Saarland University
Center for Bioinformatics
Master's Program for Bioinformatics

Master's Thesis in Bioinformatics

Correlation of Differential Gene Expression and Histone Modifications in Transcription Factor and microRNA Co–Regulatory Motifs

submitted by

Markus Hollander

on August 14, 2019

Supervisor
Professor Dr. Volkhard Helms

Reviewers
Professor Dr. Volkhard Helms
Professor Dr. Tobias Marschall

Hollander, Markus
Correlation of Differential Gene Expression and Histone Modifications in Transcription Factor and microRNA Co–Regulatory Motifs
Master's Thesis in Bioinformatics
Saarland University
Saarbrücken, Germany
August 14, 2019

Declaration

I hereby confirm that this thesis is my own work and that I have documented all sources used.

I hereby declare that the submitted digital and hardcopy versions of this thesis correspond to each other. I give permission to the Saarland University to duplicate and publish this work.

Saarbrücken, on August 14, 2019

Markus Hollander

Abstract

Gene expression plays a crucial role in the development and function of the human body. An intricate system of regulatory mechanisms is necessary to ensure that each cell is able to perform its role in the survival and health of the overall organism. Dysregulation has been shown to cause potentially fatal developmental disorders as well as mental and physical diseases. Histone modifications and variants are fundamentally involved in gene expression regulation and their individual and combinatorial effects on gene expression are an active field of research. Similarly, motifs comprised of transcription factor and microRNA pairs co–regulating shared target genes modulate gene expression in many processes. This thesis explored the relationship of differential gene expression and differential histone marks, specifically H3K4me2, H3K4me3, H3K27ac and H2A.Z, in these co–regulatory motifs by examining six human cell differentiation transitions. After differential gene expression and differential histone mark analysis, the expression and histone mark correlation for different motif types was statistically compared to the genome–wide correlation, as well as to motifs in randomised gene regulatory networks, and an annotation enrichment analysis was performed. Observed correlation patterns genome–wide and in co–regulatory motifs were generally consistent with previous characterisations of individual histone marks. Motif correlations were overall stronger than genome–wide correlations, significantly so for some motif types in specific cell differentiation transitions. Some motif types were significantly enriched compared to randomised networks, and motif correlation was unusually strong in comparison to motifs in randomised networks in some instances. Process annotations of motifs were enriched in regulation of transcription, development, cell differentiation and cell proliferation, while functional annotations showed enrichment of regulatory molecular binding, including to modifiers of epigenetic states. Additionally, there were annotation and correlation pattern differences between different motif types. Further and more in–depth study of histone marks in transcription factor and microRNA co–regulatory motifs seems warranted.

Acknowledgments

I am very grateful to my supervisor Professor Dr. Volkhard Helms who offered me the chance to write this thesis at the Chair of Computational Biology at Saarland University and whose door was always open when I ran into questions. Furthermore, I would like to express my gratitude to him and Professor Dr. Tobias Marschall for reading and evaluating my work. Lastly, my thanks also goes to Dr. Maryam Nazarieh and Aditi Jain for updating the gene regulatory database used in this thesis.

Contents

List of Tables

List of Figures

List of Abbreviations

BAM	binary sequence alignment/map ... 18
bp	base pairs ... 5
ChIP–seq	chromatin immunoprecipitation DNA–sequencing ... 17
coA	coenzyme A ... 7
COV	covariance ... 32
DNA	deoxyribonucleic acid ... 1
ENCODE	Encyclopedia of DNA Elements ... 9
EPD	Eukaryotic Promoter Database ... 16
ESC	embryonic stem cell ... 5
FDR	false discovery rate ... 24
FFL	feedforward loop ... 1
FWER	family–wise error rate ... 28
GO	Gene Ontology ... 38
GRC	Genome Reference Consortium ... 15
HAT	histone acetyltransferase ... 7
HDAC	histone deacetylase ... 7
HDM	histone demethylase ... 8
hESC	human embryonic stem cell ... 18
GRCh37	GRC Human Build 37 ... 15
GRCh38	GRC Human Build 38 ... 15
HMDD	Human microRNA Disease Database ... 10
HMM	hidden Markov model ... 24
HMT	histone methyltransferase ... 7
IHW	independent hypothesis weighting ... 24
iPSC	induced pluripotent stem cell ... 18
KEGG	Kyoto Encyclopedia of Genes and Chromosomes ... 38
miRNA	microRNA ... 1
MPC	myeloid progenitor cell ... 18
mRNA	messenger RNA ... 1
MSC	mesenchymal stem cell ... 18

