are not available as usable software. Furthermore, as all methods rely on peak-calls, they use either a peak number or a peak distance cut-off to filter candidate regulatory sites. This is another parameter that needs to be set by the user and can not be easily determined in a data driven way. We aimed at designing a method that can be easily applied to user-specific data sets and that does not depend on fixed distance cut-offs.

8.2 Brief summary and outline of this chapter

Our novel method StitchIt fulfills these aforementioned goals. It is an easy-to-use, peak caller independent method that can be applied to user-defined genomic intervals around a gene of interest to suggest candidate REMs. The major difference to any existing method that either links or identifies REMs is that StitchIt solves the combined task of identifying potential REMs and linking them to their putative target genes at the same time.

Basically, StitchIt solves a classification problem by segmenting a large genomic area around the specified target gene. The resulting segmentation highlights regions exhibiting epigenetic signal variance, which is linked to the expression of the analyzed gene. Thus, StitchIt belongs to the class of association based methods delineated by Yao *et al.* [Y+15]. However, StitchIt extends the typical function of this class of methods as it does not require the precise location of candidate REMs as input. As illustrated in Figure 8.2, StitchIt is designed to improve on peak-based approaches by providing a higher resolution and greater accuracy in pinpointing REMs.

In the scope of this chapter as well as in Schmidt *et al.* [S+19], we use DNaseI-seq data to model regulatory activity. We emphasize that the proposed methods are not limited to that input, but can also be applied to other chromatin accessibility assays as discussed Chapter 6, or for example to HM ChIP-seq data.

Within Section 8.3, we sketch relevant related methods to link REMs to their targets. Three approaches against which we compare our StitchIt approach in various validation scenarios are introduced in Section 8.4. Details on StitchIt are provided in Section 8.5, a toy-example illustrating the function of the algorithm is provided in Section 8.6. In Section 8.9 we present an application of StitchIt to various IHEC datasets and analyze the performance of our method with various validation scenarios. Section 8.10 concludes this chapter with a discussion of our method and by pointing out limitations of our approach.

8.3 Related methods linking REMs to genes

Here, we describe several approaches that have been used by others to link putative regulatory elements to their target genes.

Cao *et al.* propose to integrate predicted REMs into cell type specific interaction networks [C+17]. For this purpose, they developed the Jeme method to associate

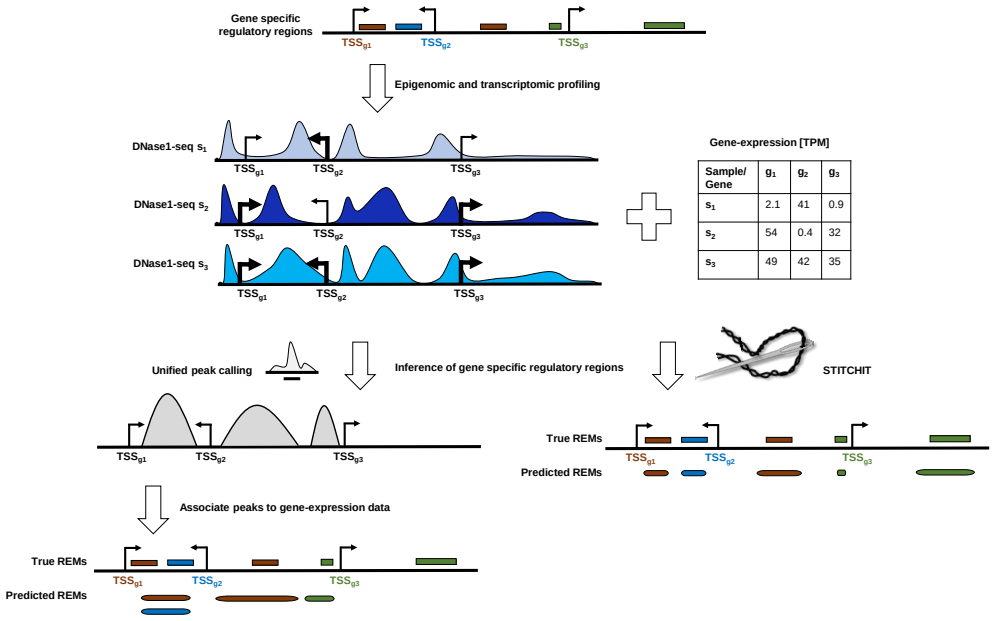


Figure 8.2: Here three genes (g_1, g_2, g_3) and their colour coded REMs are depicted. Their DNaseI-seq and RNA-seq profiles have distinct patterns. One the left, REMs are derived based on correlation tests between called peaks and the expression of the target genes. As shown, the peak based linkage is inaccurate: the neighbouring REMs for g_1 and g_2 are not distinguished accurately due to insufficient resolution during peak calling. Also, a distal REM for g_1 is predicted to be larger than it is and an intragenic REM of g_3 is missed. StitchIt however, uses the chromatin accessibility and gene-expression data in a unified statistical framework to define gene-specific REMs with greater accuracy, as shown on the right hand site. Figure from Schmidt *et al.* [S$^+$19].

enhancers to their target genes using many samples. Briefly, Jeme first extracts candidate enhancer elements in a 1 mb window around all annotated TSSs. These enhancer elements are obtained from ChromHMM segmentations computed for 127 samples from ENCODE and Roadmap. Cao *et al.* "*took the union of the predicted enhancers from all samples, removed those larger than* 2500 *bp, merged the remaining enhancers that overlapped and removed the ones larger than* 2500 *bp again after merging*" [C$^+$17], leading to 489581 candidate enhancers. Next, for each TSS i, a linear model using all its candidate enhancers and the signal of H3K4me1, H3K27me3, H3K27ac and DNaseI-seq computed within those enhancers is trained to predict the expression measurements linked to the TSS i. The model uses Lasso regularization to avoid over-fitting. On the basis of these models that utilize the full spectrum of all candidate enhancers, the importance of the individual enhancers can be computed in terms of error terms er, by learning individual regression models

using only one type of epigenetic signal within one enhancer as a feature at a time. This allows to rank enhancer-TSS pairs *et* by the individual error terms er_{et}.

JEME uses these enhancer-TSS error terms together with the distance information of enhancers and their assigned TSSs as well as the epigenetic signal within the enhancers in Random Forest(RF) models to learn about sample specific enhancer-TSS importance across all enhancer-TSS pairs within one sample. The RFs are trained using data on chromatin contacts such as ChIA-PET data.

Note that within JEME, the definition of enhancers strongly influences the initial set of candidate regions. For instance, the definition could be strictly neglecting regulatory events at the promoter of a gene, or could be more lenient and considering those sites as well. Within STITCHIT we do not require any preselected candidate regions, thereby circumventing this problem in the first place. Our methods simply highlights any region that shows epigenetic signal variance that is associated with expression changes.

The methodology proposed by Hait *et al.* within their method FOCS [H$^+$18b] has some similarities to the strategy implemented in JEME. FOCS requires the user to provide the DNaseI signal within candidate enhancer sites and uses this signal to predict gene-expression. FOCS automatically selects the n closest candidate enhancers for each promoter and builds an ordinary least squares model from that data. In Hait *et al.* the 10 closest enhancers are considered. The importance of individual enhancers is determined in a leave-one-cell-type-out cross-validation procedure. Next, models with poor quality are discarded. Whether a model is of poor quality or not is decided according to the p-value of the Spearman correlation between the predicted and the measured promoter activity. The p-values were computed using the Benjamini-Yekutieli correction [BY01]. Models with a corrected p-value ≤ 0.1 are considered for a more detailed analysis. Specifically, linear models with elastic net regularization and a fixed value for α, the parameter controlling the trade of between lasso and ridge, are used to determine the final coefficients of each candidate enhancer.

Gonzales *et al.* called DHSs on Roadmap DNaseI-seq data for six different cell types using MACS. Only peaks that are reproducible across all replicates of one cell type are used in an iterative heuristic approach to generate an atlas of DHSs that is comprised of DHSs from all six tissues. Gonzales *et al.* suggest to neglect the non-overlapping part of two overlapping peaks if the overlap between the two is $> 75\%$. If the overlap is $\leq 75\%$, the overlapping area is removed and two individual peaks are added to the atlas. Next, all DHSs of the atlas are linked to their target genes by simply assigning each peak to its closest gene in terms of distance to TSS or TTS, depending on which is closer [G$^+$15b].

Shooshtari *et al.* used regulatory sites derived from chromatin accessibility data together with Genome-Wide Association Studies (GWAS) to better pinpoint regulatory events in autoimmune and inflammatory diseases [Sho17]. In more detail, they called DHSs in 350 samples obtained from Roadmap. To compare the chromatin accessibility landscape across samples, Shooshtari *et al.* computed the overlap between neighboring peaks per sample. Using a clustering approach, they combined

overlapping peaks into one merged peak that is defined by its extreme positions. Next, a cluster is termed to be active in a sample, if it is overlapping with at least one DHS site within this sample. Active clusters have been used by Shooshtari *et al.* to prioritize GWAS hits that are likely to be influenced by mutations in regulatory regions [Sho17].

In the FANTOM5 consortium, putative REMs have been linked to their target genes by associating enhancer activity to gene-expression [A+14]. More specifically, they compute the correlation between enhancer activity derived from CAGE data and gene-expression measurements across 808 samples considering primary cells, cell lines and tissues. Unlike the other methods that have been introduced so far, this is the only correlation based method that uses CAGE data and does not rely on DNaseI-seq data.

These examples of recent methods that link REM to their genes require a pre-selected set of regulatory regions and at the same time are often based on peaks, typically DHSs. Also, prior knowledge on regulatory regions is used, e.g. in JEME.

8.4 Alternative approaches used to assess the performance of STITCHIT

To mimic common strategies of REM linkage approaches pursuit in the methods introduced in the previous section and to compare them to our STITCHIT software, we devise three general strategies to identify and to link REMs to their target genes. They are depicted in Figure 8.3. We use an unsupervised, window based aggregation of DHSs per gene and per sample, representing purely distance based approaches. Secondly, we generate the union of DHSs across all samples (UNIFIEDPEAKS) and thirdly we consider known REMs from the GENEHANCER database. Command line arguments along with further details on how to produce the respective scores are provided in Section B.6.

Unsupervised integration of peaks per sample

Similar to earlier work [G+15b, SS18], we determine for each gene g in each sample i how many DHSs c_i^g are located within a distinct window w, how wide the accessible regions l_i^g are and we aggregate the signal intensity within the selected DHSs s_i^g. The contribution of each DHS p is also weighted by its distance $\text{dist}(p, g)$ to the TSS of gene g following an exponential decay.

As introduced in Chapter 3 in the context of Schmidt *et al.* [SS18], these quantities

are computed as:

$$c_i^g = \sum_{p \in P_{w,g}} I(p) \cdot \exp\left(\frac{-\text{dist}(p,g)}{d_0}\right), \tag{8.1}$$

$$l_i^g = \sum_{p \in P_{w,g}} |p| \cdot \exp\left(\frac{-\text{dist}(p,g)}{d_0}\right), \tag{8.2}$$

$$s_i^g = \sum_{p \in P_{w,g}} s(p) \cdot \exp\left(\frac{-\text{dist}(p,g)}{d_0}\right). \tag{8.3}$$

Here, I is the indicator function, $P_{w,g}$ refers to all DHSs p that overlap the window w around gene g,d_0 is a constant set to 5000, $|p|$ is the genomic length of p and $s(p)$ refers to the DNaseI-seq signal within p. In this study, we considered three different instances for the window w: a 5kb window centered at the 5′TSS of g, a 50kb window centered at the 5′TSS of g and a 2.5kb of the TSS of g as well as the entire gene body of g.

As the region-specific view on the data is lost within this scoring methodology, it is not considered for interpretation purposes at later stages of this Chapter.

Unified peaks

Here, we generate aggregations of DHSs across all samples under consideration. Overlapping DHSs are merged using the BEDTools [QH10] merge command. Thereby, we obtain a set of regions representing all accessible sites within a dataset. Using the bigwig files generated with DEEPTools [R$^+$14a] and the libBigWig library (`https://zenodo.org/record/45278`), we compute the DNaseI-seq signal within the merged peaks for each sample. Next, we test for all candidate peaks within a distinct window w, here $w = 25$kb, upstream of a genes TSS and downstream of its TTS, whether there is a significant correlation ($p \leq 0.05$) between the DNaseI-seq signal within the peak and the expression of the gene. All merged peaks passing this test ($\mathcal{U}$) are considered to be a candidate regulatory element. We refer to this as the UnifiedPeaks approach. This methodology is conceptually similar to the peak aggregation approaches suggested by Hait *et al.* [H$^+$18b] and Shooshtari *et al.* [Sho17] introduced above.

GeneHancer

For all REMs obtained from the GeneHancer database, we calculate the sample-specific DNaseI-seq signal within each region for each gene using the libBigWig library. Note that a window or distance cut-off is not required because each region is already assigned to its putative target gene. Considering that the GeneHancer database is comprised of REMs originating from many different sources identified with a plethora of assays and molecular signatures, we perform the same correlation-based test as with the UnifiedPeaks approach to identify a subset ($\mathcal{G}$) of regions with sufficient correlation between the DNaseI-seq signal and the gene-expression of the respective target gene.

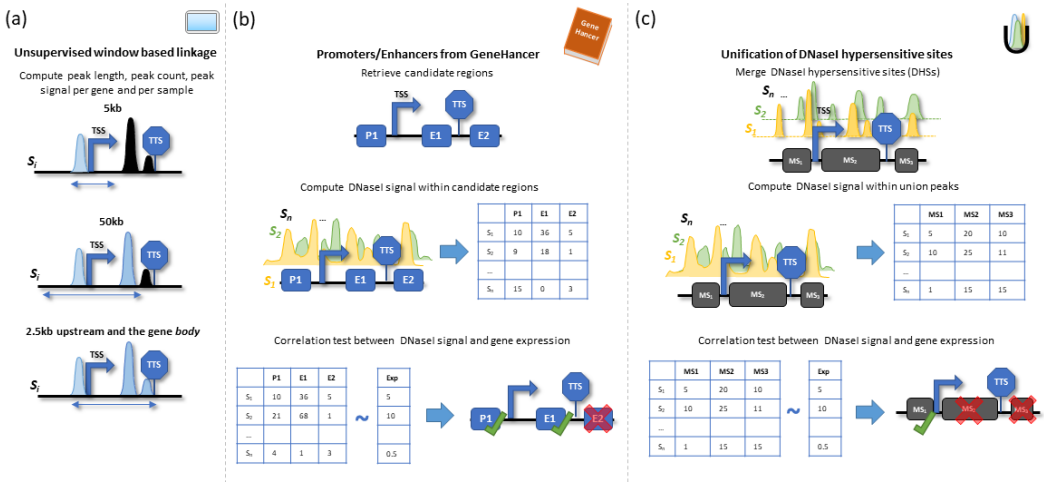


Figure 8.3: (a) Following a window based approach, as normally applied in per sample models, we assess length, count and signal of DNase hypersensitive sites (DHSs) per gene and per sample using three different windows: 5kb, 50kb and a gene body window. (b) Also, we consider a curated set of promoters and enhancers contained in the GeneHancer database. Next, we select regions whose DNaseI-seq signal correlate with the expression of the tested target gene. Another approach is depicted in (c). Here, we identify the union of all DHSs and select regions that exhibit DNaseI-seq signals that correlate with the expression of the target gene. Figure from Schmidt *et al.* [S+19].

8.5 Overall workflow of STITCHIT

Conceptually, we pursue the idea of identifying regions in large genomic intervals around a gene of interest that can be associated with the gene's expression variation across many samples. To identify such regions we utilize paired epigenetics and gene-expression data. The StitchIt algorithm uses the actual signal of the epigenetics data to highlight segments of the data showing signal variation that can be used to separate samples according to the expression of the target gene. Thus, the peak-calling step can be omitted and the two tasks of identifying regulatory sites and their linkage to targets are solved simultaneously. To refine the list of putative REMs identified by StitchIt, we apply a two-level learning approach that is detailed below in Section 8.7. The two-level learning enables us to judge the explanatory power of the found regions for gene-expression and to obtain a p-value for the significance of each identified site. The workflow of the proposed methodology is depicted in Figure 8.2.

In the following we are given a dataset D_g with m rows, corresponding to the samples and n columns representing the epigenetic signal at base pair resolution around the target gene g. Further, we assign a class label to each row indicating whether the corresponding sample is associated with a high, medium, or low ex-

pression value ($C = 0, 1, 2$). Depending on the distribution of the gene-expression values, also a two-level classification was used here ($C = 0, 1$).

StitchIt can be used with any number of distinct class labels that is smaller or equal to the number of samples. With C_k we relate to all rows to which we assigned class label $k \in C$.

A segment s has a start point i and an end point j, where $1 \leq i \leq j \leq n$. We call S_g a segmentation of D_g, if it contains a set of non-overlapping segments that covers the entire range from 1 to n. There are two trivial segmentations: Firstly, a segmentation consisting of only a single segment with start point $i = 1$ and end point $j = n$ and secondly the segmentation containing n segments, where each segment contains only a single column that is the DNaseI-seq signal at single base. The former would contain no information about the class labels, while the latter would consist of a large set of noisy segments. Our goal is to provide a small set of robust features for the learning step. We achieve this by joining adjacent base pairs to segments such that the variance between the epigenetic signals of base pairs that are contained in a segment is low, w.r.t. to the class labels representing the discrete gene-expression state. The optimal segmentation according to the score we define below finds a trade-off between the number of segments and the variance.

To score a segmentation we propose an information theoretic score based on the Minimum Description Length (MDL) principle [Grü07b]. MDL is a practical instantiation of Kolmogorov complexity [Kol68] and thus belongs to the class of compression-based scores. Formally, given a model class $\mathcal{M}$, MDL identifies the best model $M \in \mathcal{M}$ for data D as the one minimizing

$$L(D, M) = L(M) + L(D \mid M) \,, \tag{8.4}$$

where $L(M)$ is the length of the description of the model M in bits and $L(D \mid M)$ is the length of the description of the data D given M in bits. This is known as two-part, or crude MDL. In essence, we try to find the simplest model that can explain the data well. We follow the convention that all logarithms are base two, since the length of the encoding relates to bits and define $0/\log 0 = 0$. In this work, we use MDL to balance our segmentation between having too few segments and running at risk of missing structure in the data and finding too many segments, which contain spurious information and make the post-processing infeasible.

From now on, we consider the model class of segmentations $\mathcal{S}$ from which we want to find the optimal segmentation S_g^{opt} that is

$$S_g^{opt} = \arg\min_{S_g \in \mathcal{S}} (L(S_g) + L(D_g \mid S_g)) \tag{8.5}$$

In particular, we encode a segmentation S_g as follows

$$L(S_g) = L_{\mathbb{N}}(|S_g|) + |S_g||C| \log_2 \left(\frac{|\max(D_g) - \min(D_g)|}{\tau} \right) + \log_2 \binom{n-1}{|S_g|-1}, \tag{8.6}$$

where $|S_g|$ denotes the number of segments, $\min(D_g)$ refers to the minimum value occurring in D_g, $\max(D_g)$ refers to the maximum value occurring in D_g, $|C|$ is the

number of class labels, τ is the data resolution. The value of τ is smaller or equal to one. The smaller it is, the more accurate is the representation of floating-point numbers. $L_{\mathbb{N}}$ is the universal prior for integer numbers [Grü07b], which is defined recursively:

$$L_{\mathbb{N}}(x) = \begin{cases} \log_2(2,865064), & \text{if } (x \leq 0), \\ L_{\mathbb{N}}(\log_2(x)) + \log_2(x), & \text{otherwise.} \end{cases} \tag{8.7}$$

In summary, we first compute the costs to encode the number of all segments ($L_{\mathbb{N}}(|S_g|)$), then for each segment per category we calculate the associated mean signal value of the considered epigenetic assay by assuming it lies between the minimum and the maximum value in the data ($|S_g||C| \log_2 \left(\frac{|\max(D_g) - \min(D_g)|}{\tau} \right)$)and lastly we determine model complexity to select $|S_g|$ segments from possible n segments ($\log_2 \binom{n-1}{|S_g|-1}$).

To compute $L(D_g \mid S_g)$, the costs for representing the data given a segmentation, we simply sum over the costs per segment

$$L(D_g \mid S_g) = \sum_{s \in S_g} \sum_{k \in C} \frac{1}{|C_k|} L(D_g \mid s, k) \ , \tag{8.8}$$

where $|C_k|$ corresponds to the number of rows associated with class label k and $L(D_g \mid s, k)$ refers to the cost for gene g's segment s being assigned to class k. To encode the costs for a specific segment and the data associated with class k, we encode the error assuming a Gaussian distribution. Using $\hat{\sigma}^k$ as the standard deviation of the data corresponding to segment s and class label k, we get (compare [Grü07b])

$$L(D_g \mid s, k) = \frac{|s||C_k|}{2} \left(\frac{1}{ln(2)} + \log_2(2\pi \hat{\sigma}^{k^2}) \right) + |s||C_k| \log \tau, \tag{8.9}$$

$$\hat{\sigma}^k_{i,j} = \sqrt{\frac{1}{n^k_{i,j}} \left((SS_k[j] - SS_k[i-1]) - \frac{1}{n^k_{i,j}(S_k[j] - S_k[i-1])^2} \right)}, \tag{8.10}$$

$$n^k_{i,j} = (j + 1 - i) \cdot |C_k|, \tag{8.11}$$

$$S_k[j] = S_k[j-1] + \sum_{i \in C_k} D_{g_{i,j}}, \tag{8.12}$$

$$SS_k[j] = SS_k[j-1] + \sum_{i \in C_k} D^2_{g_{i,j}}. \tag{8.13}$$

with $|s|$ being the length of the segment. To find the optimal segmentation S^{opt}_g, we use dynamic programming [Bel54]. In essence, we start with a segmentation containing only one single segment. Then we iteratively compute the best segmentation containing i segments based on the best segmentation containing $i-1$ segments for $i \in \{2, \ldots, n\}$. Lastly, we select S^{opt}_g among the optimal segmentations for each possible number of segments. The runtime complexity of this algorithm is $\mathcal{O}(n^2)$. By selecting a minimum segment size of β and partitioning the search space

into l chunks, we can run each chunk in parallel and the total runtime complexity reduces to $\mathcal{O}(\frac{n^2}{l\beta^2})$. In our experiments, we use $\beta = 10$ and set l to $\lceil \frac{n}{5000} \rceil$, which makes the algorithm feasible to be applied to large genomic intervals. Here, we have considered 25kb upstream of a gene's Transcription Start Site (TSS) and 25kb downstream of a gene's Transcription Termination Site (TTS), although also larger areas can be chosen at the users convenience. As shown in Figure 8.4, the runtime of StitchIt is still feasible even if a windows of 1mb is considered.

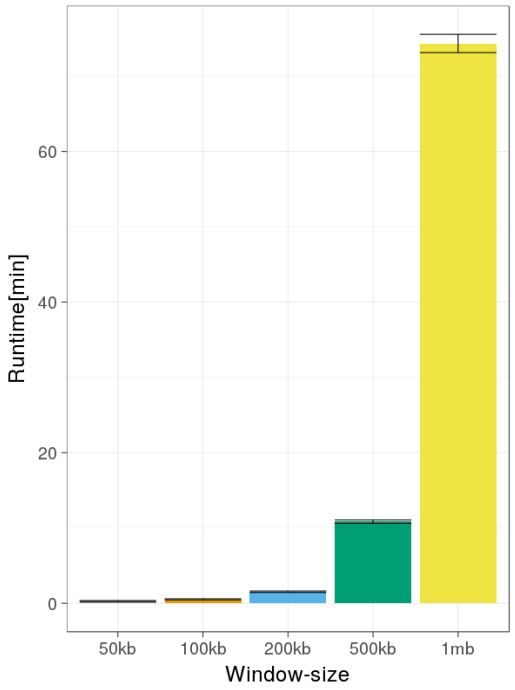


Figure 8.4: Runtime [min] of StitchIt depending on the window size that is the sum of the up- and downstream extension of the considered search window. Figure from Schmidt *et al.* [S+19].

Upon completion of the StitchIt algorithm, those segments that are associated with the observed expression changes need to be extracted from S_g^{opt}. Thus, for all segments $s \in S_g^{opt}$ we compute both Pearson and Spearman correlation between the epigenetic signal in s across all samples m and the continuous expression values of the target gene g. We select all segments that achieve either a Pearson or Spearman correlation with a significance threshold of $p \leq 0.05$.

8.6 A toy example illustrating STITCHIT

The following example showcases a run of the STITCH algorithm on a data matrix $\mathcal{D}$ consisting of six samples with six initial segments ($\beta = 1$). The example tries to find the segmentation that best explains the expression difference for two classes, labelled 0 and 1, respectively. Each class has been assigned to three samples.

The data matrix $\mathcal{D}_{i,j}$ holding raw counts of the data that should be segmented is given as:

	b1	**b2**	**b3**	**b4**	**b5**	**b6**	**exp**
s1	10	9	1	1	7	8	0
s2	11	10	2	1	7	8	0
s3	10	8	1	0	9	8	0
s4	2	4	25	24	0	1	1
s5	3	5	23	22	1	0	1
s6	2	6	26	25	0	1	1

Using $\mathcal{D}$, we compute the sum S and the sum of squares SS for each column of $\mathcal{D}$ per class k according to

$$S_k[j] = S_k[j-1] + \sum_{i \in C_k} D_{i,j}, \tag{8.14}$$

$$SS_k[j] = SS_k[j-1] + \sum_{i \in C_k} D_{i,j}^2. \tag{8.15}$$

$S_0[1]$	$S_0[2]$	$S_0[3]$	$S_0[4]$	$S_0[5]$	$S_0[6]$	$S_1[1]$	$S_1[2]$	$S_1[3]$	$S_1[4]$	$S_1[5]$	$S_1[6]$
31	58	62	64	87	111	7	22	96	167	168	170

$SS_0[1]$	$SS_0[2]$	$SS_0[3]$	$SS_0[4]$	$SS_0[5]$	$SS_0[6]$	$SS_1[1]$	$SS_1[2]$	$SS_1[3]$	$SS_1[4]$	$SS_1[5]$	$SS_1[6]$
321	566	572	574	753	945	17	94	1924	3609	3610	3612

To calculate the data costs $w_{i,j}^k$ for each hypothetical segment from position i to j, with $i \leq j$, we need to compute the empirical standard deviation $\hat{\sigma}_{i,j}^k$ as well as the number of data points within a bin $n_{i,j}^k$. This is possible in constant time using the precomputed vectors S_k and SS_k [E06], where

$$\hat{\sigma}_{i,j}^k = \sqrt{\frac{1}{n_{i,j}^k}\left((SS_k[j] - SS_k[i-1]) - \frac{1}{n_{i,j}^k (S_k[j] - S_k[i-1])^2}\right)}, \tag{8.16}$$

$$n_{i,j}^k = (j + 1 - i) \cdot |C_k|, \tag{8.17}$$

where $|C_k|$ denotes the number of samples related to class k.
It holds that $n_{i,j}^0 = n_{i,j}^1$:

i\j	1	2	3	4	5	6
1	3	6	9	12	15	18
2	0	3	6	9	12	15
3	0	0	3	6	9	12
4	0	0	0	3	6	9
5	0	0	0	0	3	6
6	0	0	0	0	0	3

The standard deviation for class 0 $\hat{\sigma}^0_{i,j}$ evaluates to

i\j	1	2	3	4	5	6
1	0.471405	0.942809	4.012327	4.403282	4.069398	3.804237
2	0	0.816497	3.890873	3.829708	3.771236	3.627059
3	0	0	0.471405	0.57735	3.224137	3.47511
4	0	0	0	0.471405	3.578485	3.435472
5	0	0	0	0	0.942809	0.687184
6	0	0	0	0	0	0

and for class 1 $\hat{\sigma}^1_{i,j}$ we have

i\j	1	2	3	4	5	6
1	0.471405	1.490712	10	10.34777	10.73437	10.55789
2	0	0.816497	9.889669	9.113821	10.92748	11.02643
3	0	0	1.247219	1.34371	11.29186	11.87668
4	0	0	0	1.247219	11.7047	10.9522
5	0	0	0	0	0.471405	0.5
6	0	0	0	0	0	0.471405

From $\hat{\sigma}^k_{i,j}$ and $n^k_{i,j}$, we compute the cost matrices $w^k_{i,j}$ following the MDL encoding for normally distributed data points [Grü07b]. For notational convenience, we write s to denote to the segment ranging from i to j. The data costs for segment s and class k can be computed as

$$L(D|s,k) = w^k_{i,j} = \frac{n^k_{i,j}}{2}\left(\frac{1}{ln(2)} + \log_2(2\pi\hat{\sigma}^k_{i,j}\hat{\sigma}^k_{i,j})\right) + n^k_{i,j}\log_2\tau. \tag{8.18}$$

Here τ is the data resolution. Because we consider only integers in this example, we can set $\tau = 1$. According to Equation 8.14, we obtain for $w^0_{i,j}$

i\j	1	2	3	4	5	6
1	5.705249	17.4105	44.92036	61.50349	75.17291	88.45775
2	0	8.082693	29.68083	44.31552	58.821	72.68268
3	0	0	5.705249	13.16539	42.08063	57.40525
4	0	0	0	5.705249	28.95637	42.90498
5	0	0	0	0	8.705249	14.67289
6	0	0	0	0	0	0

and for $w^1_{i,j}$

i\j	1	2	3	4	5	6
1	5.705249	21.37628	56.77776	76.29552	96.16316	114.9653
2	0	8.082693	37.75581	55.57291	77.2392	96.7441
3	0	0	9.916281	20.4776	58.35531	78.68126
4	0	0	0	9.916281	39.21437	57.95874
5	0	0	0	0	5.705249	11.92027
6	0	0	0	0	0	5.705249

The total cost matrix $\mathcal{W}$ is defined as

$$\mathcal{W}_{i,j} = \sum_{l=1}^{k} \frac{w^k_{i,j}}{|C_k|}. \tag{8.19}$$

In our example, we get for $\mathcal{W}$

i\j	1	2	3	4	5	6
1	3.803499	12.92893	33.89937	45.93301	57.11203	67.80769
2	0	5.388462	22.47888	33.29614	45.3534	56.47559
3	0	0	5.207177	11.21433	33.47865	45.36217
4	0	0	0	5.207177	22.72358	33.62124
5	0	0	0	0	4.803499	8.864386
6	0	0	0	0	0	1.90175

Using dynamic programming [NV15], we compute the costs for all possible segmentations according to the recursion formula

$$c_{i,j} = \min_{k=1,\ldots,j-1}(c_{i-1,k} + W_{k+1,j}), \tag{8.20}$$

yielding the cost matrix $\mathcal{C}$.

i\j	1	2	3	4	5	6
1	3.803499	12.92893	33.89937	45.93301	57.11203	67.80769
2	0	9.191961	18.1361	24.14326	46.40757	54.79739
3	0	0	14.39914	20.40629	28.94675	33.00764
4	0	0	0	19.60631	25.20979	29.27068
5	0	0	0	0	24.40981	28.4707
6	0	0	0	0	0	26.31156

After computing $L(D|S)$, we need to compute the costs of $L(S)$. These are obtained according to

$$L(S) = L_{\mathbb{N}}(|S|) + |S||C| \log_2 \left(\frac{|\max(D) - \min(D)|}{\tau} \right) + \log_2 \binom{n-1}{|S|-1}, \tag{8.21}$$

where $|C|$ is the number of possible class labels and $L_{\mathbb{N}}(|S|)$ refers to the *optimal encoding for natural numbers* as defined above [Grü07b, p.100].

Thereby, we obtain $L(S)$ as:

Level 1	10.91945
Level 2	23.64225
Level 3	35.29254
Level 4	45.44401
Level 5	54.66348
Level 6	62.33327

For the combined costs $L(S)+L(D|S)$ we get:

Level 1	78.72714
Level 2	78.43965
Level 3	**68.30019**
Level 4	74.71469
Level 5	83.13418
Level 6	88.64484

As demonstrated above, the smallest costs are obtained for the segmentation at level 3. We perform backtracking through the dynamic programming matrix to identify the corresponding segmentation. Below, the backtracking path is highlighted in blue.

i\j	1	2	3	4	5	6
1	3.803499	12.92893	33.89937	45.93301	57.11203	67.80769
2	0	9.191961	18.1361	24.14326	46.40757	54.79739
3	0	0	14.39914	20.40629	28.94675	33.00764
4	0	0	0	19.60631	25.20979	29.27068
5	0	0	0	0	24.40981	28.4707
6	0	0	0	0	0	26.31156

We obtain the final segmentation $b_{1,2}$-$b_{3,4}$-$b_{5,6}$, which merges two consecutive bins into one segment.

8.7 A two-level learning approach to refine suggested regulatory elements

StitchIt provides a matrix X holding the epigenetic signal for all selected segments $s \in S_{opt}$. The m rows of X denote the samples, the n columns refer to the regions selected by StitchIt. To further refine the suggested regions for a distinct gene g, we first train a linear model using elastic net regularization, as implemented in the glmnet R-package [FHT10]. Here, we are utilizing the DNaseI-seq signal within candidate REMs (X) to predict the expression of g. We use the same learning scheme as introduced in Section 3.4.1.

Significance of the correlation between predicted and measured gene-expression is adjusted using the Benjamini-Yekutieli correction [BY01], which is designed to account for dependency between the tests. This has been used before in Hait *et al.* as well [H$^+$18b]. Only models with a q-value ≤ 0.05 are considered for interpretation of the selected regions. The q-value is the adjusted p-value obtained from the Benjamini-Yekutieli correction. For those models, we refer to all features with a median non-zero regression coefficient across the outer folds by X_{NZ}.

In a second learning step, similar to Hait *et al.* [H$^+$18b], we train an Ordinary Least Squares model (OLS) on the pre-selected features X_{NZ} predicting y and report the regression coefficients β_{OLS} as well as the p-values per feature for downstream analysis:

$$y = X_{NZ}\beta_{OLS}. \tag{8.22}$$

The OLS model allows for a simple comparison of regression coefficients β_{OLS} across genes, as there is no bias introduced by the regularization as in elastic net. Also, OLS provides a straight forward way to compare individual REMs. Therefore, we invert the order of using the elastic net and the simple OLS model compared to Hait *et al.*. All regions and model coefficients used for interpretation and validation are obtained from the OLS models.

8.8 Implementation & usability

We have implemented the StitchIt algorithm, the UnifiedPeaks approach and a linking using previously defined regions (e.g. from GeneHancer) using C++. Each linkage method is available as a separate executable in our repository. The code can be easily build using cmake (version $\geq$ 3.1) and requires a C++11 compiler supporting openmp for parallel execution of StitchIt. We have thoroughly tested StitchIt using googletest. Scripts for the unsupervised peak linkage are available at `www.github.com/schulzlab/TEPIC`. All the other code is available at `www.github.com/schulzlab/STITCHIT`.

8.9 Application of STITCHIT to IHEC data sets

We applied StitchIt to a large collection of paired, uniformly reprocessed DNaseI-seq and RNA-seq samples from Blueprint, ENCODE and Roadmap to determine gene-specific REMs. These datasets are very different. The Blueprint dataset is rather homogeneous representing a wide spectrum of the haematopoietic lineage, the ENCODE dataset is composed of few diverse samples and the Roadmap dataset is a large, highly diverse, heterogeneous dataset. Thus, these three datasets are ideal to test the capabilities of StitchIt.

8.9.1 Data and processing

Paired DNaseI-seq and RNA-seq data were downloaded from the ENCODE data portal for 41 ENCODE and 110 Roadmap samples. Upon granted access, we obtained 56 paired DNaseI-seq and RNA-seq samples from Blueprint. An overview is provided in Table 8.1 on sample numbers and tissue/cell type diversity. In Section B.6 all data accession numbers are listed. Paired samples are required as they are expected to have a better correlation between chromatin structure and gene-expression, because both samples originate from the same donor. Details on data processing as well as used command calls are provided in Section B.6.

Table 8.1: Overview of the data used in this chapter.

	Blueprint	**Roadmap**	**ENCODE**
#paired samples	56	110	41
#different cell types	13	33	25
primary cells only	Yes	No	No

Further, we obtained H3K27ac data in *wig* format from the Blueprint data portal for four samples (C0011IH1, S00C0JH1, S00XUNH1, C0010KH1). REMs contained in the GeneHancer database were obtained from the GeneLoc website [P+17b].

Files generated within Schmidt *et al.* [S+19] are available at Zenodo (10.5281/zenodo.2547384). The genome annotation file from GenCode [H+12a] as well as the

candidate REMs from the GeneHancer database are included in the StitchIt repository.

8.9.2 Performance of gene-specific expression models

Prior to a biological evaluation of the suggested REMs, we investigated their general characteristics. Here, REMs computed using StitchIt, the UnifiedPeaks approach, the GeneHancer database and using the aggregation of individual DHSs for the datasets originating from Blueprint, ENCODE and Roadmap are investigated more closely.

As illustrated in Fig. 8.5a both StitchIt and UnifiedPeaks identify more candidate regions per gene than GeneHancer. Simultaneously, the regions retrieved by StitchIt and UnifiedPeaks are shorter than those extracted from GeneHancer (Fig. 8.5b). The same observation is made using Pearson correlation as a measure to filter candidate regions (data not shown). This suggests that although StitchIt predicts more individual segments, the total genomic space covered by those might not be larger than that of UnifiedPeaks regions. As shown in Table 8.2, the UnifiedPeaks regions indeed cover a larger fraction of the genome than StitchIt and GeneHancer regions.

Figure 8.5c depicts the number of genes for which a model could be learned per consortia and linkage method. StitchIt and UnifiedPeaks segments lead to more statistically significant models than GeneHancer segments, while StitchIt has a slight quantitative advantage over UnifiedPeaks.

In Fig. 8.5d, the Spearman correlation of elastic net models predicting gene-expression from the DNaseI-seq signal within the identified REMs is depicted. For ease of comparison, we only show model performance for genes that are covered by each tested method. For each gene, the boxes labeled *Individual peaks* show only the best performing model based on either the 5kb, 50kb, or the *gene body* window. Across all datasets, we observe that models based on StitchIt regions achieve a significantly better correlation ($p \leq 0.0001$) than models based on the other approaches. This is independent of the correlation measure used for the initial filtering of REMs within StitchIt, UnifiedPeaks and GeneHancer. In a gene-to-gene comparison, as shown in Fig. 8.5e exemplary for Roadmap data, StitchIt shows a favorable performance, too.

In case of Blueprint data, we observe that the absolute performance difference between StitchIt and UnifiedPeaks is less pronounced than for the other two datasets. This is also reflected by the number of selected regions and their length. The median length of selected regions is more similar for Blueprint data between StitchIt and UnifiedPeaks than for ENCODE and Roadmap data (Fig. 8.5b). At the same time, their average length is almost identical, in contrast to the segments identified in the other two datasets (Fig. 8.5a). In terms of total genomic coverage, as indicated in Table 8.2, the UnifiedPeaks approach covers about 1.65 times the space covered by StitchIt on Roadmap data, whereas the difference on Blueprint data is far less (1.02). Further, our results clearly indicate that the supervised generation of REMs outperforms the unsupervised selection considerably, as

different window sizes used with the unsupervised approach do not generalize well across different genes (Figure B.6). We found that using Spearman correlation for the internal filtering leads to a better model performance and thus, we decided to use Spearman correlation for all remaining experiments in the manuscript [S+19].

Table 8.2: Total genomic space covered by the predicted REM in various data sets

	Blueprint	**ENCODE**	**Roadmap**
StitchIt	162,453,061	33,229,288	165,805,230
UnifiedPeaks	165,590,118	70,571,718	273,259,579
GeneHancer	137,722,391	109,432,986	155,548,667

8.9.3 STITCHIT generates an extensive catalogue of REMs

To better understand the nature of StitchIt REMs, we have conducted several statistical analyses. Not only the number and the length of REMs is different between datasets (Figure 8.5a,b), we also find that the overlap in terms of genes for which a model could be learned, is generally less than 50% between two datasets (Figure B.7, independent of the method used for computation. Specifically for StitchIt, only 12.7% (4477) of all gene-specific models are shared between Blueprint, Roadmap and ENCODE. Roughly 23% (8214) of all genes could be exclusively modeled using Blueprint data, about 2% (6917) with Roadmap and 6% (2175) with ENCODE.

To gain a better understanding of the characteristics of the suggested regions learned with StitchIt, we overlapped them with the Ensembl Regulatory Build (ERB) [Z+15] (release 86). We considered the terms: Open chromatin, Promoter, Promoter Flanking Region, TF binding site and Enhancer to compare predicted REMs to an established regulatory annotation of the genome.

Interestingly, although the absolute numbers differ, the relative trend among the different datasets is similar: most StitchIt regions do not overlap with an annotated region (Figure 8.6b-d). Thus, they are labelled as *Unknown*. Only in about a quarter of all cases an overlap is found with a state annotated as Promoter, Promoter flanking region, TF binding site, Enhancer, or Open chromatin. The question arises whether the remaining regions are simply noise or whether they reflect REMs that have not been annotated so far. To investigate whether these unknown REMs are performing regulatory functions, we assessed the H3K27ac signal in four randomly chosen Blueprint samples within windows of size 1kb centered in the middle of the considered REMs. The signal was calculated within the top $10,000$ regions per class contained in the ERB as well as in a randomly shuffled set of the same size. As indicated in Figure 8.6e the strongest H3K27ac signal occurs within *Promoter* and *Promoter Flanking Regions*. Importantly, the signal of the randomly distributed regions is the lowest. The signal of the *Unknown* regions is similar to that of *TF binding sites* and *Open Chromatin* suggesting that these regions do have a regulatory effect.

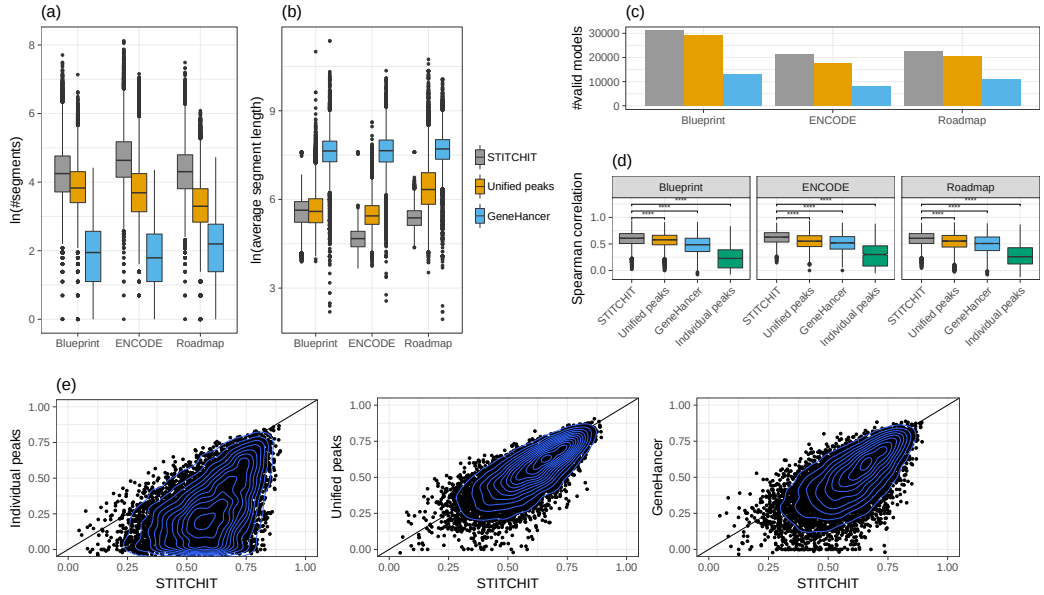


Figure 8.5: (a) The natural logarithm of the number of segments selected by StitchIt, UnifiedPeaks and GeneHancer is shown for each dataset, respectively, whereas in (b), the average length of the selected segments is depicted. The number of learned models is shown in (c), separately per consortia and method. (d) Boxplots showing Spearman correlation between predicted and measured gene-expression using linear regression with elastic net penalty considering all regions identified by StitchIt, the UnifiedPeaks approach, GeneHancer and individual peak aggregation, respectively for Blueprint, ENCODE and Roadmap data. Within StitchIt, UnifiedPeaks and GeneHancer Spearman correlation was used for the initial filtering of candidate regions. Within each consortia, the same set of genes is displayed to allow comparability (Blueprint: 11140 , ENCODE: 2057 , Roadmap: 9102). As indicated by a two-sided t-test, StitchIt regions achieve the best model performance (****: $p \leq 0.0001$). The estimated values for the variances are: 0.018, 0.017 ,0.029, 0.038 (Blueprint), 0.018, 0.024, 0.032, 0.041 (Roadmap), 0.016, 0.021 ,0.026, 0.048 (ENCODE), for StitchIt, UnifiedPeaks, GeneHancer and the peak aggregation, respectively. (e) Scatter plots comparing the performance of StitchIt (x-axis) against the individual peak aggregation, UnifiedPeaks and GeneHancer regions (y-axis) on Roadmap data. Each plot shows 9102 genes. A single dot represents the performance of gene-expression models for a distinct gene. Figure from Schmidt *et al.* [S+19].

Additionally, the trends between the datasets are varying, e.g. we find most associations with StitchIt on the Roadmap dataset, whereas in case of the Uni-

fiedPeaks approach, we find most associations on Blueprint data.

In Figure 8.7, we sketch the distribution of StitchIt regions around a gene. We find that most StitchIt regions are located in intragenic positions that is within the gene body of their respective target gene. Importantly, we do not observe an enrichment in identified sites upstream of the considered genes' TSSs indicating that we do not focus on promoter specific features. The histograms of Fig. 8.6f illustrate how the number of associated REMs are distributed among their target genes. As expected, the distribution for GeneHancer is different from that of UnifiedPeaks and StitchIt. While the latter predict up to 20-30 REMs per gene depending on the dataset, GeneHancer reaches the optimum at 1-4 predicted sites per gene. Our results also indicate that StitchIt tends to find more sites per gene than the UnifiedPeaks approach.

8.9.4 Validation of suggested REMs with external data

All details on the methods used for the following analyses can be found in Section B.6.3.

As StitchIt can be used to learn interactions with distant sites, we compare the learned interactions to ChIA-PET data for K562 and MCF-7 cells as well as to Promoter-Capture Hi-C data for GM12878 cells. On Blueprint and Roadmap data, about one third of all possible interactions overlap with predictions by StitchIt and UnifiedPeaks, on ENCODE data about one sixth of all ChIA-PET contacts are retrieved (Figure 8.8a). While GeneHancer constantly finds the smallest overlap, StitchIt segments overlap with more chromosomal contacts using Blueprint and Roadmap data, while the UnifiedPeaks approach finds more on ENCODE data. There is no clear advantage for any method on the Promoter-Capture data (Figure 8.9d).

Another approach to assess the reliability of our predictions is to compute the number of recovered interactions from GeneHancer. As shown in Figure 8.8b, about $32 - 36\%$ of GeneHancer REMs are retrieved in case of Blueprint and Roadmap, respectively and about 18% using ENCODE data. In the latter case the UnifiedPeaks approach finds marginally more overlapping regions, whereas StitchIt finds more known associations with Blueprint and Roadmap datasets.

The COSMIC database is a collection of known somatic mutations in cancer. Especially mutations occurring in the non-coding part of the genome overlapping REMs might affect TF binding sites in those regions and thereby cause changes in gene-expression that contribute to the progression of cancer. By overlapping our predictions with all non-coding mutations stored in COSMIC, we suggest to which gene the non-coding somatic mutation can be linked (Supplementary material of Schmidt *et al.* [S$^+$19]). In total, we find $1,006,848$ associations between $883,111$ somatic mutations and $22,588$ distinct genes. StitchIt is especially well suited for this task, as its overall enrichment of regulatory sites is higher compared to UnifiedPeaks and GeneHancer regions (Figure 8.8c) while more mutations can be linked (Figure 8.9b). Interestingly, randomly selected regions overlap more mutations than predicted REMs. This suggests that the sequence in the predicted

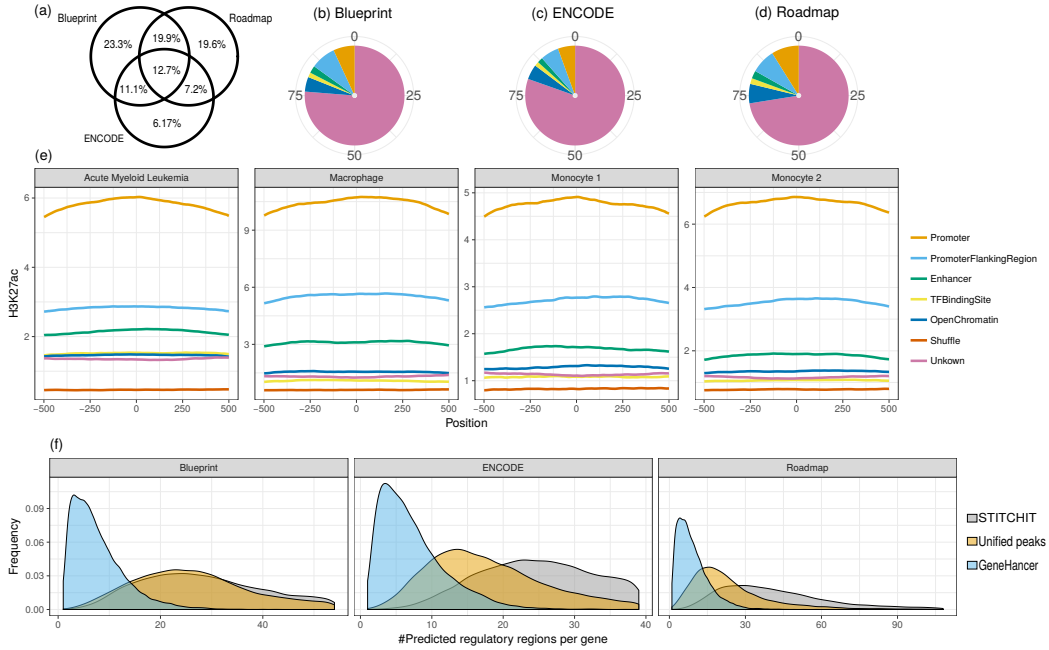


Figure 8.6: (a) The Venn diagram illustrates the overlap of valid StitchIt models between the datasets. (b-d) The distribution of a mapping of StitchIt sites to the ERB is shown. (e) H3K27ac signal shown for randomly shuffled regions, as well as StitchIt regions split according to the ERB categories. H3K27ac signal is shown for four Blueprint samples in a window of 1kb centered in the middle of the putative REMs. StitchIt regions overlapping *Promoter* or *Promoter Flanking Regions* show the highest H3K27ca signal, while the signal in randomly determined regions is the lowest. Most regions that were labeled as unknown have a similar signal intensity as sites labeled as a *TF binding site* or *Open Chromatin*. (f) The density plots delineate the number of predicted REMs per gene, shown separately for the used datasets and tested methods. Figure from Schmidt *et al.* [S+19].

REMs is conserved and thus likely to be functionally relevant (Figure 8.9b).

Similarly, we overlape GWAS sites from the EMBL-EBI GWAS catalog with our predictions suggesting to which gene(s) GWAS hits might be associated (Supplementary material of Schmidt *et al.* [S+19]). Overall we find 4697 associations with Blueprint, 888 with ENCODE and 4588 with Roadmap data covering 3394, 753 and 3366 genes, respectively using StitchIt. Similar to above, StitchIt yields a better enrichment score than the other two methods (Fig. 8.8d). Compared to a random setting, all actual regions yield significantly more associations and obtain a significantly better enrichment score (Figure 8.9a).

Expression quantitative trait loci (eQTLs) are distinct genomic loci that are linked to the expression of genes. We obtained eQTL data from the ExSNP database

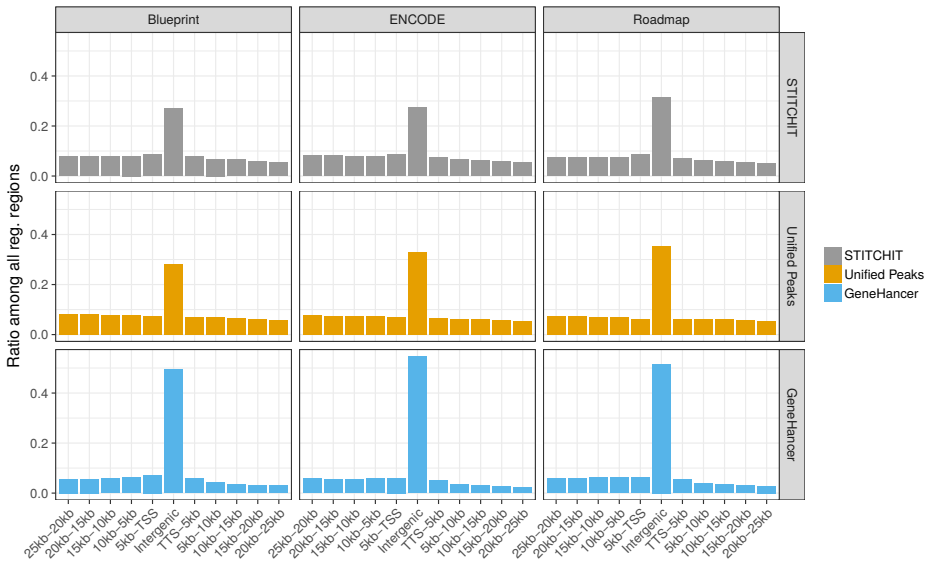


Figure 8.7: Distribution of PEIs around genes for all datasets and all methods. We observe a peak in the intergenic part of the gene and a decline downstream of the gene.

and overlaid it with our predictions by computing how many eQTL links are correctly overlapping our predictions. Fig. 8.8e shows that StitchIt regions obtain the best enrichment score for the identification of eQTL overlaps, supporting the accuracy of StitchIt predictions. Furthermore, compared to randomization experiments, we see an enrichment of eQTL overlaps using StitchIt REMs (Figure 8.9c).

8.9.5 CRISPR-Cas9 validated enhancers for *ERBB2* are accurately retrieved with STITCHIT

Klann *et al.* validated several REMs for ERBB2 using CRISPR-Cas 9 experiments [K$^+$17b], leading to a set comprising 11 validated regions influencing the expression of *ERBB2*. We assembled a list of these validated regions (c.f. Supplementary Table S2) and generated a bed file containing all validated DHSs in SKBR3 breast cancer cells using the original DHS calls from Klann *et al*, obtained from the GEO [GSE96876] [K$^+$17b].

We have calculated the overlap between all DHSs identified in SKBR3 cells and predicted REMs using BEDTools across all datasets. If more than one predicted region overlapped a DHS site, only the most significant one is kept in the intersection.

We obtained REMs for *ERBB2* using StitchIt, UnifiedPeaks and GeneHancer. Next, we pooled the REMs learned across the three datasets by inter-

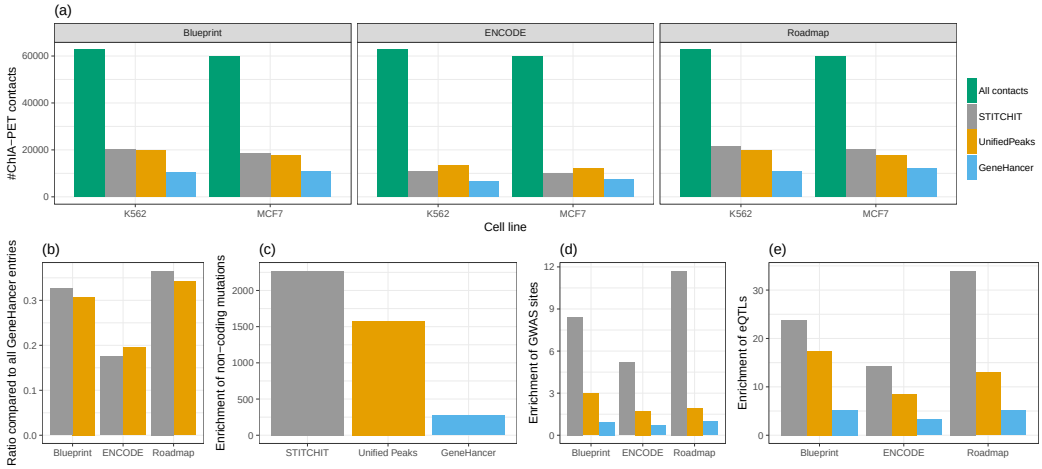


Figure 8.8: (a) The bar plot indicates how many ChIA-PET contacts are matching the associations of REMs to their target genes across all three datasets and linkage approaches for the K562 and MCF-7 cell lines. With the exception of the ENCODE dataset, StitchIt retrieves more interactions than both UnifiedPeaks and GeneHancer. (b) Here, the ratio of recovered entries from the entire GeneHancer database is shown for StitchIt and UnifiedPeaks. While UnifiedPeaks retrieves slightly more entries than StitchIt on ENCODE data, StitchIt retrieves more known sites on Blueprint and Roadmap data. (c-e) Length-normalized enrichment of non-coding mutations (c), of GWAS sites (d) and eQTLs (e). With the exception of (c), which considers only Blueprint data, all other analysis are performed on all datasets and indicate that StitchIt regions achieve the best score. Figure from Schmidt *et al.* [S+19].

secting them with all potential SKBR3 DHSs (see B.6.3 for details).

As depicted in Figure 8.10, StitchIt finds 9 of the 11 regions, the UnifiedPeaks approach retrieves 7 regions and GeneHancer retrieves 6 regions. As an additional support of the predictions, we found that ChIA-Pet data (4DGenome) is linking REMs around *GRB7* to *ERBB2* [W+15].

As shown in Fig. 8.10, StitchIt regions are very sharp and rather a subset of the validated DHSs, especially for those within the gene body of *ERBB2*, while UnifiedPeaks and GeneHancer segments are very broad, not providing a good resolution on the regulatory landscape of the chromatin around *ERBB2*. For instance, 6 validated DHSs are covered by only 2 GeneHancer segments, rendering the GeneHancer segments less useful in an exploratory analysis of gene-expression, because the actual sites of importance are not pinpointed precisely. The UnifiedPeaks approach performs better in terms of this ratio, as only in one instance an identified site overlaps 2 validated DHS, but the mean length of the UnifiedPeaks regions (1449bp) is still longer than the mean length of StitchIt regions

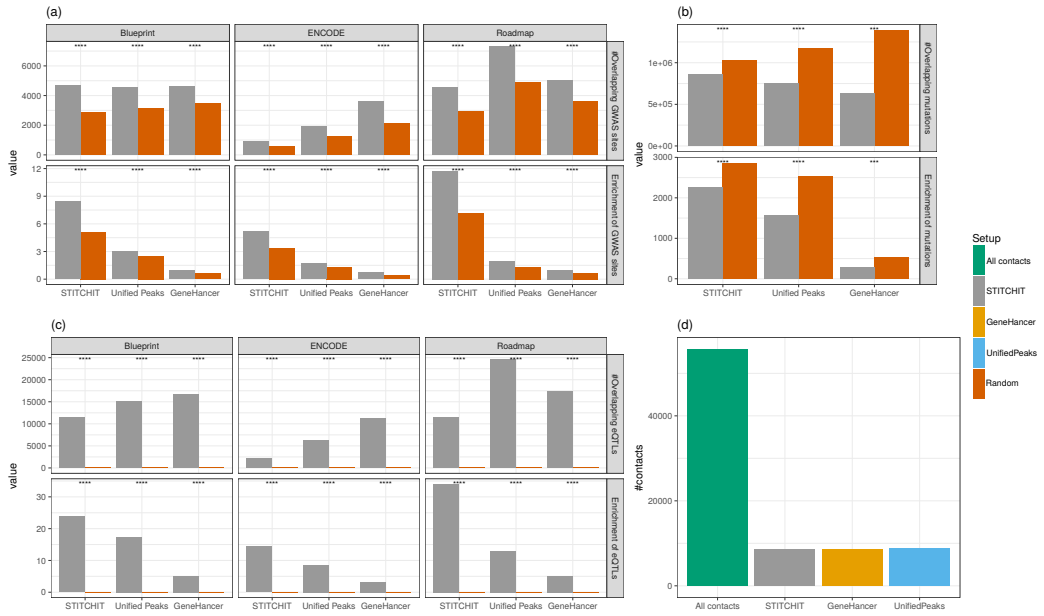


Figure 8.9: (a) Random comparison to overlapping GWAS sites indicating the total number of overlapping sites as well as the enrichment score. (b) Total number of overlapping mutations from the COSMIC database and the enrichment score compared between randomized and original data. (c) Random analysis of the eQTL data from expSNP showing the total number of correctly overlapping eQTLs as well as the enrichment score. (d) Total number as well as the number of REM overlapping contacts from a Capture Hi-C experiment. Significance in a)-c) is indicated by a two-sided t-test (****: $p \leq 0.0001$). Figure from Schmidt *et al.* [S+19].

(214bp) and longer than the validated SKBR3 DHSs (562bp). Especially in downstream applications such as TF-binding predictions or foot-print calling, the more fine grained resolution of StitchIt is beneficial and can avoid false-positive calls.

8.9.6 STITCHIT can be used to segment large regulatory elements

From the overlap between UnifiedPeaks ($\mathcal{U}$) and StitchIt ($\mathcal{S}$) regions it can be computed how many StitchIt segments s a peak $p \in \mathcal{U}$ is split into. We refer to the segmentation of p into several segments s as a *split event*, with the additional constraint that the new sub-regions should not be linked to the same gene as the original peak. The *degree of a split event* denotes the number of StitchIt segments s a peak $p \in \mathcal{U}$ is divided into. Within this counting procedure we also impose that any s overlapping p needs to be linked to a different gene g than p, while any StitchIt segment s can be assigned to the same target gene g' as long as $g' \neq g$. In addition, we quantify how many *split events* are supported by conformation data.

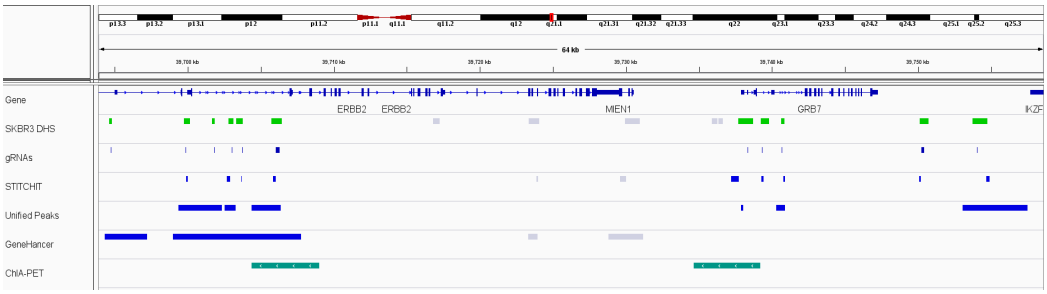


Figure 8.10: This genome-browser visualization depicts in the gene track the genomic locus of *ERBB2*. The second track shows DHSs identified in SKBR3 cells and sites highlighted in green have been validated using CRISPR-Cas9 experiments [K+17b], indicated by the gRNA binding sites shown in the track below. The track labeled ChIA-PET shows an interaction obtained from the 4DGenome database. The StitchIt track contains all StitchIt regions identified across the three datasets. A blue color indicates that the region overlaps a validated DHS site. The tracks for UnifiedPeaks and GeneHancer are generated analogously. StitchIt retrieved 9 sites, UnifiedPeaks identified 7 and GeneHancer 6. Figure from Schmidt *et al.* [S+19].

To this end, for each *split event*, we assess how many StitchIt segments overlap a matching genomic contact obtained from ChIA-Pet or Capture Hi-C data. If all StitchIt regions are supported, we call a split *fully supported*, if not all but at least one region is supported we call it *partially supported*.

As shown above, the UnifiedPeaks approach produces longer candidate regions than StitchIt. As depicted in Figure 8.11a *split events* do occur frequently. Note that for illustration purposes, split events of degree > 10 and smaller than < 2 are not displayed.

An example for a *split event* is provided in Figure 8.11b. Here, a peak is linked exclusively to *TMEM14B* by the UnifiedPeaks method. The peak itself is located around the promoter of *TMEM14C* and covers a total genomic range of 2497*bp*. StitchIt divides that peak into segments linked to PAK1P1, to *TMEM14C* itself and to *TMEM14B*. ChIA-PET data obtained from K562 cells support the long range interactions to *PAK1P1* and *TMEM14B*. Together with the analysis presented in Figure 8.11a, this example underlines the ability of StitchIt to precisely pinpoint regions of regulatory potential and suggests segmentations of large REMs into more refined segments to reveal their regulatory interactions

8.9.7 Exploratory analysis of *EGR1* regulation

To better understand the functional advantage of StitchIt over UnifiedPeaks, we investigate the regulatory landscape of *EGR1* in more detail. For *EGR1*, the Spearman correlation achieved by the UnifiedPeaks REMs in gene-expression

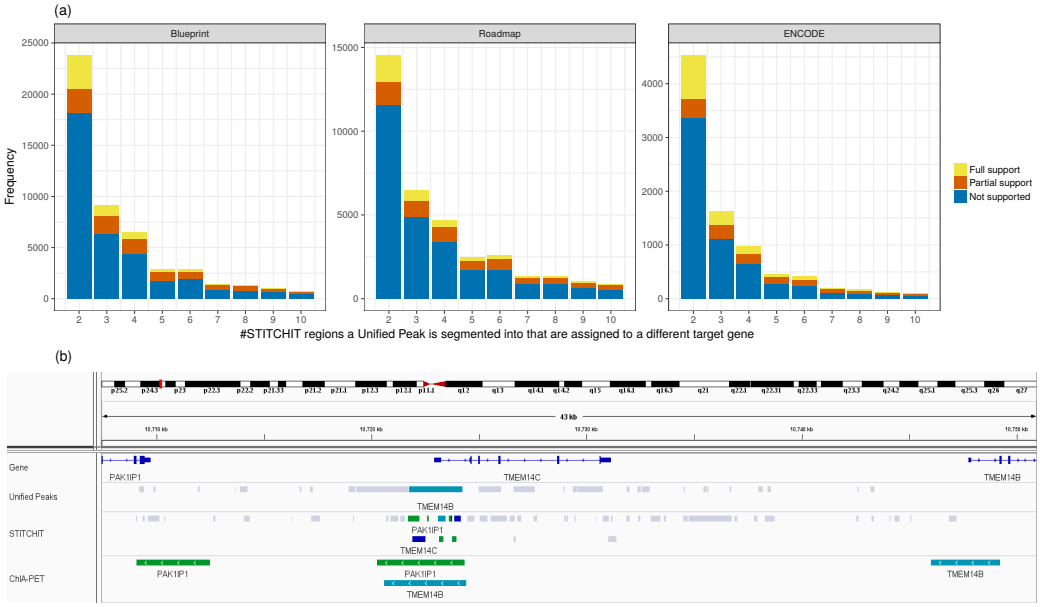


Figure 8.11: (a) The x-axis of the bar plots indicate the magnitude of a *Split event* that is the number of differently linked StitchIt segments a peak is split into. The y-axis holds the frequency for the individual counts. The color code indicates whether StitchIt associations are fully, partially or not at all supported by conformation data. (b) Example of a *split event* at the *TMEM14C* locus. At the promoter of *TMEM14C*, a peak that is linked to *TMEM14B* is split into several StitchIt segments. These are associated with PAK1P1, *TMEM14C* itself and *TMEM14B*. All StitchIt associations shown here are supported by ChIA-PET data. Figure from Schmidt *et al.* [S+19].

modelling is 0.55, while StitchIt regions achieve a correlation of 0.72. Here, we test whether this difference in model performance is also reflected by an improved interpretability of the identified regions regarding the regulation of *EGR1*. In Figure 8.12a, we show the identified candidate regions ranked by the absolute value of the regression coefficients per site (Table B.24).

A striking difference between StitchIt and UnifiedPeaks is that the latter identifies one large segment (U1: 8970bp) covering 2842bp upstream of *EGR1*, the entire *EGR1* gene as well as 2304bp downstream of *EGR1* TTS. This segment is split up into two regions using StitchIt: a region downstream of *EGR1* TTS (S1) and into a region within the first exon of *EGR1* (S2). As shown by the DNaseI-seq signal tracks in Figure 8.12a, StitchIt region S1 and S2 do overlap DNaseI-seq signal in sample *C0010KB*, in which *EGR1* is expressed, whereas they lack signal in *C005VG11*, where *EGR1* is not expressed. This difference between StitchIt and UnifiedPeaks is likely the main reason for the observed performance difference.

Another interesting association can be observed for S3 and S8, which are both

overlapping with U2. S3 has the strongest negative regression coefficient identified by StitchIt for *EGR1* and indeed this region (as well as S8) shows signal in *C005VG11* but not in *C0010KB*, supporting the role of the regions as an active repressor of *EGR1*. The link of S3 to *EGR1* is further supported by ChIA-PET data.

While these examples provide insights on the level of individual samples, we have considered the DNaseI-seq signal within all identified StitchIt regions and used it to cluster the Blueprint samples (Figure 8.12b). Using only the signal within the candidate regulatory sites, an almost perfect clustering into samples according to *EGR1* expression levels was obtained. The clustering can be used to assess the cell type specificity of the suggested regions.

We further hypothesized that regions with strong regression coefficients should be functionally active, e.g. are containing TF binding sites. We use Fimo [G+11] to predict TF binding in REM S1 (Table B.25) and REM S3 (Table B.26), the regions with the strongest positive and negative association predicted by StitchIt, respectively. For the top hits (ranked by Fimo q-value), we checked ENCODE for TF ChIP-seq data. Indeed, we found a peak of TEAD4 and BHLHE40 in S1, which are ranked as first and fourth hit by Fimo. For the second and third TF, namely SP8 and BCL6, as well as for most TFs binding in S3 no ChIP-seq data was available at ENCODE. However, EGR1 ChIP-seq data, which was predicted to bind in S3, is available. Surprisingly, EGR1 does not only bind to S3, but also to S1, S2 and S10, a region uniquely identified using StitchIt that is located about 15kb upstream of the *EGR1* gene. Overall, the ChIP-seq analysis not only suggests that the regions identified with StitchIt are functionally relevant, it also suggests a potential self-regulatory role of EGR1 by binding REMs associated with its own gene.

8.9.8 STITCHIT retrieves REMs related to doxorubicin resistance

CRISPR screens are performed routinely on a genome-wide scale, yet most studies and gRNA libraries focus on protein-coding regions, leading to limited availability of data containing sequences of intergenic or non-coding origin.

We performed a genome-wide doxorubicin CRISPR-Cas9 resistance screen, which led to 332 non-coding target sites of 226 validated gRNA sequences. Experimental details are provided in Section B.6.4. In total, we found 111 putative REMs overlapping with a non-coding gRNA binding site (TableB.27). Importantly, StitchIt obtains significantly more REMs overlapping the gRNA binding sites than randomly selected regions (Figure 8.13).

While several of the predicted REMs have additional support using ChIA-PET or GeneHancer data, StitchIt identified 78 new putative genes via regulatory interactions, among them 24 having a p-value smaller than 0.1. Two genes targeted by gRNAs via regulatory element interactions have been previously reported to be associated with doxorubicin resistance: *MMP9* and *ATF3* (p-values 0.03 and 0.08, respectively). Doxorubicin is a chemotherapy medication used to treat different forms of cancer. It induces double-strand DNA breaks and triggers DNA damage

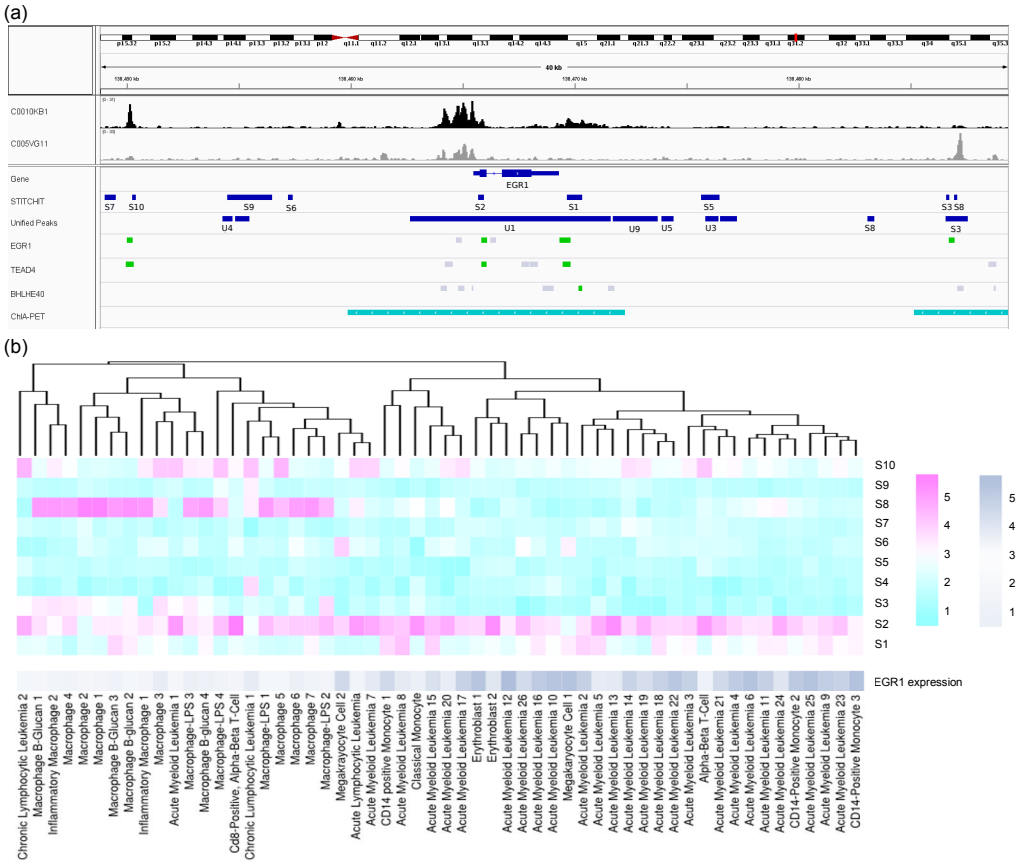


Figure 8.12: (a) Genome browser tracks describing the regulation of *EGR1*. Track *C0010KB1* (black) exemplifies the DNaseI-seq signal for a sample where *EGR1* is expressed, whereas track *C005VG11* (gray) illustrates the case where *EGR1* is not expressed. For the last three tracks, a green segment indicates that the respective TF ChIP-seq peak overlaps with a StitchIt region. (b) Heat map that is clustered according to the DNaseI-seq signal in the candidate REMs $S1, ..., S10$ identified by StitchIt. The gene-expression of *EGR1* is not used for the clustering itself and shown for illustration purposes only. The data has been log transformed with a pseudo-count of 1. Two major clusters can be observed corresponding to samples where *EGR1* is expressed and to those samples where *EGR1* is not expressed. The heatmap shows the $\log_2$ of read counts for DNaseI-seq and $\log_2$ of TPM for gene-expression, respectively. Figure from Schmidt *et al.* [S+19].

associated cell cycle arrest and apoptosis pathways, for example via MAPK/ERK pathway [T+84, C+10a].

An increase of Myocardial Metallo Proteinases (MMPs) expression through in-

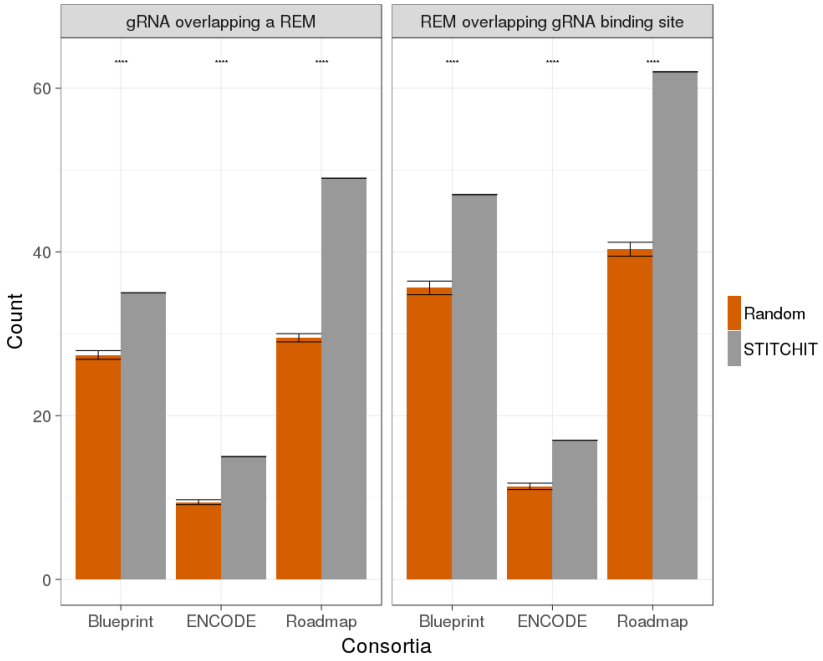


Figure 8.13: Illustration of the number of gRNAs overlapping a REM and of the number of REMs overlapping gRNA binding sites for StitchIt REMs and randomly sampled regions.

creasing ROS formation induced by doxorubicin has been observed in myocytes [S+06]. At position chr20:45993823, overlaps between a gRNA hit and StitchIt predictions suggest a regulatory site of MMP9, being in agreement with the finding of Spallarossa *et al.* [S+06]. ATF3 is a TF known to be involved in cellular stress response and is enriched in cells exposed to stress signals [H+99]. Under doxorubicin treatment, it has been reported that ATF3 affects cell death and cell cycle progression, however it is unclear whether the factor acts as a negative or positive regulator [N+02, P+12]. Nobori *et al.* claim that ATF3 plays a pivotal role as transcriptional regulator in the process of doxorubicin-induced cytotoxicity via an ERK-dependent pathway [N+02]. StitchIt identified a REM of *ATF3* in proximity of position chr1:212621320, a gRNA target site. Both *MMP3* and *ATF3* are also supported by GeneHancer data as well as ChIA-PET interactions.

StitchIt is capable to reveal interactions between genetic loci. In particular, StitchIt can link one gRNA perturbation to several genetic targets of regulatory elements. We observed that in 33% of all cases, intronic or exonic gRNA-target regions of a gene function as enhancer for upstream or downstream located genes, as is the case for *APOL5* and *FANCA* introns, serving as enhancers for *RBFOX2* and *ZNF276*, respectively. Moreover, StitchIt proposes single REMs being linked to multiple genes, demonstrating that one gRNA can lead to multiple correlated interrogations and functional interactions. For instance, the gRNA target site chr3:147409323 was linked to altered gene-expression of *ZIC1* as well

as *ZIC4*. An enhancer targeting the genomic location chr3:147407359-147412176 has also been identified by GeneHancer (GRCh38/hg38, GH03J147407). In an extreme example, StitchIt indicated one gRNA target site chr9:92116400 to be linked to four pseudogenes simultaneously (MTATP6P29, LINC00475, AL354751.3 and AL354751.1). In summary, the combination of StitchIt and CRISPR-Cas9 screens has the potential to prioritize non-coding gRNA hits in known and unknown REMs.

8.10 Limitations of the STITCHIT approach

Although StitchIt considerably improves over existing methods to identify and to link REMs, it has some limitations to be aware of. First of all, we emphasize that the performance of predictive models alone does not proof that the identified regions truly play a role in gene regulation. Additional checks, for instance, the demonstrated overlap with H3K27ac data, should be performed to ensure the validity of the predicted REMs. Furthermore, we stress that StitchIt associations do not imply causation. Thus, we can not distinguish whether the accessibility of certain regions is driving expression of a gene, or whether it is a consequence of that gene being expressed. Also indirect associations, which could be caused by co-regulation of genes, can not be avoided. Therefore, it is especially important to characterize the predicted REMs further.

Another limitation concerns the usability of StitchIt: The current implementation requires discrete gene-expression data, which forces the user to perform an additional pre-processing step.

Throughout this study, we observed that especially on large heterogeneous datasets such as the Roadmap dataset, the peak-independent generation of REMs shows clear advantages over the peak-based strategies. This is particularly obvious comparing our results obtained for Blueprint and Roadmap data. While the Blueprint dataset is composed of primary cells related to the hematopoietic lineage, the Roadmap dataset is more diverse and also comprised of tissue samples. On the more homogeneous Blueprint data, StitchIt and UnifiedPeaks identify almost the same number of segments with similar length (Figure 8.5b,c). In contrast to that, on Roadmap data, StitchIt selects more, but shorter REMs than UnifiedPeaks (Figure 8.5b,c, Figure8.6a). This difference is also reflected by the performance of the gene-expression models. The overall difference between StitchIt and UnifiedPeaks is more pronounced for Roadmap than for Blueprint data (Figure 8.5a). The most likely explanation for this behavior is that due to the high variance in the Roadmap data, merging peaks introduces a loss of specificity by removing the information of the exact genomic location of accessible chromatin (Figure 8.2). On the less heterogeneous Blueprint dataset, this seems to be less of an issue. StitchIt however, is able to resolve the sample and tissue specific variance, therefore obtaining better results on Roadmap data compared to the UnifiedPeaks method. Importantly, we note that StitchIt is not able to outperform the UnifiedPeaks approach on ENCODE data in terms of recall from the GeneHancer database as

well as in the overlap with ChIA-PET data, which is likely to be due to the low number of samples contained in the ENCODE dataset. Compared to Blueprint and Roadmap data, this might also explain why much fewer REMs have been predicted in total (Figure 8.6a). Taken together, this shows how closely the analyzed data is linked to model quality and reliability.

8.11 Future work and applications of STITCHIT

In the future, we plan to apply StitchIt to the joint dataset of all IHEC consortia. On a small scale, we will first utilize the paired gene-expression and DNaseI-seq data introduced in this chapter for this purpose to see how StitchIt can deal with the batch effects included in the data (Figure 7.9).

In a second step, we will utilize StitchIt in the scope of the `EpiMap` project of the IHEC integrative analysis working group, which is currently working on a uniform joint processing of all available epigenomic datasets produced under the umbrella of the IHEC consortium. The dataset generated by `EpiMap` will be the largest uniformly processed epigenomic dataset that is currently available and is thus a rich resource for StitchIt.

Furthermore, as part of the `EpiReg` project funded by BMBF, we plan to set up a database that holds all StitchIt predicted REMs and provides an easy-to-use API for users to query the database, e.g. to retrieve regulators for a distinct gene, or to retrieve regulators that overlap a certain site in the genome. With this service, we attempt to provide the community an easy way to utilize our predictions.

8.12 Contributions of all researchers involved in the described project

The following people contributed to this project: Florian Schmidt (Saarland University), Alexander Marx (Saarland University), Marie Hebel (Göthe University Frankfurt), Martin Wegner (Göthe University Frankfurt), Nina Baumgarten (Göthe University Frankfurt), Manuel Kaulich (Göthe University Frankfurt), Jonathan Göke (Genome Institut of Singapore), Jilles Vreeken (Helmholtz Center for Cyper Security) and Marcel H. Schulz (Saarland University, currently at Göthe University Frankfurt).

Specifically, Alexander Marx, together with Jilles Vreeken, developed the MDL based segmentation algorithm used within StitchIt. The algorithm was optimized to be usable in genomics experiments by Florian Schmidt and Alexander Marx. Florian Schmidt performed the uniform data reprocessing, conducted all computational experiments presented in this chapter and wrote the computational framework embedding the segmentation algorithm, the UnifiedPeaks and the GeneHancer approach. The doxorubicin resistance screen (Section 8.9.8) was performed by Marie Hebel, Martin Wegner and Manuel Kaulich. Marie Hebel performed the literature review to check for known REMs and genes associated with doxorubicin resistance.

Nina Baumgarten assisted in the validation analysis by computing the FIMO predictions used in Section 8.9.7 to analyze the regulation of *EGR1*. Marcel H. Schulz designed the study and advised Florian Schmidt, together with Jonathan Göke. Further, Marcel H. Schulz, Florian Schmidt and Jonathan Göke wrote a BMBF proposal that funded parts of the `EpiReg` project, which STITCHIT is a part of.

9 Summary, Discussion and Outlook

9.1 Software created in the scope of this thesis

In this Section, we recapitulate all software tools written in the scope of this thesis. We sketch their functions, limitations and outline possible future improvements.

9.1.1 The TEPIC framework

The TEPIC framework is a versatile software tool to elucidate transcriptional regulation through TFs. It is freely available on github (`www.github.com/Schulzlab/TEPIC`).

Its core module predicts TF affinities, which is a quantitative, biophysically motivated measure of TF binding, using TRAP [R+07]. In a comparison against TF ChIP-seq data, we have shown that these predictions are accurate and outperform hit-based approaches like FIMO [G+11]. To obtain the predictions, TF motifs in form of position specific energy matrices are used. The TEPIC repository contains a maintained set of such matrices for various species including data from JASPAR [K+18c] and Hocomoco [K+18d]. As TEPIC is not depending on the assay used to prioritize candidate TFBS, it can be used flexibly. In light of the conclusions drawn from the comparison of chromatin accessibility assays in Chapter 6, this point becomes even more relevant. The TFBS prediction module runs extremely fast compared to related approaches and can be easily accessed either via a stand-alone Linux tool or via the REGULATORTRAIL webserver [S+18b, K+17a].

Nevertheless, there is still room for improvement. As discussed in Chapter 3, aside from position weight matrices and related TF motif representations, there are more sophisticated models that capture dependencies within nucleotides, for instance, Slim models [KG15]. Replacing TRAP or modifying it to consider such TF motif representations could improve our TFBS predictions considerably. Aside from the TF motif representation, the scoring of TRAP could be further improved by including the true activity of a TF modelled by the factors concentration in a cell. This could be inferred either from gene-expression or more accurately from proteomics data and would result in a more accurate prediction with respect to the activity of the TFs in a cell.

In our comparison of TFBS predictions between peaks and TF-footprints, we have already mentioned that some TFs do not cause TF footprints in the chromatin accessibility profile due to their short residence time on DNA. It would be

worthwhile to systematically analyze the residence time of TFs in tissues to build up a database informing users when it is recommendable to use TF footprints for TFBS predictions and when peaks should be used.

Another aspect of TF binding is only marginally considered by TEPIC, which is the formation of TF complexes. We have illustrated in Chapter 3 that TF complexes can be revealed and discovered by ChIP-seq assays, however, their prediction using sequence based methods is more challenging. We have contributed a random forest model to predict TFBS genome wide for the ENCODE DREAM in vivo transcription factor binding site prediction challenge (not discussed in this thesis) that learns associations between the occurrence of various TFs from ChIP-seq data and utilizes this information for TF predictions [A$^+$18]. However, we do not model TF complexes in TEPIC aside from a few dimers included in the JASPAR database. It might be possible to use, for instance, the String database to generate a score for the binding of distinct TF complexes [S$^+$17c].