NCBI National Center for Biotechnology Information 15
NGS next–generation sequencing 17
NPC neural progenitor cell 18
NSPC neural stem progenitor cell 18
PIC pre–initiation–complex 7
pre–miRNA precursor–microRNA 10
pre–mRNA precursor mRNA 4
pri–miRNA primary–microRNA 10
RISC RNA–induced silencing complex 10
RNA ribonucleic acid 1
RNA–seq RNA–sequencing 15
SD standard deviation 32
SE standard error 37
TF transcription factor 1
TSS transcription start site 4

Chapter 1

Introduction

Humans develop from a single, fertilised egg cell (zygote) to a complex organism with an estimated number of 37.2 trillion ($3.72 \cdot 10^{13}$) cells [1]. During development, the zygote multiplies and differentiates into ever more specialised cell types that form tissues, organs, organ systems and finally the entire human body [2].

All of these cells share the almost same genome, providing the genetic blueprint for cellular components and processes encoded as genes on deoxyribonucleic acid (DNA) [2]. However, in each cell only a subset of genes is actively transcribed and translated into functional products (expressed) at any given time, giving rise to the distinct morphology and function of different cell types, and allowing individual cells to communicate with their neighbours and to adapt to changing environmental conditions [2].

Gene expression is governed by an intricate system of regulatory mechanisms that tightly control each step of the process to ensure proper development and function of single cells as well as the entire organism [2, 3]. Consequently, genetic dysregulation, depending on the nature and extent, has the potential to cause developmental disorders, mental and physical diseases, which can prove fatal. For instance, dysregulation has been implicated in the development of intellectual disability [4], schizophrenia [5] and cancer [4, 6, 7]. Studying gene expression regulation is thus crucial for better understanding the human body, its components, systems and development, as well as gaining knowledge about disease processes and identifying effective interventions.

Well–studied regulators of gene expression are small proteins called transcription factors (TFs) that can facilitate or inhibit transcription, and thus expression, of their target genes in a variety of ways, such as preparing or blocking the gene for transcription, recruiting components of the transcriptional machinery [3]. Similarly, microRNAs (miRNAs), which are small ribonucleic acid (RNA) transcripts that are not translated into proteins, have been shown to repress gene expression by participating in the degradation of protein–coding messenger RNAs (mRNAs) or by interfering with their translation into proteins [8, 9]. Both TFs and miRNAs are regulatory actors in many biological processes, including cell proliferation, cell differentiation transitions, cell signalling or apoptosis [3, 10–14].

In addition to their individual roles, TFs and miRNAs can co–regulate shared targets genes and form feedforward loops (FFLs), which can be iden-

tified by examining gene regulatory networks [7, 15]. Depending on their precise nature, these TF–miRNA co–regulatory motifs can contribute to rapid and strong regulatory responses, fine–tuning of gene expression and buffering against stochastic noise and transient signals [7, 16–18]. Similar to their individual components, TF–miRNA co–regulatory motifs have been implicated in the cell cycle, cell differentiation and development, as well as disease processes [7, 15, 19, 20].

Another active field of research studies the effect of variants or modifications of histones, which are the core building blocks of DNA packaging into chromatin and chromosomes in the cell nucleus, on gene expression [6, 21, 22]. Histone modifications or variants can affect chromatin structure, and thus the accessibility of the DNA for components of the transcriptional machinery, and can serve as recognition sites for TFs and other effectors [4, 22–26]. Due to the large number of different histone modifications and variants, their specific effects on gene expression varies, as does their involvement in processes such as DNA replication and repair, stress response, cell proliferation and differentiation [4, 6, 24, 25, 27, 28].

In this thesis, the aim was to explore the relationship of gene expression and epigenetic histone modifications and variants in human TF–miRNA co–regulatory motifs. It was examined in cell differentiation transitions specifically, since the motifs as well as the histone modifications or variants are involved in cell differentiation, and allowed to study how a change in modifications or variants coincided with changes in gene expression. Differential gene expression analysis and differential histone modification or variant analysis was performed for six cell differential transitions, and corresponding gene regulatory networks were constructed and searched for TF–miRNA co–regulatory motifs. On that basis, the correlation between differential gene expression and differential histone modifications or variants was computed for genes involved in these motifs and compared to the genome–wide correlation baseline. In addition, the motif correlation was assessed in comparison with motifs in randomised regulatory networks. Furthermore, enrichment analysis was performed for function and process annotations.