In the literature, it was shown that TF binding does not only depend on chromatin accessibility but can also depend on DNA methylation. For instance, the TF CTCF binds preferably to unmethylated DNA [BF00]. As the dependence on DNA methylation can not be easily modelled with classical TF motifs, we are currently working together with Jan Grau from the University of Halle(Saale) to develop a DNA methylation dependent representation of TF binding.

Furthermore, we acknowledge that the window based aggregation of TF affinities or TF ChIP-seq scores to the gene-level is a strong simplification of the regulatory landscape. However, in case of limited data in terms of biological samples and diversity, we have shown that the window based aggregation is the best approximation. We also note that TEPIC is one of the few tools that provides this gene-centric view of TF scores as a build-in feature. As TEPIC is not limited to a distinct species, this is a valuable asset for the bioinformaticians toolbox.

In further *in silico* experiments, we have illustrated that additional data such as histone modifications in form of CHROMHMM state segmentations can help to reduce noise while aggregating TFBS predictions to the gene-level. At the same time, we have shown that the unsupervised integration of current chromatin conformation capture data such as Hi-C leads to noisy associations of genes to their putative enhancer regions. We believe that this can be improved especially by enhancements of the chromatin conformation capture assays in the future. Also, a different way of modelling the Hi-C data, for example a binned approach based that takes into account the genome-wide distribution of Hi-C read counts, might be helpful. However, a sensible simplification would be to consider TAD boundaries in the TF-gene score generations to avoid that the considered windows cross the TADs. With our STITCHIT approach [S$^+$19], we have suggested an alternative *in silico* solution for the problem of identifying gene-specific regulatory elements introduced in Chapter 8 and discussed below in Section 9.1.3. In the TF-gene score generation one could also consider all gene-specific regulatory elements inferred by STITCHIT and reweigh them using the chromatin accessibility signal for the sample of interest. Thereby, the simplifying window based aggregation could be omitted.

Alternatively, the linkage of regulatory elements to genes could be obtained from the GeneHancer database [P+17b].

Aside from the prediction of TFBS and the generation of TF-gene scores, the Tepic framework enables the user to directly carry out two machine learning pipelines to elucidate the regulatory role of TFs: INVOKE and DYNAMITE. We provide easy-to-use and well documented pipelines for both approaches. Figures to judge the performance and reliability of the models and to interpret them, are generated automatically.

In the INVOKE pipeline we learn a linear model using elastic net regularization predicting gene-expression to infer a tissue specific coefficient matrix that informs about the general importance of TFs for transcriptional regulation across all genes within the considered tissue. We have illustrated in using several DEEP and ENCODE data sets that the coefficient vectors can be easily interpreted in case of predicted TFBS and in case of TF ChIP-seq data. The INVOKE method has been used in several DEEP projects and is also included in the RegulatorTrail web service.

This functionality was complemented by the DYNAMITE method described in Chapter 4 which can be used to determine regulators linked to expression changes occurring between samples. Within this approach a logistic regression classifier is used to prioritize regulators. Just as for INVOKE, the coefficient vector of the model can be interpreted by the user. To simplify the interpretation, we provide an additional script that generates plots specifically tailored to visualize the coefficient values in the differential setup. Furthermore, one of our predictions has been experimentally validated within DEEP sub-project 41 regarding the role of FOXP1 in the differentiation of T-cells [D+16e].

Yet another use case is covered by the Epic-Drem workflow. This approach combines a binary TF-gene score matrix computed by Tepic, which is new Tepic feature added for this application, with the Drem software to identify regulators that are associated with gene-expression changes measured over time. As demonstrated in Chapter 5, the time-point specific features used in Epic-Drem perform better than previous approaches using static input features. Also, several of our predictions could be validated using wet-lab experiments. We remind the reader that the Epic-Drem approach is rather a workflow and not a one click pipeline, as this would have required major changes to the Drem software and could not be done in the scope of this project. As the original Drem output can be overwhelming for interpretation purposes, we devised a network visualization strategy to help biologists deriving meaningful insights from the analyzed data.

While all three approaches have been shown to provide reliable predictions, an obvious drawback is the linearity assumption of the underlying statistical models. The linear relationship between TF binding and gene-expression might be an oversimplified assumption, although it did perform well in practice. An argument supporting the use of linear models is that they can be easily interpreted opposed to non-linear methods.

We have been frequently asked why we do not explicitly consider the expression

of TFs neither in INVOKE nor in DYNAMITE. We do acknowledge that including the expression of TFs into the models might help to emphasize TFs that are enriched in a tissue. However, it was shown in the literature that post-translational modifications are a major workhorse of the cell to regulate TF activity [W+13d]. By neglecting the expression information, we allow the model to infer the importance of TFs in an unbiased way. As shown in Chapter 3, the expression level of the inferred TFs is higher than that of the non-selected features supporting that the models do not depend on the prior knowledge of TF expression. In the Epic-Drem project, we did remove TFs from consideration that are not expressed in at least one considered time point.

Using elastic net regularization in INVOKE and in DYNAMITE ensures that TFs that exhibit similar binding behaviour are all kept in the model if they are relevant for gene-expression. This allows the modelling of co-binding effects of TFs to DNA. However, it does not help in modelling TF complexes if factors bind to other proteins and not to DNA. In the current implementation of Tepic it is not possible to constrain the regression coefficients such that prior knowledge on TFs working jointly can be incorporated. However, as this would require a tissue specific database of TF-TF interactions, which cannot be easily obtained in most instances, we believe that such a feature would not have a strong practical benefit.

We do see that our models could benefit from a multitask approach if several similar samples, e.g. biological replicates, are considered. The multitasking would likely result in less noisy predictions although there is the risk that some biological variance occurring within the individual samples might not be resolved accurately. Nevertheless, the multitask models are expected to have a higher generalizability across unseen replicates of the same tissue than the single task models. This has been shown, for instance, by Jain *et al.* in reconstruction regulatory networks [J+14]. Especially in light of the vast amount of up-coming single cell data, it might be worthwhile to reconsider per-sample that is per-cell models in a multi-task regression setup to account for the noise present in single cell omics data.

In applications of INVOKE, DYNAMITE and Epic-Drem on bulk data the performance of the models also depends on several biological aspects. One of the most important points is the cell-cycle stage of the analyzed samples. To remove variation that is due to heterogeneity caused by different cell-cycle stages within the cell pool, it is essential that the cells are synchronized.

We are convinced and have shown in this thesis that Tepic and the machine learning pipelines contribute to a better understanding of transcriptional regulation and help to interpret the vast amount of epigenomic data sets being produced. With ongoing improvements in modelling, our methods will also be applicable to single-cell data sets that will help elucidating gene regulation on a more detailed level.

9.1.2 Ontology Scoring

During the development of the StitchIt method we were faced with the integration of several IHEC data sets. While trying out various batch effect adjustment methods (BEA), as described in Chapter 7, we realized that there is a lack of methods

that objectively quantify the quality of BEA. To fill this gap, we developed a novel approach that utilizes the Cell Ontology to derive a gold-standard of expected sample similarity. We have illustrated the applicability of our approach using various simulated examples as well as real RNA-seq data sets.

One of the major limitations of our method is that it requires a mapping of CO terms to the samples at hand. If this is not done by the researchers producing the data, the end-users will have to perform this mapping, which is a potentially error prone endeavour. Furthermore, the accuracy of the ontology itself is influencing our score. Despite these two downsides, the ontology score is a easy-to-use, highly interpretable and flexible tool to judge BEA methods. Unlike other approaches the consideration of sample similarities between all samples in a data set, across tissue and cell types, makes it applicable to very heterogeneous data set, which is a unique feature of our approach.

We note that our score can not only be applied to judge BEA methods, but also for comparing other data normalization approaches, e.g. RNA-seq normalization. Besides, although we have tested our score only on bulk RNA-seq data, it is not limited to that. It could also be applied to single-cell RNA-seq batch correction approaches, for instance, a method by John Marioni's lab [H$^+$18a]. To avoid additional complexity introduced by handling single-cell RNA-seq data itself we did not consider such data sets in our project. However, for example in the scope of the Human Cell Atlas [R$^+$17b], our method might be useful to compare various BEA approaches for data integration purposes.

9.1.3 STITCHIT

With our StitchIt method we provide a novel software that is able to infer regulatory regions such as enhancers or repressors on a gene-specific level. It utilizes paired DNaseI-seq and RNA-seq data sets obtained from IHEC in a segmentation approach that links variation in DNaseI-seq signal to gene-expression changes of a single gene. Thereby, the location of candidate regulatory elements is pinpointed. Importantly, this approach does not depend on peak calls. As illustrated in Chapter 8, StitchIt outperforms traditional peak based approaches and known curated regulatory elements from GeneHancer in various validation scenarios. An advantage of using the minimum description length principle in the underlying optimization problem is that the model inferred by StitchIt can not over-fit by construction.

Our method requires large and heterogeneous data sets, ideally containing several biological replicates of the same tissue or cell type. In the scope of the EpiMap project in IHEC, an even larger data set than the one considered in this thesis will be generated by the IHEC integrative analysis working group. We believe that applying StitchIt to this data set will boost the value of our predictions further. While developing our method we ensured that the tool is well tested and can be easily applied to large novel data sets making the analysis of new data sets easily possible. StitchIt can be obtained from github: `www.github.com/schulzlab/STITCHIT`. Although we have tested StitchIt on DNaseI-seq data, our method is not specifically tailored to that. Other chromatin accessibility assays or even HM

ChIP-seq data could be used as input, too.

It is important to realize that our predicted sites are based on correlation, not on causation. In other words, StitchIt does not perform causal discovery. To draw causal conclusions, biological experiments such as gene-editing using CRISPR-cas9 need to be performed. Alternatively, it might be possible, but computationally challenging to generate a causal regulatory network from all StitchIt predictions using the PC algorithm [S$^+$00b]. Learning the network is computationally challenging not only because the algorithm's runtime is exponential in the number of considered variables, but also the tissue and cell type specificity of the regulatory interactions need to be considered. However, especially for applications in precision medicine, the establishment of causal links is essential.

In the current version of StitchIt we have omitted to consider TFs in our model. As we have shown in Chapter 3 that chromatin accessibility alone is highly predictive for gene-expression, this was solely a reasonable simplification to identify regulatory elements. In future work, we will integrate TFBS predictions as well as TF ChIP-seq data into the model.

An example for a future extension of Stitchit is the EpiReg database which is currently under development in our group. EpiReg will contain all of StitchIt's predictions in a database that can be queried via a web interface. It will allow users to retrieve regulatory elements for a distinct gene, overlap known regulatory regions with other genomic data sets such as mutations or differentially methylated loci. Additionally, the database will contain an annotation of the predicted regulatory elements with TF binding sites. With the advancement of IHECs EpiMap project, our database will contain more and more regulatory elements and thus might be a future standard resource to retrieve gene-specific regulatory elements. Furthermore, the associations contained in EpiReg could also be easily augmented using single-cell data to generate cell specific regulatory maps.

Here, I also like to mention that StitchIt was a collaborative effort developed with the group of Jilles Vreeken from the Cluster of Excellence. The project "was born" during a Cluster retreat. Thus it is a nice example for how useful scientific exchange, even within one institute, can be.

9.2 General challenges and Outlook

It is safe to say that we have reached the area of biological data science. Within DEEP, we were already faced with several data integration issues and challenges, although only few samples were considered. Within the IHEC consortium we are faced with even bigger challenges. In Chapter 6, we have illustrated how differences can arise from using different chromatin accessibility assays in profiling the chromatin landscape of identical samples. This very detailed and specific analysis illustrates how challenging it is to distinguish biological signal from technical and methodological biases.

With the advancement of deep learning approaches, especially in precision medicine and cancer research [R$^+$18, NS$^+$19], it is essential to be aware of potential biases to

ensure that these hardly interpretable models do not over-fit to confounders. Also in single cell analysis, where the signal to noise ratio is considerably worse compared to bulk assays, data integration and processing are future challenges for the field.

The methods introduced in this thesis utilize biological data in an interpretable manner and, in case of TEPIC, could even be applied to single-cell data. With STITCHIT we introduced a novel and innovative approach that might have a strong influence on the understanding of gene-specific regulatory events and will help the community to advance in the field.

With the joint advancements in biology, physics, computational resources and data science, we believe that we are at the edge of achieving invaluable insights in molecular biology and medicine which were unimaginable a decade ago. Hopefully the work presented in this thesis can contribute to this endeavour.

A Nomenclature

Abbreviations

3C Chromosome Conformation Capture

acc Accuracy

ADLD autosomal dominant adult onset demyelinating leukodystrophy

AHR Aryl Hydrocarbon Receptor

AML Acute Myeloid Leukemia

asRNA antisense RNA

ATAC Assay for Transposase-Accessible Chromatin

BAC Bacterial Artificial Chromosomes

BEA Batch Effect Adjustment

CAF-1 Chromatin Assembly Factor 1

cDNA complemenarty DNA

CEBPA CCAAT/enhancer-binding protein alpha

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

ChIP-seq Chromatin Immuno Precipitation followed by DNA-Sequencing

CL Cell Ontology

CRISPR clustered regularly interspaced short palindromic repeats

CRISPR-ko CRISPR-knock out

CTCF CCCTC-binding factor

DEEP Deutsches Epigenom Programm

DHS DNaseI hypersensitive site

DNA Deoxyribonucleic acid

DNMT DNA methyltransferase

DNMT1 DNA cytosine-5-methyltransferase 1

DREM Dynamic Regulatory Events Miner

DWM Dinucleotide Weight Matrix

eIF eukaryotic Initiation Factor

EPE Estimated Prediction Error

eRNA enhancer RNA

FAIRE-seq Formaldehyde Assisted Isolation of Regulatory Elements

FDR False Discovery Rate

FISH Fluorescence In Situ Hybridization

FN False Negative

FP False Positive

FPKM Fragments Per Kilobase of transcript per Million mapped reads

FPR False Positive Rate

GLIS1 Glis Family Zinc Finger 1

GMEB1 Glucocorticoid Modulatory Element Binding Protein 1

gRNA guideRNA

GTEx Genotype-Tissue Expression

GWAS Genome-Wide Association Studies

HATs Histone-Acetyl-Transferases

HDAC Histone deacetylase

HDACs Histone-Deacetylases

HM Histone Modification

HMM Hidden Markov Model

HP1 Heterochromatin Protein 1

IHEC International Human Epigenomics Consortium

IOHMM Input Output Hidden Markov Model

iPS induced Pluripotent Stem Cells

Klf4 Kruppel-like factor 4

LAR Least Angle Regression

LCS Longest Common Subsequence

LDA Linear Discriminant Analysis

LMNB1 Lamin B1

lncRNA long non-coding RNA

LOLA Locus Overlap Analysis

Lpl Lipoprotein Lipase

MDL Minimum Description Length

MDL Minimum Description Length

MDL Minimum Description Length

miRNA micro RNA

mRNA messenger RNA

MSE Mean Squared Error

NDR Nucleosome Depleted Regions

NFR Nucleosome Free Regions

NGS Next Generation Sequencing

NIH National Institute of Health

NRF1 Nuclear respiratory factor 1

Oct3 octamer-binding transcription factor 3

Oct4 Octamer-binding transcription factor 4

ORF open reading frame

p-TEFb positive transcription elongation factor b

PBM Protein Binding Microarray

PC Principal Component

PCNA proliferating cell nuclear antigen

PCR Polymerase Chain Reaction

PEI Promoter-Enhancer-Interactions

PFM Position Specific Frequency Matrix

PPM Position Specific Probability Matrices

pre Precision

pre-mRNA precursor mRNA

PSEM Position Specific Energy Matrix

PTM Post-Translational Modification

rec Recall

REM Regulatory element

RF Random Forest

RNA Ribonucleic acid

RNAi RNA interference

ROC Receiver Operating Characteristic

RPKM Reads Per Kilobase of Transcripts per Million mapped reads

rRNA ribosomal RNA

RSS Residual Sum of Squares

RTI Regulator-Target Interaction

RUV Removing unwanted variation

SE Super-enhancer

SELEX Systematic Evolution of Ligands by Exponential Enrichment

siRNA short interfering RNA

SNP Single Nucleotide Polymorphism

SNP Single nucleotide polymorphism

SNV Single Nucleotide Variation

Sox2 Sex determining region Y-box 2

SVA Surrogate variable analysis

SVD Singular Value Decomposition

TAD Topologically Associated Domain

TAF1 Transcription initiation factor TFIID subunit 1

TBP TATA binding protein

TCGA The Cancer Genome Atlas

TCM Central Memory T-Cells

TEM Effector memory cells

TF Transcription factor

TFBS Transcription Factor binding site

TFII Transcription factor for RNA polymerase II

TN Naive T-cells

TN True Negative

TP True Positive

TPM Transcripts Per Million

tRNA transfer RNA

TSS Transcription Start Site

ZBTB7B Zinc Finger And BTB Domain Containing 7B

RNA-Pol II RNA-polymerase II

Glossary

β-thalassemia Heritable disease affecting red blood cells by a reduced production of haemoglobin beta

bed-file Tab delimited file containing genomic positions and annotations for those. See www.ensembl.org/info/website/upload/bed.html for details

Cardiomyocytes These are muscle cells with a high mitochondrial density forming the heart

Cell type A cell type is defined by a cells morphology and function. There are about 200 different cell types known in human [A$^+$05]

Chaperone Proteins that assist in the correct folding of other proteins and/or in the correct (dis)assembly of protein complexes

Chromatin Chromatin refers to the complex of DNA, histone proteins and other non-histone DNA binding proteins

Deletions Removal of a DNA nucleotide from the sequence

Eigenvector A vector v is called an eigenvector of a $n \times n$ matrix A if there exists a λ such that $Av = \lambda v$. The parameter λ is called eigenvalue.

Enhancers Genomic regions harbouring binding sites for TFs. Enhancers can be found several kilobases away from their target gene(s). Distal enhancer can be brought into spacial proximity to their target gene via chromatin looping

Epigenesis $\sim$ describes the emergence of new structures from non-structured matter [AP15]

Epigenetics $\sim$ is the study of changes in gene function that are mitotically and/or meiotically heritable and that do not entail a change in DNA sequence [WM01]

Erythrocytes Red blood cells transporting oxygen

Ester bond Chemical bond between an carboxylic acid and an alcohol

Euchromatin A transcriptionally active, loosely compacted form of chromatin that is accessible for DNA binding proteins.

FAIRE-seq A sample is cross-linked with formaldehyde. As the cross linking is not efficient in nucleosome free regions, the non-cross linked sites can be extracted, sequenced and used to determine chromatin accessibility

FDR The false disovery rate is the ratio of false rejections of H_0 to all rejections of H_0

Fibroblast Fibroblasts are the most common cell type in connective tissue, they produce, among other products, collagen which is essential for the extracellular matrix

Frobeniusnorm The Frobeniusnorm of a matrix is the square root of the sum of the squared absolute values at each matrix position

Haematopoiesis Development of cells of the blood cell lineage

Heterochromatin A transcriptionally silent, highly condensed form of chromatin that is inaccessible for DNA binding proteins, except for pioneering TF.

Hydrogen bond Interaction between a hydrogen atom from a molecule XâĂŞH in which X is more electronegative than H and an atom or a group of atoms in the same or a different molecule [A^+11]

Hypomethylation Opposite to hypermethylation, refers to a decrease of DNA methylation

Insertions Addition of a DNA nucleotide to the genome

Monoallelic gene-expression This refers to the setting mRNA is only generated from one allele of a gene, while the second copy of the gene on the sister chromosome is not transcribed

Multipotent Multipotent cells are precursor cells for other cell types of a lineage, e.g. all cell types of the haematopoietic linage

Mutations Change of the DNA nucleotide at a distinct position

Myelodysplastic syndrome Disease of the bone marrow preventing the maturation of blood cells

N-glycosidic bond Bond between a carbohydrate and another group

ncRNAs non-coding RNAs

Occam's razor $\sim$ states that in explaining a thing no more assumptions should be made than necessary

Oil Red O staining A method to highlight triglycerides with a red color

Omics The term omics refers to all fields in biology containing the suffix omics in the name, such as genomics, proteomics, metabolomics, or epigenomics

Osteogenesis Development of osteoblasts

Peptide bond Bond between two amino acids. Generated using a condensation that is the elimination of water while formation of the bond

Per-gene learning Machine learning approach that is applied to one-gene using many observations for the gene of interest

Per-sample learning Machine learning approach that generalizes across all genes within one sample

Phenotype The set of an organism's observable traits

Pioneer TF TFs with the ability to bind to heterochromatin mediating chromatin remodelling

Pluripotent A pluripotent cell can give rise to any of the three germ layers: endoderm, mesoderm, or ectoderm

Polymerase Chain Reaction An automated reaction used to amplify double stranded DNA fragments

Promoters Genomic region upstream the transcription start site of a gene. It is bound by the transcriptional machinery to initiate transcription

Proteasome Protein complex that degrades other proteins

Proteomics Proteomics elucidates the present proteins in a analyzed samples, typically via Mass spectrometry

Reverse transcription Process of generating DNA from a RNA template

S-phase Phase of the cell cycle. During S-phase, the DNA is replicated

Stochastic process A stochastic process is a referring to an indexed collection of random variables

B

Supplementary Information

B.1 Appendix Chapter 3

Details on data processing presented in Sections B.1.1 to B.1.4 are taken from the Supplementary Material of Schmidt *et al.* [S+17a].

B.1.1 Experimental processing of DEEP DNaseI-seq data

DEEP DNaseI-sequencing was performed according to a publicly available ENCODE protocol with some modifications. Briefly, 1 x 107 cells (HepG2, liver hepatocytes (LiHe)) were resuspended in $2ml$ buffer A ($60mM$ KCl, $15mM$ Tris-HCl (pH 8.0), $15mM$ NaCl, $1mM$ EDTA (pH 8.0), $0.5mM$ EGTA (pH 8.0), $0.5mM$ spermidine free base), combined with a protease inhibitor cocktail (Roche, Basel, Switzerland). Nuclei were extracted by adding an equal volume of buffer A with 0.1% -0.2% of NP-40 and incubation on ice for 10-15min. Nuclei were washed in buffer A and aliquots of 2.3 x 106 nuclei were digested for $3min$ at $37°C$ in DNase buffer ($13.5mM$ Tris-HCl pH 8.0, $88.5mM$ NaCl, $54mM$ KCl, $6mM$ CaCl2, $0.9mM$ EDTA, $0.45mM$ EGTA, $0.45mM$ spermidine) with different amounts of DNaseI (Roche; $40 - 80U/ml$). Using equal volumes of stop buffer, the reactions were stopped ($50mM$ Tris-Cl (pH 8.0), $100mM$ NaCl, 0.1% SDS, $100mM$ EDTA (pH 8.0), $1mM$ spermidine and $0.3mM$ spermine) with proteinase K ($50g/ml$) and incubated at $55°C$ for $1h$. DNA was then purified using phenol chloroform extraction and quality controls (agarose gels, qPCRs) were performed to determine the optimal digestion level per sample. Double-hit fragments of $100bp$-$500bp$ were selected using either gel-electrophoresis followed by electro-elution (HepG2, LiHe1, LiHe2) or sequential purifications with Agencourt AMPure XP Beads (Beckman Coulter, Brea, USA; LiHe3). The sequencing libraries were prepared from $8ng$ of purified DNA using the TruSeq ChIP Library Preparation kit (Illumina, San Diego, USA) according to the manufacturer's protocol and sequenced on HiSeq v3 paired-end flow cells (HiSeq2500 system).

B.1.2 Computational processing of DEEP and ENCODE DNaseI-seq data

DEEP DNaseI-seq bam files were created according to the DEEP GAL v1 process (http://doi.org/10.17617/1.2W). Alignments were produced with BWA, sorted with SAMTOOLS, and duplicated reads were marked with PICARD TOOLS.

DHSs have been called with JAMM using default parameters for both ENCODE and DEEP data. All peaks passing the JAMM filtering step have been used for further analysis. Additionally, MACS2 peaks have been called using the options `-keep-dup all -genomesize 2900000000 -nomodel -shift -100 -extsize 200 -qvalue 0.05`.

B.1.3 Experimental and computational processing of DEEP RNA-seq data

Pooled samples for RNAseq were homogenized in QIAzol Lysis (Qiagen) using a 21G needle. Total RNA was extracted from the aqueous phase by using the miRNeasy Micro Kit (Qiagen) following the manufacturers recommendations, upon addition of chloroform and phase separation using centrifugation. Starting from 2x500ng totalRNA of RIN 9.6, one stranded total RNA as well as mRNA libraries were prepared according to the manufacturer's instructions (Illumina). Both libraries were sequenced for 2x101nt on an Illumina HiSeq 2000, resulting in $\sim$ 100 million paired-end reads for each library.

RNA-Seq bam files were generated with TopHat version 2.0.11 [T$^+$09], and Bowtie version 2.2.1 [LS12], using NCBI build 37.1 in `-library-type fr-firststrand` and `-b2-very- sensitive` setting. Gene-expression quantification was performed using Cufflinks version 2.0.2 [T$^+$10] using hg19 as the reference genome and enables program settings `frag-bias-correct`, `multi-read-correct`, and `compatible-hits-norm`.

B.1.4 Experimental and computational processing of DEEP NOMe-seq data

Nuclei from formaldehyde fixed cells were extracted using nuclei extraction buffer ($60mM$ KCl; $15mM$ Tris-HCl, pH 8.0; $15mM$ NaCl; $1mM$ EDTA, pH 8.0; $0.5mM$ EGTA, pH 8.0; $0.5mM$ spermidine free base) complemented with a protease inhibitor cocktail from Roche as well as 0.1% NP40 from Sigma-Aldrich. They were incubated for 30min on ice and dounced 10 to 20 during that time using a douncing pistil from Qiagen. After incubation, the nuclei were centrifuged ($500g$, $4°C$, $8min$) and the pellet was washed using the same buffer as described before but without NP-40. Next, the pellet was gently re-suspended in 90μl of 1x GpC-buffer from NEB after another centrifugation round. This was followed by the addition of 70μl of NOMe reaction mix ($7l$ 10x GpC buffer (NEB), $1.5l$ of $32mM$ SAM (NEB), 45μl of $1M$ sucrose and $60U$ of M. CviPI (NEB). The reaction was incubated $3h$ at $37°C$. After one and two hours from the reaction start, 0.5μl of SAM were added. Using 16μl of NOMe stop buffer ($20mM$ Tris-HCl, pH 8.0; $600mM$ NaCl; 1% SDS, $10mM$ EDTA) plus $10l$ proteinase K ($20mg/ml$, Sigma-Aldrich), the reaction was stopped and genomic DNA (classic phenol chloroform purification) was extracted. Next, using EZ DNA Methylation-Gold kit (Zymo, Irvine , USA) $100ng$ of DNA were bisulfite converted and a NGS library was prepared using the TruSeq DNA

Methylation Kit (Illumina, San Diego, USA), following the manufacturer's protocol. The presence of adapter dimers was tested for all libraries. The same holds for the fragment distribution, which was tested on a Bioanalyzer HS chip from Agilent Technologies. Sequencing was conducted on a HiSeq2500 system using 1-2 lanes of a HiSeq v3 paired-end flow cell. Read adapters were trimmed using TRIM GALORE! and mapped with BWA [LD09]. Duplicate reads were removed after mapping. BISSNP [L$^+$12b] was used to call DNA methylation levels at CpGs and GpCs. DNA methylation i a GCG context was not considered. The efficiency of the bisulfite conversion was ensured by genome-wide considering cytosines in an HCH-context.

B.1.5 Peak-calling on NOMe-seq data

This Section is based on the Methods section of Nordström *et al.* [N$^+$19].

Using a Hidden Markov Model (HMM), all cytosines with methylation values M_i, with $i \in [1, m]$, where m is the total number of cytosines in a GCH context within the sequence, are segmented into one of two states. The cytosine is either in an open chromatin region, a nucleosome free (or depleted) region, or in heterochromatic region, that is occupied by nucleosomes. Considering that the DNA methylation varies in the interval $[0, 1]$, it can be easily modelled with a binomial distribution in each HMM state. As described in Section 2.2.5, the Baum-Welch algorithm can be used to fit an HMM. Here, an implementation of this algorithm from the R package HIDDENMARKOV [Har15] was used to fit chromosome specific HMMs in parallel using the SNOW package. The algorithm is stopped either after 1000 iterations are completed or if the likelihood between two consecutive rounds falls below 10^{-3}. Each GC nucleotide is predicted to be either open or closed based on its posterior decoding using the fitted binomial HMM. Several consecutive GCs that are predicted as accessible form a peak. Using a one-sided Fisher's exact test, p-values were computed setting the number of methylated cytosines in GCH context in relation to that of unmethylated cytosines in GCH context between a query and a background region. As background, the closest 4kb of closed chromatin up- and down-stream of the tested region was chosen. They are further ranked by statistical significance based on empirical false discovery rates and the corresponding q-values [ST03]. The false discovery rate at significance threshold t, $FDR(t)$ can be computed as the expected value of the ratio of false discoveries $f(t)$ made at threshold t and the total number of significant discoveries $s(t)$ at threshold t:

$$FDR(t) = E[f(t)/s(t)] \tag{B.1}$$

$$FDR(t) \approx \frac{E[f(t)]}{E[s(t)]} \tag{B.2}$$

Using the computed p-values, a value for $E[f(t)]$ can be estimated by counting peaks with a p-value smaller than or equal to t. To estimate $E[s(t)]$, we permuted the methylation levels of the original input data and segmented the permuted data leading to a p-value distribution of regions in-corrected labelled as open.

To account for the influence of read coverage on the tests, the suggested peak callers performs an automated stratification procedure based on non-parametric mixture model that performs clustering using the MCLUST R-package ([FR02]). The assumption is that regions with deviating copy numbers are the exception, and the median represents the common coverage. Consequently, the clustering tries to minimize the number of clusters showing a mean coverage that is below the median. After each loci is assigned to a cluster, the false discovery rates are estimated separately for each cluster. The peak caller is available online at `https://github.com/karl616/gNOMePeaks`.

B.1.6 ENCODE TF ChIP-seq data

TF ChIP-seq data was obtained from ENCODE for several TFs for K562, GM12878, HepG2, and H1-hESCs in narrow peak format listed in Tables B.1. No further filtering or processing was performed.

B.1.7 Runtime and TF ChIP-seq comparison

Details on the data used

The used DNaseI-seq data from ENCODE and DEEP is listed in Table B.2. DHS sites are identified using JAMM as described in Section B.1.2. TF footprints used in the TF ChIP-seq comparison for GM12878, HepG2, H1-hESCs, and K562 have been called using HINT-BC and are available online (http://costalab.org/publications-2/dh-hmm/).

Command lines

TEPIC1

bash TEPIC.sh -g hs37d5.fa -b JAMM/41/LiHe/01/peaks/filtered.peaks.narrowPeak -o Time_Asses_Hf01 -p pwm_vertebrates_jaspar_uniprobe_converted.txt -a gencode.v19.protein_coding_only.gtf -c 16

TEPIC2

bash TEPIC.sh -g hs37d5.fa -b JAMM/41/LiHe/01/peaks/filtered.peaks.narrowPeak -o Time_Asses_Hf01 -p pwm_vertebrates_jaspar_uniprobe_converted.txt -a gencode.v19.protein_coding_only.gtf -c 16

PIQ

Execute the provided shell script adapted to the datasets at hand:

bash PIQ_1_3/testers/runall.k562.sh

Fimo-Prior

/create-priors sequences_50000.fa 41Hf01.wig –parse-genomic-coord –oc Priors/41 Hf01_50000

and

/fimo –oc T41_01_50000_1 –psp Priors/41Hf01_50000/priors.wig –prior-dist Priors/41Hf01_50000/priors.dist pwm_vertebrates_jaspar_uniprobe_converted.meme sequences_50000.fa –max-stored-scores 200000

Table B.1: ENCODE TF ChIP-seq data used to assess the performance of TFBS prediction methods.

ENCODE Accession number	TF ChIP-seq in K562
ENCSR000BRQ	CEBPB
ENCSR000DWE	CTCF
ENCSR000BLI	E2F6
ENCSR000BNE	EGR1
ENCSR000BMD	ELF1
ENCSR000BKQ	ETS1
ENCSR000BMV	FOSL1
ENCSR000BLO	GABPA
ENCSR000BKM	GATA2
ENCSR000EFV	MAX
ENCSR000BNV	MEF2A
ENCSR000BMW	REST
ENCSR000BKO	SP1
ENCSR000BGW	SPI1
ENCSR000BLK	SRF
ENCSR000BRR	STAT5A
ENCSR000BKT	USF1
ENCSR000BKU	YY1
ENCSR000BKF	ZBTB33
	TF ChIP-seq in GM12878
ENCFF002CGQ	BATF
ENCFF002CGU	CEBPB
ENCFF002CGV	EBF1
ENCFF002CGW	EGR1
ENCFF002CGX	ELF1
ENCFF002CGY	ETS1
ENCFF002CGZ	FOXM1
ENCFF002CHA	GABPA
ENCFF939TZS	JUNB
ENCFF002CHC	MEF2A
ENCFF002CHH	REST
ENCFF002CHT	RXRA
ENCFF002CHV	SP1
ENCFF002CHQ	SPI1
ENCFF002CHW	SRF
ENCFF002CHX	STAT5A
ENCFF002CHZ	TCF12
ENCFF002CIA	TCF3
ENCFF144PGS	TCF7
ENCFF002CIB	USF1
ENCFF002CIC	YY1
ENCFF694OTE	ZBED1
ENCFF002CID	ZBTB33
ENCFF002CIE	ZEB1
	TF ChIP-seq in HepG2
ENCSR000BID	BHLHE40
ENCFF002CTU	BRCA1
ENCFF002CTV	CEBPB
ENCSR000DUG	CTCF
ENCSR000BMZ	ELF1
ENCFF002CUA	ESRRA

ENCSR000BHP	FOSL2
ENCSR000BMO	FOXA1
ENCSR000BNI	FOXA2
ENCSR000BJK	GABPA
ENCSR000BLF	HNF4A
ENCSR000BNJ	HNF4G
ENCFF002CUD	HSF1
ENCSR000BGK	JUND
ENCFF002CUG	MAFF
ENCFF002CUI	MAFK
ENCFF002CUJ	MAX
ENCFF002CUY	NR2C2
ENCFF002CUM	NRF1
ENCSR000BOT	REST
ENCFF002CUT	RFX5
ENCSR00BHU	RXRA
ENCSR000BJX	SP1
ENCSR000BOU	SP2
ENCFF002CUV	SREBF1
ENCFF001VLB	SREBF2
ENCSR000BLV	SRF
ENCFF002CUW	TBP
ENCSR200BJG	TCF12
ENCFF002CUX	TCF7L2
ENCSR000BGM	USF1
ENCFF002CUZ	USF2
ENCSR000BHR	ZBTB33
	TF ChIP-seq in H1-hESC
ENCFF002CQQ	BRCA1
ENCFF002CQR	CEBPB
ENCFF002CIU	CTCF
ENCFF002CIV	EGR1
ENCFF002CIW	FOSL1
ENCFF002CIX	GABPA
ENCFF002CQU	JUN
ENCFF002CQY	JUND
ENCFF002CQZ	MAFK
ENCFF002CRA	MAX
ENCFF002CRC	NRF1
ENCFF002CJB	REST
ENCFF002CRE	RFX5
ENCFF002CJH	RXRA
ENCFF002CJK	SP1
ENCFF002CJL	SP2
ENCFF002CJN	SRF
ENCFF002CRH	TBP
ENCFF002CJQ	TCF12
ENCFF002CJS	USF1
ENCFF002CRI	USF2
ENCFF002CJT	YY1

Peak-Calling with JAMM:

bash JAMM.sh -s 41_Hf01.bed -o Peaks/41_Hf01 -g hg19.genomseSize.txt -f 1 -p 16

Note that the provided sample ID acts as a placeholder for all considered samples.

Table B.2: DNaseI-seq data used for the runtime experiments as well as the TF ChIP-seq comparisons

DEEP Sample ID	**Thesis Sample ID**
01_HepG2_LiHG_Ct1	HepG2
41_Hf01_LiHe_Ct	LiHe1
41_Hf02_LiHe_Ct	LiHe2
41_Hf03_LiHe_Ct	LiHe3
DEEP File	**Data Type**
01_HepG2_LiHG_Ct1_DNase_S_1.bwa.20140719.bam	Dnase-1 seq
41_Hf01_LiHe_Ct_DNase_S_1.bwa.20131216.bam	Dnase-1 seq
41_Hf02_LiHe_Ct_DNase_S_1.bwa.20131216.bam	Dnase-1 seq
41_Hf03_LiHe_Ct_DNase_S_1.bwa.20150120.bam	Dnase-1 seq
ENCFF000SVN	DNase -1 seq of K562
ENCFF000SKV	DNase -1 seq of GM12878
ENCFF000SKW	DNase -1 seq of GM12878
ENCFF000SKZ	DNase -1 seq of GM12878
ENCFF000SLB	DNase -1 seq of GM12878
ENCFF000SLD	DNase -1 seq of GM12878
ENCFF000SOA	DNase-1 seq of H1-hESC
ENCFF000SOC	DNase-1 seq of H1-hESC

B.1.8 Data used in gene-expression models

The data used for the experiments delineated in Section 3.4.2 is shown in Table B.3. ChIP-seq data used in the context of [S$^+$17a] is shown in Table B.1.

B.1.9 Overview of TF-gene score matrices used to assess the stability of TF-gene scores

Within this section, we detail all TF-gene scores introduced in Section 3.3.2.

ChIP-seq TF features (C)

ChIP-seq TF features ($\mathcal{C}$) are an aggregated version of ENCODE TF ChIP-seq peak scores. The feature matrix is exemplified in Table B.4.

ChIP-seq TF features normalized (CN)

ChIP-seq TF features normalized ($\mathcal{CN}$) are an aggregated version of ENCODE TF ChIP-seq peak scores, normalized according to the overall number of peaks c_g^C. The content of the feature matrix is shown in Table B.5.

ChIP-seq peak features (CPF)

ChIP-seq peak features ($\mathcal{CPF}$) quantify the number of ChIP-seq peaks in the vicinity of a genes TSS as well as the length of these peaks. $\mathcal{CPF}$ scores are exemplified in Table B.6.

Table B.3: Overview of the epigenetics and transcriptomics data used within Schmidt *et al.* [S+17a].

DEEP Sample ID	Thesis Sample ID
01_HepG2_LiHG_Ct1	HepG2
41_Hf01_LiHe_Ct	LiHe1
41_Hf02_LiHe_Ct	LiHe2
41_Hf03_LiHe_Ct	LiHe3
51_Hf03_BlTN_Ct	T1
51_Hf04_BlTN_Ct	T2
51_Hf03_BlCM_Ct	T3
51_Hf04_BlCM_Ct	T4
51_Hf03_BLEM_Ct	T5
51_Hf04_BLEM_Ct	T6
DEEP File ID	**Data type**
01_HepG2_LiHG_Ct1_mRNA_K_1.LXPv1.20150508_genes.fpkm_tracking	Quantified mRNA
41_Hf01_LiHe_Ct_mRNA_K_1.LXPv1.20150530_genes.fpkm_tracking	Quantified mRNA
41_Hf02_LiHe_Ct_mRNA_K_1.LXPv1.20150530_genes.fpkm_tracking	Quantified mRNA
41_Hf03_LiHe_Ct_mRNA_K_1.LXPv1.20150530_genes.fpkm_tracking	Quantified mRNA
51_Hf03_BlCM_Ct_mRNA_M_1.LXPv1.20150708_genes.fpkm_tracking	Quantified mRNA
51_Hf04_BlCM_Ct_mRNA_M_1.LXPv1.20150708_genes.fpkm_tracking	Quantified mRNA
51_Hf03_BlEM_Ct_mRNA_M_1.LXPv1.20150708_genes.fpkm_tracking	Quantified mRNA
51_Hf04_BlEM_Ct_mRNA_M_1.LXPv1.20150708_genes.fpkm_tracking	Quantified mRNA
51_Hf03_BlTN_Ct_mRNA_M_1.LXPv1.20150708_genes.fpkm_tracking	Quantified mRNA
51_Hf04_BlTN_Ct_mRNA_M_1.LXPv1.20150708_genes.fpkm_tracking	Quantified mRNA
01_HepG2_LiHG_Ct1_DNase_S_1.bwa.20140719.bam	Dnase-1 seq
41_Hf01_LiHe_Ct_DNase_S_1.bwa.20131216.bam	Dnase-1 seq
41_Hf02_LiHe_Ct_DNase_S_1.bwa.20131216.bam	Dnase-1 seq
41_Hf03_LiHe_Ct_DNase_S_1.bwa.20150120.bam	Dnase-1 seq
51_Hf03_BlCM_Ct_NOMe_S_1.NCSv2.20150513.GRCh37.cpg.filtered.GCH.peaks.fdr001.bed	NOMe signal
51_Hf04_BlCM_Ct_NOMe_S_1.NCSv2.20150609.GRCh37.cpg.filtered.GCH.peaks.fdr001.bed	NOMe signal
51_Hf03_BlEM_Ct_NOMe_S_1.NCSv2.20150513.GRCh37.cpg.filtered.GCH.peaks.fdr001.bed	NOMe signal
51_Hf04_BlEM_Ct_NOMe_S_1.NCSv2.20150609.GRCh37.cpg.filtered.GCH.peaks.fdr001.bed	NOMe signal
51_Hf03_BlTN_Ct_NOMe_S_1.NCSv2.20150513.GRCh37.cpg.filtered.GCH.peaks.fdr001.bed	NOMe signal
51_Hf04_BlTN_Ct_NOMe_S_1.NCSv2.20150729.GRCh37.cpg.filtered.GCH.peaks.fdr001.bed	NOMe signal

Table B.4: ChIP-seq TF features ($\mathcal{C}$)

	Chipped TF 1	...	Chipped TF n
Gene 1	$a^C_{1,1}$		$a^C_{1,n}$
...			
Gene m	$a^C_{m,1}$		$a^C_{m,n}$

Epi-Decay (E) and Epi-Decay-Scaled (ES)

Epi-Decay ($\mathcal{E}$) and Epi-Decay-Scaled ($\mathcal{ES}$) are an aggregated version of predicted TFBS using Tepic. In case of $\mathcal{ES}$ scores, they are additionally scaled with the epigenetic signal within the candidate binding site of the TF. TF ChIP-seq peak scores, normalized according to the overall number of peaks c^C_g. The composition of $\mathcal{E}$, respectively $\mathcal{ES}$, feature matrices is outlined in Table B.7.

Table B.5: ChIP-seq TF features normalized ($\mathcal{CN}$)

	Chipped TF 1	...	Chipped TF n
Gene 1	$\bar{a}^{C}_{1,1}$		$\bar{a}^{C}_{1,n}$
...			
Gene m	$\bar{a}^{C}_{m,1}$		$\bar{a}^{C}_{m,n}$

Table B.6: ChIP-seq peak features ($\mathcal{CPF}$)

	ChIP-seq peak count	ChIP-seq peak length
Gene 1	c^{C}_{1}	l^{C}_{1}
...		
Gene m	c^{C}_{m}	l^{C}_{m}

Table B.7: Epi-Decay ($\mathcal{E}$) and Epi-Decay-Scaled ($\mathcal{ES}$) features

	Predicted TF 1	...	Predicted TF n
Gene 1	$a^{E(S)}_{1,1}$		$a^{E(S)}_{1,n}$
...			
Gene m	$a^{E(S)}_{m,1}$		$a^{E(S)}_{m,n}$

Epi-Decay normalized (EN)

$\mathcal{EN}$ scores are the normalized version of $\mathcal{E}$ scores where TF-gene scores are normalized for the genomic length of the aggregated candidate TFBS per gene. In addition to the normalized TF-gene scores, also the number of candidate TFBS and their length is considered in the feature matrix, shown in Table B.8.

Table B.8: Epi-Decay normalized ($\mathcal{EN}$)

	Predicted TF 1	...	Predicted TF n	Number of TFBS	Length of TFBS
Gene 1	$\bar{a}^{E}_{1,1}$		$\bar{a}^{E}_{1,n}$	c^{E}_{1}	l^{E}_{1}
...					
Gene m	$\bar{a}^{E}_{m,1}$		$\bar{a}^{E}_{1,n}$	c^{E}_{m}	l^{E}_{m}

Epi peak-features (EPF)

Epi peak-features ($\mathcal{EPF}$) quantify the number of TFBS in the vicinity of a genes TSS as well as the length of those regions. Table B.9 holds an example of $\mathcal{EPF}$ scores.

Table B.9: Epi peak-features ($\mathcal{EPF}$)

	TFBS count	TFBS length
Gene 1	c_1^E	l_1^E
...		
Gene m	c_m^E	l_m^E

Epi peak-features and signal (EPFS)

Epi peak-features and signal ($\mathcal{EPFS}$) extends the $\mathcal{EPF}$ features by an additional column holding the epigenetic signal within the aggregated TFBS. $\mathcal{EPFS}$ features are illustrated in Table B.10.

Table B.10: Epi peak-features and signal($\mathcal{EPFS}$)

	TFBS count	TFBS length	Epigenetic signal
Gene 1	c_1^E	l_1^E	f_1^E
...			
Gene m	c_m^E	l_m^E	f_m^E

Epi-Decay-Scaled normalized (ESN)

Epi-Decay-Scaled normalized ($\mathcal{ESN}$) extends the $\mathcal{EN}$ features by an additional column holding the epigenetic signal within the aggregated TFBS. $\mathcal{ESN}$ features are illustrated in Table B.11.

Table B.11: Epi-Decay-Scaled normalized ($\mathcal{ESN}$)

	Predicted TF 1	...	Predicted TF n	Number of TFBS	Length of TFBS	Epigenetic signal
Gene 1	$\bar{a}_{1,1}^E$		$\bar{a}_{1,n}^E$	c_1^E	l_1^E	f_1^E
...						
Gene m	$\bar{a}_{m,1}^E$		$\bar{a}_{1,n}^E$	c_m^E	l_m^E	f_m^E

B.1.10 Example for feature matrix permutation

In Schmidt *et al.* [SS18] we follow the permutation strategy suggested by Bessiere *et al.* [B$^+$18a]. They suggested to randomize the feature matrix independently for each row, i.e. per gene. Thereby, TF specific signal would be lost, but confounders that affect all TF-genes scores for a distinct gene would be preserved. In Figure B.1 the effect of the permutation is illustrated.

Figure B.1: Illustration of the feature matrix permutation suggested by Bessiere *et al.* [B+18a].

B.1.11 TF ChIP-seq data used for expression models to assess model reliability

The TF ChIP-seq used in Schmidt *et al.* [SS18], to learn predictive gene-expression models is shown in Table B.12.

B.1.12 Gold standard set used for primary human hepatocytes

According to the human protein atlas [U+15], the following TFs that are available in the latest TF-motif collection of Tepic (version 2.1) are expressed by at least 5FPKM in primary human hepatocytes:

ID2, E2F4, MAX, CEBPB, SREBF2, NR3C1, CEBPZ, TOPORS, GATA4, ELK1, TBX15, SRF, ETS1, ARNT, MAZ, HERPUD1, HSF1, ZBTB18, CENPB, TGIF1, YY1, NFIX, SMAD2, CDC5L, ESR1, HES1, CEBPD, RFX5, SPI1, ELF2, NR4A1, HMGN3, CTCF, NFATC3, SOX5, SP3, IRF8, FUBP1, NR1D1, CCNT2, RARA, ELK4, NR2F6, USF1, SP1, TFDP1, PBX2, RAD21, IRF1, FOSL2, ZBED1, MEF2A, ESRRA, PBX3, GATA6, SETDB1, STAT6, RXRA, FOXO1, NFE2L2, KLF4, NR4A3, HMGA1, GTF2I, MYC, TCF12, JUNB, ZFX, NFKB2, BACH1, NR1H2, HBP1, CREB1, NR5A2, FOXO3, ZNF410, PPARG, PPARA, FOSB, PTEN, STAT3, BHLHE40, GABPA, HNF4A, ELK3, MBD2, ETS2, THRB, ATF4, JUND, RELA, DBP, FOXJ3, EPAS1, KLF6, CCDC6, ERF, JDP2, NFYA, NFIA, EGR1, NR1H4, SMAD4, HDAC2, TP53, HNF4G, CREB3, MLX, IRF2, NFYC, STAT2, HLX, HNF1A, CUX2, ZNF263, ELF3, SMAD3, HIF1A, TRIM28, NFE2L1, MEF2C, USF2, AR, FOS, SNAI2, DDIT3, NFIB, CHD2, KLF12, SREBF1, HLF, ZEB1, ELF1, AHR, SMC3, ARHGEF12, FOXA1, REST, NFKB1, RORA, TCF4, NR1I2, NR2F2, MXI1, NR2F1, IRF9, NFIC, RREB1, CREM, BPTF, IRF6, SIN3A, CREB3L2, JUN, CEBPA, ZNF143, XBP1, SMAD1, ZNF384, ZBTB16, BCL6, TEAD1, NFIL3, MLXIPL, STAT1, FOXA3, BBX, SP100, ATF1, ATF7, TFCP2, TEF, FOXA2, NR1I3, EP300, PATZ1, CEBPG, HLTF, NR4A2, ATF3, ONECUT1, MAF, ZBTB14, ITGB2, NFYB, ZBTB7B, MAFF, ZBTB33, FOXP1, ATF2, ZNF281, ZNF691, PROX1, CUX1, MAFB, TCF7L2, GRHL1, IRF3, RBPJ, ATF6.

B.1.13 Data used for Hi-C models

The Hi-C data used in Section 3.4.5 is listed in Table B.13. In addition to DNaseI-seq, RNA-seq, and TF ChIP-seq experiments from Table B.2 and B.12, the data

Table B.12: TF ChIP-seq data used in Schmidt *et al.* [SS18].

ENCODE Accession number	TF ChIP-seq in K562
ENCSR000BNU	ATF3
ENCSR000BRT	CBX3
ENCSR000BRQ	CEBPB
ENCSR077DKV	CREM
ENCSR000DWE	CTCF
ENCSR000BLI	E2F6
ENCSR000BNE	EGR1
ENCSR000BMD	ELF1
ENCSR000BKQ	ETS1
ENCSR000BMV	FOSL1
ENCSR000BLO	GABPA
ENCSR000BKM	GATA2
ENCSR000EFV	MAX
ENCSR000BNV	MEF2A
ENCSR000BRS	NR2F2
ENCSR000BQY	PML
ENCSR000BKV	RAD21
ENCSR000BMW	REST
ENCSR920BLG	SIN3A
ENCSR000BGX	SIX5
ENCSR000FCD	SMAD5
ENCSR000BKO	SP1
ENCSR000BGW	SPI1
ENCSR000BLK	SRF
ENCSR000BRR	STAT5A
ENCSR000BKS	TAF1
ENCSR863KUB	TCF7
ENCSR000BRK	TEAD4
ENCSR000BNN	THAP1
ENCSR000BKT	USF1
ENCSR000BKU	YY1
ENCSR000BKF	ZBTB33
ENCSR000BME	ZBTB7A
	TF ChIP-seq in HepG2
ENCFF002CTS	ARID3A
ENCSR000BID	BHLHE40
ENCFF002CTU	BRCA1
ENCFF002CTV	CEBPB
ENCSR000DUG	CTCF
ENCSR000BMZ	ELF1
ENCFF002CUA	ESRRA
ENCSR000ARI	EZH2
ENCSR000BHP	FOSL2
ENCSR000BMO	FOXA1
ENCSR000BNI	FOXA2
ENCSR000BJK	GABPA
ENCSR000BMC	HDAC2
ENCSR000BLF	HNF4A
ENCSR000BNJ	HNF4G
ENCFF002CUD	HSF1
ENCFF002CTY	JUN
ENCSR000BGK	JUND
ENCFF002CUG	MAFF
ENCFF002CUI	MAFK
ENCFF002CUJ	MAX
ENCSR000BQX	NFIC
ENCFF002CUY	NR2C2

ENCFF002CUM	NRF1
ENCSR000BOT	REST
ENCFF002CUT	RFX5
ENCSR00BHU	RXRA
ENCSR000BJX	SP1
ENCSR000BOU	SP2
ENCFF002CUV	SREBF1
ENCFF001VLB	SREBF2
ENCSR000BLV	SRF
ENCSR000BJN	TAF1
ENCFF002CUW	TBP
ENCSR200BJG	TCF12
ENCFF002CUX	TCF7L2
ENCSR000BGM	USF1
ENCFF002CUZ	USF2
ENCSR000BHR	ZBTB33
	TF ChIP-seq in H1-hESC
ENCFF002CIR	ATF2
ENCFF002CIS	ATF3
ENCFF002CQP	BACH1
ENCFF002CIT	BCL11A
ENCFF002CQQ	BRCA1
ENCFF002CQR	CEBPB
ENCFF002CQS	CHD1
ENCFF002CQT	CHD2
ENCFF002CQW	CTBP2
ENCFF002CIU	CTCF
ENCFF002CIV	EGR1
ENCFF002CJC	EP300
ENCFF002CDT	EZH2
ENCFF002CIW	FOSL1
ENCFF002CIX	GABPA
ENCFF002CQX	GTF2F1
ENCFF002CIY	HDAC2
ENCFF002CQU	JUN
ENCFF002CQY	JUND
ENCFF002CDU	KDM5A
ENCFF002CQZ	MAFK
ENCFF002CRA	MAX
ENCFF002CRB	MXI1
ENCFF002CQV	MYC
ENCFF002CJA	NANOG
ENCFF002CRC	NRF1
ENCFF002CJE	POLR2A
ENCFF002CJF	POU5F1
ENCFF002CRD	RAD21
ENCFF002CJG	RAD21
ENCFF002CDV	RBBP5
ENCFF002CJB	REST
ENCFF002CRE	RFX5
ENCFF002CJH	RXRA
ENCFF002CRF	SIN3A
ENCFF002CJJ	SIX5
ENCFF002CJK	SP1
ENCFF002CJL	SP2
ENCFF002CJM	SP4
ENCFF002CJN	SRF

ENCFF002CRG	SUZ12
ENCFF002CJO	TAF1
ENCFF002CJP	TAF7
ENCFF002CRH	TBP
ENCFF002CJQ	TCF12
ENCFF002CJR	TEAD4
ENCFF002CJS	USF1
ENCFF002CRI	USF2
ENCFF002CJT	YY1
ENCFF002CRJ	ZNF143
	TF ChIP-seq in GM12878
ENCFF002CGO	ATF2
ENCFF002CGP	ATF3
ENCFF002CGQ	BATF
ENCFF002CGR	BCL11A
ENCFF002CGS	BCL3
ENCFF002CGT	BCLAF1
ENCFF809BIO	CBFB
ENCFF002CGU	CEBPB
ENCFF804OVD	CREM
ENCFF002CGV	EBF1
ENCFF515PNJ	EED
ENCFF002CGW	EGR1
ENCFF002CGX	ELF1
ENCFF002CHI	EP300
ENCFF002CGY	ETS1
ENCFF191HSP	ETV6
ENCFF002CGZ	FOXM1
ENCFF002CHA	GABPA
ENCFF002CHB	IRF4
ENCFF939TZS	JUNB
ENCFF002CHC	MEF2A
ENCFF002CHD	MEF2C
ENCFF002CHE	MTA3
ENCFF002CHF	NFATC1
ENCFF002CHG	NFIC
ENCFF002CHJ	PAX5
ENCFF002CHK	PAX5
ENCFF002CHL	PBX3
ENCFF002CHM	PML
ENCFF002CHO	POLR2A
ENCFF002CHP	POU2F2
ENCFF002CHR	RAD21
ENCFF002CHH	REST
ENCFF002CHS	RUNX3
ENCFF002CHT	RXRA
ENCFF002CHU	SIX5
ENCFF374VLY	SMAD5
ENCFF002CHV	SP1
ENCFF002CHQ	SPI1
ENCFF002CHW	SRF
ENCFF002CHX	STAT5A
ENCFF002CHY	TAF1
ENCFF002CHZ	TCF12
ENCFF002CIA	TCF3
ENCFF144PGS	TCF7
ENCFF002CIB	USF1
ENCFF002CIC	YY1
ENCFF694OTE	ZBED1
ENCFF002CID	ZBTB33
ENCFF002CIE	ZEB1

from Table B.14 has been used as well. DNaseI-seq data has been processed as described in Section B.1.2.

Table B.13: Hi-C data obtained from Rao *et al.*. [R$^+$14b] used in Section 3.4.5, available at GEO under ID GSE63525.

Supplement Identifier	Cell-line	Available resolutions
GSE63525_GM12878_primary_HiCCUPS_looplist.txt.gz	GM12878	10kb
GSE63525_HUVEC_HiCCUPS_looplist.txt.gz	HUVEC	5kb, 10kb, 25kb
GSE63525_HeLa_HiCCUPS_looplist.txt.gz	HeLa	5kb, 10kb, 25kb
GSE63525_IMR90_HiCCUPS_looplist.txt.gz	IMR90	5kb, 10kb
GSE63525_K562_HiCCUPS_looplist.txt.gz	K562	5kb, 10kb, 25kb

B.2 Appendix Chapter 4

The NOMe-seq and and RNA-seq expression data used in Chapter 4 is identical to the one used in Chapter 3, listed in Table B.3. Differential gene-expression information was obtained from the DEEP RNA-seq analysis pipelines, and is available in the Supplementary Material of Durek *et al.* [D$^+$16e]. No additional data has been considered.

B.3 Appendix Chapter 5

The text presented here describing the methods used in our manuscript Gérard *et al.* [GSo18] is based on the articles method section.

B.3.1 Data generated in scope of the project

For all experiments, the ST2 mouse cell line was used. It is a bone marrow stromal cell line, generated from Whitlock-Witte type long-term bone marrow cultures of BC8 mice. In the controlled environment with a temperature of 37°C and 5% CO_2 concentration, the cells were cultivated using a Roswell Park Memorial Institute 1640 medium (Gibco, Life Technologies, 32404014), which *"was supplemented with* 10% *fetal bovine serum (FBS) (Gibco, Life Technologies, 10270-106, lot #41F8430K) and* 1% *L-Glutamine (Lonza, BE17-605E)"* [GSo18] within 10 cm^2 dishes. The experiments were performed with cells that passaged less than 10 times. ST2 cells were seeded 4 days before the differentiation into adipocytes and osteoblasts and reached 100% confluency after 48h. They were further maintained for 48h post-confluency. The addition of differentiation medium I, which consists of *"growth medium,* 0.5*mM isobutylmethylxanthine (IBMX) (Sigma-Aldrich, I5879),* 0.25μ*M dexamethasone (DEXA) (Sigma-Aldrich, D4902) and* 5μ*g/ml insulin (Sigma-Aldrich, I9278)"* [GSo18], caused the initiation of adipogenesis. From

Table B.14: Additional ENCODE data used in Section 3.4.5

ENCODE accession number	Data Type
ENCFF000DJU	Quantified mRNA of IMR90
ENCFF000SOC	DNase-1 seq of IMR90
ENCFF000DNW	Quantified mRNA of HeLa
ENCFF000SPR	DNase-1 seq of HeLa
ENCFF000DUQ	Quantified mRNA of HUVEC
ENCFF001DNS	DNase-1 seq of HUVEC
ENCFF916QPX	ChromHMM states for K562
ENCFF869GUF	ChromHMM states for GM12878
ENCFF147PPH	ChromHMM states for IMR90
ENCFF654HNG	ChromHMM states for HeLa
ENCFF3970PB	ChromHMM states for HUVEC
	TF ChIP-seq in HUVEC
ENCFF001XSL	CTCF
ENCFF221BNG	FOS
ENCFF222CSK	GATA2
ENCFF001VLI	JUN
ENCFF001VLK	MAX
ENCFF001USZ	MYC
	TF ChIP-seq in IMR90
ENCFF940YUU	BHLHE40
ENCFF001VLN	CEBPB
ENCFF001VLM	CHD1
ENCFF0700IO	CTCF
ENCFF306SXM	ELK1
ENCFF714YWI	FOS
ENCFF886HLJ	MAFK
ENCFF001VLR	MAZ
ENCFF001VLS	MXI1
ENCFF917EWZ	NFE2L2
ENCFF001VLU	RAD21
ENCFF001VLO	RCOR1
ENCFF001VLV	RFX5
ENCFF551KOG	SMC3
ENCFF882BWU	USF2
	TF ChIP-seq in HeLa
ENCFF001VHU	BDP1
ENCFF001VHV	BRCA1
ENCFF001VHW	BRF1
ENCFF002CRY	BRF2
ENCFF002CSA	CEBPB
ENCFF843RWM	CHD1
ENCFF001VIB	CHD2
ENCFF001USV	CTCF
ENCFF304NNT	DEK
ENCFF817LEL	E2F1
ENCFF001VIG	E2F4
ENCFF002CSI	E2F6
ENCFF002CSJ	ELK1
ENCFF001VII	ELK4
ENCFF001VIZ	EP300
ENCFF002CDX	EZH2
ENCFF001VHZ	FOS
ENCFF002CJX	GABPA
ENCFF001VIL	GTF2F1
ENCFF001VJN	GTF3C2

ENCFF002CSM	HA-E2F1
ENCFF001VIN	HCF1
ENCFF001VIP	IRF3
ENCFF002CSP	JUND
ENCFF002CSD	JUN
ENCFF001VIK	KAT2A
ENCFF910DMQ	MAFF
ENCFF796KTU	MAFK
ENCFF002CSR	MAX
ENCFF002CSS	MAZ
ENCFF0022CST	MXI1
ENCFF002DAT	MYC
ENCFF866SGX	NEF2L2
ENCFF002CSU	
ENCFF002CSV	NFYB
ENCFF002CTM	NR2C2
ENCFF001VIY	NRF1
ENCFF002DAV	POLR2A
ENCFF002CSZ	POLR2AphosphoS2
ENCFF002CTD	POLR3A
ENCFF001VJC	PRDM1
ENCFF002CTB	RAD21
ENCFF001VIF	RCOR1
ENCFF002CJY	REST
ENCFF002CTC	RFX5
ENCFF002CRZ	SMARCA4
ENCFF002CSN	SMARCB1
ENCFF002CRT	SMARCC1
ENCFF001VHT	SMARCC2
ENCFF002CTE	SMC3
ENCFF479OQC	SREBF2
ENCFF001VJI	STAT1
ENCFF001VJJ	STAT3
ENCFF001VJG	SUPT20H
ENCFF002CKA	TAF1
ENCFF002CTI	TBP
ENCFF002CTK	TCF7L2
ENCFF922BGZ	UBTF
ENCFF002CTN	USF2
ENCFF738BNS	ZHX1
ENCFF002CTO	ZKSCAN1
ENCFF002CTP	ZNF143
ENCFF002CTQ	ZNF274
ENCFF002CTR	ZZZ3

day 2 until the end of the differentiation, the differentiation medium II, which consists of *"growth medium,* 500*nM rosiglitazone (RGZ) (Sigma-Aldrich, R2408) and* 5*µg/ml insulin (Sigma-Aldrich, I9278) was added and replaced in a* 2 *day-cycle. The growth medium supplemented with* 100*ng/ml bone morphogenetic protein-4 (BMP-4) (PeproTech, 315-27)"* [GSo18] was used to trigger osteoblastogenesis. As in case of adipogenesis, the media was replaced in a 2 day-cycle. Real-time quantitative polymerase chain reaction of well known marker genes was used to verify successful differentiation of ST2 cells into adipocytes and osteoblasts. Furthermore, the cell morphology and the results of Oil Red O staining informed about the cellular differentiation.

Using 1000μl of TRIsure reagents (Bioline, BIO-38033) and 200μl of chloroform (Carl Roth, 6340.1), total RNA was extracted and separated from DNA as well as from proteins. The RNA was precipitated from the aqueous phase, by adding 400μl of 100% isopropanol (Carl Roth, 6752.4) and incubation at -20°C. 1μg of total RNA was used in reverse transcription into *"cDNA using* 0.5*mM dNTPs (ThermoFisher Scientific, R0181),* 2.5*µM oligo dT-primer (Eurofins MWG GmbH, Germany),* 1*U/µl Ribolock RNase inhibitor (ThermoFisher Scientific, EO0381) and* 1*U/µl M-MulV Reverse transcriptase (ThermoFisher Scientific, EP0352) for* 1 *h at* 37°*C."* [GSo18] Alternatively, also 5 U/μl RevertAid Reverse transcriptase for 1h at 42°C have been used. In either case, the PCR was terminated by raising the temperature to 70°C for 10 min. RNA-seq was performed at the Genomics Core Facility at EMBL Heidelberg, using an Illumina NextSeq machine with single-end and unstranded reads.

To profile Histone modifications, chromatin was cross-linked with 1% formaldehyde (Sigma-Aldrich, F87759-25ML) in the culture media for 8 minutes at room temperature. The cross-linking was stopped with 125mM glycine (Carl Roth, 3908.3), which was active for 5 minutes at room temperature as well, before the reagents were removed and the cells were washed two times using ice-cold PBS (Lonza, BE17-516F) which contained the complete^{TM} mini Protease Inhibitor (PI) Cocktail (Roche, 11846145001).

After washing, *"cells were lysed in* 1.7*ml of ice-cold lysis buffer [*5*mM* 1, 4-*Piperazine-diethanesulfonic acid (PIPES) pH* 8.0 *(Carl Roth, 9156.3);* 85*mM potassium chloride (KCl) (PanReac AppliChem, A2939);* 0.5% *4-Nonylphenyl-polyethylene glycol (NP-40) (Fluka Biochemika, 74385)]"* [GSo18], which contained PI as well. After that, cells were incubated on ice for 30 minutes before they were centrifuged at 660 x g for 10 min at 7°C. The remaining *"pellet was resuspended in* 400*µl of ice-cold shearing buffer [*50 *mM Tris Base pH* 8.1 *(Carl Roth, 4855.2);* 10 *mM ethylenediamine tetraacetic acid (EDTA) (Carl Roth, CN06.3);* 0.1% *SDS (PanReac Applichem, A7249);* 0.5% *Sodium deoxycholate (Fluka Biochemika, 30970)]"* [GSo18], that contained PI.

Using a UCD-200TM-EX sonicator, the chromatin was sheared. The sonication went on for 20 cycles, where a cycle is composed of a 30seconds break and 30seconds of sonication. For osteoblasts after 9 days of differentiation, 25 of such sonication cycles were performed. The sheared chromatin was centrifuged at 20817

x g for 10 min at 7°C and diluted in a ratio of 1 : 10 using an enhanced *"RIPA buffer [*140*mM NaCl (Carl Roth, 3957.2);* 10*mM Tris pH* 7.5 *(Carl Roth, 4855.2);* 1*mM EDTA (Carl Roth, CN06.3);* 0.5*mM ethylene glycol-bis(β-amino-ethyl ether)-N,N,N',N'-tetraacetic acid (EGTA) (Carl Roth, 3054.3);* 1% *Triton X-100 (Carl Roth, 3051.2);* 0.01% *SDS (PanReac Applichem, A7249);* 0.1% *sodium deoxycholate (Fluka Biochemika, 30970)]"* [GSo18] containing PI. To perform the immunoprecipitation of H3K4me3 10μg of the sheared chromatin were used. For H3K27ac as well as for H3K36me3 15μg of sheared chromatin were used. The input consisted of 4μg. The antibodies were incubated overnight with the chromatin samples. For H3K4me3, a Millipore antibody 17-614 was used, for H3K27ac and H3K36me3 Abcam antibodies were used, that is ab4729 and ab9050, respectively. The antibodies were captured with 25μl of PureProteome Protein A Magnetic (PAM) Bead System from Millipore. The reaction took place on a rotating wheel for 2*h* at 4°C. The next day, the antibodies were captured using 25μl of PureProteome Protein A Magnetic (PAM) Bead System (Millipore, LSKMAGA10) for 2h at 4°C on a rotating wheel. A DynaMag-2 magnetic stand developed by Life Technologies (12321D) was used to catch the PAM beads. After discarding the supernatant, the beads were washed two times with *"*800μ*l of Immunoprecipitation wash buffer* 1 *(IPWB1) [*20 *mM Tris, pH* 8.1 *(Carl Roth, 4855.2);* 50 *mM NaCl (Carl Roth, 3957.2);* 2 *mM EDTA (Carl Roth, CN06.3);* 1% *Triton X-100 (Carl Roth, 3051.2);* 0.1% *SDS (PanReac Applichem, A7249)], once with* 800μ*l of Immunoprecipitation wash buffer* 2 *(IPWB2) [*10*mM Tris, pH* 8.1 *(Carl Roth, 4855.2);* 150*mM NaCl (Carl Roth, 3957.2);* 1*mM EDTA (Carl Roth, CN06.3),* 1% *NP-40 (Fluka Biochemika, 74385),* 1% *sodium deoxycholate (Fluka Biochemika, 30970),* 250*mM of lithium chloride (LiCl) (Carl Roth, 3739.1)] and twice with* 800μ*l of Tris-EDTA (TE) buffer [*10*mM Tris, pH* 8.1 *(Carl Roth, 4855.2);* 1*mM EDTA (Carl Roth, CN06.3), pH* 8.0*] and incubated with* 100μ*l of ChIP elution buffer [*0.1 *M sodium bicarbonate (NaHCO3) (Sigma-Aldrich, S5761);* 1% *SDS (PanReac Applichem, A7249)]."* [GSo18] The addition of 10μg of RNase A (ThermoFisher, EN0531) and 20μg of proteinase K (ThermoFisher, EO0491) at 65°C overnight, caused the cross-linking to be suspended. Chromatin purification was performed using a MinElute Reaction Cleanup Kit from Qiagen (28206).

As for RNA-seq, the ChIP-seq data has been sequenced on an Illumina HiSeq 2000 machine, using single-end, unstranded reads in the Genomics Core Facility in EMBL Heidelberg resulting in 979.572.918 raw reads. An analysis of read quality using FASTQC version 0.11 [And] showed the presence of adapters, which were cleaned using version 1.5 of AdapterRemoval [Lin12]

Basic processing computational processing of the `fastq` files was done using the PALEOMIX pipeline [S$^+$14a] (version1.0.1). We required reads to have a minimum length of 25 bp and Phred scores > 2. Filtering according to these criteria removed 31.909.435 reads and left us with 947.663.483, which were aligned with BWA [LD09] (version v0.7.10) against the mouse genome GRCm38.p3 (mm10).

For validating, merging BAM files, and marking duplicates, we used the suite tool Picard (version 1.119) [Ins18].

Duplicate reads in the BAM files were marked with PICARD tools (version 1.119) [Ins18], but were not removed. However, reads with a mapping quality < 30 were removed, thus only 661.364.143 reads were used for peak-calling. For peak-calling different tools were used for different histone marks. In detail MACS [Z+08a] version 2.1.0 was used for H3K4me3, HOMER [H+10] was applied to H3K27ac, and SICER [Z+09] version 1.1 was used to call H3K36me3 peaks.

Raw data can be obtained from the European Nucleotide Archive with the accession number PRJEB20933. Ready to use ChIP-seq tracks are available at the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTracks?hubUrl=https://biostat2.uni.lu/dgerard/hub.txt&genome=mm10).

B.3.2 TF ChIP-seq data used for the TF affinity binarization experiments

We obtained TF-ChIP-seq data from ENCODE for K562, GM12878 and HepG2 as listed in Table B.15. In addition, we downloaded H3K27ac data from ENCODE, specifically ENCFF001SWK and ENCFF805KGN for HepG2, ENCFF301TVL and ENCFF001SZE for K562 as well as ENCFF001SUG and ENCFF804NCH for GM12878.

B.3.3 Identification of super-enhancers from H3K27ac data

As mentioned before, version 4.7.2 of HOMER was used to call peaks for H3K27ac. Using the `genomeCoverageBed` command of BEDTOOLS [QH10] version 2.24.0, the coverage for individual SEs was computed and combined into a single region utilizing the commands `unionBedGraphs` and `mergeBed`. The coverage within these merged super-enhancers was computed using the script `annotatePeaks.pl` with the parameters `-size` and `-noann`. In the end, the merged super-enhancers were clustered with STEM, version 1.3.8 to identify temporal profiles of super-enhancer activity.

B.3.4 Experimental validation of suggested regulators

Generation of stable cell lines

ST2 cells were transduced with lentiviral particles (Sirion Biotech) with a MOI of 2, including a reverse tetracycline transactivator controlled by the mouse cytomegalovirus promoter to integrate the CopGFP, Ahr, and Glis1 genes regulated by a Tet-On 3G promoter. Effectively transduced cells were chosen by the addition of 1μg/ml of puromycin to the growth medium. The induced genes were activated by the addition of 1μg/ml doxycycline (Takara, 631311).

Gene silencing

Undifferentiated ST2 cells (day-1) were transfected with Lipofectamine RNAiMAX (Life Technologies, 13778150) according to manufacturer's instructions using 50nM

Table B.15: ENCODE TF ChIP-seq data used to assess the influence of the p-value threshold in the binarziation of TF affinities.

ENCODE accession number	Data Type
ENCFF000DJU	Quantified mRNA of IMR90
ENCFF000SOC	DNase-1 seq of IMR90
ENCFF000DNW	Quantified mRNA of HeLa
ENCFF000SPR	DNase-1 seq of HeLa
ENCFF000DUQ	Quantified mRNA of HUVEC
ENCFF001DNS	DNase-1 seq of HUVEC
ENCFF916QPX	ChromHMM states for K562
ENCFF869GUF	ChromHMM states for GM12878
ENCFF147PPH	ChromHMM states for IMR90
ENCFF654HNG	ChromHMM states for HeLa
ENCFF3970PB	ChromHMM states for HUVEC
	TF ChIP-seq in HUVEC
ENCFF001XSL	CTCF
ENCFF221BNG	FOS
ENCFF222CSK	GATA2
ENCFF001VLI	JUN
ENCFF001VLK	MAX
ENCFF001USZ	MYC
	TF ChIP-seq in IMR90
ENCFF940YUU	BHLHE40
ENCFF001VLN	CEBPB
ENCFF001VLM	CHD1
ENCFF0700IO	CTCF
ENCFF306SXM	ELK1
ENCFF714YWI	FOS
ENCFF886HLJ	MAFK
ENCFF001VLR	MAZ
ENCFF001VLS	MXI1
ENCFF917EWZ	NFE2L2
ENCFF001VLU	RAD21
ENCFF001VLO	RCOR1
ENCFF001VLV	RFX5
ENCFF551KOG	SMC3
ENCFF882BWU	USF2
	TF ChIP-seq in HeLa
ENCFF001VHU	BDP1
ENCFF001VHV	BRCA1
ENCFF001VHW	BRF1
ENCFF002CRY	BRF2
ENCFF002CSA	CEBPB
ENCFF843RWM	CHD1
ENCFF001VIB	CHD2
ENCFF001USV	CTCF
ENCFF304NNT	DEK
ENCFF817LEL	E2F1
ENCFF001VIG	E2F4
ENCFF002CSI	E2F6
ENCFF002CSJ	ELK1
ENCFF001VII	ELK4
ENCFF001VIZ	EP300
ENCFF002CDX	EZH2
ENCFF001VHZ	FOS
ENCFF002CJX	GABPA
ENCFF001VIL	GTF2F1
ENCFF001VJN	GTF3C2
ENCFF002CSM	HA-E2F1

ENCFF001VIN	HCF1
ENCFF001VIP	IRF3
ENCFF002CSP	JUND
ENCFF002CSD	JUN
ENCFF001VIK	KAT2A
ENCFF910DMQ	MAFF
ENCFF796KTU	MAFK
ENCFF002CSR	MAX
ENCFF002CSS	MAZ
ENCFF0022CST	MXI1
ENCFF002DAT	MYC
ENCFF866SGX	NEF2L2
ENCFF002CSU	NFYA
ENCFF002CSV	NFYB
ENCFF002CTM	NR2C2
ENCFF001VIY	NRF1
ENCFF002DAV	POLR2A
ENCFF002CSZ	POLR2AphosphoS2
ENCFF002CTD	POLR3A
ENCFF001VJC	PRDM1
ENCFF002CTB	RAD21
ENCFF001VIF	RCOR1
ENCFF002CJY	REST
ENCFF002CTC	RFX5
ENCFF002CRZ	SMARCA4
ENCFF002CSN	SMARCB1
ENCFF002CRT	SMARCC1
ENCFF001VHT	SMARCC2
ENCFF002CTE	SMC3
ENCFF479OQC	SREBF2
ENCFF001VJI	STAT1
ENCFF001VJJ	STAT3
ENCFF001VJG	SUPT20H
ENCFF002CKA	TAF1
ENCFF002CTI	TBP
ENCFF002CTK	TCF7L2
ENCFF922BGZ	UBTF
ENCFF002CTN	USF2
ENCFF738BNS	ZHX1
ENCFF002CTO	ZKSCAN1
ENCFF002CTP	ZNF143
ENCFF002CTQ	ZNF274
ENCFF002CTR	ZZZ3

of gene-specific siRNAs against mouse Ahr (siAhr) (Dharmacon, M-044066-01-0005), Glis1 (siGlis1) (Dharmacon, M-065576-01-0005) or 50nM of a negative control siRNA duplexes (siControl). Using 50nM of siRNAs from Dharmacon, specifically designed against Ahr (M-044066-01-0005), Glis1 (M-065576-01-0005) and a negative control of siRNA duplexes (D-001206-14-05), the respective genes were silenced. Lipofectamine RNAiMAX (13778150) from Life Technologies was used to transfect the cells. 48h after transfection, cells were collected.

RT-qPCR

An Applied Biosystems 7500 Fast Real-Time PCR System was used to perform RT-qPCR experiments, together with a Thermo Scientific Absolute Blue qPCR SYBR Green Low ROX Mix (AB4322B). For each RT-qPCR run, 5μl of cDNA, 5μl of primer pairs (2μM) and 10μl of the Absolute Blue qPCR mix were used and the reactions took place under the following conditions: 95°C for 15 min followed by 40 cycles of 95°C for 15s, 55°C for 15s and 72°C for 30s. Gene-expression levels were calculated according to the $2(\Delta\Delta Ct)$ method where

$$\Delta\Delta Ct = (\Delta Ct(tg) - \Delta Ct(hg))_{test} - (\Delta Ct(tg) - \Delta Ct(hg))_{control}. \quad \text{(B.3)}$$

Here, tg is the target gene, hg is a housekeeping gene, which in this study was Rpl13a. As controls, the control samples indicated above have been used.

B.4 Appendix Chapter 6

This Section is a slightly adapted from the Methods Section of Nordström *et al.* [N$^+$19].

B.4.1 Sequencing and pre-processing of NGS data

Fastq files generated from HepG2 were trimmed for Adapter sequence and low quality tails (Q< 20) were trimmed from Hepg2 fastq files with TrimGalore!, and mapped to the human reference genome hg19 [C$^+$15b]. To map WGBS as well as NOMe-seq data, *GSNAP* [WN10] was used. DNaseI-seq and ATAC-seq data, was mapped with GEM [MS$^+$12].

B.4.2 WGBS and NOMe-seq

Reads that could not be mapped have been removed with samtools [L$^+$09]. For further processing, the Bis-SNP pipeline [L$^+$12c] was used. Remapping of the reads was focused on regions nearby known SNPs, provided by the Single Nucleotide Polymorphisms database (dbSNP), Build ID: 138 [S$^+$00a]. Picard tools [Ins18] was used to mark duplicates, overlapping sections between two paired reads were removed with bamUtils ([B$^+$13b]). Methylation levels were assessed for all cytosines and extracted with a modified version of the Bis-SNP *vcf2bed.pl* helper-script. Bed files

containing methylation counts for NOMe-seq data were generated for all cytosines in a GCH and HCG context, for WGBS data in a CG context.

B.4.3 DNaseI-seq and ATAC-seq

As for WGBS and NOMe-seq reads, duplicates were with PICARD tools. Accessible regions were identified with MACS2 [Liu18].

In contrast to ChIP-seq peak calling, cutting with DNaseI and the inclusion of adapters with the TN5 transposase focuses on the start and end of fragments. To account for that, the following MASC2 parameters were set: `-shift -100`, `-extsize 200`, `-nomodel` and `-keep-dup all`.

B.4.4 Finding open chromatin regions with NOMe data

NOMe-seq peak calling was performed as explained in Section B.1.5.

B.4.5 Processing of RNA-seq data

In this project, the same RNA-seq processing pipelines was used as introduced in Section B.1.3.

B.4.6 Access to the HepG2 data sets used in this study

The HepG2 data generated in this study can be obtained via EGAD00001002527. The corresponding IDs are:

- EGAX00001422533 for NOMe-seq,
- EGAX00001422534 for DNaseI-seq,
- EGAX00001422548 for ATAC-seq.

Due to brevity, the external data IDs used for validation of the sequence bias, and shape predictions as well as for the methylation signal assessment are not provided here. We refer the reader to the Supplement of Nordström *et al.* [N+19], were these details are provided.

B.4.7 Motif, shape and methylation analysis on additional data sets

Karl Nordström has analysed additional DEEP and Blueprint DNaseI-seq, ATAC-seq, and NOMe-seq samples. The results, shown in Figures B.2, B.3, B.4, B.5, indicate that the observed signatures are not specific to the DEEP HepG2 sample, but are rather assay specific.

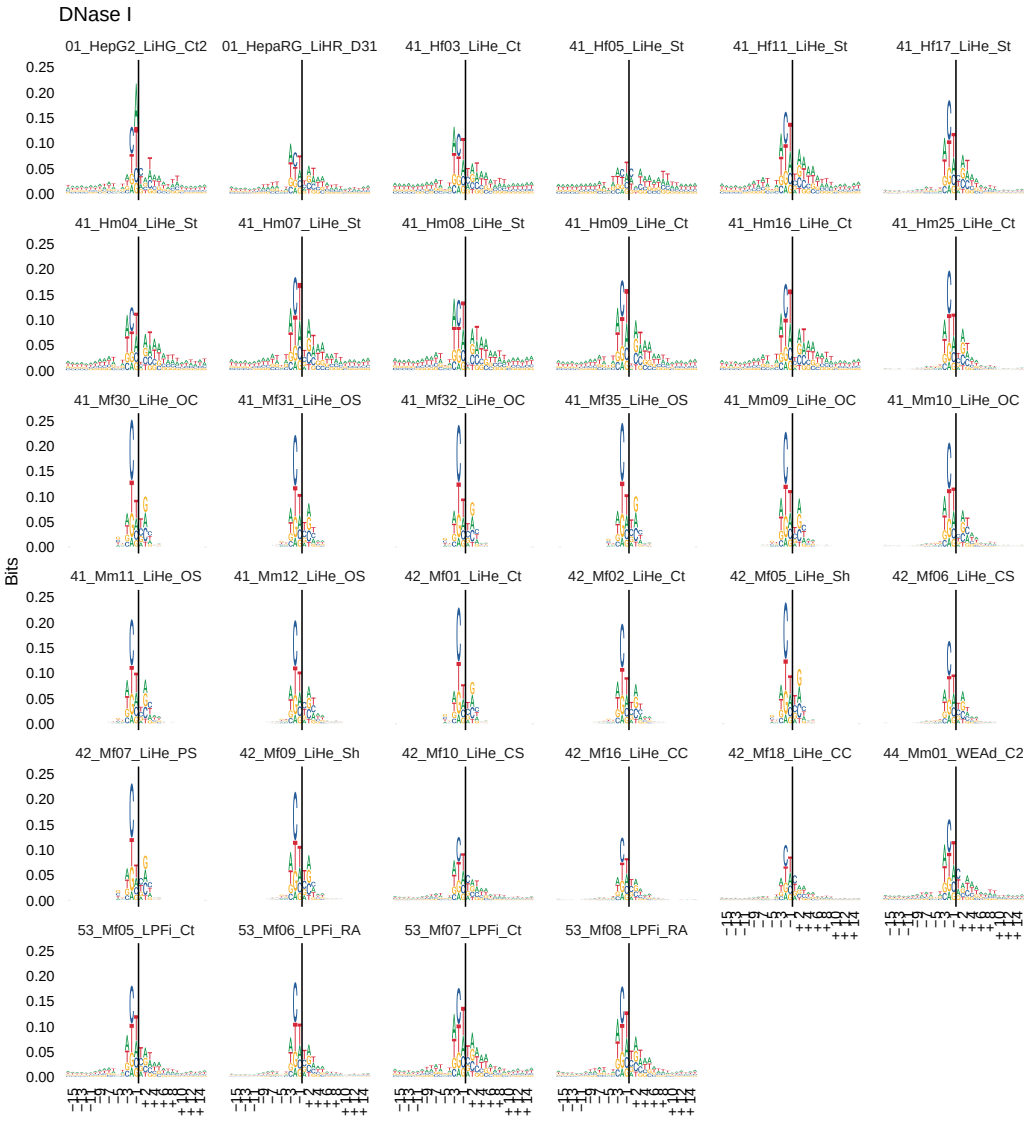


Figure B.2: Motifs of the sequence bias of the DNaseI enzyme computed for additional DNaseI-seq samples from DEEP. Sample IDs are provided in the Figure. All samples were processed identically to the HepG2 data discussed in Chapter 6. Figure from Nordström *et al.* [N$^+$19].