The following chapter introduces further background information on cell differentiation and human development, DNA packaging and gene expression, and regulatory mechanisms such as histone modifications and variants, as well as transcription factors and miRNAs and the different types of co–regulatory motifs. Subsequently, Chapter 3 lists and describes the gene expression and histone modification and variant data sets, along with the used gene model and gene regulatory information. In Chapter 4, the different pre–processing and analysis steps are explained and implementation details are given. The results of the various correlation analyses are presented, interpreted and discussed in Chapter 5. Finally, Chapter 6 provides a brief summary of this work and concluding remarks.

Chapter 2

Background

This chapter presents the biological background and context needed to understand the purpose, methodology and results of this thesis. The goal was to investigate the relationship between differential gene expression and differential histone modifications or histone variants in transcription factor and microRNA co–regulatory motifs in *Homo sapiens*. First, gene expression and its role in the differentiation of different cell types and human development is explained. Then, histones are introduced as the basic building blocks of DNA packaging into chromatin, and it is explained how histone modifications and histone variants function are regulatory mechanisms of gene expression. This is followed by an introduction of TFs and miRNAs as gene expression regulators, and lastly describes the regulatory function of co–regulatory TF–miRNA motifs.

2.1 Gene Expression and Cell Differentiation

The body of an adult human is estimated to be comprised of 37.2 trillion ($3.72 \cdot 10^{13}$) eukaryotic cells of various types [1] that developed from a single, fertilised egg cell and each possess the same genome, despite their different morphologies and functions [2]. This section introduces gene expression and its role in the development of different cell types, and was summarised from *Life: The Science of Biology* (2008) [2] unless otherwise indicated.

The genome of eukaryotic cells is stored in the cell nucleus and contains the genetic blueprint for development and function encoded in deoxyribonucleic acid (DNA). The DNA is a double–helix consisting of two strands of nucleotides, each of which is composed of one of the four nitrogenous bases adenine, guanine, thymine and cytosine, as well as a deoxyribose sugar and a phosphate group. The sugar–phosphates form the negatively charged backbone of the double–helix, while the nitrogenous bases form hydrogen bonds with the corresponding base on the other strand. One end of each strand ends in a phosphate group, the other with a hydroxyl–group, which are called the 5'– and 3'–end, respectively. In the double helix, the 5'–end of one strand is paired with the 3'–end of the other. The pairing of these nitrogenous bases is complementary, whereby the purine base guanine only pairs with the pyrimidine base cytosine, and the purine base adenine only forms bonds with the pyrimidine base thymine. Consequently, one strand is sufficient to reconstruct the other.

Genes are 5'– to 3'–oriented segments on one of the two strands of the DNA double–helix that are transcribed into ribonucleic acid (RNA), which is a single–stranded molecule similar to DNA, where the deoxyribose sugar is replaced by ribose and thymine by uracil. RNA is distinguished into non–coding RNA and coding messenger RNA (mRNA), the latter of which can be translated by ribosomes into proteins, which are molecules consisting of one or several amino acid chains.

Directly upstream (on the 5'–side) of the gene body is the transcription start site (TSS) and the proximal promoter region, as well as enhancer regions with distal control elements much further upstream or downstream (3') of the gene. During transcription, a protein called the RNA polymerase binds to the proximal promoter and, starting with the TSS, produces a single–stranded primary transcript corresponding to the DNA sequence of the gene body that is then further processed, depending on the type of RNA.

In the case of eukaryotic protein–coding genes, the gene body as well as the corresponding primary transcript, the precursor mRNA (pre–mRNA), are comprised of protein–coding exons with non–coding introns in between. The pre–mRNA is refined into mature, protein–coding mRNA in a process called RNA splicing, in which exons are concatenated after removal of the introns (Figure 2.1). Hereby it is possible that a single pre–mRNA can yield several different mature mRNAs by joining varying numbers and combinations of exons. The result of this alternative splicing is that a single gene can produce different proteins with different functions. Additionally, after transcription, multiple adenosine–monophosphates are added to the 3'–end of the mRNA [29]. This polyadenylation enables export of the mature mRNA from the nucleus to the cytoplasm and stabilises the mRNA for translation into proteins [29].