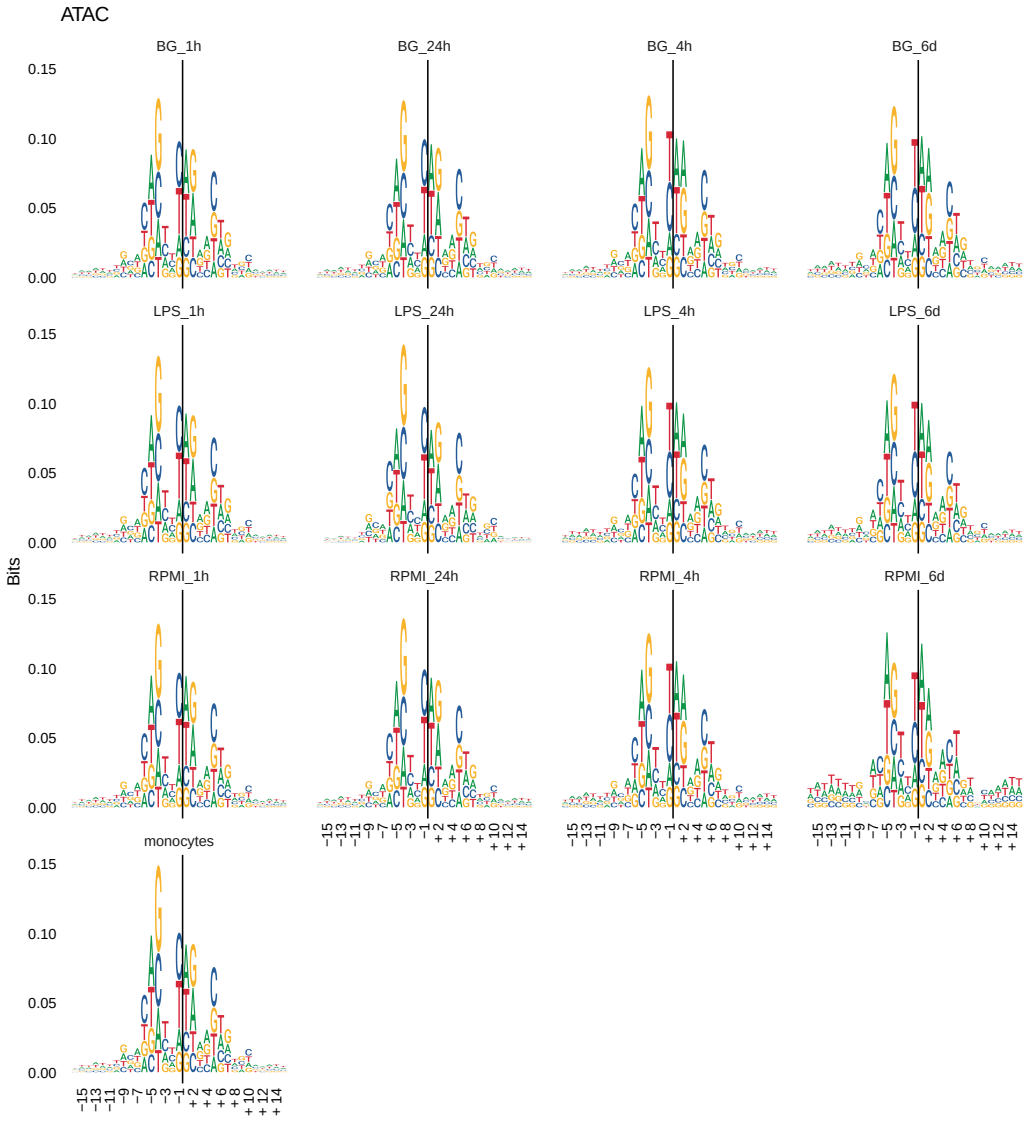


Figure B.3: Sequence bias for additional ATAC-seq samples from Blueprint. The used samples are: BG_1h(GSM2325679), BG_24h (GSM2325681), BG_4h (GSM2325680), BG_6d (GSM2325682), LPS_1h (GSM2325683), LPS_24h (GSM2325685), LPS_4h (GSM2325684), LPS_6d (GSM2325686), monocytes (GSM2325687), RPMI_1h (GSM2325688), RPMI_24h (GSM2325690), RPMI_4h (GSM2325689), RPMI_6d (GSM2325691). All samples were processed identically to the HepG2 data discussed in Chapter 6. Figure from Nordström *et al.* [N+19].

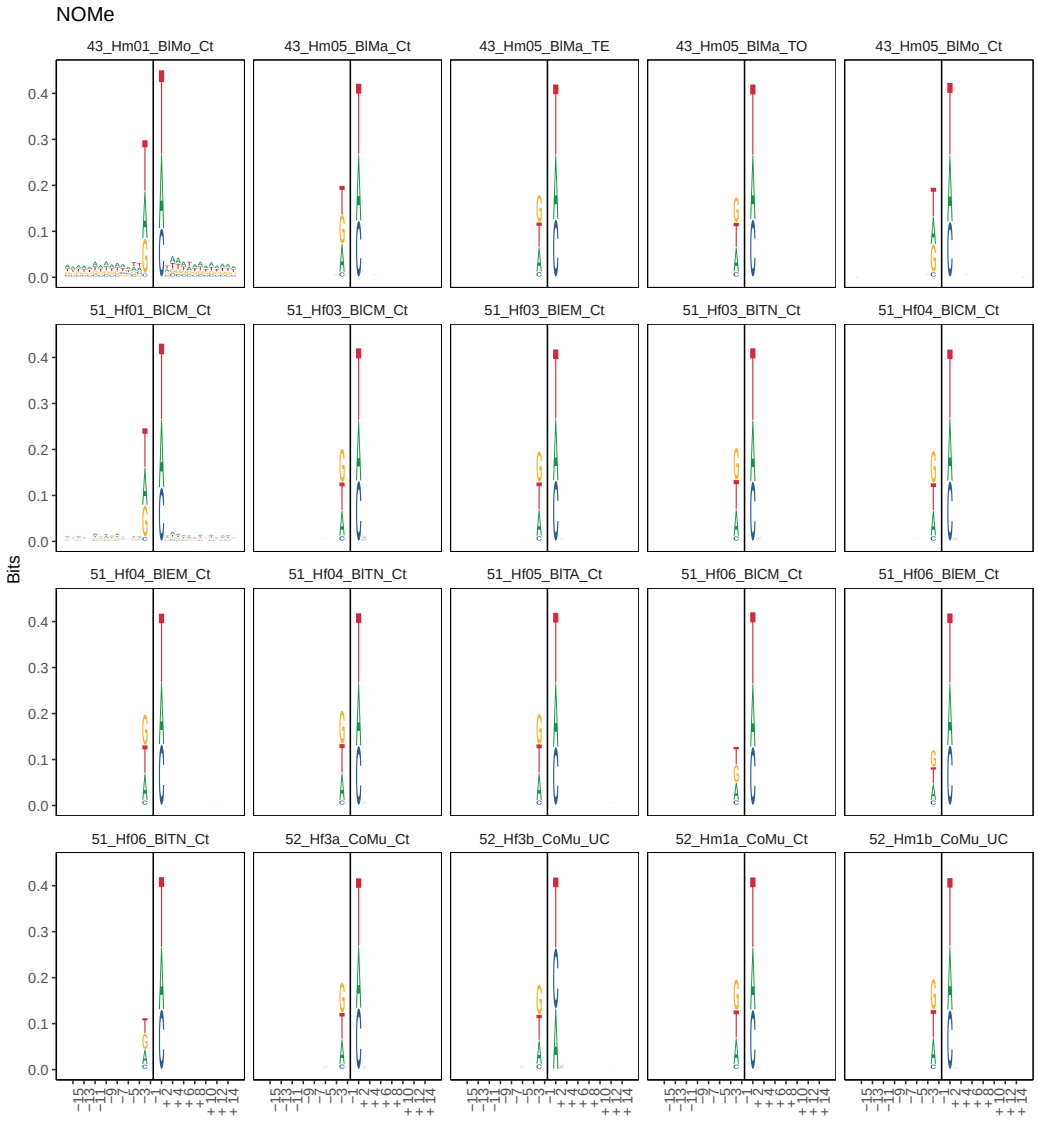


Figure B.4: Sequence bias for additional NOMe-seq samples from DEEP. Sample IDs are provided in the Figure. All samples were processed identically to the HepG2 data discussed in Chapter 6. Figure from Nordström *et al.* [N+19].

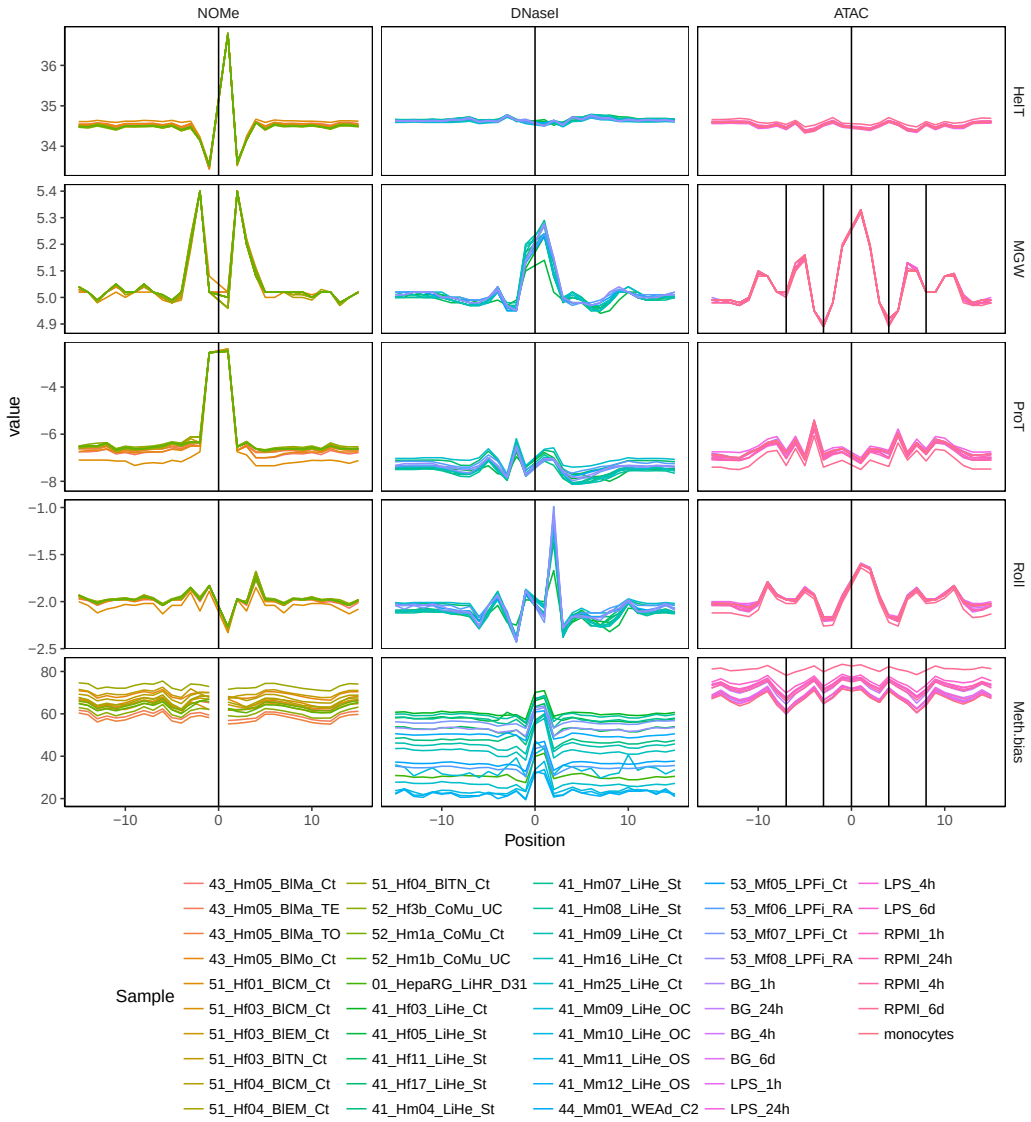


Figure B.5: Prediction of DNA shape and DNA methylation patterns for additional samples. DEEP sample IDs are provided in the legend. Additionally the Blueprint samples: BG_1h(GSM2325679), BG_24h (GSM2325681), BG_4h (GSM2325680), BG_6d (GSM2325682), LPS_1h (GSM2325683), LPS_24h (GSM2325685), LPS_4h (GSM2325684), LPS_6d (GSM2325686), monocytes (GSM2325687), RPMI_1h (GSM2325688), RPMI_24h (GSM2325690), RPMI_4h (GSM2325689), RPMI_6d (GSM2325691) have been used. Figure from Nordström *et al.* [N+19].

B.5 Appendix Chapter 7

B.5.1 IHEC data IDs and CL mapping

DEEP data is listed in Table B.16, ENCODE data is provided in Table B.17, Blueprint data is contained in Table B.18, and Roadmap data is listed in Table B.19 and .

Table B.16: DEEP data IDs and CL mapping

Sample/Experiment ID	Consortia	Tissue	Cell Ontology Term
41_Hf01_LiHe_Ct	DEEP	Hepatocyte	CL:0000182166
41_Hf02_LiHe_Ct	DEEP	Hepatocyte	CL:0000182166
41_Hf03_LiHe_Ct	DEEP	Hepatocyte	CL:0000182166
41Hm09_LiHe_Ct	DEEP	Hepatocyte	CL:0000182166
41_Hm16_LiHe_Ct	DEEP	Hepatocyte	CL:0000182166
41_Hm25_LiHe_Ct	DEEP	Hepatocyte	CL:0000182166
41_Hf05_LiHe_St	DEEP	Hepatocyte	CL:0000182166
41_Hf11_LiHe_St	DEEP	Hepatocyte	CL:0000182166
41_HF14_LiHe_St	DEEP	Hepatocyte	CL:0000182166
41_Hf17_LiHe_St	DEEP	Hepatocyte	CL:0000182166
41_Hm07_LiHe_St	DEEP	Hepatocyte	CL:0000182166
41_Hm08_LiHe_St	DEEP	Hepatocyte	CL:0000182166

B.5.2 Quantification of IHEC RNA-seq data

Gene-expression was quantified using Salmon, version 0.8.2., the Gencode transcript index v26, and the Gencode genome annotation v26.

For **single end** reads, we used the command:

./salmon quant -i gencode.v26.transcripts.index/ -l A -r <Sample>_R1.fastq.gz -p 12 -o quants/<Sample> –seqBias –gcBias -g gencode.v26.annotation.gtf

For **paired end** reads, we used the command:

./salmon quant -i gencode.v26.transcripts.index/ -l A -1 <Sample>_R1.fastq.gz -2 <Sample>_R2.fastq.gz -p 12 -o quants/<Sample> –seqBias –gcBias -g gencode.v26.annotation.gtf

B.5.3 GTEx and TCGA data and CL mapping

An overview of sample counts, and CL IDs for GTEx and TCGA data is provided in Table B.20. Due to data sharing restrictions, we can not provide the distinct IDs for the considered RNA-seq samples. The processed gene-expression data was provided to us by the Genome Institute of Singapore. The data was processed as described in Section 7.4.2 by Engin Cukuroglu.

Table B.17: ENCODE data IDs and CL mapping

Sample/Experiment ID	Consortia	Tissue	Cell Ontology Term
ENCSR000EYS	ENCODE	Endothelial cell of umbilical vein	CL:00026182022
ENCSR000CPK	ENCODE	keratinocyte	CL:0000312254
ENCSR000CPI	ENCODE	keratinocyte	CL:0000312254
ENCSR000EYT	ENCODE	keratinocyte	CL:0000312254
ENCSR000COO	ENCODE	fibroblast of lung	CL:00025531958
ENCSR000CPM	ENCODE	fibroblast of lung	
ENCSR000COP	ENCODE	Foreskin fibroblast	CL:10016081592
ENCSR000CTV	ENCODE	B cell	CL:0000236210
ENCSR000CUC	ENCODE	CD14-positive monocyte	CL:00010541126
ENCSR444WHQ	ENCODE	Skeletal muscle myoblast	CL:0000515430
ENCSR000CUA	ENCODE	Hematopoietic multipotent progenitor cell	CL:0000837754
ENCSR797BPP	ENCODE	Fibroblast of arm	CL:20000151188
ENCSR233IJT	ENCODE	Astrocyte	CL:0000127114
ENCSR276MMH	ENCODE	Adrenal gland	CL:10016011602
ENCSR801MKV	ENCODE	Adrenal gland	CL:10016011602
ENCSR954PZB	ENCODE	Adrenal gland	CL:10016011602
ENCSR532LJV	ENCODE	Thyroid gland	CL:00022581668
ENCSR023ZXN	ENCODE	Thyroid gland	CL:00022581668
ENCSR653ZJF	ENCODE	Transverse colon	CL:1000283774
ENCSR630VJN	ENCODE	Transverse colon	CL:1000283774
ENCSR800WIY	ENCODE	Transverse colon	CL:1000283774
ENCSR967JPI	ENCODE	Gastrocnemius medialis	CL:0000188171
ENCSR071ZLM	ENCODE	Uterus	CL:00021491532
ENCSR113HQM	ENCODE	Uterus	CL:00021491532
ENCSR042GYH	ENCODE	Ovary	CL:00020941469
ENCSR029KNZ	ENCODE	Testis	CL:00022381649
ENCSR701TST	ENCODE	Prostate gland	CL:20000591211
ENCSR968WKR	ENCODE	Bipolar spindle neuron	CL:000010390
ENCSR908ZAS	ENCODE	Hepatocyte	CL:0000182166
ENCSR828TEI	ENCODE	Myotube	CL:00023721780
ENCSR244ISQ	ENCODE	Neural progenitor cell	CL:000004739
ENCSR000EYP	ENCODE	H1-hESC	CL:000003429
ENCSR000COU	ENCODE	H1-hESC	CL:000003429
ENCSR000COW	ENCODE	H1-hESC	CL:000003429
ENCSR000COV	ENCODE	H1-hESC	CL:000003429
ENCSR490SQH	ENCODE	H7-hESC	CL:000003429

Table B.18: Blueprint data IDs and CL mapping

Sample/Experiment ID	Consortia	Tissue	Cell Ontology Term
C0066P12	Blueprint	CD8-positive, alpha-beta T cell	CL:00006255
C005PS12	Blueprint	CD14-positive, CD16-negative classical monocyte	CL:00020571427
S00DFM11	Blueprint	Acute Lymphocytic Leukemia	CL:00020921467
S00HSH11	Blueprint	macrophage - T=6days LPS	CL:0000235209
S00JRB11	Blueprint	macrophage - T=6days LPS	CL:0000235209
S00BYT11	Blueprint	macrophage - T=6days LPS	CL:0000235209
S00CS011	Blueprint	macrophage - T=6days LPS	CL:0000235209
C006NSB1	Blueprint	CD34-negative, CD41-positive, CD42-positive megakaryocyte cell	CL:00020051368
S004BT	Blueprint	CD34-negative, CD41-positive, CD42-positive megakaryocyte cell	CL:00020051368
S008H111	Blueprint	CD4-positive, alpha-beta T cell	CL:0000624532
S002R512	Blueprint	erythroblast	CL:0000765677
S002S312	Blueprint	erythroblast	CL:0000765677
S001S714	Blueprint	Macrophage	CL:0000235209
S001MJ12	Blueprint	Inflammatory macrophage	CL:0000863793
S0022I14	Blueprint	Inflammatory macrophage	CL:0000863793
S00HRJ11	Blueprint	macrophage - T=6days untreated	CL:0000235209
S00BXV11	Blueprint	macrophage - T=6days untreated	CL:0000235209
S00CR211	Blueprint	macrophage - T=6days untreated	CL:0000235209
S00JQD11	Blueprint	macrophage - T=6days untreated	CL:0000235209
S013M311	Blueprint	Acute Myeloid Leukemia	CL:0000766678
S00D0F11	Blueprint	Acute Myeloid Leukemia	CL:0000766678
S00D6311	Blueprint	Acute Myeloid Leukemia	CL:0000766678
S005EJ11	Blueprint	Acute Myeloid Leukemia	CL:0000766678
S013QW11	Blueprint	Acute Myeloid Leukemia	CL:0000766678
S005FH11	Blueprint	Acute Myeloid Leukemia	CL:0000766678
S00XXH11	Blueprint	Acute Myeloid Leukemia	CL:0000766678
S00CXR11	Blueprint	Acute Myeloid Leukemia	CL:0000766678
S013N111	Blueprint	Acute Myeloid Leukemia	CL:0000766678
S00CYP11	Blueprint	Acute Myeloid Leukemia	CL:0000766678
S00D5511	Blueprint	Acute Myeloid Leukemia	CL:0000766678
S00D3911	Blueprint	Acute Myeloid Leukemia	CL:0000766678
S00XUN11	Blueprint	Acute Myeloid Leukemia	CL:0000766678
S00XYF11	Blueprint	Acute Myeloid Leukemia	CL:0000766678
S013PY11	Blueprint	Acute Myeloid Leukemia	CL:0000766678
S00Y1311	Blueprint	Acute Myeloid Leukemia	CL:0000766678
S00XWJ11	Blueprint	Acute Myeloid Leukemia	CL:0000766678
S013RU11	Blueprint	Acute Myeloid Leukemia	CL:0000766678
S00D1D11	Blueprint	Acute Myeloid Leukemia	CL:0000766678
S00Y0511	Blueprint	Acute Myeloid Leukemia	CL:0000766678
S00D4711	Blueprint	Acute Myeloid Leukemia	CL:0000766678
S00XVL11	Blueprint	Acute Myeloid Leukemia	CL:0000766678
S013SS11	Blueprint	Acute Myeloid Leukemia	CL:0000766678
S00Y6U11	Blueprint	Acute Myeloid Leukemia	CL:0000766678
S00CWT11	Blueprint	Acute Myeloid Leukemia	CL:0000766678
S00Y4Y11	Blueprint	Acute Myeloid Leukemia	CL:0000766678
S0022I12	Blueprint	Macrophage	CL:0000235209
C005VG11	Blueprint	Macrophage	CL:0000235209
S00B0N11	Blueprint	Chronic Lymphocytic Leukemia	CL:00025431948
S00B2J11	Blueprint	Chronic Lymphocytic Leukemia	CL:00025431948
S00C0J11	Blueprint	macrophage - T=6days B-glucan	CL:0000235209
S00HTF11	Blueprint	macrophage - T=6days B-glucan	CL:0000235209
S00JS911	Blueprint	macrophage - T=6days B-glucan	CL:0000235209
S00CTZ11	Blueprint	macrophage - T=6days B-glucan	CL:0000235209
C0010KB1	Blueprint	CD14-positive, CD16-negative classical monocyte	CL:00010541126
C001UYB4	Blueprint	CD14-positive, CD16-negative classical monocyte	CL:00010541126
C0011IB1	Blueprint	CD14-positive, CD16-negative classical monocyte	CL:00010541126

Table B.19: Roadmap data IDs and CL mapping

Sample/Experiment ID	Consortia	Tissue	Cell Ontology Term
ENCSR637GBV	Roadmap	Skin fibroblast	CL:00026202024
ENCSR655XQF	Roadmap	Skin fibroblast	CL:00026202024
ENCSR022MON	Roadmap	Skin fibroblast	CL:00026202024
ENCSR982VYI	Roadmap	Skin fibroblast	CL:00026202024
ENCSR361DRG	Roadmap	Fibroblast of skin of abdomen	CL:20000131190
ENCSR681ALA	Roadmap	Fibroblast of skin of abdomen	CL:20000131190
ENCSR762CJN	Roadmap	Trophoblast cell	CL:0000351287
ENCSR406YML	Roadmap	Muscle of arm	CL:0000188171
ENCSR364IBB	Roadmap	Muscle of arm	CL:00001881
ENCSR317LMH	Roadmap	Muscle of arm	CL:00001881
ENCSR620ZNQ	Roadmap	Muscle of arm	CL:00001881
ENCSR305NXN	Roadmap	Muscle of arm	CL:00001881
ENCSR677MYO	Roadmap	Muscle of arm	CL:00001881
ENCSR990LHE	Roadmap	Muscle of arm	CL:00001881
ENCSR922VBO	Roadmap	Stomach	CL:1000313832
ENCSR721HDG	Roadmap	Stomach	CL:1000313832
ENCSR702IGQ	Roadmap	Stomach	CL:1000313832
ENCSR549DVY	Roadmap	Stomach	CL:1000313832
ENCSR783BUO	Roadmap	Stomach	CL:1000313832
ENCSR951NPS	Roadmap	Stomach	CL:1000313832
ENCSR123ZCX	Roadmap	Stomach	CL:1000313832
ENCSR774SEX	Roadmap	Stomach	CL:1000313832
ENCSR729ZII	Roadmap	Muscle of back	CL:0000188171
ENCSR806ESH	Roadmap	Muscle of back	CL:0000188171
ENCSR995ORR	Roadmap	Muscle of back	CL:0000188171
ENCSR891JVD	Roadmap	Muscle of back	CL:0000188171
ENCSR652AWW	Roadmap	Muscle of back	CL:0000188171
ENCSR027EJD	Roadmap	Muscle of back	CL:0000188171
ENCSR576UKA	Roadmap	Muscle of back	CL:0000188171
ENCSR094RGI	Roadmap	Muscle of back	CL:0000188171
ENCSR239BBI	Roadmap	Muscle of back	CL:0000188171
ENCSR522XTV	Roadmap	Muscle of back	CL:0000188171
ENCSR719HRO	Roadmap	Small intestine	CL:10015981554
ENCSR621FYE	Roadmap	Small intestine	CL:10015981554
ENCSR150JIX	Roadmap	Small intestine	CL:10015981554
ENCSR446RKD	Roadmap	Small intestine	CL:10015981554
ENCSR523EDD	Roadmap	Small intestine	CL:10015981554
ENCSR096USV	Roadmap	Muscle of leg	CL:0000188171
ENCSR860DST	Roadmap	Muscle of leg	CL:0000188171
ENCSR144UVO	Roadmap	Muscle of leg	CL:0000188171
ENCSR545WAC	Roadmap	Muscle of leg	CL:0000188171
ENCSR174ESD	Roadmap	Muscle of leg	CL:0000188171
ENCSR086DZF	Roadmap	Muscle of leg	CL:0000188171
ENCSR561WEX	Roadmap	Muscle of leg	CL:0000188171
ENCSR447UE	Roadmap	Muscle of leg	CL:0000188171
ENCSR286KWP	Roadmap	Large intestine	CL:1000320881
ENCSR859KGW	Roadmap	Large intestine	CL:1000320881
ENCSR777ONH	Roadmap	Large intestine	CL:1000320881
ENCSR930URM	Roadmap	Large intestine	CL:1000320881
ENCSR857VKL	Roadmap	Large intestine	CL:1000320881
ENCSR363BVC	Roadmap	Large intestine	CL:1000320881
ENCSR861SOG	Roadmap	Left lung	CL:00020621433
ENCSR733MWN	Roadmap	Left lung	CL:00020621433
ENCSR592EZK	Roadmap	Left lung	CL:00020621433
ENCSR499NEL	Roadmap	Left lung	CL:00020621433
ENCSR222IGR	Roadmap	Left lung	CL:00020621433
ENCSR572FXC	Roadmap	Left lung	CL:00020621433

ENCSR907KDH	Roadmap	Kidney	CL:10004971165
ENCSR212AMA	Roadmap	Kidney	CL:10004971165
ENCSR896QPD	Roadmap	Kidney	CL:10004971165
ENCSR495UXA	Roadmap	Kidney	CL:10004971165
ENCSR554KBK	Roadmap	Right lung	CL:00020621433
ENCSR074APH	Roadmap	Right lung	CL:00020621433
ENCSR560MDQ	Roadmap	Right lung	CL:00020621433
ENCSR176WMG	Roadmap	Right lung	CL:00020621433
ENCSR044JAQ	Roadmap	Right lung	CL:00020621433
ENCSR367QHR	Roadmap	Thymus	CL:00022931702
ENCSR158XIJ	Roadmap	Thymus	CL:00022931702
ENCSR069CMT	Roadmap	Thymus	CL:00022931702
ENCSR175CNQ	Roadmap	Thymus	CL:00022931702
ENCSR047LIJ	Roadmap	Heart	CL:00024941900
ENCSR863BUL	Roadmap	Heart	CL:00024941900
ENCSR328PVI	Roadmap	Renal cortex interstitium	CL:10005961200
ENCSR899SWV	Roadmap	Renal cortex interstitium	CL:10005961200
ENCSR436ZKE	Roadmap	Renal cortex interstitium	CL:10005961200
ENCSR335GET	Roadmap	Adrenal gland	CL:10016011602
ENCSR120NEA	Roadmap	Adrenal gland	CL:10016011602
ENCSR688YOZ	Roadmap	Adrenal gland	CL:10016011602
ENCSR7400OPV	Roadmap	Adrenal gland	CL:10016011602
ENCSR424TSZ	Roadmap	Renal Pelvis	CL:10004971165
ENCSR204XBB	Roadmap	Renal Pelvis	CL:10004971165
ENCSR929KRW	Roadmap	Renal Pelvis	CL:10004971165
ENCSR702IMR	Roadmap	Left Kidney	CL:10004971165
ENCSR015EMF	Roadmap	Left renal cortex interstitium	CL:10005961200
ENCSR125NGM	Roadmap	Left renal cortex interstitium	CL:10005961200
ENCSR759WPF	Roadmap	Left renal cortex interstitium	CL:10005961200
ENCSR413LXW	Roadmap	Left renal cortex interstitium	CL:10005961200
ENCSR029FTY	Roadmap	Left renal pelvis	CL:10004971165
ENCSR321ROU	Roadmap	Left renal pelvis	CL:10004971165
ENCSR410DUZ	Roadmap	Left renal pelvis	CL:10004971165
ENCSR160UAZ	Roadmap	Left renal pelvis	CL:10004971165
ENCSR552YAE	Roadmap	Right renal pelvis	CL:10004971165
ENCSR352GCS	Roadmap	Right renal pelvis	CL:10004971165
ENCSR543TQW	Roadmap	Right renal pelvis	CL:10004971165
ENCSR928CEQ	Roadmap	Right renal pelvis	CL:10004971165
ENCSR899NLW	Roadmap	Spinal cord	CL:00050002081
ENCSR333FZW	Roadmap	Spinal cord	CL:00050002081
ENCSR822AOE	Roadmap	Right renal cortex interstitium	CL:10005961200
ENCSR884EVS	Roadmap	Right renal cortex interstitium	CL:10005961200
ENCSR400DJE	Roadmap	Right renal cortex interstitium	CL:10005961200
ENCSR265NZF	Roadmap	Spleen	CL:000265120
ENCSR817TLH	Roadmap	Psoas muscle	CL:0000188171
ENCSR531RKI	Roadmap	Muscle of trunk	CL:0000188171
ENCSR727VTD	Roadmap	Ovary	CL:00020941469
ENCSR725TPW	Roadmap	Ovary	CL:00020941469
ENCSR629VMZ	Roadmap	Pancreas	CL:10015991552
ENCSR571BML	Roadmap	Pancreas	CL:10015991552
ENCSR755LFM	Roadmap	Testis	CL:00022381649
ENCSR711NGL	Roadmap	Forelimb muscle	CL:0000188171
ENCSR516VDS	Roadmap	Hindlimb muscle	CL:0000188171
ENCSR911GQI	Roadmap	H1-hESC	CL:000003429
ENCSR844HLP	Roadmap	H1-hESC	CL:000003429

Table B.20: Sample counts per tissue and consortia as well as CL terms for TCGA and GTEx data

Tissue	Consortia	Counts	Cell Ontology Term
Thyroid	TCGA	59	CL:0000452
Thyroid	GTEx	355	CL:0000452
Liver	TCGA	50	CL:0000182
Liver	GTEx	176	CL:0000182
Kidney	TCGA	72	CL:1000497
Kidney	GTEx	36	CL:1000497
Colon	TCGA	41	CL:1001588
Colon	GTEx	376	CL:1001588
Prostate	TCGA	52	CL:0002231
Prostate	GTEx	119	CL:0002231

B.6 Appendix Chapter 8

B.6.1 Data used within the project

Within the scope of the StitchIt project, we utilized paired DNaseI-seq and RNA-seq data from ENCODE, Blueprint, and Roadmap, listed in Tables B.21 to B.23. We utilized all fastq files available using the provided data accession IDs. The processed datasets are available at Zenodo (10.5281/zenodo.2547384).

RNA-seq processing

The same RNA-seq processing pipeline as introduced in Section B.5 has been used in context of this project as well: Gene-expression was quantified using Salmon, version 0.8.2., the Gencode transcript index v26, and the Gencode genome annotation v26.

For **single end** reads, we used the command:
./salmon quant -i gencode.v26.transcripts.index/ -l A -r <Sample>_R1.fastq.gz -p 12 -o quants/<Sample> –seqBias –gcBias -g gencode.v26.annotation.

For **paired end** reads, we used the command:
./salmon quant -i gencode.v26.transcripts.index/ -l A -1 <Sample>_R1.fastq.gz -2 <Sample>_R2.fastq.gz -p 12 -o quants/<Sample> –seqBias –gcBias -g gencode.v26.annotation.gtf

Discretized gene-expression values were computed using the POE method [GP03].

Table B.21: Blueprint and internal sample IDs of paired DNaseI-seq and RNA-seq data

Matched Sample	Internal Sample ID	Description
C0066P12	B_C0066P12	CD8-positive, alpha-beta T cell
C005PS12	B_C005PS12	CD14-positive, CD16-negative classical monocyte
S00DFM11	B_S00DFM11	Acute Lymphocytic Leukemia
S00HSH11	B_S00HSH11	macrophage - T=6days LPS
S00JRB11	B_S00JRB11	macrophage - T=6days LPS
S00BYT11	B_S00BYT11	macrophage - T=6days LPS
S00CS011	B_S00CS011	macrophage - T=6days LPS
C006NSB1	B_C006NSB1	CD34-negative, CD41-positive, CD42-positive megakaryocyte cell
S004BT	B_S004BT12	CD34-negative, CD41-positive, CD42-positive megakaryocyte cell
S008H111	B_S008H111	CD4-positive, alpha-beta T cell
S002R512	B_S002R512	erythroblast
S002S312	B_S002S312	erythroblast
S001S714	B_S001S714	macrophage
S001MJ12	B_S001MJ12	inflammatory macrophage
S0022I14	B_S0022I14	inflammatory macrophage
S00HRJ11	B_S00HRJ11	macrophage - T=6days untreated
S00BXV11	B_S00BXV11	macrophage - T=6days untreated
S00CR211	B_S00CR211	macrophage - T=6days untreated
S00JQD11	B_S00JQD11	macrophage - T=6days untreated
S013M311	B_S013M311	Acute Myeloid Leukemia
S00D0F11	B_S00D0F11	Acute Myeloid Leukemia
S00D6311	B_S00D6311	Acute Myeloid Leukemia
S005EJ11	B_S005EJ11	Acute Myeloid Leukemia
S013QW11	B_S013QW11	Acute Myeloid Leukemia
S005FH11	B_S005FH11	Acute Myeloid Leukemia
S00XXH11	B_S00XXH11	Acute Myeloid Leukemia
S00CXR11	B_S00CXR11	Acute Myeloid Leukemia
S013N111	B_S013N111	Acute Myeloid Leukemia
S00CYP11	B_S00CYP11	Acute Myeloid Leukemia
S00D5511	B_S00D5511	Acute Myeloid Leukemia
S00D3911	B_S00D3911	Acute Myeloid Leukemia
S00XUN11	B_S00XUN11	Acute Myeloid Leukemia
S00XYF11	B_S00XYF11	Acute Myeloid Leukemia
S013PY11	B_S013PY11	Acute Myeloid Leukemia
S00Y1311	B_S00Y1311	Acute Myeloid Leukemia
S00XWJ11	B_S00XWJ11	Acute Myeloid Leukemia
S013RU11	B_S013RU11	Acute Myeloid Leukemia
S00D1D11	B_S00D1D11	Acute Myeloid Leukemia
S00Y0511	B_S00Y0511	Acute Myeloid Leukemia
S00D4711	B_S00D4711	Acute Myeloid Leukemia
S00XVL11	B_S00XVL11	Acute Myeloid Leukemia
S013SS11	B_S013SS11	Acute Myeloid Leukemia
S00Y6U11	B_S00Y6U11	Acute Myeloid Leukemia
S00CWT11	B_S00CWT11	Acute Myeloid Leukemia
S00Y4Y11	B_S00Y4Y11	Acute Myeloid Leukemia
S0022I12	B_S0022I12	macrophage
C005VG11	B_C005VG11	macrophage
S00B0N11	B_S00B0N11	Chronic Lymphocytic Leukemia
S00B2J11	B_S00B2J11	Chronic Lymphocytic Leukemia
S00C0J11	B_S00C0J11	macrophage - T=6days B-glucan
S00HTF11	B_S00HTF11	macrophage - T=6days B-glucan
S00JS911	B_S00JS911	macrophage - T=6days B-glucan
S00CTZ11	B_S00CTZ11	macrophage - T=6days B-glucan
C0010KB1	B_C0010KB1	CD14-positive, CD16-negative classical monocyte
C001UYB4	B_C001UYB4	CD14-positive, CD16-negative classical monocyte
C0011IB1	B_C0011IB1	CD14-positive, CD16-negative classical monocyte

Table B.22: Roadmap IDs and internal ID of paired DNaseI-seq and RNA-seq data

Sample ID	Internal Sample ID	Description
ENCBS336CDQ	R_ENCBS336CDQ	skin fibroblast
ENCBS890NFL	R_ENCBS890NFL	skin fibroblast
ENCBS180EDA	R_ENCBS180EDA	skin fibroblast
ENCBS405WVO	R_ENCBS405WVO	fibroblast of skin of abdomen
ENCBS599YIE	R_ENCBS599YIE	fibroblast of skin of abdomen
ENCBS048TNH	R_ENCBS754TWW_ENCBS048TNH	IMR-90
ENCBS376RZJ	R_ENCBS150HBC_ENCBS376RZJ	trophoblast cell
ENCBS090AGL	R_ENCBS090AGL	muscle of arm
ENCBS516CJQ	R_ENCBS516CJQ	muscle of arm
ENCBS054WKY	R_ENCBS054WKY	muscle of arm
ENCBS586CPQ	R_ENCBS586CPQ	muscle of arm
ENCBS261IIB	R_ENCBS261IIB	muscle of arm
ENCBS180RZG	R_ENCBS180RZG	muscle of arm
ENCBS892WJE	R_ENCBS892WJE	muscle of arm
ENCBS578HBL	R_ENCBS578HBL	stomach
ENCBS441WEO	R_ENCBS441WEO	stomach
ENCBS220QDW	R_ENCBS220QDW	stomach
ENCBS246MHN	R_ENCBS246MHN	stomach
ENCBS291NHF	R_ENCBS291NHF	stomach
ENCBS878MRX	R_ENCBS878MRX	stomach
ENCBS159PIU	R_ENCBS159PIU	stomach
ENCBS716QQK	R_ENCBS716QQK	stomach
ENCBS384LIR	R_ENCBS384LIR	muscle of back
ENCBS174IGM	R_ENCBS174IGM	muscle of back
ENCBS345TTL	R_ENCBS345TTL	muscle of back
ENCBS136SDO	R_ENCBS136SDO	muscle of back
ENCBS645HGQ	R_ENCBS645HGQ	muscle of back
ENCBS020XIW	R_ENCBS020XIW	muscle of back
ENCBS897YOR	R_ENCBS897YOR	muscle of back
ENCBS479AOA	R_ENCBS479AOA	muscle of back
ENCBS825MQT	R_ENCBS825MQT	muscle of back
ENCBS136EGD	R_ENCBS136EGD	muscle of back
ENCBS853LFM	R_ENCBS853LFM	small intestine
ENCBS615YKY	R_ENCBS615YKY	small intestine
ENCBS529UES	R_ENCBS529UES	small intestine
ENCBS623YHX	R_ENCBS623YHX	small intestine
ENCBS133LAN	R_ENCBS133LAN	small intestine
ENCBS611ZBY	R_ENCBS611ZBY	muscle of leg
ENCBS517FUR	R_ENCBS517FUR	muscle of leg
ENCBS023IXF	R_ENCBS023IXF	muscle of leg
ENCBS947JRD	R_ENCBS947JRD	muscle of leg
ENCBS011TVS	R_ENCBS011TVS	muscle of leg
ENCBS099OIO	R_ENCBS099OIO	muscle of leg
ENCBS143XQJ	R_ENCBS143XQJ	muscle of leg
ENCBS984JKS	R_ENCBS984JKS	muscle of leg
ENCBS997WGU	R_ENCBS997WGU	large intestine
ENCBS588ZWT	R_ENCBS588ZWT	large intestine
ENCBS445IVN	R_ENCBS445IVN	large intestine
ENCBS699KFK	R_ENCBS699KFK	large intestine
ENCBS867ILV	R_ENCBS867ILV	large intestine
ENCBS383OVQ	R_ENCBS383OVQ	large intestine
ENCBS078XUR	R_ENCBS078XUR	left lung
ENCBS574MIZ	R_ENCBS574MIZ	left lung
ENCBS859ASH	R_ENCBS859ASH	left lung
ENCBS143LJK	R_ENCBS143LJK	left lung
ENCBS516MKG	R_ENCBS516MKG	left lung
ENCBS117CVU	R_ENCBS117CVU	left lung

ENCBS478OZL	R_ENCBS478OZL	kidney
ENCBS263ZZU	R_ENCBS263ZZU	kidney
ENCBS434EOI	R_ENCBS434EOI	kidney
ENCBS034SKE	R_ENCBS034SKE	kidney
ENCBS917VNB	R_ENCBS917VNB	right lung
ENCBS993PWO	R_ENCBS993PWO	right lung
ENCBS467PJB	R_ENCBS467PJB	right lung
ENCBS421OZO	R_ENCBS421OZO	right lung
ENCBS122USS	R_ENCBS122USS	right lung
ENCBS484BGT	R_ENCBS484BGT	thymus
ENCBS948PMG	R_ENCBS948PMG	thymus
ENCBS054CPR	R_ENCBS054CPR	thymus
ENCBS198CXJ	R_ENCBS198CXJ	thymus
ENCBS172XKB	R_ENCBS172XKB	heart
ENCBS407ALA	R_ENCBS407ALA	heart
ENCBS620YJZ	R_ENCBS620YJZ	renal cortex interstitium
ENCBS448HVV	R_ENCBS448HVV	renal cortex interstitium
ENCBS026RJE	R_ENCBS026RJE	renal cortex interstitium
ENCBS232ONZ	R_ENCBS232ONZ	adrenal gland
ENCBS660CJK	R_ENCBS660CJK	adrenal gland
ENCBS200XAZ	R_ENCBS200XAZ	adrenal gland
ENCBS119WRO	R_ENCBS119WRO	adrenal gland
ENCBS262QPN	R_ENCBS262QPN	renal pelvis
ENCBS142XDW	R_ENCBS142XDW	renal pelvis
ENCBS785KTZ	R_ENCBS785KTZ	renal pelvis
ENCBS610XAP	R_ENCBS610XAP	left kidney
ENCBS376PWL	R_ENCBS376PWL	left renal cortex interstitium
ENCBS674SNK	R_ENCBS674SNK	left renal cortex interstitium
ENCBS281CNH	R_ENCBS281CNH	left renal cortex interstitium
ENCBS636QOC	R_ENCBS636QOC	left renal cortex interstitium
ENCBS055ULH	R_ENCBS055ULH	left renal pelvis
ENCBS257BTU	R_ENCBS257BTU	left renal pelvis
ENCBS226ZND	R_ENCBS226ZND	left renal pelvis
ENCBS754ANY	R_ENCBS754ANY	left renal pelvis
ENCBS145EYH	R_ENCBS145EYH	right renal pelvis
ENCBS935VKR	R_ENCBS935VKR	right renal pelvis
ENCBS855RFN	R_ENCBS855RFN	right renal pelvis
ENCBS827OFK	R_ENCBS827OFK	right renal pelvis
ENCBS373RUA	R_ENCBS373RUA	spinal cord
ENCBS300UPT	R_ENCBS300UPT	spinal cord
ENCBS100DZU	R_ENCBS100DZU	right renal cortex interstitium
ENCBS183TBX	R_ENCBS183TBX	right renal cortex interstitium
ENCBS818WCN	R_ENCBS818WCN	right renal cortex interstitium
ENCBS599HUG	R_ENCBS599HUG	spleen
ENCBS008QPC	R_ENCBS008QPC	psoas muscle
ENCBS992XAC	R_ENCBS992XAC	muscle of trunk
ENCBS645JEU	R_ENCBS645JEU	ovary
ENCBS341OKA	R_ENCBS341OKA	ovary
ENCBS507RPJ	R_ENCBS507RPJ	pancreas
ENCBS914JTX	R_ENCBS914JTX	pancreas
ENCBS796DWQ	R_ENCBS796DWQ	testis
ENCBS988KQJ	R_ENCBS988KQJ	forelimb muscle
ENCBS105LQM	R_ENCBS105LQM	hindlimb muscle
ENCBS945MCY	R_ENCBS559QNR_ENCBS568FYY_ENCBS945MCY	H1-hESC

Table B.23: ENCODE IDs and internal IDs of paired DNaseI-seq and RNA-seq data

Experiment ID	Internal Sample ID	Description
ENCSR000EKF	E_112ENC_124GGK_774AAA_719AAA_743IPG	Endothelial cell of umbilical vein
ENCSR000EPQ	E_589ENC_591ENC_818IBE_567ENC_565ENC_569ENC_564ENC_563ENC_586ENC	keratinocyte
ENCSR000ELY		fibroblast of lung
ENCSR000EPR	E_340AAA_936GPP_612ENC_613ENC	fibroblast of lung
ENCSR000EME	E_074ENC_911DVL_753AAA_754AAA	foreskin fibroblast
ENCSR000EMJ	E_852WTL_483ENC	B cell
ENCSR000ELE	E_477CHZ_628ENC_626ENC	CD14-positive monocyte
ENCSR000EOO	E_328AAA	skeletal muscle myoblast
ENCSR000EOO	E_665DRD	skeletal muscle myoblast
ENCSR000EMK	E_485ENC	Hematopoietic multipotent progenitor cell
ENCSR217TAW	E_367AAA	fibroblast of arm
ENCSR217TAW	E_372AAA	fibroblast of arm
ENCSR000EPM	E_021ENC	astrocyte
ENCSR000EPM	E_0052WQU	astrocyte
ENCSR191FOV	E_371OZD_227VDO	adrenal gland
ENCSR865ICK	E_423TBO_548ULT	adrenal gland
ENCSR848XIY	E_942WBM_724UYV	adrenal gland
ENCSR158VAT	E_658GLE_376ASB	thyroid gland
ENCSR902XFY	E_624WYQ_702NUN	thyroid gland
ENCSR979ZJS	E_174MRL_767MGS	transverse colon
ENCSR790FIS	E_409AIP_668MJW	transverse colon
ENCSR504WYA	E_767POE_974EYF	transverse colon
ENCSR171ETY	E_921WWM_472LWI	gastrocnemius medialis
ENCSR209TXI	E_020UWI_887WTT	uterus
ENCSR237WJY	E_310EYM_210YNQ	uterus
ENCSR855DJV	E_279OPH_711CPB	ovary
ENCSR475SYH	E_197JOA_315DHM	testis
ENCSR456SOX	E_524RBS_291PVB	prostate gland
ENCSR626RVD	E_369AAA	bipolar spindle neuron
ENCSR626RVD	E_374AAA	bipolar spindle neuron
ENCSR364MFN	E_077RUJ	hepatocyte
ENCSR364MFN	E_520VFV	hepatocyte
ENCSR000EOP	E_526EMC	myotube
ENCSR000EOP	E_236AFP_140MNV	myotube
ENCSR000EPD	E_234AAA	LHCN-M2
ENCSR000EPD	E_869RCC_231OMQ	LHCN-M2
ENCSR963ALV	E_018TPT	neural progenitor cell
ENCSR963ALV	E_044KWE	neural progenitor cell
ENCSR000EJN	E_111ENC_780AAA_716AAA_051SJH_731AAA_734AAA_733AAA_732AAA	H1-hESC
ENCSR000EMZ	E_293AAA	H7-hESC
ENCSR000EMZ	E_297CQV_291AAA_624XJG	H7-hESC

DNaseI-seq processing

DNaseI-seq reads were aligned to the hg38 reference genome with BOWTIE, version 1.2.1.1, and SAMTOOLS version 1.2.
For **single end** reads, this was done using the command call:
.bowtie –threads 10 -S GRCh38_no_alt/GC_A_000001405.15_GRCh38_no_alt _analysis_set <Sample>_R1.fastq.gz | samtools view -b -o <Sample>.bam -) 2 > Sample > bowtie_statistics.txt

and for **paired end** reads using:
./bowtie –threads 10 -S GRCh38_no_alt/GCA_000001405.15_GRCh38_no_alt_ analysis_set -1 <Sample>_R1.fastq.gz -2 <Sample>_R2.fastq.gz | samtools view -b -o <Sample>.bam -) 2> <Sample> .bowtie_statistics.txt

In case that multiple fasta files exist for one sample, they were concatenated using the linux CAT command prior to alignment, if applicable, separately per strand. DNase Hypersensitive Sites were identified with JAMM, version 1.0.7.5 and SAMTOOLS version 1.2.
For **single end** reads, we used the commands:
bedtools bamtobed -i <Sample>.bam > JAMM-Input/<Sample>.bed
bash JAMM.sh -s JAMM-Input -g hg38_chrSize.txt -o <Sample>_peaks -f 1 -p 8
rm -r JAMM-Input/
For **paired end** reads, we run the commands:
samtools sort -n -O bam -@10 -T Bam-Sort-Pre <Sample>.bam | samtools view -bf 0x2 - | bedtools bamtobed -bedpe -i stdin > JAMM-Input/<Sample>.bed
bash JAMM.sh -s JAMM-Input -g hg38_chrSize.txt -o <Sample>_peaks -f 1 -p 8 -t paired
rm -r JAMM-Input/
As described in the JAMM github, we computed the enrichment of DNaseI-seq signal within the peak by dividing column 7 by column 9 of the JAMM output files:
awk 'print $1 $2 $3 $7 $9' <Sample>/peaks/filtered.peaks.narrowPeak > <Sample>/peaks/filtered.peaks.narrowPeak.adoptedCoverage

Bigwig files were generated using the DEEPTOOLS bamCoverage program, version 2.4.2 with the command
bamCoverage -b <Sample>.bam.sorted.bam -o <Sample>.bw -p 4 -normalizeUsingRPKM

B.6.2 Details on executing the tested methods

STITCHIT

To run STITCHIT, the user needs to provide discretized (-d) and original expression data (-o), a gene annotation file (-a), a chromosome size file (-s), as well as big wig

files with the epigenetic signal to consider (-b). Using the command:

./build/core/STITCHIT -b <Consortium>/DNase_bw/ -a ../../../nobackup/References/gencode.v26.annotation.gtf -d <Consortium>/_Discretised_Complet.txt -o <Consortium>_expression.txt -s data/hg38_chrSize.txt -w 25000 -c 12 -p 0.05 -g <geneID> -z 10 -f ../../../archive00/Segmentation_<Consortium>/ -r 500000 -t 2000

a call to STITCHIT can be invoked.

The parameter -w denotes the size of the window extension up an downstream of the gene, -c denotes the number of used CPU, -p is the significance threshold for the correlation test, -g is the parameter to denote the target gene ID, -z indicates the width of the initial binning, -f denotes the output path, -r is the maximum size of the entire search region and -t refers to the maximum size of a segment.

Unsupervised peak based assignment

We compute these quantities using the TEPIC [12] tool using the following commands

5kb:

bash TEPIC.sh -g hg38.noPrefix.masked.fa -b <Sample>_peaks/peaks/filtered.peaks.narrowPeak.adoptedCoverage -o <Sample>_5kb -p ../PWMs/human_jaspar_hoc_kellis.PSEM -c 12 -n 4 -w 5000 -a gencode.v26.annotation.gtf -f gencode.v26.annotation.gtf -q TRUE

50kb:

bash TEPIC.sh -g hg38.noPrefix.masked.fa -b <Sample>_peaks/peaks/filtered.peaks.narrowPeak.adoptedCoverage -o <Sample>_50kb -p ../PWMs/human_jaspar_hoc_kellis.PSEM -c 12 -n 4 -w 50000 -a gencode.v26.annotation.gtf -f gencode.v26.annotation.gtf -q TRUE

gene body:

bash TEPIC.sh -g hg38.noPrefix.masked.fa -b <Sample>_peaks/peaks/filtered.peaks.narrowPeak.adoptedCoverage -o <Sample>_ -p ../PWMs/human_jaspar_hoc_kellis.PSEM -c 12 -n 4 -w 5000 -a gencode.v26.annotation.gtf -f gencode.v26.annotation.gtf -q TRUE -y

B.6.3 Details on various STITCHIT validation experiments

Integration of GeneHancer elements

Note that this approach does not depend on any size or region cut-offs. We provide a script to parse the GeneHancer tsv file into a simple tab delimited format: chr tab start tab end tab ENSG-ID tab gene name The script can be used with the command:

python rewriteGeneHancer.py <Original GeneHancer dump> ENSGIds_GeneName.txt > <Destination.txt>

Subsequently, candidate regions can be computed via:

./build/core/GENEHANCER -b <Consortium>/DNase_bw/ -a /gencode.v26.annotation.gtf -o <Consortium>_expression.txt -s data/hg38_chrSize.txt

-w 25000 -p 0.05 -g <geneID> -f GeneHancer_<Consortium> -r 500000 -k genehancer_database_06.07.2018_EnsembleMatch.sorted.bed

Merging DHS sites across samples

We used the linux cat and sort commands together with the BEDTools merge command

cat <Consortium>/*/peaks/filtered.peaks.narrowPeak.adoptedCoverage
sort -s -k1,1 -k2,2n Merged_<Consortium>.bed >
Merged_<Consortium>.sorted.bed
bedtools merge -i Merged_<Consortium>.sorted.bed >
Merged_<Consortium>.sorted.merged.bed

to generate the merged DHS sites and subsequently generated the feature matrices using:

./build/core/UNIFIED_PEAKS -b <Consortium>/DNase_wig_Normalized/ -a gencode.v26.annotation.gtf -o <Consortium>_expression.txt -s data/hg38_chrSize.txt -w 25000 -p 0.05 -g <geneID> -f UnifiedPeaks_<Consortium> -r 500000 -k Merged_<Consortium>_Peaks.sorted.merged.bed

Overlap with GeneHancer

Using BEDTools *intersect* we computed the overlap between all candidate regulatory sites identified with StitchIt and the two-level learning with all unique entries contained in the GeneHancer database that are within the searched $25kb$ search window and downstream of each gene ($193,298$ distinct regions). The same is done for regions based on the UnifiedPeaks approach, thereby assessing how many known REMs from GeneHancer can be recovered.

Overlap with non-coding mutations from the COSMIC database

The COSMIC database is a vast collection of somatic mutations occurring in cancer. We assembled a collection $\mathcal{M}_C$ containing non-coding mutations, extracted from the file *CosmicNCV.tsv.gz*. Using $\mathcal{M}_C$, we compute a length normalized score e_C describing the enrichment of mutations in the REMs as

$$e_C = \frac{O_M(\mathcal{M}_C, \mathcal{R}) \cdot O_R(\mathcal{M}_C, R)}{L_R(\mathcal{M}_C, \mathcal{R})}, \tag{B.4}$$

where $O_M(\mathcal{M}_C, R)$ is the number of mutations in $m \in \mathcal{M}_C$ overlapping a candidate REM $r \in \mathcal{R}$, $O_R(\mathcal{M}_C, R)$ is the number of regions $r \in \mathcal{R}$ overlapping a mutation $m \in \mathcal{M}_C$, and $L_R(\mathcal{M}_C, \mathcal{R})$ is the total genomic space covered by all regions $r \in \mathcal{R}$ overlapping a mutation $m \in \mathcal{M}_C$. Normalizing by $\frac{O_R(\mathcal{M}_C,R)}{L_R(\mathcal{M}_C,\mathcal{R})}$ is necessary to account for the length difference between StitchIt, UnifiedPeaks, and GeneHancer segments. This normalization factor, which we call resolution, is large, if $O_R(\mathcal{M}_C, R)$ is big, that is there are many overlapping REMs, and

$L_R(\mathcal{M}_C, \mathcal{R})$ is small, that is the covered genomic space is small. The resolution is small, if $O_R(\mathcal{M}_C, R)$ is small, that is there are only a few overlapping REMs, and $L_R(\mathcal{M}_C, \mathcal{R})$ is big, that is the covered genomic space is large. Thus, the normalization adjusts the number of retrieved mutations such that if two methods identify the same number of mutations $O_M(\mathcal{M}_C, R)$, the method with a better resolution, i.e. there are many distinct REMs covering only a small part of the genome, is preferred.

Here, the score e is computed for all REMs suggested by all methods as well as for ten randomly shuffled region sets containing the same number of regions as the original sets, respectively. The COSMIC analysis was only performed on Blueprint data due to the large number of included acute myeloid leukemia samples.

GWAS hits

We compiled a collection $\mathcal{M}_G$ comprising all GWAS sites contained in the EMBL-EBI GWAS Catalog [M$^+$16a]. Using $\mathcal{M}_G$, we compute a length normalized score e_G denoting the enrichment of GWAS hits in candidate regulatory sites as above:

$$e_G = \frac{O_M(\mathcal{M}_G, \mathcal{R}) \cdot O_R(\mathcal{M}_G, R)}{L_R(\mathcal{M}_G, \mathcal{R})}, \tag{B.5}$$

where $O_M(\mathcal{M}_G, \mathcal{R})$ is the number of mutations in $m \in \mathcal{M}_G$ overlapping a candidate REM $r \in \mathcal{R}$ and $O_R(\mathcal{M}_G, \mathcal{R})$ refers to the number of regions $r \in \mathcal{R}$ overlapping overlap a GWAS hit $m \in \mathcal{M}_G$.

eQTL analysis

We obtained all eQTLS $\mathcal{Q}$ contained in the ExSNP database ([Y$^+$16]), which we mapped to hg38 using dbSNP [S$^+$01]. To assess how many of those eQTLs overlap regulatory sites that are assigned to the same target gene as the eQTL, we compared the gene-locus assignment from all $q \in \mathcal{Q}$ with our predictions in terms of a length-normalized enrichment score e_Q:

$$e_Q = \frac{TP \cdot O_R(\mathcal{Q}, R)}{L_R(\mathcal{Q}, \mathcal{R})}, \tag{B.6}$$

where TP refers to true positives, i.e. eQTLs $q \in \mathcal{Q}$ that overlap a suggested REM that is linked to the same gene as q itself, $O_R(\mathcal{Q}, R)$ refers to the number of regions $r \in \mathcal{R}$ that overlap any eQTL site $q \in \mathcal{Q}$ and $L_R(\mathcal{Q}, \mathcal{R})$ denotes the entire genomic space covered by overlapping REMs.

ChIA-Pet & Capture Hi-C data

ChIA-Pet data for K562 and MCF-7 was downloaded from the 4DGenome database [W$^+$15] and lifted to hg38 using the UCSC liftover tool. Capture Hi-C data for GM12878 was obtained from Mifsud *et al.* [M$^+$15a] and also lifted to hg38. The

conformation data allows us to calculate how many contacts captured by the ChIA-Pet or Promoter Capture data are matching to the associations inferred by the approaches tested in this study. To match chromatin interaction data to our suggested REMs, we consider the entire gene-body of the linked gene as the second coordinates. We count a REM as contained in the conformation datasets if either the gene or the coordinate of the associated REM overlaps one coordinate of the verified interaction and the second coordinate of the interaction site overlaps the remaining coordinate of the association. Interactions that could not be detected by any of the tests, due to an exceeding genomic distance to the target gene or due to the absence of any DNaseI-seq signal in the cell line related to the sample, are excluded from consideration.

B.6.4 Generation of a CRISPR-Cas9 library for Doxorubicin resistance

We use a recently published genome-wide CRISPR perturbation library consisting of partially randomized degenerated oligonucleotides (5'-NNDNNNNNHNNNNHD-HNVVR-3') with flanking 3Cs homology regions, that was created using ssDNA of template-plasmids and site-specific mutagenesis targeting coding and non-coding regions of the human genome in hTERT-RPE1 cells from ATCC (CRL-4000) [W$^+$19b]. In brief, pooled gRNAs with oligonucleotide diversity of $7.3 \cdot 10^{10}$ targeting coding and non-coding regions of the human genome were used for a doxorubicin resistance screen in hTERT-RPE1 cells. Specifically, a total of $5.5 \cdot 10^8$ immortalized hTERT-RPE1 cells with doxycycline inducible Cas9 expression were transduced with lentiviral particles with a multiplicity of infection (MOI) of 1. The experiment was performed in three independent replicates. Cells were cultured for 7 days in standard media with $1\mu M$ doxocycline and $10\mu g/ml$ puromycin. Doxorubicin resistance selection took place from day 7 on by addition of $1\mu M$ doxorubicin. Fresh media and doxorubicin were supplemented every 4 days and cells surviving doxorubicin treatment were harvested after 3 weeks. Genomic DNA was extracted using PureLinkâĎĆ Genomic DNA Mini Kit and gRNA sequences were PCR-amplified and high-throughput sequenced on an Illumina NextSeq500 sequencer according to the manufacturer's protocol. Illumina sequencing data was processed with BCL2FASTQ v2.17 and CUTADAPT v1.15 and custom python scripts. 4232 overlapping gRNAs were found in all three replicates and experimentally validated with a new 3Cs-gRNA library (4232 gRNAs only) and a repeated CRISPR screen under established conditions (coverage 1000, MOI 0.5). An enrichment of at least two-fold after 21 days of doxorubicin treatment (compared to untreated control) was considered as a hit. 795 gRNAs were further investigated regarding target sites in the human genome using Cas-OFFinder v2.4 and GRCh38.86 with the limitation to find up to 2 mismatches [C$^+$14c, H$^+$13a, P$^+$13b]. Overall, 226 unique gRNAs could be mapped to the coding and non-coding part of the genome, resulting in 332 unique genomic target sites. In order to link putative regulatory sites detected by the gRNAs to genes, the 332 distinct genomic targets sites were extended by a window of 100bp up and downstream of the gRNA binding site. The extended windows are intersected with STITCHIT regions. All predicted non-coding interactions as well as

additional ChIA-Pet and GeneHancer support are shown in Table B.27.

B.6.5 Additional Figures and Tables

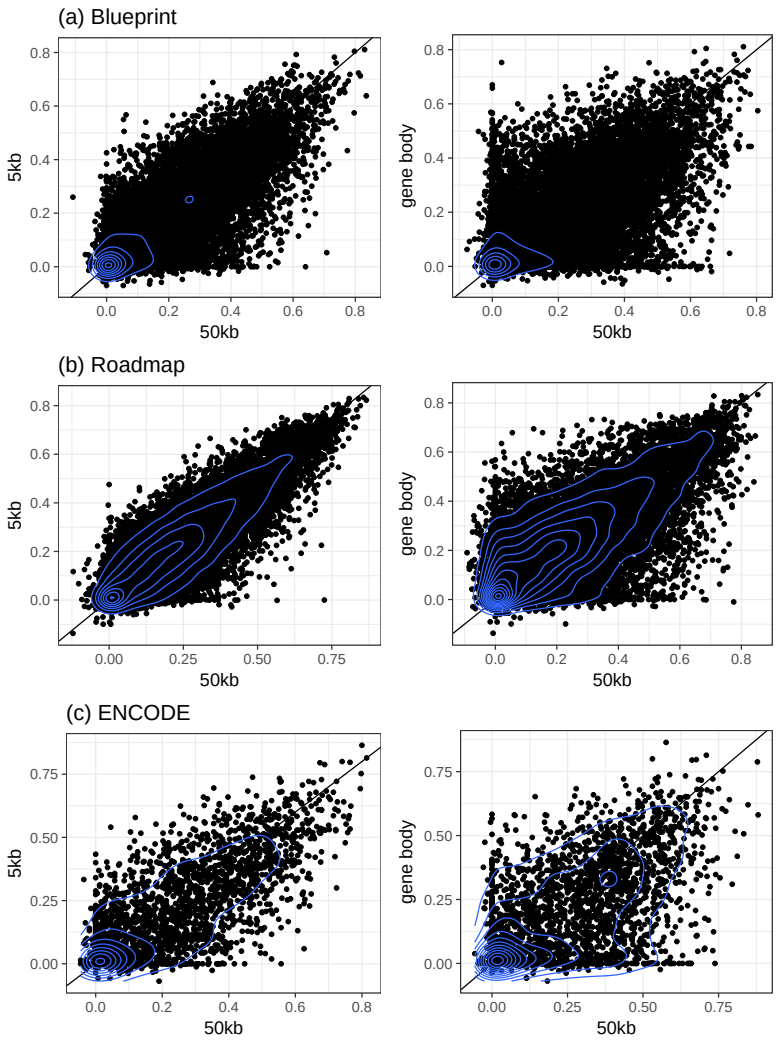


Figure B.6: Scatterplots contrasting the performance of gene-expression models using either a 5kb, a 50kb, or a gene-body window, on Blueprint (a), Roadmap (b) and ENCODE (c) data.

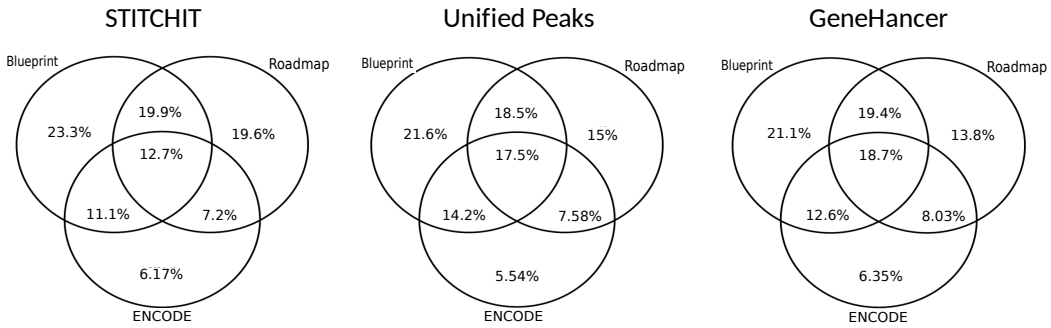


Figure B.7: Venn Diagramms indicating the overlap in terms of covered genes for STITCHIT, Unified Peaks, and GeneHancer data respectivel

Table B.24: Regression coefficients and OLS p-values for regulatory sites of *EGR1* identified with StitchIt (S1-S10) and the UnifiedPeaks (U1-U11)

chr	start	end	Reem ID	coefficient	OLS p-value
chr5	138442350	138442489	S4	0.12347378	0.13772486
chr5	138449000	138449499	S7	0.06534721	0.52094523
chr5	138450250	138450399	S10	-0.03976734	0.65534283
chr5	138454490	138456489	S9	0.061487	0.55857386
chr5	138457200	138457409	S6	0.07735484	0.40351668
chr5	138465700	138465949	S2	0.21150264	0.05986547
chr5	138469650	138470349	S10	0.45303257	0.00021482
chr5	138475650	138476489	S5	0.11132624	0.22930495
chr5	138486600	138486749	S3	-0.15254059	0.15120576
chr5	138486950	138487099	S8	-0.06192892	0.56018674
chr5	138443499	138444336	U11	-0.02633325	0.84467668
chr5	138446871	138447822	U10	0.04075988	0.76654673
chr5	138454281	138454728	U4	0.19843172	0.11272934
chr5	138454842	138455474	U7	0.07477763	0.58821766
chr5	138462649	138471619	U1	0.4100908	0.00147926
chr5	138471717	138473719	U9	-0.04813992	0.77152777
chr5	138473888	138474432	U5	0.11320651	0.45458859
chr5	138475851	138476436	U3	0.21427014	0.12540462
chr5	138476499	138477251	U6	0.10563739	0.4268543
chr5	138483081	138483386	U8	0.00735699	0.95430046
chr5	138486580	138487561	U2	-0.25307334	0.09496258

Table B.25: TFBS predictions computed using FIMO in STITCHIT REM S9 for EGR1

Motif	position	score	p-value	q-value
BHLHE40	chr5:138469650-138470349	15.3654	1.97E-06	0.00254
SP8	chr5:138469650-138470349	15.7237	2.18E-06	0.0028
BCL6	chr5:138469650-138470349	15.7097	2.22E-06	0.00302
TEAD1	chr5:138469650-138470349	13.4355	2.92E-06	0.00402
TEAD4	chr5:138469650-138470349	12.9516	2.92E-06	0.00403
STAT4	chr5:138469650-138470349	15	4.65E-06	0.00636
STAT3	chr5:138469650-138470349	14.898	5.56E-06	0.00764
BHLHE41	chr5:138469650-138470349	14.3846	1.20E-05	0.00799
BHLHE41	chr5:138469650-138470349	13.9038	1.35E-05	0.00799
STAT3	chr5:138469650-138470349	14.0816	1.25E-05	0.00857
STAT1	chr5:138469650-138470349	15.2727	6.68E-06	0.00915
E2F6	chr5:138469650-138470349	14.7759	7.24E-06	0.00918
SP3	chr5:138469650-138470349	13.7561	8.16E-06	0.0103
SOX3	chr5:138469650-138470349	14.6724	8.14E-06	0.011
HES2	chr5:138469650-138470349	12.6984	9.83E-06	0.0131
KLF16	chr5:138469650-138470349	13.7241	1.16E-05	0.0136
ID2	chr5:138469650-138470349	14.3443	1.21E-05	0.0148
TFE3	chr5:138469650-138470349	12.7347	1.48E-05	0.0174
NR5A2	chr5:138469650-138470349	13.3448	1.40E-05	0.0174
STAT1	chr5:138469650-138470349	12.4909	2.57E-05	0.0176
BACH1::MAFK	chr5:138469650-138470349	9.75	2.87E-05	0.018
BACH1::MAFK	chr5:138469650-138470349	10.2105	2.42E-05	0.018
SOX6	chr5:138469650-138470349	13.8714	1.34E-05	0.0182
KLF12	chr5:138469650-138470349	12.1475	1.42E-05	0.0182
TFAP2B(VAR.3)	chr5:138469650-138470349	12.3934	2.63E-05	0.0187
TFAP2B(VAR.3)	chr5:138469650-138470349	11.6885	3.56E-05	0.0187
TFAP2C	chr5:138469650-138470349	11.5102	5.45E-05	0.0197
TFAP2C	chr5:138469650-138470349	12.0612	3.64E-05	0.0197
TFAP2C	chr5:138469650-138470349	11.8061	4.42E-05	0.0197
TFAP2C	chr5:138469650-138470349	11.0204	7.52E-05	0.0204
NFKB2	chr5:138469650-138470349	11.2576	2.33E-05	0.0204
NFKB2	chr5:138469650-138470349	10.3636	3.25E-05	0.0204
ARNTL	chr5:138469650-138470349	13.1639	1.83E-05	0.0208
ARNTL	chr5:138469650-138470349	12.541	3.11E-05	0.0208
ZNF263	chr5:138469650-138470349	12.2857	1.64E-05	0.0211
RREB1	chr5:138469650-138470349	8.16514	1.64E-05	0.0212
KLF13	chr5:138469650-138470349	9.2931	1.60E-05	0.0213
TFAP2C	chr5:138469650-138470349	10.5714	0.000101	0.022
SP2	chr5:138469650-138470349	12.9828	1.78E-05	0.022
KLF1	chr5:138469650-138470349	13.3636	1.70E-05	0.022
TFAP2C(VAR.3)	chr5:138469650-138470349	11.1897	4.16E-05	0.0224
TFAP2C(VAR.3)	chr5:138469650-138470349	11.0517	4.41E-05	0.0224
BHLHE40	chr5:138469650-138470349	13.1538	3.53E-05	0.0227
KLF14	chr5:138469650-138470349	11.7931	1.99E-05	0.0229
TFAP2A(VAR.3)	chr5:138469650-138470349	9.65574	8.86E-05	0.0237
TFAP2A(VAR.3)	chr5:138469650-138470349	9.54098	9.20E-05	0.0237
TFAP2A(VAR.3)	chr5:138469650-138470349	11.2459	4.61E-05	0.0237
TFAP2A(VAR.3)	chr5:138469650-138470349	10.5082	6.20E-05	0.0237
FOXH1	chr5:138469650-138470349	13.6923	1.81E-05	0.0249
TFAP2B	chr5:138469650-138470349	10.8909	8.38E-05	0.0251
TFAP2B	chr5:138469650-138470349	10.7636	9.03E-05	0.0251
TFAP2B	chr5:138469650-138470349	11.7455	4.81E-05	0.0251
TFAP2B	chr5:138469650-138470349	11.7273	4.87E-05	0.0251
TFAP2B	chr5:138469650-138470349	10.3818	0.000114	0.0254
STAT5A::STAT5B	chr5:138469650-138470349	12.3878	3.71E-05	0.0255
STAT5A::STAT5B	chr5:138469650-138470349	13.1429	2.45E-05	0.0255
TFEC	chr5:138469650-138470349	12.8171	1.99E-05	0.0262
NR5A2	chr5:138469650-138470349	11.2241	4.22E-05	0.0263

SP2	chr5:138469650-138470349	11.2759	4.49E-05	0.0276
TCF7L2	chr5:138469650-138470349	13.1273	2.04E-05	0.0278
TFAP2B	chr5:138469650-138470349	9.89091	0.000151	0.028
SP4	chr5:138469650-138470349	7.56364	2.32E-05	0.0286
REL	chr5:138469650-138470349	12.8283	2.11E-05	0.0291
KLF13	chr5:138469650-138470349	6.01724	4.53E-05	0.0301
TFEB	chr5:138469650-138470349	12.7692	2.25E-05	0.0302
EGR1	chr5:138469650-138470349	9.51923	9.51E-05	0.031
EGR1	chr5:138469650-138470349	9.90385	8.01E-05	0.031
EGR1	chr5:138469650-138470349	10.9423	4.92E-05	0.031
EGR1	chr5:138469650-138470349	10.6731	5.61E-05	0.031
PPARG::RXRA	chr5:138469650-138470349	12.3153	2.48E-05	0.0313
EGR1	chr5:138469650-138470349	8.84615	0.000127	0.0329
TFAP2C	chr5:138469650-138470349	9.59184	0.000182	0.033
HEY1	chr5:138469650-138470349	11.7455	5.63E-05	0.0341
HEY1	chr5:138469650-138470349	11.4727	7.01E-05	0.0341
TFAP2A(VAR.2)	chr5:138469650-138470349	10.4	0.000115	0.0346
TFAP2A(VAR.2)	chr5:138469650-138470349	10.1636	0.000132	0.0346
TFAP2A(VAR.2)	chr5:138469650-138470349	11.6182	5.38E-05	0.0346
TFAP2A(VAR.2)	chr5:138469650-138470349	10.8727	8.68E-05	0.0346
ZNF263	chr5:138469650-138470349	8.61224	0.000111	0.0359
ZNF263	chr5:138469650-138470349	9.63265	6.77E-05	0.0359
ZNF263	chr5:138469650-138470349	8.69388	0.000107	0.0359
KLF5	chr5:138469650-138470349	10.8776	8.24E-05	0.0381
KLF5	chr5:138469650-138470349	10.3265	9.83E-05	0.0381
KLF5	chr5:138469650-138470349	11.0204	7.50E-05	0.0381
KLF5	chr5:138469650-138470349	9.69388	0.00012	0.0381
STAT4	chr5:138469650-138470349	11.6552	5.63E-05	0.0385
USF2	chr5:138469650-138470349	11.6939	5.47E-05	0.0386
USF2	chr5:138469650-138470349	11.5306	6.01E-05	0.0386
SP1	chr5:138469650-138470349	12.5	3.08E-05	0.0388
TFAP2C(VAR.3)	chr5:138469650-138470349	7.15517	0.000151	0.0394
TFAP2C(VAR.3)	chr5:138469650-138470349	7.05172	0.000155	0.0394
TFE3	chr5:138469650-138470349	11.0612	7.33E-05	0.0431
TFAP2A(VAR.2)	chr5:138469650-138470349	9.29091	0.000213	0.0443
TFAP2A(VAR.2)	chr5:138469650-138470349	8.96364	0.000252	0.0443
HEY2	chr5:138469650-138470349	10.7759	7.56E-05	0.0459
HEY2	chr5:138469650-138470349	10.7241	7.68E-05	0.0459
ZFX	chr5:138469650-138470349	10.5	7.65E-05	0.0461
ZFX	chr5:138469650-138470349	10.4545	7.78E-05	0.0461
TCFL5	chr5:138469650-138470349	9.8871	8.31E-05	0.0467
TCFL5	chr5:138469650-138470349	9.8871	8.31E-05	0.0467
TFAP2B(VAR.3)	chr5:138469650-138470349	6.91803	0.000164	0.0475
TFAP2B(VAR.3)	chr5:138469650-138470349	6.44262	0.000181	0.0475
NR5A2	chr5:138469650-138470349	8.94828	0.000115	0.0479
SP1	chr5:138469650-138470349	10.6346	7.65E-05	0.0482
HIC2	chr5:138469650-138470349	11.4182	3.72E-05	0.0483

Table B.26: TFBS predictions computed using FIMO in STITCHIT REM S3 for EGR1

motif	position	score	p-value	q-value
ZNF263	chr5:138486600-138486749	16.4694	1.18E-06	0.000244
ZNF263	chr5:138486600-138486749	14.7755	3.66E-06	0.000272
ZNF263	chr5:138486600-138486749	14.6531	3.96E-06	0.000272
ZNF263	chr5:138486600-138486749	13.0408	1.06E-05	0.000365
ZNF263	chr5:138486600-138486749	13.2041	9.64E-06	0.000365
ZNF263	chr5:138486600-138486749	13.4694	8.23E-06	0.000365
ZNF263	chr5:138486600-138486749	11.9388	2.00E-05	0.000446
ZNF263	chr5:138486600-138486749	11.9388	2.00E-05	0.000446
ZNF263	chr5:138486600-138486749	11.7959	2.16E-05	0.000446
ZNF263	chr5:138486600-138486749	12.0816	1.84E-05	0.000446
ZNF263	chr5:138486600-138486749	10.4286	4.51E-05	0.000846
ZNF263	chr5:138486600-138486749	10.2449	4.96E-05	0.000853
ZNF263	chr5:138486600-138486749	9.61224	6.84E-05	0.00102
ZNF263	chr5:138486600-138486749	9.59184	6.91E-05	0.00102
ZNF263	chr5:138486600-138486749	9.06122	8.98E-05	0.00124
ZNF263	chr5:138486600-138486749	8.08163	1.43E-04	0.00185
ZNF263	chr5:138486600-138486749	7.91837	1.54E-04	0.00188
ZNF263	chr5:138486600-138486749	7.71429	1.70E-04	0.00195
ZNF263	chr5:138486600-138486749	7.10204	2.24E-04	0.00237
ZNF263	chr5:138486600-138486749	7.04082	2.30E-04	0.00237
ZNF263	chr5:138486600-138486749	6.63265	2.75E-04	0.0027
ZNF263	chr5:138486600-138486749	6.02041	3.57E-04	0.00335
PRDM1	chr5:138486600-138486749	12.6727	1.96E-05	0.00434
ZNF263	chr5:138486600-138486749	4.87755	5.71E-04	0.00512
ZNF263	chr5:138486600-138486749	4.7551	5.99E-04	0.00516
ZNF263	chr5:138486600-138486749	4.4898	6.65E-04	0.0055
ZNF263	chr5:138486600-138486749	4.18367	7.49E-04	0.00578
ZNF263	chr5:138486600-138486749	4.16327	7.55E-04	0.00578
ZNF263	chr5:138486600-138486749	3.97959	8.10E-04	0.00598
ZNF263	chr5:138486600-138486749	3.61224	9.31E-04	0.00601
ZNF263	chr5:138486600-138486749	3.65306	9.17E-04	0.00601
ZNF263	chr5:138486600-138486749	3.71429	8.96E-04	0.00601
ZNF263	chr5:138486600-138486749	3.61224	9.31E-04	0.00601
ZNF263	chr5:138486600-138486749	3.30612	1.04E-03	0.00653
ZNF263	chr5:138486600-138486749	2.97959	1.18E-03	0.00694
ZNF263	chr5:138486600-138486749	2.97959	1.18E-03	0.00694
ZIC1	chr5:138486600-138486749	8.65957	3.09E-05	0.00812
ZNF263	chr5:138486600-138486749	2.22449	0.00154	0.00883
ZNF263	chr5:138486600-138486749	1.97959	1.68E-03	0.009
ZNF263	chr5:138486600-138486749	1.93878	1.70E-03	0.009
ZNF263	chr5:138486600-138486749	2	1.66E-03	0.009
ZNF263	chr5:138486600-138486749	1.26531	2.14E-03	0.011
IRF1	chr5:138486600-138486749	9.41935	5.63E-05	0.0118
ZIC4	chr5:138486600-138486749	11.0182	4.77E-05	0.0125
IRF1	chr5:138486600-138486749	6.98387	1.43E-04	0.0149
PRDM1	chr5:138486600-138486749	6.27273	4.05E-04	0.0158
PRDM1	chr5:138486600-138486749	8.54545	1.60E-04	0.0158
PRDM1	chr5:138486600-138486749	7.8	2.20E-04	0.0158
PRDM1	chr5:138486600-138486749	6.47273	3.75E-04	0.0158
PRDM1	chr5:138486600-138486749	5.74545	4.93E-04	0.0158
PRDM1	chr5:138486600-138486749	5.70909	5.00E-04	0.0158
ZNF263	chr5:138486600-138486749	-0.142857	3.35E-03	0.0169
NFAT5	chr5:138486600-138486749	10.9516	7.18E-05	0.0177
IRF1	chr5:138486600-138486749	5.20968	0.000264	0.0184
TFAP2C	chr5:138486600-138486749	10.7041	9.34E-05	0.0196
TFAP2C	chr5:138486600-138486749	10.0204	1.43E-04	0.0196
ZIC3	chr5:138486600-138486749	7.7	7.91E-05	0.0199
NFATC2	chr5:138486600-138486749	12.2022	7.93E-05	0.021

ZNF263	chr5:138486600-138486749	-1.06122	4.42E-03	0.0218
ZNF263	chr5:138486600-138486749	-1.18367	4.59E-03	0.022
SPZ1	chr5:138486600-138486749	9.97273	0.000149	0.0222
SPZ1	chr5:138486600-138486749	9.7	1.71E-04	0.0222
PRDM1	chr5:138486600-138486749	4.2	8.47E-04	0.0234
SP2	chr5:138486600-138486749	6.05172	3.42E-04	0.0275
SP2	chr5:138486600-138486749	6.41379	3.06E-04	0.0275
SP2	chr5:138486600-138486749	8.13793	1.71E-04	0.0275
TFAP2B	chr5:138486600-138486749	10.5455	1.03E-04	0.0285
TFAP2B	chr5:138486600-138486749	9.29091	2.10E-04	0.0289
ZNF263	chr5:138486600-138486749	-2.28571	6.27E-03	0.0294
PRDM1	chr5:138486600-138486749	3.07273	1.22E-03	0.0298
ZNF263	chr5:138486600-138486749	-2.4898	0.00663	0.0304
SPZ1	chr5:138486600-138486749	8.26364	0.000371	0.0322
SP1	chr5:138486600-138486749	4.32692	3.89E-04	0.0325
SP1	chr5:138486600-138486749	3.34615	4.53E-04	0.0325
SP1	chr5:138486600-138486749	2.38462	0.00052	0.0325
SP1	chr5:138486600-138486749	6	0.000288	0.0325
SPI1	chr5:138486600-138486749	0.381818	1.27E-04	0.0327
CDX2	chr5:138486600-138486749	10.0339	1.43E-04	0.0337
IRF1	chr5:138486600-138486749	2.29032	0.000653	0.0341
ELF3	chr5:138486600-138486749	6.37931	1.29E-04	0.0345
ZNF263	chr5:138486600-138486749	-3.06122	0.00772	0.0347
TEAD4	chr5:138486600-138486749	10.4677	1.44E-04	0.0347
SPIC	chr5:138486600-138486749	9.50704	1.46E-04	0.0378
SP2	chr5:138486600-138486749	2.58621	8.62E-04	0.0383
SP2	chr5:138486600-138486749	2.51724	0.000876	0.0383
SP2	chr5:138486600-138486749	2.15517	9.51E-04	0.0383
RARA	chr5:138486600-138486749	-14.1515	1.62E-04	0.039
EGR1	chr5:138486600-138486749	8.28846	1.59E-04	0.0392
ELF5	chr5:138486600-138486749	10.2909	1.51E-04	0.0393
EGR1	chr5:138486600-138486749	6.19231	0.000345	0.0425
FOSL2	chr5:138486600-138486749	8.41818	0.000159	0.0427
PRDM1	chr5:138486600-138486749	1.09091	2.16E-03	0.0433
PRDM1	chr5:138486600-138486749	1.21818	0.00208	0.0433
TBP	chr5:138486600-138486749	9.76389	0.00019	0.0464
IRF1	chr5:138486600-138486749	0.193548	1.17E-03	0.0488
GATA4	chr5:138486600-138486749	9.17241	1.90E-04	0.0495

Table B.27: Details to StitchIt REMs overlapping with gRNA binding sites and indications whether the interaction is supported by ChIA-Pet or GeneHancer data

chr	REM start	REM end	GeneID	coef	p-value	ChIA-Pet	GeneHancer	gRNA	start	end
chr3	147409152	147409401	ENSG00000152977	0.378326275	1.74E-005		X	CCGGCACCCATTCATCAAGG	147409223	147409423
chr11	43949847	43950306	ENSG00000244953	-0.274825376	2.95E-005			TTATAGAGTTTCCTTTGCCA	43950223	43950423
chr17	28404050	28404239	ENSG00000265254	0.481973601	0.001400099			TTTAAGAACATATTAATCGG	28404183	28404383
chr3	134378494	134378603	ENSG00000114019	0.328670581	0.002697551		X	CTGCCTGTTGCCAAAAGCAG	134378350	134378550
chr3	134378494	134378603	ENSG00000114019	0.328670581	0.002697551		X	CTGCCTGTTGCCAAGAGCAG	134378350	134378550
chr17	42093046	42093155	ENSG00000187595	0.187034942	0.002857962			TGGCAGAATTATTTGTAACA	42093081	42093281
chr17	48740854	48740893	ENSG00000244514	-0.598451082	0.004260338			CCATTTTTTCCCCCGATAGA	48740701	48740901
chr14	23008203	23008892	ENSG00000279656	-0.375730282	0.008893199	X		AATAGGGCTCCTATAACACA	23008590	23008790
chr3	114219101	114219600	ENSG00000174255	0.322885945	0.009562447			TCAAGTAAAGGTGTGCACAG	114219495	114219695
chr6	149919000	149920999	ENSG00000131015	-0.169969304	0.01740522			GCACTTGTAGCGGCGTTAAA	149920916	149921116
chr6	149230595	149230744	ENSG00000283608	-0.580879058	0.018230986			GGGAGCCTCCTTGCGCTACA	149230624	149230824
chr6	35460847	35461346	ENSG00000007866	-0.166890804	0.021030043			GTGCACATATGGGCTAGGCG	35461211	35461411
chr15	93351647	93351956	ENSG00000257060	-0.145040624	0.026855206			GCTTACTGCACCAATCACCA	93351822	93352022
chr20	45993548	45993757	ENSG00000100985	-0.167059874	0.03118279	X	X	CCTACTCCTCAATTTCCCCA	45993723	45993923
chrX	19894200	19894299	ENSG00000173681	-0.122213972	0.033450487			CCTACCCCACCAAAATCCCA	19894067	19894267
chr20	50212698	50212807	ENSG00000277449	0.271475446	0.037002049			CCTCTCGCCAGCCATCTCAA	50212667	50212867
chr10	73358503	73358702	ENSG00000233144	0.23456791	0.037403606			CGGAGGCCTCCGACTAGAAG	73358365	73358565
chr9	99202096	99202205	ENSG00000119523	-0.209406396	0.038562862	X		CATATCCTTAAGAATTAGCA	99201985	99202185
chr3	147409146	147409405	ENSG00000174963	0.377582252	0.042256958		X	CCGGCACCCATTCATCAAGG	147409223	147409423
chr9	92116298	92116447	ENSG00000232179	-0.296081695	0.057251487	X		TGGCAGAATTATTTGTAACA	92116300	92116500
chr9	130742706	130743055	ENSG00000224797	-0.159784083	0.069104178			AAATCAAGATCTTAACAAAA	130742909	130743109
chr18	31917003	31917152	ENSG00000153339	0.166626727	0.081261789			CCATCAATATATATGTCACA	31916875	31917075
chr1	212621054	212622333	ENSG00000162772	0.853205857	0.083075455	X	X	TGAATTTCCTCGAATTTGCA	212621220	212621420
chr15	48651001	48651350	ENSG00000273925	-0.164491956	0.089737634			ATTCATAAAATCAAAAAGAG	48650897	48651097
chr10	73358499	73358698	ENSG00000156042	0.53519506	0.092722936		X	CGGAGGCCTCCGACTAGAAG	73358365	73358565
chr9	92116448	92116507	ENSG00000232179	0.189450866	0.106388303	X		TGGCAGAATTATTTGTAACA	92116300	92116500
chr5	140090947	140091606	ENSG00000185129	-0.157041833	0.116161307			GCAGTAAATGTGGTGAGAGG	140091078	140091278
chr16	30020947	30021106	ENSG00000149926	0.475787434	0.120981011			AATGCATCAGCGCAGCACAG	30020911	30021111
chr10	113850401	113850600	ENSG00000196865	-0.219384309	0.1481812			CTGCCTGCTGGCTCAAAGCA	113850460	113850660
chr8	120445460	120445499	ENSG00000172167	-2.588460612	0.155609711		X	CCTGCTCCTGGTGATATGGG	120445381	120445581
chr3	186615799	186616898	ENSG00000090512	0.152998161	0.155877699			CCTTCCCCACTTCCTCGACA	186616458	186616658
chr8	120445450	120445459	ENSG00000172167	3.125332421	0.15810417			CCTGCTCCTGGTGATATGGG	120445381	120445581
chr10	89411649	89412148	ENSG00000152779	-0.426390776	0.159552393			GCGACTCTCCAAAAGCGCAA	89411571	89411771
chr17	81635351	81635400	ENSG00000182612	0.24330982	0.162623163		X	GAACCTAGCATTTATTTGAA	81635398	81635598
chr3	186616353	186616852	ENSG00000145192	0.180161833	0.164895397		X	CCTTCCCCACTTCCTCGACA	186616458	186616658
chr9	92116398	92116457	ENSG00000225511	-0.159616831	0.169251307			TGGCAGAATTATTTGTAACA	92116300	92116500
chr20	49830125	49831194	ENSG00000237788	0.188372895	0.176663252			CTATCTCCCTCTTCATCACA	49831024	49831224
chr17	65181452	65181551	ENSG00000108370	0.165611915	0.176843689	X	X	CAAGGTATTGAAGCTAAAAG	65181462	65181662
chr8	93022850	93022899	ENSG00000205133	0.195155507	0.184574517			ACTATTCCTTACAAGTAAAG	93022700	93022900
chr8	93022850	93022899	ENSG00000205133	0.195155507	0.184574517			ACTTTTCCTTACCAGTAAAG	93022700	93022900
chr8	93022850	93022899	ENSG00000205133	0.195155507	0.184574517			TCTATTCCTTACCAGTAAAG	93022700	93022900
chr12	102127943	102128052	ENSG00000075188	-0.121815411	0.187684234			TCTAAGTCCACACAGTCAAA	102127745	102127945
chr10	73358351	73358550	ENSG00000138279	0.214164144	0.190596661			CGGAGGCCTCCGACTAGAAG	73358365	73358565
chr16	30021099	30021208	ENSG00000149927	0.181801372	0.19075819	X	X	AATGCATCAGCGCAGCACAG	30020911	30021111
chr17	28404306	28404455	ENSG00000076351	0.25608548	0.1930928		X	TTTAAGAACATATTAATCGG	28404183	28404383
chr16	30021001	30021100	ENSG00000149929	-0.115797749	0.198426634			AATGCATCAGCGCAGCACAG	30020911	30021111
chr7	142929543	142929602	ENSG00000165131	-0.083241641	0.211633889			CTTCCGCCCTCGGCTTGGCA	142929574	142929774
chr8	120445350	120445449	ENSG00000172167	-1.026563442	0.225818351			CCTGCTCCTGGTGATATGGG	120445381	120445581
chr17	28404357	28404646	ENSG00000258924	0.196362209	0.260259001			TTTAAGAACATATTAATCGG	28404183	28404383
chr2	64089353	64089652	ENSG00000228079	0.255300707	0.263143276			TAAGAATCCCTCCTGATGAA	64089573	64089773
chr4	139651558	139651647	ENSG00000085871	0.120460039	0.264610385		X	CTGCCTGCCCGGTCTCCCAA	139651536	139651736
chr18	24627802	24627951	ENSG00000265485	-0.072310988	0.27634866			TATATTAATACCATATAGAG	24627862	24628062
chr7	140331753	140332692	ENSG00000157800	0.09489202	0.282261166			CTGGCCCCCAAATCAATACA	140331748	140331948
chr7	128938897	128939156	ENSG00000275106	-0.310745846	0.296485428	X		GGACCCATAACTACTCGGGG	128939109	128939309
chr17	42093194	42093353	ENSG00000108771	0.150817299	0.297490601			TGGCAGAATTATTTGTAACA	42093081	42093281
chrX	40943902	40943981	ENSG00000216866	-0.239128373	0.299841717			GCTGGGAACCTGGCTATAAA	40943875	40944075

chr20	50212804	50212903	ENSG00000172216	0.122891307	0.301708705	X		CCTCTCGCCAGCCATCTCAA	50212667	50212867
chr17	80399898	80400827	ENSG00000263069	0.092610397	0.308724059			CCTGCCGTCATTATACACCA	80400448	80400648
chr2	64089653	64089902	ENSG00000228079	0.162027634	0.315780086			TAAGAATCCCTCCTGATGAA	64089573	64089773
chr22	35728996	35729045	ENSG00000100320	-0.072352143	0.318897319			TGAATGCCAAAGGCACCAGG	35728951	35729151
chr19	11111302	11111451	ENSG00000130164	0.119137874	0.319741622	X		CCTGACGTCATTATTCACCA	11111224	11111424
chr1	202938350	202939879	ENSG00000199471	-0.095500771	0.330872399			CCGCTAGTATAAGAAAAGGG	202938430	202938630
chr1	40036653	40036742	ENSG00000131236	0.133279466	0.33534476		X	CTATCTTGCTTCCCTTCCAA	40036552	40036752
chr17	77894027	77894456	ENSG00000204283	0.063671599	0.3376392		X	CTGAGGAGTTGCTCGAGACA	77894331	77894531
chr18	3632150	3633249	ENSG00000262001	0.163875499	0.358751563		X	TAAAGCCCATAACTTCCCCA	3632948	3633148
chrX	1352046	1352755	ENSG00000185291	0.451293933	0.363103106			CGACAAACTTATCTGTGCAG	1352368	1352568
chr9	92116248	92116557	ENSG00000234537	0.119489106	0.371706339			TGGCAGAATTATTTGTAACA	92116300	92116500
chr16	4045102	4045201	ENSG00000263159	-0.186393535	0.372966401			GCTTTAGACAAAGTTCTGAA	4044989	4045189
chr13	87671045	87671194	ENSG00000165300	0.105043238	0.37437256		X	CGTGACAGCAGCATACTGAA	87671002	87671202
chr18	26135692	26136551	ENSG00000154611	-0.091288986	0.381219399			CTATAAGAAAAGCATCAACA	26136043	26136243
chr17	48740203	48741002	ENSG00000159184	0.132337474	0.382191782			CCATTTTTTCCCCCGATAGA	48740701	48740901
chr18	3632800	3632949	ENSG00000266401	0.107568527	0.391771083			TAAAGCCCATAACTTCCCCA	3632948	3633148
chr8	133596053	133596182	ENSG00000261220	-0.427016955	0.393951973			GTACTACCTGATGTGCAGGA	133595904	133596104
chr5	58872901	58873220	ENSG00000152932	0.056573482	0.406340934			CATTGACCCTTACTGTTCAA	58872898	58873098
chr3	75585049	75587048	ENSG00000272710	0.148841707	0.408538957			CATGTTTATACTGTACACAA	75585784	75585984
chr9	5402973	5403052	ENSG00000107020	0.163367021	0.409134247			CAATTTATTTCACCAATACG	5403047	5403247
chr9	92116205	92116444	ENSG00000275756	0.116715855	0.4091644			TGGCAGAATTATTTGTAACA	92116300	92116500
chr1	119894503	119894552	ENSG00000134250	-0.101041335	0.438858365			CCTTGGTCAGGTATTCCCAG	119894425	119894625
chr17	28404106	28404205	ENSG00000004139	-0.070742446	0.529379867		X	TTTAAGAACATATTAATCGG	28404183	28404383
chr7	100337602	100337851	ENSG00000214300	0.079248513	0.534001349			CCTACCCCACCAAAATCCCA	100337475	100337675
chr1	89013053	89013142	ENSG00000137944	-1.084471206	0.55306902			CCTTTAGTTCTAACAATGAA	89013039	89013239
chr5	54204945	54206704	ENSG00000185305	0.334802559	0.558465875			ACATAAAACAAACCAAACAG	54205969	54206169
chr15	48650851	48650950	ENSG00000259705	-0.050045296	0.569712888			ATTCATAAAATCAAAAAGAG	48650897	48651097
chr5	80110701	80111360	ENSG00000251675	0.057448286	0.58134695			CTGCCTGTTGCCAAAAGCAG	80110714	80110914
chr5	168344651	168344750	ENSG00000113645	0.058299594	0.59243178			CGGGTTCATATTCCAACGGG	168344628	168344828
chr3	186615655	186616594	ENSG00000283149	-0.059454239	0.612820317			CCTTCCCCACTTCCTCGACA	186616458	186616658
chr7	64682453	64683002	ENSG00000196247	-0.064595919	0.614693193	X		CAGCCTCGATAACAGAGGCG	64682966	64683166
chr14	22562545	22563054	ENSG00000129562	-0.113703126	0.624236905		X	TAAGACCCTATTCTTAAAGG	22562628	22562828
chr3	136056184	136057903	ENSG00000227267	0.076150156	0.630676323			TGAAAAGAACATCTACAGAG	136056679	136056879
chr3	120004050	120004099	ENSG00000239835	-0.079857132	0.650189715			CTGCCTGTTGCCAAGAGCAG	120003957	120004157
chr1	105967694	105968103	ENSG00000237480	-0.042757719	0.65098277			CTGCCTGTTGCCAAAAGCAG	105967567	105967767
chr6	143494051	143494100	ENSG00000001036	-0.130859432	0.656522987			CCAGGGATACAAATGCCCAG	143493996	143494196
chr7	53810509	53810748	ENSG00000205628	-0.135075699	0.659476595			ATTTATAAAATCAAAAAGAG	53810447	53810647
chr17	69512649	69513048	ENSG00000267653	-0.06683863	0.68457981			GCAGACCTTTGGTCTTCAAG	69512872	69513072
chr9	92116298	92116357	ENSG00000225511	0.050525643	0.694387813			TGGCAGAATTATTTGTAACA	92116300	92116500
chr17	42092854	42093093	ENSG00000108771	0.139988791	0.698998094			TGGCAGAATTATTTGTAACA	42093081	42093281
chrX	56546449	56546548	ENSG00000188021	0.056264059	0.732447423			TATATTGATACCATATAGAG	56546443	56546643
chr2	14517151	14517250	ENSG00000237261	-0.073140366	0.754529835			ATTTATAAAATCAAAAAGAG	14517042	14517242
chr5	134925755	134925854	ENSG00000279799	-0.388412602	0.771901659			TGGGGCCCTCTATTGCTGAG	134925638	134925838
chr3	186616546	186616855	ENSG00000275696	0.031154766	0.793696223			CCTTCCCCACTTCCTCGACA	186616458	186616658
chr17	28404146	28404255	ENSG00000076351	-0.207575338	0.809089792		X	TTTAAGAACATATTAATCGG	28404183	28404383
chr16	89746000	89747399	ENSG00000158805	-0.040492179	0.846043817			CCACAGTCATGCTCTCAGGA	89746101	89746301
chr8	61642000	61642099	ENSG00000222898	-0.018021703	0.861076186			GGAAAATAAAATCTACGGCA	61641927	61642127
chr5	134925653	134925752	ENSG00000113621	0.040834188	0.911680302			TGGGGCCCTCTATTGCTGAG	134925638	134925838
chr8	120445500	120445649	ENSG00000172167	-0.035488991	0.91619706			CCTGCTCCTGGTGATATGGG	120445381	120445581
chr7	100337297	100337776	ENSG00000201913	-0.013467722	0.936116363			CCTACCCCACCAAAATCCCA	100337475	100337675
chr13	22795106	22795195	ENSG00000237952	-0.008976307	0.940049237			CGTGACAGCAGCATACTGAA	22795183	22795383
chr10	95880848	95881147	ENSG00000270099	-0.012167852	0.946159672			CTTCCTTTACCCAATTCAGA	95880698	95880898
chr10	50601356	50601745	ENSG00000198964	0.04615091	0.946432153	X	X	GTTCCAAGCATGCTGCAAAG	50601626	50601826
chr11	90142515	90143154	ENSG00000077616	-0.003950993	0.947656188			ACATAAAACAAACCAAACAG	90142646	90142846
chr14	22562706	22563205	ENSG00000277734	0.007592207	0.956742813		X	TAAGACCCTATTCTTAAAGG	22562628	22562828

C Publications

- D Stöckel*, **F Schmidt***, P Trampert , HP Lenhof, CausalTrail: Testing hypothesis using causal Bayesian networks [version1; referees: 2 approved], F1000 Research, 2015
- HG Stunnenberg, The International Human Epigenome Consortium (including **F Schmidt**, MH Schulz), Martin Hirst, The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery, Cell, 2016
- P Durek, K Nordstrom, G Gasparoni, A Salhab, C Kressler, and M de Almeida, K Bassler, T Ulas, **F Schmidt**, J Xiong, P Glazar, F Klironomos, A Sinha, S Kinkley, X Yang, L Arrigoni, AD Amirabad, FB Ardakani, L Feuerbach, O Gorka, P. Ebert, F Müller, N Li, S Frischbutter, S Schlickeiser, C Cendon, S Frohler, B Felder, N Gasparoni, CD Imbusch, B Hutter, G Zipprich, Y Tauchmann, S Reinke, G Wassilew, U Hoffmann, AS Richter, L Sieverling, HD Chang, U Syrbe, U Kalus, J Eils, B Brors, T Manke, J Ruland, T Lengauer, N Rajewsky, W Chen, J Dong, B Sawitzki, HR Chung, P Rosenstiel, MH Schulz, JL Schultze, A Radbruch, J Walter, A Hamann, JK Polansky, Epigenomic Profiling of Human CD4+ T Cells Supports a Linear Differentiation Model and Highlights Molecular Regulators of Memory Development, Immunity, 2016
- **F Schmidt**, N Gasparoni, G Gasparoni, K Gianmoena, C Cadenas, JK Polansky, P Ebert, KJV Nordström, M Barann, A Sinha, S FrÃűhler, J Xiong, A Dehghani Amirabad, F Behjati Ardakani, B Hutter, G Zipprich, B Felder, EJ Eils, B Brors, W Chen, JG Hengstler, A Hamann, T Lengauer, P Rosenstiel, J Walter, MH Schulz, Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction, Nucleic Acids Research, 2016
- T Kehl, L Schneider, **F Schmidt**, D Stöckel, N Gerstner, C Backes, E Meese, A Keller, MH Schulz, HP Lenhof, RegulatorTrail: a web service for the identification of key transcriptional regulators, Nucleic Acids Research, 2017
- **F Schmidt***, M List*, E Cukuroglu, S Köhler, J Göke, MH Schulz, An ontology-based method for assessing batch effect adjustment approaches in heterogeneous datasets, Bioinformatics, 2018

- **F Schmidt**, MH Schulz, On the problem of confounders in modeling gene expression, Bioinformatics, 2018
- **F Schmidt**, F Kern, P Ebert, N Baumgarten, MH Schulz, TEPIC 2.0 - An extended framework for transcription factor binding prediction and integrative epigenomic analysis, Bioinformatics, 2018
- D Gérard, **F Schmidt**, A Ginolhac, M Schmitz, R Halder, P Ebert, MH Schulz, T Sauter, L Sinkkonen, Temporal epigenomic profiling identifies AHR and GLIS1 as super-enhancer controlled regulators of mesenchymal multipotency, Nucleic Acids Research, 2018
- F Behjati Ardakani*, **F Schmidt***, MH Schulz, Predicting transcription factor binding using ensemble random forest models [version1; referees: 2 approved with reservations] F1000Research, 2018
- K Nordström, **F Schmidt***, N Gasparoni*, A Salhab*, G Gasparoni, K Kattler, F Müller, P Ebert, IG Costa, DEEP consortium, N Pfeifer, T Lengauer, MH Schulz, J Walter, Unique and assay specific features of NOMe-, ATAC- and DNase1-seq data, Nucleic Acids Research, 2019
- **F Schmidt***, A Marx, M Hebel, M Wegner, N Baumgarten, M Kaulich, J Göke, J Vreeken, MH Schulz, Integrative analysis of epigenetics data identifies gene-specific regulatory elements, bioRxiv, 2019

* indicates equal contribution

Bibliography

[A+05] B. Alberts et al. *Lehrbuch der Molekularen Zellbiologie*. Wiley-VCH, 2005.

[A+08] S. Arora et al. Egr1 regulates the coordinated expression of numerous EGF receptor target genes as identified by ChIP-on-chip. *Genome Biol.*, 9(11):R166, 2008.

[A+11] E Arunan et al. Definition of the hydrogen bond (iupac recommendations 2011). *PURE AND APPLIED CHEMISTRY*, 83(8):1637–1641, 2011.

[A+12] D. Adams et al. BLUEPRINT to decode the epigenetic signature written in blood. *Nat. Biotechnol.*, 30(3):224–226, Mar 2012.

[A+14] R. Andersson et al. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461, March 2014.

[A+15a] N. Abe et al. Deconvolving the recognition of DNA shape from sequence. *Cell*, 161(2):307–318, Apr 2015.

[A+15b] B. Alipanahi et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, 33(8):831–838, Aug 2015.

[A+17a] T. H. Ambrosi et al. Adipocyte Accumulation in the Bone Marrow during Obesity and Aging Impairs Stem Cell-Based Hematopoietic and Bone Regeneration. *Cell Stem Cell*, 20(6):771–784, 06 2017.

[A+17b] F. Ardito et al. The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy (Review). *Int. J. Mol. Med.*, 40(2):271–280, Aug 2017.

[A+18] FB. Ardakani et al. Predicting transcription factor binding using ensemble random forest models[version 1; referees: 2 approved with reservations]. *F1000Res*, 7:1603, 2018.

[Aga06] P. K. Agarwal. Enzymes: An integrated view of structure, dynamics and function. *Microb. Cell Fact.*, 5:2, Jan 2006.

[AH10] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biol.*, 11(10):R106, 2010.

[Ahl02] P. Ahlquist. RNA-dependent RNA polymerases, viruses, and RNA silencing. *Science*, 296(5571):1270–1273, May 2002.

[And] S. Andrews. FastQC A Quality Control tool for High Throughput Sequence Data.

[And81] S. Anderson. Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Res.*, 9(13):3015–3027, Jul 1981.

[And05] J. T. Anderson. RNA turnover: unexpected consequences of being tailed. *Curr. Biol.*, 15(16):R635–638, Aug 2005.

[Ann08] A. Annunziato. DNA Packaging: Nucleosomes and Chromatin. *Nature Education*, 1(1):26, 2008.

[Ao98] D. L. Alexander and other. Aryl-hydrocarbon receptor is an inhibitory regulator of lipid synthesis and of commitment to adipogenesis. *J. Cell. Sci.*, 111 (Pt 22):3311–3322, Nov 1998.

[AP15] Aristotle and A. Platt. *On the Generation of Animals*. The University of Adelaide, 2015.

[AT15] B. L. Allen and D. J. Taatjes. The Mediator complex: a central integrator of transcription. *Nat. Rev. Mol. Cell Biol.*, 16(3):155–166, Mar 2015.

[AZ+12] B. Akhtar-Zaidi et al. Epigenomic enhancer profiling defines a signature of colon cancer. *Science*, 336(6082):736–739, May 2012.

[B+81] J. Banerji et al. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, 27(2 Pt 1):299–308, Dec 1981.

[B+90] C. I. Brannan et al. The product of the H19 gene may function as an RNA. *Mol. Cell. Biol.*, 10(1):28–36, Jan 1990.

[B+98] C. A. Bewley et al. Minor groove-binding architectural proteins: structure, function, and DNA recognition. *Annu Rev Biophys Biomol Struct*, 27:105–131, 1998.

[B+05a] G. Balazsi et al. Topological units of environmental signal processing in the transcriptional regulatory network of Escherichia coli. *Proc. Natl. Acad. Sci. U.S.A.*, 102(22):7841–7846, May 2005.

[B+05b] J. Bard et al. An ontology for cell types. *Genome Biol.*, 6(2):R21, 2005.

[B+06a] N. A. Balatsos et al. Inhibition of mRNA deadenylation by the nuclear cap binding complex (CBC). *J. Biol. Chem.*, 281(7):4517–4522, Feb 2006.

[B+06b] M. Beckstette et al. Fast index based algorithms and software for matching position specific scoring matrices. *BMC Bioinformatics*, 7:389, Aug 2006.

[B+06c] A. W. Bruce et al. The transcriptional repressor REST is a critical regulator of the neurosecretory phenotype. *J. Neurochem.*, 98(6):1828–1840, Sep 2006.

[B+08] A. P. Boyle et al. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322, Jan 2008.

[B+10a] Y. Barash et al. Deciphering the splicing code. *Nature*, 465(7294):53–59, May 2010.

[B+10b] B. E. Bernstein et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.*, 28(10):1045–1048, Oct 2010.

[B+13a] E. Bianconi et al. An estimation of the number of cells in the human body. *Ann. Hum. Biol.*, 40(6):463–471, 2013.

[B+13b] M.R. Breese et al. NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. *Bioinformatics (Oxford, England)*, 29(4):494–6, feb 2013.

[B+13c] V. N. Budhavarapu et al. How is epigenetic information maintained through DNA replication? *Epigenetics Chromatin*, 6(1):32, Oct 2013.

[B+13d] J. D. Buenrostro et al. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12):1213–1218, dec 2013.

[B+14a] S. Bhattacharya et al. Structural and functional insight into TAF1-TAF7, a subcomplex of transcription factor II D. *Proc. Natl. Acad. Sci. U.S.A.*, 111(25):9103–9108, Jun 2014.

[B+14b] M. J. Boland et al. Epigenetic regulation of pluripotency and differentiation. *Circ. Res.*, 115(2):311–324, Jul 2014.

[B+14c] C. Bornstein et al. A negative feedback loop of transcription factors specifies alternative dendritic cell chromatin States. *Mol. Cell*, 56(6):749–762, Dec 2014.

[B+15a] D. M. Budden et al. Predictive modelling of gene expression from transcriptional regulatory elements. *Brief. Bioinformatics*, 16(4):616–628, Jul 2015.

[B+15b] J. D. Buenrostro et al. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol*, 109:1–9, Jan 2015.

[B+16a] A. R. Barutcu et al. C-ing the Genome: A Compendium of Chromosome Conformation Capture Methods to Study Higher-Order Chromatin Organization. *J. Cell. Physiol.*, 231(1):31–35, Jan 2016.

[B+16b] S. A. Bhat et al. Long non-coding RNAs: Mechanism of action and functional utility. *Noncoding RNA Res*, 1(1):43–50, Oct 2016.

[B+16c] M. J. Borok et al. Unique functions of Gata4 in mouse liver induction and heart development. *Dev. Biol.*, 410(2):213–222, Feb 2016.

[B+16d] N. L. Bray et al. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, 34(5):525–527, 05 2016.

[B+16e] D. Bujold et al. The International Human Epigenome Consortium Data Portal. *Cell Syst*, 3(5):496–499, 11 2016.

[B+17] Daniel J Benjamin et al. Redefine statistical significance, Jul 2017.

[B+18a] C. Bessiere et al. Probing instructions for expression regulation in gene nucleotide compositions. *PLoS Comput. Biol.*, 14(1):e1005921, Jan 2018.

[B+18b] J. D. Buenrostro et al. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell*, 2018.

[Bai11] T. L. Bailey. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, 27(12):1653–1659, Jun 2011.

[Bal70] D. Baltimore. RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature*, 226(5252):1209–1211, Jun 1970.

[Bel54] R. Bellman. The theory of dynamic programming. *Bull. Amer. Math. Soc.*, 60(6):503–515, 11 1954.

[BF95] Yoshua Bengio and Paolo Frasconi. An input output hmm architecture. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, pages 427–434. MIT Press, 1995.

[BF00] A. C. Bell and G. Felsenfeld. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*, 405(6785):482–485, May 2000.

[BH72] C. Bresch and R. Hausmann. *Klassische und molekulare Genetik*. Springer, 1972.

[BH95] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*, 57(1):289–300, 1995.

[Bio18] Nature Reviews Molecular Cell Biology. Translation and protein quality control. `https://www.nature.com/collections/qmwjqzqcrb`, 2018. Accessed: 2018-12-26.

[BK98] E. M. Blackwood and J. T. Kadonaga. Going the distance: a current view of enhancer action. *Science*, 281(5373):60–63, Jul 1998.

[BK02] J. E. Butler and J. T. Kadonaga. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev.*, 16(20):2583–2592, Oct 2002.

[BK11] A. J. Bannister and T. Kouzarides. Regulation of chromatin by histone modifications. *Cell Res.*, 21(3):381–395, Mar 2011.

[Blo15] RNA-Seq Blog. RPKM, FPKM and TPM, clearly explained. `https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/`, 2015. Accessed: 2018-12-26.

[BM14] R. Barrangou and L. A. Marraffini. CRISPR-Cas systems: Prokaryotes upgrade to adaptive immunity. *Mol. Cell*, 54(2):234–244, Apr 2014.

[BM17] N. Bobola and S. Merabet. Homeodomain proteins in action: similar DNA binding preferences, highly variable connectivity. *Curr. Opin. Genet. Dev.*, 43:1–8, Apr 2017.

[BM+18] M. Ben Maamar et al. Epigenetic Transgenerational Inheritance of Altered Sperm Histone Retention Sites. *Sci Rep*, 8(1):5308, Mar 2018.

[BS04] A. M. Burger and A. K. Seth. The ubiquitin-mediated protein degradation pathway in cancer: therapeutic implications. *Eur. J. Cancer*, 40(15):2217–2229, Oct 2004.

[Bul07] M. L. Bulyk. Protein binding microarrays for the characterization of DNA-protein interactions. *Adv. Biochem. Eng. Biotechnol.*, 104:65–85, 2007.

[Bum13] R. Bumgarner. Overview of DNA microarrays: types, applications, and their future. *Curr Protoc Mol Biol*, Chapter 22:Unit 22.1., Jan 2013.

[BvH87] O. G. Berg and P. H. von Hippel. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, 193(4):723–750, Feb 1987.

[BY01] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Ann. Statistics*, 29(4):1165–1188, 2001.

[C+03] R. H. Costa et al. Transcription factors in liver development, differentiation, and regeneration. *Hepatology*, 38(6):1331–1347, Dec 2003.

[C+08] L. C. Chao et al. Inhibition of adipocyte differentiation by Nur77, Nurr1, and Nor1. *Mol. Endocrinol.*, 22(12):2596–2608, Dec 2008.

[C+09] TH. Cormen et al. *Introduction To Algorithms (3ed)*. MIT Press, 2009.

[C+10a] S. Cagnol et al. ERK and cell death: Mechanisms of ERK-induced cell death-apoptosis, autophagy and senescence. *The FEBS Journal*, 277(1):2–21, 2010.

[C+10b] M. P. Creyghton et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U.S.A.*, 107(50):21931–21936, Dec 2010.

[C+11a] C. Chen et al. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS One*, 6(2):e17238, February 2011.

[C+11b] I. G. Costa et al. Predicting gene expression in T cell differentiation from histone modifications and transcription factor binding affinities by linear mixture models. *BMC Bioinformatics*, 12 Suppl 1:S29, Feb 2011.

[C+12a] C. Cheng et al. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome research*, 22(9):1658–1667, 2012.

[C+12b] C. Cheng et al. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res.*, 22(9):1658–1667, Sep 2012.

[C+12c] M. Ciofani et al. A validated regulatory network for Th17 cell specification. *Cell*, 151(2):289–303, Oct 2012.

[C+14a] E. I. Campos et al. Epigenetic inheritance: histone bookmarks across generations. *Trends Cell Biol.*, 24(11):664–674, Nov 2014.

[C+14b] W. P. Cawthorn et al. Bone marrow adipose tissue is an endocrine organ that contributes to increased circulating adiponectin during caloric restriction. *Cell Metab.*, 20(2):368–375, Aug 2014.

[C+14c] SW. Cho et al. Analysis of off-target effects of crispr/cas-derived rna-guided endonucleases and nickase. *Genome Res*, 24:132–141, 2014.

[C+15a] K. Chen et al. The Overlooked Fact: Fundamental Need for Spike-In Control for Virtually All Genome-Wide Analyses. *Mol. Cell. Biol.*, 36(5):662–667, Dec 2015.

[C+15b] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[C+15c] GTEx Consortium et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[C+15d] J. Crocker et al. Low affinity binding site clusters confer hox specificity and regulatory robustness. *Cell*, 160(1-2):191–203, Jan 2015.

[C+16a] T. P. Chiu et al. DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics*, 32(8):1211–1213, 04 2016.

[C+16b] S. R. Choudhury et al. CRISPR-dCas9 mediated TET1 targeting for selective DNA demethylation at BRCA1 promoter. *Oncotarget*, 7(29):46545–46556, Jul 2016.

[C+16c] A. Conesa et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.*, 17:13, Jan 2016.

[C+17] Q. Cao et al. Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. *Nature Genetics*, 49(10):1428–1436, September 2017.

[C+18a] J. Cao et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, 361(6409):1380–1385, 2018.

[C+18b] M. R. Corces et al. The chromatin accessibility landscape of primary human cancers. *Science*, 362(6413):eaav1898, 2018.

[CF10] S. W. Choi and S. Friso. Epigenetics: A New Bridge between Nutrition and Health. *Adv Nutr*, 1(1):8–16, Nov 2010.

[Cle04] C. V. Clevenger. Roles and regulation of stat family transcription factors in human breast cancer. *Am. J. Pathol.*, 165(5):1449–1460, Nov 2004.

[Com18] Wikipedia Commons. Aminoacids_table.svg. `https://de.wikipedia.org/wiki/Datei:Aminoacids_table.svg#/media/File:Aminoacids_table.svg`, 2018. Accessed: 2018-12-25.

[Con18] The DEEP Consortium. Welcome to DEEP? `http://www.deutsches-epigenom-programm.de/`, 2018. Accessed: 2019-01-02.

[CP+12] G. Cuellar-Partida et al. Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics*, 28(1):56–62, Jan 2012.

[Cri66] F. H. Crick. Codon–anticodon pairing: the wobble hypothesis. *J. Mol. Biol.*, 19(2):548–555, Aug 1966.

[Cri70] F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, Aug 1970.

[CS14] T. R. Cech and J. A. Steitz. The noncoding RNA revolution-trashing old rules to forge new ones. *Cell*, 157(1):77–94, Mar 2014.

[CZ10] K.-B. Chen and Y. Zhang. A varying threshold method for ChIP peak-calling using multiple sources of information. *Bioinformatics*, 26(18):i504–i510, 09 2010.

[D+75] SUSAN J. DEVLIN et al. Robust estimation and outlier detection with correlation coefficients. *Biometrika*, 62(3):531–545, 1975.

[D+02] J. Dekker et al. Capturing chromosome conformation. *Science*, 295(5558):1306–1311, Feb 2002.

[D+03] B. Dorigo et al. Chromatin fiber folding: requirement for the histone H4 N-terminal tail. *J. Mol. Biol.*, 327(1):85–96, Mar 2003.

[D+12a] A. Diaz et al. Normalization, bias correction, and peak calling for ChIP-seq. *Statistical applications in genetics and molecular biology*, 11:Article 9, 01 2012.

[D+12b] I. Dunham et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012.

[D+14] N. H. Dryden et al. Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Res.*, 24(11):1854–1868, Nov 2014.

[D+16a] K. Diamanti et al. Maps of context-dependent putative regulatory regions and genomic signal interactions. *Nucleic Acids Res.*, 44(19):9110–9120, Nov 2016.

[D+16b] A. D. Diehl et al. The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J Biomed Semantics*, 7(1):44, 07 2016.

[D+16c] J. R. Dixon et al. Chromatin Domains: The Unit of Chromosome Organization. *Mol. Cell*, 62(5):668–680, 06 2016.

[D+16d] A. Drazic et al. The world of protein acetylation. *Biochim. Biophys. Acta*, 1864(10):1372–1401, 10 2016.

[D+16e] P. Durek et al. Epigenomic Profiling of Human CD4+ T Cells Supports a Linear Differentiation Model and Highlights Molecular Regulators of Memory Development. *Immunity*, 45(5):1148–1161, 11 2016.

[D+17] H. Dana et al. Molecular Mechanisms and Biological Functions of siRNA. *Int J Biomed Sci*, 13(2):48–57, Jun 2017.

[D+18] J. Ding et al. iDREM: Interactive visualization of dynamic regulatory networks. *PLoS Comput. Biol.*, 14(3):e1006019, 03 2018.

[Dav11] E.H. Davidson. *Introduction to Proteins: Structure, Function, and Motion*. Chapman and Hall/CRC Mathematical and Computational Biology, 2011.

[Die98] A. M. Diehl. Roles of CCAAT/enhancer-binding proteins in regulation of liver regenerative growth. *J. Biol. Chem.*, 273(47):30843–30846, Nov 1998.

[dLG03] W. de Laat and F. Grosveld. Spatial organization of gene expression: the active chromatin hub. *Chromosome Res.*, 11(5):447–459, 2003.

[DM92] W. H. Day and F. R. McMorris. Critical comparison of consensus methods for molecular sequences. *Nucleic Acids Res.*, 20(5):1093–1099, Mar 1992.

[Doe18] J. G. Doench. Am I ready for CRISPR? A user's guide to genetic screens. *Nat. Rev. Genet.*, 19(2):67–80, Feb 2018.

[E+04a] Bradley Efron et al. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

[E+04b] S. Eyquem et al. The Ets-1 transcription factor is required for complete pre-T cell receptor function and allelic exclusion at the T cell receptor beta locus. *Proc. Natl. Acad. Sci. U.S.A.*, 101(44):15712–15717, Nov 2004.

[E+05] E. S. Emison et al. A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. *Nature*, 434(7035):857–863, Apr 2005.

[E06] Terzi E. *Problems and algorithms for sequence segmentations*. Uninversity of Helsinki, 2006.

[E+07] J. Ernst et al. Reconstructing dynamic regulatory maps. *Mol. Syst. Biol.*, 3:74, 2007.

[E+11] J. Ernst et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, March 2011.

[E+14a] R. Eggeling et al. On the value of intra-motif dependencies of human insulator protein CTCF. *PLoS ONE*, 9(1):e85629, 2014.

[E+14b] J. Erler et al. The role of histone tails in the nucleosome: a computational study. *Biophys. J.*, 107(12):2911–2922, Dec 2014.

[E+17] J. R. Edwards et al. DNA methylation and DNA methyltransferases. *Epigenetics Chromatin*, 10:23, 2017.

[Edu14] Nature Education. SNP. `https://www.nature.com/scitable/definition/single-nucleotide-polymorphism-snp-295`, 2014. Accessed: 2018-12-25.

[EK12] J. Ernst and M. Kellis. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, 9(3):215–216, Feb 2012.

[Erl03] L. Erlbau. *Research Design and Statistical Analysis (2nd ed.)*. Routledge, 2003.

[ES90] A. D. Ellington and J. W. Szostak. In vitro selection of RNA molecules that bind specific ligands. *Nature*, 346(6287):818–822, Aug 1990.

[F+92] M. Frommer et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U.S.A.*, 89(5):1827–1831, Mar 1992.

[F+04] M. C. Frith et al. Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res.*, 32(4):1372–1381, 2004.

[F+09] M. J. Fullwood et al. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, 462(7269):58–64, Nov 2009.

[F+10a] C. E. Forristal et al. Hypoxia inducible factors regulate pluripotency and proliferation in human embryonic stem cells cultured at reduced oxygen tensions. *Reproduction*, 139(1):85–97, Jan 2010.

[F+10b] J. Friedman et al. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*, 33(1):1–22, 2010.

[F+17a] S. A. Forbes et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.*, 45(D1):D777–D783, 01 2017.

[F+17b] X. Fu et al. Genomic and molecular control of cell type and cell type conversions. *Cell Regen (Lond)*, 6:1–7, Dec 2017.

[Far06] S. R. Farmer. Transcriptional control of adipocyte formation. *Cell Metab.*, 4(4):263–273, Oct 2006.

[FHT10] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[For73] G. D. Forney. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, March 1973.

[FR02] C. Fraley and A.E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 2002.

[Fra08] T. F. Franke. Intracellular signaling by Akt: bound to be specific. *Sci Signal*, 1(24):pe29, Jun 2008.

[FT+09] K. Fejes-Toth et al. Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature*, 457(7232):1028–1032, Feb 2009.

[Fur12] T. S. Furey. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat. Rev. Genet.*, 13(12):840–852, Dec 2012.

[G+00] M. Gao et al. Interaction between a poly(A)-specific ribonuclease and the 5' cap influences mRNA deadenylation rates in vitro. *Mol. Cell*, 5(3):479–488, Mar 2000.

[G+04] P. H. Giangrande et al. A role for E2F6 in distinguishing G1/S- and G2/M-specific transcription. *Genes Dev.*, 18(23):2941–2951, Dec 2004.

[G+11] C. E. Grant et al. FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, Apr 2011.

[G+12a] Gregor D. Gilfillan et al. Limitations and possibilities of low cell number chip-seq. *BMC Genomics*, 13(1):645, Nov 2012.

[G+12b] S. Guibert et al. CTCF-binding sites within the H19 ICR differentially regulate local chromatin structures and cis-acting functions. *Epigenetics*, 7(4):361–369, Apr 2012.

[G+13] W. Guo et al. BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. *BMC Genomics*, 14:774, Nov 2013.

[G+14a] A. Ghaffarizadeh et al. Modeling and visualizing cell type switching. *Computational and mathematical methods in medicine*, 2014:293980, 04 2014.

[G+14b] N. Gosalia et al. Architectural proteins CTCF and cohesin have distinct roles in modulating the higher order structure and expression of the CFTR locus. *Nucleic Acids Res.*, 42(15):9612–9622, Sep 2014.

[G+14c] S. Groschel et al. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell*, 157(2):369–381, Apr 2014.

[G+14d] E. G. Gusmao et al. Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. *Bioinformatics*, 30(22):3143–3151, Nov 2014.

[G+15a] E. Giorgio et al. A large genomic deletion leads to enhancer adoption by the lamin B1 gene: a second path to autosomal dominant adult-onset demyelinating leukodystrophy (ADLD). *Hum. Mol. Genet.*, 24(11):3143–3154, Jun 2015.

[G+15b] A. J. Gonzalez et al. Early enhancer establishment and regulatory locus complexity shape transcriptional programs in hematopoietic differentiation. *Nat. Genet.*, 47(11):1249–1259, Nov 2015.

[G+15c] J. Grau et al. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics*, 31(15):2595–2597, Aug 2015.

[G+16a] E. G. Gusmao et al. Analysis of computational footprinting methods for DNase sequencing experiments. *Nat. Methods*, 13(4):303–309, Apr 2016.

[G+16b] E.G. Gusmao et al. Analysis of computational footprinting methods for DNase sequencing experiments. *Nature Methods*, 13(4):303–309, Apr 2016.

[G+17a] M. Gialitakis et al. Activation of the Aryl Hydrocarbon Receptor Interferes with Early Embryonic Development. *Stem Cell Reports*, 9(5):1377–1386, 11 2017.

[G+17b] W.W.B. Goh et al. Why Batch Effects Matter in Omics Data, and How to Avoid Them. *Trends Biotechnol.*, 35(6):498–507, June 2017.

[G+17c] K. D. N. Gomes et al. Doxycycline induces bone repair and changes in Wnt signalling. *Int J Oral Sci*, 9(3):158–166, 09 2017.

[GB18] N. R. Genuth and M. Barna. Heterogeneity and specialized functions of translation machinery: from genes to organisms. *Nat. Rev. Genet.*, 19(7):431–452, Jul 2018.

[GBS12a] J. A. Gagnon-Bartsch and T. P. Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, Jul 2012.

[GBS12b] J.A. Gagnon-Bartsch and T.P. Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, July 2012.

[GKN05] Emden R. Gansner, Yehuda Koren, and Stephen North. Graph drawing by stress majorization. In János Pach, editor, *Graph Drawing*, pages 239–250, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[GP03] Garret and G. Parmigiani. *POE: Statistical Methods for Qualitative Analysis of gene-expression*, chapter 16, pages 362–387. Springer, 2003.

[GPGK13] J. Grau, S. Posch, I. Grosse, and J. Keilwagen. A general approach for discriminative de novo motif discovery from high-throughput data. *Nucleic Acids Res.*, 41(21):e197, Nov 2013.

[Gru07a] P.D. Gruenwald. *The Minimum Description Length Principle.* MIT Press, 2007.

[Grü07b] P. D. Grünwald. *The minimum description length principle.* MIT press, 2007.

[GS17] J. D. Gessaman and E. U. Selker. Induction of H3K9me3 and DNA methylation by tethered heterochromatin factors in Neurospora crassa. *Proc. Natl. Acad. Sci. U.S.A.*, 114(45):E9598–E9607, 11 2017.

[GSo18] D. Gerard, F. Schmidt, and others. Temporal enhancer profiling of parallel lineages identifies AHR and GLIS1 as regulators of mesenchymal multipotency. *Nucleic Acids Res.*, Dec 2018.

[H+99] T. Hai et al. ATF3 and Stress Responses. *Gene Expr*, 7(4-5-6):321–335, 1999.

[H+04] M. Q. Hassan et al. Dlx3 transcriptional regulation of osteoblast differentiation: temporal recruitment of Msx2, Dlx3, and Dlx5 homeodomain proteins to chromatin of the osteocalcin gene. *Mol. Cell. Biol.*, 24(20):9248–9261, Oct 2004.

[H+05] S. C. Ha et al. Crystal structure of a junction between B-DNA and Z-DNA reveals two extruded bases. *Nature*, 437(7062):1183–1186, Oct 2005.

[H+06] T. Hastie et al. *The Elements of Statistical Learning*. Springer, 2006.

[H+07] N. D. Heintzman et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, 39(3):311–318, Mar 2007.

[H+08] T. Horie et al. TG-interacting factor is required for the differentiation of preadipocytes. *J. Lipid Res.*, 49(6):1224–1234, Jun 2008.

[H+09] G. C. Hon et al. Predictive chromatin signatures in the mammalian genome. *Hum. Mol. Genet.*, 18(R2):195–201, Oct 2009.

[H+10] S. Heinz et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, 38(4):576–589, May 2010.

[H+12a] J. Harrow et al. Gencode: The reference human genome annotation for the encode project. *Genome Research*, 22(9):1760–1774, 2012.

[H+12b] J. W. Hershey et al. Principles of translational control: an overview. *Cold Spring Harb Perspect Biol*, 4(12), Dec 2012.

[H+12c] E. Hervouet et al. Kinetics of DNA methylation inheritance by the Dnmt1-including complexes during the cell cycle. *Cell Div*, 7:5, Feb 2012.

[H+13a] PD. Hsu et al. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol*, 31:827–832, 2013.

[H+13b] M. R. Hubner et al. Chromatin organization and transcriptional regulation. *Curr. Opin. Genet. Dev.*, 23(2):89–95, Apr 2013.

[H+14a] A. M. Hakelien et al. The regulatory landscape of osteogenic differentiation. *Stem Cells*, 32(10):2780–2793, Oct 2014.

[H+14b] D. J. Hazelett et al. Comprehensive functional annotation of 77 prostate cancer risk loci. *PLoS Genet.*, 10(1):e1004102, Jan 2014.

[H+14c] H. H. He et al. Refined dnase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nature methods*, 11(1):73, 2014.

[H+14d] Y. He et al. Long noncoding RNAs: Novel insights into hepatocelluar carcinoma. *Cancer Lett.*, 344(1):20–27, Mar 2014.

[H+14e] S. Heerboth et al. Use of epigenetic drugs in disease: an overview. *Genet Epigenet*, 6:9–19, 2014.

[H+14f] H. M. Herz et al. Enhancer malfunction in cancer. *Mol. Cell*, 53(6):859–866, Mar 2014.

[H+15a] I. B. Hilton et al. Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat. Biotechnol.*, 33(5):510–517, May 2015.

[H+15b] M. A. Hume et al. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, 43(Database issue):D117–122, Jan 2015.

[H+17] A. S. Hansen et al. CTCF and cohesin regulate chromatin loop stability with distinct dynamics. *Elife*, 6, 05 2017.

[H+18a] L. Haghverdi et al. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, 36(5):421–427, June 2018.

[H+18b] T. A. Hait et al. Focs: a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer–promoter map. *Genome Biology*, 19(1):56, May 2018.

[H+18c] B. Hwang et al. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.*, 50(8):96, Aug 2018.

[Har15] D. Harte. *HiddenMarkov: Hidden Markov Models*. Statistics Research Associates, Wellington, 2015.

[Hay13] W. Haynes. *Bonferroni Correction*, pages 154–154. Springer New York, New York, NY, 2013.

[HC16] J. M. Heather and B. Chain. The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1):1–8, Jan 2016.

[HdL13] S. J. Holwerda and W. de Laat. CTCF: the protein, the binding partners, the binding sites and their chromatin loops. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 368(1620):20120369, 2013.

[HH05] Z. Hui and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[HL13] N. Harmston and B. Lenhard. Chromatin and epigenetic features of long-range gene regulation. *Nucleic Acids Res.*, 41(15):7185–7199, Aug 2013.

[HL17] T. R. Hughes and S. A. Lambert. Transcription factors read epigenetics. *Science*, 356(6337):489–490, 05 2017.

[HM15] Mohammad Hossin and Sulaiman M.N. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining and Knowledge Management Process*, 5:01–11, 03 2015.

[Hol94] R. Holliday. Epigenetics: an overview. *Dev. Genet.*, 15(6):453–457, 1994.

[HT11] T. M. Hardy and T. O. Tollefsbol. Epigenetic diet: impact on the epigenome and cancer. *Epigenomics*, 3(4):503–518, Aug 2011.

[HT17] Z. Hu and W. W. Tee. Enhancers and chromatin structures: regulatory hubs in gene expression and diseases. *Biosci. Rep.*, 37(2), 04 2017.

[Hun00] T. Hunter. Signaling - 2000 and beyond. *Cell*, 100(1):113–127, Jul 2000.

[I+06] K. Iwata et al. Bisphosphonates suppress periosteal osteoblast activity independently of resorption in rat femur and tibia. *Bone*, 39(5):1053–1058, Nov 2006.

[I+10] M. Ieda et al. Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. *Cell*, 142(3):375–386, Aug 2010.

[I+11] S. Ito et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science*, 333(6047):1300–1303, Sep 2011.

[I+15] M. M. Ibrahim et al. JAMM: a peak finder for joint analysis of NGS replicates. *Bioinformatics*, 31(1):48–55, Jan 2015.

[IDZ16] M. Iwafuchi-Doi and K. S. Zaret. Cell fate control by pioneer transcription factors. *Development*, 143(11):1833–1837, 06 2016.

[Inc16] Illumina Inc. An introduction to Next-Generation Sequencing Technology. `https://www.illumina.com/documents/products/illumina_sequencing_introduction.pdf`, 2016. Accessed: 2018-12-27.

[Inc17] Illumina Inc. HiSeq X Series of Sequencing Systems. `https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet-hiseq-x-ten.pdf`, 2017. Accessed: 2018-12-27.

[Ins18] Broad Institute. Picard toolkit. `http://broadinstitute.github.io/picard/`, 2018. Accessed: 2018-12-27.

[J+07] W. E. Johnson et al. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, Jan 2007.

[J+10] A. Jolma et al. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.*, 20(6):861–873, Jun 2010.

[J+14] S. Jain et al. Multitask learning of signaling and regulatory networks with application to studying human response to flu. *PLoS Comput. Biol.*, 10(12):e1003943, Dec 2014.

[J+15] W. Jin et al. Genome-wide detection of dnase i hypersensitive sites in single cells and ffpe tissue samples. *Nature*, 528(7580):142, 2015.

[J+16a] L. Jacob et al. Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. *Biostatistics*, 17(1):16–28, Jan 2016.

[J+16b] K. Jamieson et al. Loss of HP1 causes depletion of H3K27me3 from facultative heterochromatin and gain of H3K27me2 at constitutive heterochromatin. *Genome Res.*, 26(1):97–107, Jan 2016.

[J+16c] N. Jayaram et al. Evaluating tools for transcription factor binding site prediction. *BMC Bioinformatics*, Nov 2016.

[Joh09] W. Johannsen. Elemente der exakten Erblichkeitslehre. *Gustav Fischer, Jena*, 1909.

[K+83] D. Kioussis et al. Beta-globin gene inactivation by DNA translocation in gamma beta-thalassaemia. *Nature*, 306(5944):662–666, 1983.

[K+05] E. Kaminskas et al. FDA drug approval summary: azacitidine (5-azacytidine, Vidaza) for injectable suspension. *Oncologist*, 10(3):176–182, Mar 2005.

[K+07] T. H. Kim et al. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*, 128(6):1231–1245, Mar 2007.

[K+10a] M. H. Kagey et al. Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314):430–435, Sep 2010.

[K+10b] T. K. Kim et al. Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295):182–187, May 2010.

[K+11] P. F. Kuan et al. A Statistical Framework for the Analysis of ChIP-Seq Data. *J Am Stat Assoc*, 106(495):891–903, 2011.

[K+12] T. K. Kelly et al. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res.*, 22(12):2497–2506, Dec 2012.

[K+13a] D. Kim et al. Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14(4):R36, 2013.

[K+13b] H. Koohy et al. Chromatin accessibility data sets show bias due to sequence specificity of the DNase I enzyme. *PLoS ONE*, 8(7):e69853, 2013.

[K+13c] H. Koohy et al. Chromatin accessibility data sets show bias due to sequence specificity of the dnase i enzyme. *PloS one*, 8(7):e69853, 2013.

[K+14a] S. Kelaini et al. Direct reprogramming of adult cells: avoiding the pluripotent state. *Stem Cells Cloning*, 7:19–29, 2014.

[K+14b] H. Koohy et al. A comparison of peak callers used for DNase-Seq data. *PLoS ONE*, 9(5):e96303, 2014.

[K+14c] H. Koohy et al. A comparison of peak callers used for DNase-Seq data. *PloS one*, 9(5):e96303, jan 2014.

[K+15] A. Kundaje et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, Feb 2015.

[K+17a] T. Kehl et al. RegulatorTrail: a web service for the identification of key transcriptional regulators. *Nucleic Acids Res.*, 45(W1):W146–W153, Jul 2017.

[K+17b] Tyler S. Klann et al. CRISPR-Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nature Biotechnology*, 35(6):561–568, April 2017.

[K+18a] C. Karabacak et al. Reproducible inference of transcription factor footprints in atac-seq and dnase-seq datasets via protocol-specific bias modeling. *bioRxiv*, 2018.

[K+18b] T. Kehl et al. REGGAE: a novel approach for the identification of key transcriptional regulators. *Bioinformatics*, 34(20):3503–3510, Oct 2018.

[K+18c] A. Khan et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.*, 46(D1):D260–D266, Jan 2018.

[K+18d] I. V. Kulakovskiy et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.*, 46(D1):D252–D259, Jan 2018.

[K+19] J. Keilwagen et al. Accurate prediction of cell type-specific transcription factor binding. *Genome Biol.*, 20(1):9, 01 2019.

[Ker16] F. Kern. *Integrative prediction of gene expression with open chromatin and Hi-C data.* Saarland University, Center for Bioinformatics, 2016.

[KG15] J. Keilwagen and J. Grau. Varying levels of complexity in transcription factor binding motifs. *Nucleic Acids Res.*, 43(18):e119, Oct 2015.

[KK14] P. Kheradpour and M. Kellis. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.*, 42(5):2976–2987, Mar 2014.

[KL15] J. Kahara and H. Lahdesmaki. BinDNase: a discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data. *Bioinformatics*, 31(17):2852–2859, Sep 2015.

[KM15] K. R. Kukurba and S. B. Montgomery. RNA Sequencing and Analysis. *Cold Spring Harb Protoc*, 2015(11):951–969, Apr 2015.

[Kni06] R. Knippers. *Molekulare Genetik.* Thieme, 2006.

[KO00] A. Komeili and E. K. O'Shea. Nuclear transport and transcription. *Curr. Opin. Cell Biol.*, 12(3):355–360, Jun 2000.

[Ko12] K. Kawabe and others. IL-12 inhibits glucocorticoid-induced T cell apoptosis by inducing GMEB1 and activating PI3K/Akt pathway. *Immunobiology*, 217(1):118–123, Jan 2012.

[Kol68] A. N. Kolmogorov. Three approaches to the quantitative definition of information. *International journal of computer mathematics*, 2:157–168, 1968.

[L+00a] H. Lodish et al. *Molecular Cell Biology. 4th edition, Section 1.2, The Molecules of Life.* W. H. Freeman, 2000.

[L+00b] S. C. Lu et al. Changes in methionine adenosyltransferase and S-adenosylmethionine homeostasis in alcoholic rat liver. *Am. J. Physiol. Gastrointest. Liver Physiol.*, 279(1):G178–185, Jul 2000.

[L+01] E. S. Lander et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001.

[L+07] N. K. Lee et al. Endocrine regulation of energy metabolism by the skeleton. *Cell*, 130(3):456–469, Aug 2007.

[L+08] X. Liu et al. Yamanaka factors critically regulate the developmental signaling network in mouse embryonic stem cells. *Cell Res.*, 18(12):1177–1189, Dec 2008.

[L+09] H. Li et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Aug 2009.

[L+10a] J. S. Lee et al. The language of histone crosstalk. *Cell*, 142(5):682–685, Sep 2010.

[L+10b] J.T. Leek et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.*, 11(10):733–739, October 2010.

[L+10c] J Luo et al. A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. *Pharmacogenomics J.*, 10(4):278–291, August 2010.

[L+11] F. Li et al. Coordination of DNA replication and histone modification by the Rik1-Dos2 complex. *Nature*, 475(7355):244–248, Jul 2011.

[L+12a] C. Lanzuolo et al. Concerted epigenetic signatures inheritance at PcG targets through replication. *Cell Cycle*, 11(7):1296–1300, Apr 2012.

[L+12b] Y. Liu et al. Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biol.*, 13(7):R61, Jul 2012.

[L+12c] Y. Liu et al. Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome biology*, 13(7):R61, jan 2012.

[L+13a] M. Lawrence et al. Software for computing and annotating genomic ranges. *PLoS computational biology*, 9(8):e1003118, 2013.

[L+13b] C. Lazar et al. Batch effect removal methods for microarray gene expression data integration: a survey. *Brief. Bioinform.*, 14(4):469–490, July 2013.

[L+13c] A. Lazarovici et al. Probing dna shape and methylation state on a genomic scale with dnase i. *Proceedings of the National Academy of Sciences*, 110(16):6376–6381, 2013.

[L+13d] J. Lonsdale et al. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, 45(6):580–585, Jun 2013.

[L+14a] B. D. Landry et al. Regulation of a transcription factor network by Cdk1 coordinates late cell cycle gene expression. *EMBO J.*, 33(9):1044–1060, May 2014.

[L+14b] S. Lanouette et al. The functional diversity of protein lysine methylation. *Mol. Syst. Biol.*, 10:724, Apr 2014.

[L+14c] M. I. Love et al. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, 15(12):550, 2014.

[L+17a] S. Y. Lee et al. Glis family proteins are differentially implicated in the cellular reprogramming of human somatic cells. *Oncotarget*, 8(44):77041–77049, Sep 2017.

[L+17b] J. Lever et al. Principal component analysis. *Nature Methods Online*, 14:641–42, 2017.

[L+17c] Y. Liu et al. Transcriptional landscape of the human cell cycle. *Proceedings of the National Academy of Sciences*, 114(13):3473–3478, 2017.

[L+17d] B. Lv et al. Hypoxia inducible factor 1alpha promotes survival of mesenchymal stem cells under hypoxia. *Am J Transl Res*, 9(3):1521–1529, 2017.

[L+18a] Zhijian Li et al. Identification of transcription factor binding sites using atac-seq. *bioRxiv*, 2018.

[L+18b] Y. Liu et al. The nuclear transportation routes of membrane-bound transcription factors. *Cell Commun. Signal*, 16(1):12, Apr 2018.

[LA+09] E. Lieberman-Aiden et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, Oct 2009.

[LA+14] D. Lara-Astiaso et al. Immunogenetics. Chromatin state dynamics during blood formation. *Science*, 345(6199):943–949, Aug 2014.

[LB13] J. P. Lim and A. Brunet. Bridging the transgenerational gap with epigenetic memory. *Trends Genet.*, 29(3):176–186, Mar 2013.

[LCHG15] Jeffery Li, Travers Ching, Sijia Huang, and Lana X. Garmire. Using epigenomics data to predict gene expression in lung cancer. *BMC Bioinformatics*, 16(5):S10, Mar 2015.

[LD09] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, Jul 2009.

[LH13] K. Luo and A. J. Hartemink. Using DNase digestion data to accurately identify transcription factor binding sites. *Pac Symp Biocomput*, pages 80–91, 2013.

[LI97] J. D. Lewis and E. Izaurralde. The role of the cap structure in RNA processing and nuclear export. *Eur. J. Biochem.*, 247(2):461–469, Jul 1997.

[Lin12] S. Lindgreen. AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Res Notes*, 5:337, Jul 2012.

[Liu18] T. Liu. MACS ChangeLog. `https://github.com/taoliu/MACS/blob/master/ChangeLog`, 2018. Accessed: 2019-01-09.

[LS07] J. T. Leek and J. D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, 3(9):1724–1735, Sep 2007.

[LS12] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9(4):357–359, Mar 2012.

[LS15] A.T.L. Lun and G. K. Smyth. csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. *Nucleic Acids Research*, 44(5):e45–e45, 11 2015.

[LT03] G. Legube and D. Trouche. Regulating histone acetyltransferases and deacetylases. *EMBO Rep.*, 4(10):944–947, Oct 2003.

[LY13] T. I. Lee and R. A. Young. Transcriptional regulation and its misregulation in disease. *Cell*, 152(6):1237–1251, Mar 2013.

[LY15] C. Liu and X. Yu. ADP-ribosyltransferases and poly ADP-ribosylation. *Curr. Protein Pept. Sci.*, 16(6):491–501, 2015.

[Lyk18] F. Lyko. The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. *Nat. Rev. Genet.*, 19(2):81–92, Feb 2018.

[Lys08] V. Lyssenko. The transcription factor 7-like 2 gene and increased risk of type 2 diabetes: an update. *Curr Opin Clin Nutr Metab Care*, 11(4):385–392, Jul 2008.

[M+05] Y. Ma et al. DNA CpG hypomethylation induces heterochromatin reorganization involving the histone variant macroH2A. *J. Cell. Sci.*, 118(Pt 8):1607–1616, Apr 2005.

[M+06] V. Matys et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, 34(Database issue):D108–110, Jan 2006.

[M+07] B. S. Mann et al. FDA approval summary: vorinostat for treatment of advanced primary cutaneous T-cell lymphoma. *Oncologist*, 12(10):1247–1252, Oct 2007.

[M+11] M. Maekawa et al. Direct reprogramming of somatic cells is promoted by maternal transcription factor Glis1. *Nature*, 474(7350):225–229, Jun 2011.

[M+12a] M. T. Maurano et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099):1190–1195, Sep 2012.

[M+12b] R. C. McLeay et al. Genome-wide in silico prediction of gene expression. *Bioinformatics*, 28(21):2789–2796, Nov 2012.

[M+14] M. Miyamoto et al. Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes. *BMC Genomics*, 15:699, Aug 2014.

[M+15a] B. Mifsud et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics*, 47(6):598–606, May 2015.

[M+15b] A Mora et al. In the loop: promoter - enhancer interactions and bioinformatics. *Briefings in Bioinformatics*, 2015.

[M+16a] J. MacArthur et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, 45(D1):D896–D901, 11 2016.

[M+16b] A. Mathelier et al. DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. *Cell Syst*, 3(3):278–286, 09 2016.

[M+16c] P. Milani et al. Cell freezing protocol suitable for ATAC-Seq on motor neurons derived from human induced pluripotent stem cells. *Sci Rep*, 6:25474, 05 2016.

[Mad15] P. Madrigal. On Accounting for Sequence-Specific Bias in Genome-Wide Chromatin Accessibility Experiments: Recent Advances and Contradictions. *Front Bioeng Biotechnol*, 3:144, 2015.

[Mar07] E. R. Mardis. ChIP-seq: welcome to the new frontier. *Nat. Methods*, 4(8):613–614, Aug 2007.

[MD18] A. Mayran and J. Drouin. Pioneer transcription factors shape the epigenetic landscape. *J. Biol. Chem.*, 293(36):13795–13804, Sep 2018.

[MK96] D. L. Minor and P. S. Kim. Context-dependent secondary structure formation of a designed protein sequence. *Nature*, 380(6576):730–734, Apr 1996.

[MM00] M. E. Massari and C. Murre. Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms. *Mol. Cell. Biol.*, 20(2):429–440, Jan 2000.

[MS+12] S. Marco-Sola et al. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nature Methods*, 9(12):1185–1188, 2012.

[Mue17] F. Mueller. Epigenetic regulation. `https://figshare.com/collections/Epigenetic_Regulation/3792331/1`, Oct 2017.

[MW47] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.*, 18(1):50–60, 03 1947.

[N+97] V. A. Narayan et al. Structures of zinc finger domains from transcription factor Sp1. Insights into sequence-specific protein-DNA recognition. *J. Biol. Chem.*, 272(12):7801–7809, Mar 1997.

[N+02] K. Nobori et al. Atf3 inhibits doxorubicin-induced apoptosis in cardiac myocytes: A novel cardioprotective role of atf3. *Journal of Molecular and Cellular Cardiology*, 34(10):1387–1397, 2002.

[N+12] A. Natarajan et al. Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res.*, 22(9):1711–1722, Sep 2012.

[N+16] G. Nagy et al. Motif oriented high-resolution analysis of ChIP-seq data reveals the topological order of CTCF and cohesin proteins on DNA. *BMC Genomics*, 17(1):637, Aug 2016.

[N+17] G. Nyamundanda et al. A Novel Statistical Method to Diagnose, Quantify and Correct Batch Effects in Genomic Studies. *Sci Rep*, 7(1):10849, Sep 2017.

[N+19] K. Nordstroem et al. Unique and assay specific features of NOMe-, ATAC- and DNase1-seq data. *Unpublished*, March 2019.

[NH11] B. K. Nayak and A. Hazra. How to choose the right statistical test? *Indian J Ophthalmol*, 59(2):85–86, 2011.

[NIH18] NIH. What is a genome? `https://ghr.nlm.nih.gov/primer/hgp/genome`, 2018. Accessed: 2018-12-24.

[Nob09] W. S. Noble. How does multiple testing correction work? *Nat. Biotechnol.*, 27(12):1135–1137, Dec 2009.

[NS+19] N. Noorbakhsh-Sabet et al. Artificial Intelligence Transforms the Future of Healthcare. *Am. J. Med.*, Jan 2019.

[NT09] Inc. NanoString Technologies. Reference Genes for Normalization of Expression Data. *Technical Note*, 2009.

[NV15] Hoang-Vu Nguyen and Jilles Vreeken. Flexibly mining better subgroups. *arXiv.org*, 10 2015.

[NW70] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48(3):443–453, Mar 1970.

[O+09] Z. Ouyang et al. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc. Natl. Acad. Sci. U.S.A.*, 106(51):21521–21526, Dec 2009.

[O+10] C. Oguey et al. Understanding the sequence-dependence of DNA groove dimensions: implications for DNA interactions. *PLoS ONE*, 5(12):e15931, Dec 2010.

[O+13] R. Ostuni et al. Latent enhancers activated by stimulation in differentiated cells. *Cell*, 152(1-2):157–171, Jan 2013.

[O+14] Y. Omatsu et al. Foxc1 is a critical regulator of haematopoietic stem/progenitor cell niche formation. *Nature*, 508(7497):536–540, Apr 2014.

[O+17] E. Ong et al. Ontobee: A linked ontology data server to support ontology term dereferencing, linkage, query and integration. *Nucleic Acids Res.*, 45(D1):D347–D352, 01 2017.

[OB14] T. R. O'Connor and T. L. Bailey. Creating and validating cis-regulatory maps of tissue-specific gene expression regulation. *Nucleic Acids Res.*, 42(17):11000–11010, 2014.

[OC11] C. T. Ong and V. G. Corces. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.*, 12(4):283–293, Apr 2011.

[P+08] U. J. Pape et al. Natural similarity measures between position frequency matrices with an application to clustering. *Bioinformatics*, 24(3):350–357, Feb 2008.

[P+09] C. Pesquita et al. Semantic similarity in biomedical ontologies. *PLoS Comput. Biol.*, 5(7):e1000443, Jul 2009.

[P+12] EJ. Park et al. Doxorubicin induces cytotoxicity through upregulation of pERK-dependent ATF3. *PLoS One*, 7(9):e44990, 2012.

[P+13a] D. Park et al. Widespread misinterpretable ChIP-seq bias in yeast. *PLoS ONE*, 8(12):e83506, 2013.

[P+13b] V. Pattanayak et al. High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat Biotechnol*, 31:839–843, 2013.

[P+13c] J. Piper et al. Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res.*, 41(21):e201, Nov 2013.

[P+13d] S. L. Planey et al. Post-translational modification of transcription factors: mechanisms and potential therapeutic interventions. *Curr Mol Pharmacol*, 6(3):173–182, Nov 2013.

[P+17a] R. Patro et al. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, 14(4):417–419, Apr 2017.

[P+17b] I. Plaschkes et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database*, 2017, 04 2017.

[PH11] C. P. Ponting and R. C. Hardison. What fraction of the human genome is functional? *Genome Res.*, 21(11):1769–1776, Nov 2011.

[Pot17] Sebastian Pott. Simultaneous measurement of chromatin accessibility, dna methylation, and nucleosome phasing in single cells. *Elife*, 6:e23203, 2017.

[PR+11] R. Pique-Regi et al. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.*, 21(3):447–455, Mar 2011.

[Pri17] G.J. Privitera. *Introduction to Hypothesis Testing*, page Chapter 8. SAGE Publications, 2017.

[PS13] V. Pelechano and L. M. Steinmetz. Gene regulation by antisense transcription. *Nat. Rev. Genet.*, 14(12):880–893, Dec 2013.

[Pug00] B. F. Pugh. Control of gene expression through regulation of the TATA-binding protein. *Gene*, 255(1):1–14, Sep 2000.

[PW17] P. Portin and A. Wilkins. The Evolving Definition of the Term "Gene". *Genetics*, 205(4):1353–1364, Apr 2017.

[Q+00] C. Qi et al. Peroxisome proliferator-activated receptors, coactivators, and downstream targets. *Cell Biochem. Biophys.*, 32 Spring:187–204, 2000.

[Q+04] J. P. Quivy et al. A CAF-1 dependent pool of HP1 during heterochromatin duplication. *EMBO J.*, 23(17):3516–3526, Sep 2004.

[QH10] A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, Mar 2010.

[R+84] A. Rich et al. The chemistry and biology of left-handed Z-DNA. *Annu. Rev. Biochem.*, 53:791–846, 1984.

[R+07] H. G. Roider et al. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, 23(2):134–141, Jan 2007.

[R+09] H. G. Roider et al. PASTAA: identifying transcription factors associated with sets of co-regulated genes. *Bioinformatics*, 25(4):435–442, Feb 2009.

[R+11] M. Rye et al. Clustered ChIP-Seq-defined transcription factor binding sites and histone modifications map distinct classes of regulatory elements. *BMC Biol.*, 9:80, Nov 2011.

[R+13] S. E. Reese et al. A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis. *Bioinformatics*, 29(22):2877–2883, 2013.

[R+14a] F. Ramirez et al. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.*, 42(Web Server issue):W187–191, Jul 2014.

[R+14b] S. S. Rao et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, Dec 2014.

[R+15a] P. Ramachandran et al. BIDCHIPS: bias decomposition and removal from ChIP-seq data clarifies true binding signal and its functional correlates. *Epigenetics Chromatin*, 8:33, 2015.

[R+15b] A. Rotem et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.*, 33(11):1165–1172, Nov 2015.

[R+16] A. Ramanathan et al. mRNA capping: biological functions and applications. *Nucleic Acids Res.*, 44(16):7511–7526, Sep 2016.

[R+17a] S. Rahman et al. Single-cell profiling reveals that eRNA accumulation at enhancer-promoter loops is not required to sustain transcription. *Nucleic Acids Res.*, 45(6):3017–3030, Apr 2017.

[R+17b] A. Regev et al. The Human Cell Atlas. *Elife*, 6, 12 2017.

[R+18] M. Rabbani et al. Role of artificial intelligence in the care of patients with nonsmall cell lung cancer. *Eur. J. Clin. Invest.*, 48(4), Apr 2018.

[R+19] J. Ray et al. Chromatin conformation remains stable upon extensive transcriptional changes driven by heat shock. *bioRxiv*, 2019.

[RC14] R. C. Riddle and T. L. Clemens. Insulin, osteoblasts, and energy metabolism: why bone counts calories. *J. Clin. Invest.*, 124(4):1465–1467, Apr 2014.

[RJ86] L. R. Rabiner and B. H. Juang. An introduction to hidden markov models. *IEEE ASSp Magazine*, 1986.

[S+82] G. D. Stormo et al. Characterization of translational initiation sites in E. coli. *Nucleic Acids Res.*, 10(9):2971–2996, May 1982.

[S+86] T. D. Schneider et al. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, 188(3):415–431, Apr 1986.

[S+00a] E.M. Smigielski et al. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Research*, 28(1):352–355, 01 2000.

[S+00b] P. Spirtes et al. *Causation, Prediction, and Search.* Cambridge, MA: MIT Press, 2000.

[S+01] ST. Sherry et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, 29(1):308–311, 2001.

[S+04] P. J. Sabo et al. Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. *Proc. Natl. Acad. Sci. U.S.A.*, 101(13):4537–4542, Mar 2004.

[S+06] P. Spallarossa et al. Matrix metalloproteinase-2 and -9 are induced differently by doxorubicin in H9c2 cells: The role of MAP kinases and NAD(P)H oxidase. *Cardiovascular Research*, 69(3):736–745, 2006.

[S+10] S. Stella et al. The shape of the DNA minor groove directs binding by the DNA-bending protein Fis. *Genes Dev.*, 24(8):814–826, Apr 2010.

[S+11a] K. P. Singh et al. Aryl hydrocarbon receptor-null allele mice have hematopoietic stem/progenitor cells with abnormal characteristics and functions. *Stem Cells Dev.*, 20(5):769–784, May 2011.

[S+11b] L. Song et al. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res.*, 21(10):1757–1767, Oct 2011.

[S+12a] M. A. Schaub et al. Linking disease associations with regulatory information in the human genome. *Genome research*, 22(9):1748–1759, 2012.

[S+12b] M. H. Schulz et al. DREM 2.0: Improved reconstruction of dynamic regulatory networks from time-series expression data. *BMC Syst Biol*, 6:104, Aug 2012.

[S+12c] M. H. Schulz et al. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28(8):1086–1092, Apr 2012.

[S+14a] M. Schubert et al. Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat Protoc*, 9(5):1056–1082, May 2014.

[S+14b] O. Shalem et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*, 343(6166):84–87, Jan 2014.

[S+14c] R. I. Sherwood et al. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat. Biotechnol.*, 32(2):171–178, Feb 2014.

[S+14d] M. H. Sung et al. DNase footprint signatures are dictated by factor dynamics and DNA sequence. *Mol. Cell*, 56(2):275–285, Oct 2014.

[S+15] P. Sebastiani et al. BCL11A enhancer haplotypes and fetal hemoglobin in sickle cell anemia. *Blood Cells Mol. Dis.*, 54(3):224–230, Mar 2015.

[S+16a] A. D. Schmitt et al. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Rep*, 17(8):2042–2059, 11 2016.

[S+16b] R. Singh et al. DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics*, 32(17):i639–i648, Sep 2016.

[S+16c] L. Song et al. A transcription factor hierarchy defines an environmental stress response network. *Science*, 354(6312), 11 2016.

[S+17a] F. Schmidt et al. Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Res.*, 45(1):54–66, 01 2017.

[S+17b] J. Shendure et al. DNA sequencing at 40: past, present and future. *Nature*, 550(7676):345–353, 10 2017.

[S+17c] D. Szklarczyk et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.*, 45(D1):D362–D368, 01 2017.

[S+18a] F. Schmidt et al. An ontology-based method for assessing batch effect adjustment approaches in heterogeneous datasets. *Bioinformatics*, 34(17):i908–i916, Sep 2018.

[S+18b] F. Schmidt et al. TEPIC 2 - An extended framework for transcription factor binding prediction and integrative epigenomic analysis. *Bioinformatics*, Oct 2018.

[S+19] F. Schmidt et al. Integrative analysis of epigenetics data identifies gene-specific regulatory elements. *Unpublished*, March 2019.

[SB16] Nathan C. Sheffield and Christoph Bock. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics*, 32(4):587–589, feb 2016.

[Sci18] Thermo Fischer Scientific. GeneChipâĎć Human Transcriptome Pico Assay 2.0. `https://www.thermofisher.com/order/catalog/product/902662?SID=srch-srp-902662`, 2018. Accessed: 2018-12-26.

[Sci19] Thermo Fischer Scientific. DNA Gels for Next-Generation Sequencing. `https://www.thermofisher.com/de/de/home/life-science/dna-rna-purification-analysis/nucleic-acid-gel-electrophoresis/dna-electrophoresis/agarose-gel-electrophoreis/dna-gels-ngs.html`, 2019. Accessed: 2018-12-27.

[Sha76] A. J. Shatkin. Capping of eucaryotic mRNAs. *Cell*, 9(4 PT 2):645–653, Dec 1976.

[Sha01] A. D. Sharrocks. The ETS-domain transcription factor family. *Nat. Rev. Mol. Cell Biol.*, 2(11):827–837, Nov 2001.

[Shi06] R. A. Shivdasani. MicroRNAs: regulators of gene expression and cell differentiation. *Blood*, 108(12):3646–3653, Dec 2006.

[Sho17] P.and others Shooshtari. Integrative genetic and epigenetic analysis uncovers regulatory mechanisms of autoimmune disease. *The American Journal of Human Genetics*, 101(1):75 – 86, 2017.

[Sid10] R. Siddharthan. Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. *PLoS ONE*, 5(3):e9722, Mar 2010.

[SKM01] H. Shizuya and H. Kouros-Mehr. The development and applications of the bacterial artificial chromosome cloning system. *Keio J Med*, 50(1):26–30, Mar 2001.

[Spo18] S. H. Spoel. Orchestrating the proteome with post-translational modifications. *J. Exp. Bot.*, 69(19):4499–4503, Aug 2018.

[SR09] M. D. Shoulders and R. T. Raines. Collagen structure and stability. *Annu. Rev. Biochem.*, 78:929–958, 2009.

[SS90] T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, 18(20):6097–6100, Oct 1990.

[SS16] M. Siebert and J. Soding. Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Res.*, 44(13):6055–6069, 07 2016.

[SS18] F. Schmidt and M. H. Schulz. On the problem of confounders in modeling gene expression. *Bioinformatics*, Aug 2018.

[ST03] J.D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16):9440–5, aug 2003.

[Sto00] G. D. Stormo. DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, Jan 2000.

[Stu18] H. Stunnenberg. BLUEPRINT - A BLUEPRINT of Haematopoietic Epigenomes. `http://www.blueprint-epigenome.eu/index.cfm?p=31AD6D30-9B3C-BB97-E7F81875121FEC41`, 2018. Accessed: 2019-01-02.

[T+84] KM. Tewey et al. Adriamycin-induced DNA damage mediated by mammalian DNA topoisomerase II. *Science*, 226(4673):466–468, 1984.

[T+07] V. Trevino et al. DNA microarrays: a powerful genomic tool for biomedical and clinical research. *Mol. Med.*, 13(9-10):527–541, 2007.

[T+08] J. V. Turatsinze et al. Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nat Protoc*, 3(10):1578–1588, 2008.

[T+09] C. Trapnell et al. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, May 2009.

[T+10] C. Trapnell et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, 28(5):511–515, May 2010.

[T+11] C. Tamm et al. Regulation of mouse embryonic stem cell self-renewal by a Yes-YAP-TEAD2 signaling pathway downstream of LIF. *J. Cell. Sci.*, 124(Pt 7):1136–1144, Apr 2011.

[T+12] Thurman et al. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, Sep 2012.

[T+14] Phillippa C. Taberlay et al. Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. *Genome Research*, 24(9):1421–1432, sep 2014.

[T+16] R. Thomas et al. Features that define the best ChIP-seq peak calling algorithms. *Briefings in Bioinformatics*, 18(3):441–450, 05 2016.

[Tan06] A. Tanay. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res.*, 16(8):962–972, Aug 2006.

[TC+11] M. Thomas-Chollier et al. Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. *Nat Protoc*, 6(12):1860–1869, Dec 2011.

[TM70] H. M. Temin and S. Mizutani. RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature*, 226(5252):1211–1213, Jun 1970.

[Tri04] R. C. Trievel. Structure and function of histone methyltransferases. *Crit. Rev. Eukaryot. Gene Expr.*, 14(3):147–169, 2004.

[U+02] T. Uzawa et al. Polypeptide synthesis directed by DNA as a messenger in cell-free polypeptide synthesis by extreme thermophiles, Thermus thermophilus HB27 and Sulfolobus tokodaii strain 7. *J. Biochem.*, 131(6):849–853, Jun 2002.

[U+15] M. Uhlen et al. Proteomics. Tissue-based map of the human proteome. *Science*, 347(6220):1260419, Jan 2015.

[V+01] J. C. Venter et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, Feb 2001.

[V+09a] J. M. Vaquerizas et al. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, 10(4):252–263, 04 2009.

[V+09b] A. Visel et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231):854–858, February 2009.

[vB15] A. van Bömmel. *Prediction of transcription factor co-occurence using rank based statistics.* PhD thesis, Freie Universität Berlin, 2015.

[vdMH08] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[vdVN07] H. J. van der Vliet and E. E. Nieuwenhuis. IPEX as a result of mutations in FOXP3. *Clin. Dev. Immunol.*, 2007:89017, 2007.

[vI+08] H. van Ingen et al. Structural insight into the recognition of the H3K4me3 mark by the TFIID subunit TAF3. *Structure*, 16(8):1245–1256, Aug 2008.

[Vos14] J. Voskuil. Commercial antibodies and their validation. *F1000Res*, 3:232, 2014.

[W+79] C. Wu et al. The chromatin structure of specific genes: II. Disruption of chromatin structure during gene activity. *Cell*, 16(4):807–814, Apr 1979.

[W+91] B. Wittig et al. Transcription is associated with Z-DNA formation in metabolically active permeabilized mammalian cell nuclei. *Proc. Natl. Acad. Sci. U.S.A.*, 88(6):2259–2263, Mar 1991.

[W+99] S. Williams et al. Treatment of osteoporosis with MMP inhibitors. *Ann. N. Y. Acad. Sci.*, 878:191–200, Jun 1999.

[W+02] H. M. Wain et al. Guidelines for human gene nomenclature. *Genomics*, 79(4):464–470, Apr 2002.

[W+08] L. Wang et al. The zinc finger transcription factor Zbtb7b represses CD8-lineage gene expression in peripheral CD4+ T cells. *Immunity*, 29(6):876–887, Dec 2008.

[W+09] M. C. Wahl et al. The spliceosome: design principles of a dynamic RNP machine. *Cell*, 136(4):701–718, Feb 2009.

[W+13a] S. Wang et al. Target analysis by integration of transcriptome and ChIP-seq data with BETA. *Nat Protoc*, 8(12):2502–2515, Dec 2013.

[W+13b] J.N. Weinstein et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113, 2013.

[W+13c] M. T. Weirauch et al. Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.*, 31(2):126–134, Feb 2013.

[W+13d] J. Woodsmith et al. Dual coordination of post translational modifications in human protein networks. *PLoS Comput. Biol.*, 9(3):e1002933, 2013.

[W+14] D. R. Whelan et al. Detection of an en masse and reversible B- to A-DNA conformational transition in prokaryotes in response to desiccation. *J R Soc Interface*, 11(97):20140454, Aug 2014.

[W+15] J. Wang et al. 4DGenome: a comprehensive database of chromatin interactions. *Bioinformatics*, 31(15):2560–2564, 03 2015.

[W+17] J. R. Wang et al. Correcting nucleotide-specific biases in high-throughput sequencing data. *BMC Bioinformatics*, 18(1):357, Aug 2017.

[W+19a] A. T. D. Watson et al. Evidence for Aryl hydrocarbon Receptor-Mediated Inhibition of Osteoblast Differentiation in Human Mesenchymal Stem Cells. *Toxicol. Sci.*, 167(1):145–156, Jan 2019.

[W+19b] M. Wegner et al. Circular synthesized CRISPR/Cas gRNAs for functional interrogations in the coding and noncoding genome. *Elife*, 8, Mar 2019.

[Wad57] C.H. Waddington. The strategy of the genes; a discussion of some aspects of theoretical biology. *Allen and Unwin, London*, 1957.

[Wad08] C. H. Waddington. The basic ideas of biology. *Biological Theory*, 3(3):238–253, Sep 2008.

[Wag17] O. Wagih. ggseqlogo: a versatile r package for drawing sequence logos. *Bioinformatics*, 33(22):3645–3647, 2017.

[Wal85] J. Wallis. *A Treatise of Algebra, both Historical and Practical*. Oxford, 1685.

[WB08] L. B. Wan and M. S. Bartolomei. Regulation of imprinting in clusters: noncoding RNAs versus insulators. *Adv. Genet.*, 61:207–223, 2008.

[WC53] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, Apr 1953.

[WC12] E. J. Wagner and P. B. Carpenter. Understanding the language of Lys36 methylation at histone H3. *Nat. Rev. Mol. Cell Biol.*, 13(2):115–126, Jan 2012.

[WH14] T. Will and V. Helms. Identifying transcription factor complexes and their roles. *Bioinformatics*, 30(17):i415–421, Sep 2014.

[Wil05] G. Wilcox. Insulin and insulin resistance. *Clin Biochem Rev*, 26(2):19–39, May 2005.

[WL16] Ronald L. Wasserstein and Nicole A. Lazar. The asa's statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016.

[WM01] C.-t. Wu and J. R. Morris. Genes, genetics, and epigenetics: A correspondence. *Science*, 293(5532):1103–1105, 2001.

[WM16] M. Wierer and M. Mann. Proteomics to study DNA-bound and chromatin-associated gene regulatory complexes. *Hum. Mol. Genet.*, 25(R2):R106–R114, Oct 2016.

[WN10] T.D. Wu and S Nacu. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–881, apr 2010.

[WS10] K. M. Wittkowski and T. Song. Nonparametric methods for molecular biology. *Methods Mol. Biol.*, 620:105–153, 2010.

[WZ14] H. Wu and Y. Zhang. Reversing DNA methylation: mechanisms, genomics, and biological functions. *Cell*, 156(1-2):45–68, Jan 2014.

[X+98] M. Xu et al. Cloning, characterization and expression of the gene coding for a cytosine-5-DNA methyltransferase recognizing GpC. *Nucleic Acids Res.*, 26(17):3961–3966, Sep 1998.

[X+04] Qing-Song Xu et al. Monte carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration. *Journal of Chemometrics*, 18(2):112–120, 2004.

[X+05] Z. Xu et al. Liver-specific inactivation of the Nrf1 gene in adult mouse leads to nonalcoholic steatohepatitis and hepatic neoplasia. *Proc. Natl. Acad. Sci. U.S.A.*, 102(11):4120–4125, Mar 2005.

[Y+13] J. Yan et al. Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell*, 154(4):801–813, Aug 2013.

[Y+14] G. G. Yardımcı et al. Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. *Nucleic Acids Res.*, 42(19):11865–11878, Oct 2014.

[Y+15] L. Yao et al. Demystifying the secret mission of enhancers: linking distal regulatory elements to target genes. *Crit. Rev. Biochem. Mol. Biol.*, 50(6):550–573, 2015.

[Y+16] C.H. Yu et al. Consensus genome-wide expression quantitative trait loci and their relationship with human complex trait disease. *OMICS: A Journal of Integrative Biology*, 20(7):400–414, 2016. PMID: 27428252.

[Y+18] X. Yang et al. Mitochondrial protein sulfenation during aging in the rat brain. *Biophys Rep*, 4(2):104–113, 2018.

[Z+08a] Y. Zhang et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, 9(9):R137, 2008.

[Z+08b] J. Zhao et al. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science*, 322(5902):750–756, Oct 2008.

[Z+09] C. Zang et al. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*, 25(15):1952–1958, Aug 2009.

[Z+13a] B. R. Zhou et al. Structural insights into the histone H1-nucleosome complex. *Proc. Natl. Acad. Sci. U.S.A.*, 110(48):19390–19395, Nov 2013.

[Z+13b] M. J. Ziller et al. Charting a dynamic DNA methylation landscape of the human genome. *Nature*, 500(7463):477–481, Aug 2013.

[Z+15] D. R. Zerbino et al. The ensembl regulatory build. *Genome Biology*, 16(1):56, Mar 2015.

[Z+17] J. Zhao et al. Quantifying the Impact of Non-coding Variants on Transcription Factor-DNA Binding. *Res Comput Mol Biol*, 10229:336–352, May 2017.

[Zak13] S. Zakhari. Alcohol metabolism and epigenetics changes. *Alcohol Res*, 35(1):6–16, 2013.

[Zar10] J.H. Zar. *Biostatistical Analysis 5th Edition*. Pearson Education, 2010.

[ZG15] Y. Zhao and B. A. Garcia. Comprehensive Catalog of Currently Documented Histone Modifications. *Cold Spring Harb Perspect Biol*, 7(9):a025064, Sep 2015.

[Zha12] R. F. Zhao. Genome advance of the month encode: Deciphering function in the human genome. `https://www.genome.gov/27551473`, 2012. Accessed: 2018-12-25.

[ZZ09] Ethan Zhang and Yi Zhang. *F-Measure*, pages 1147–1147. Springer US, Boston, MA, 2009.