Proteins fulfil many intra– and extracellular roles, ranging from being integral parts of cell structures to catalysing chemical reactions, signal transduction within and between cells and regulation, making them crucial actors in the metabolism of cells and the entire organism. Non–coding RNAs have various functions as well, depending on their type. For example, ribosomal RNA (rRNA) and transfer RNA (tRNA) are essential components in the translation machinery, whereas small RNAs like miRNAs play an important regulatory role (Section 2.5).

The process of synthesising functional products based on the genetic information encoded by genes is called gene expression. However, genes are not expressed all the time and which genes are expressed at any given time varies from cell to cell. This differential gene expression results in diverse cell structures and functions that can be grouped into cell types, and is a hallmark of cell differentiation.

In cell differentiation, the different cell types in the human body develop from a fertilised egg cell, the zygote. The zygote is totipotent, meaning that it can give rise to all other cell types in the embroy and fully developed body, as well as extra–embryonic cells such as those constituting the placenta. During early embryonic development, the zygote multiplies first into the morula, consisting of 16 cells, and shortly after into the blastula, both of which are still totipotent. The blastula develops into the trophoblast and hypoblast, which becomes the placenta and yolk sac, respectively, and into the epiblast, which gives rise to the

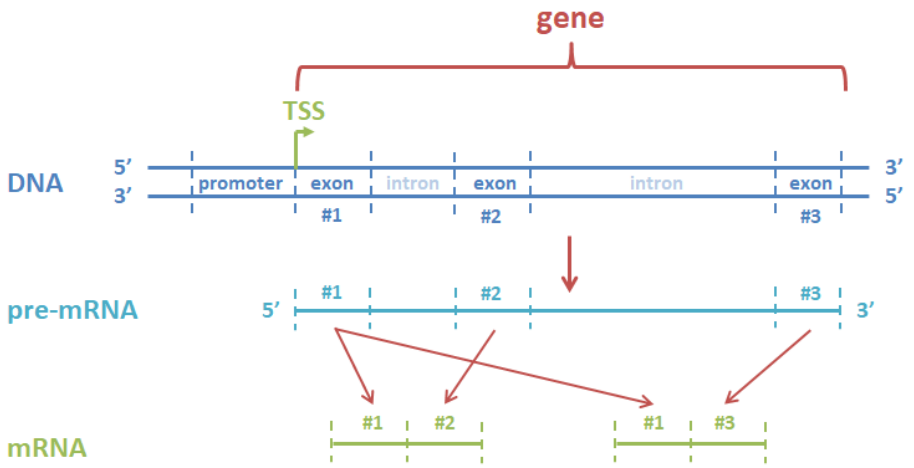

Figure 2.1: *Eukaryotic gene consisting of protein–coding exons and non–coding introns, with the promoter region upstream of the transcription start site (TSS). The gene is initially transcribed into pre–mRNA, which is subsequently spliced into mature mRNA. Alternative splicing can result in different mRNAs by combining different exons.*

embryo. The cells of the epiblast are pluripotent and as such can differentiate into all cells of the embryo, but no longer into extra–embryonic cell types.

Within the epiblast, the embryonic stem cells (ESCs) differentiate, forming the gastrula consisting of the ectoderm, mesoderm and endoderm germ layers. The cells of these layers develop into more and more specialised multipotent stem cells and progenitors that can differentiate into fewer and fewer cell types, until finally terminally differentiated cell types are developed. This can be visualised as a branching tree with the totipotent ESCs as the root and the terminally differentiated cells as leafs of the respective cell lineage branches (Figure 3.1) [30].

Gene expression is tightly controlled by several layers of regulatory mechanisms such that the right genes are expressed at the right rate in the right cells at the right time to make sure that the cells and overall organism can develop and function properly, as well as adapt to the ever changing environment. The following sections briefly introduce the regulatory mechanisms most relevant to the work conducted in this thesis.

2.2 Chromatin

In humans, the diploid genome in each cell is organised in 23 chromosome pairs, one of which is a pair of sex chromosomes (allosomes) and the other 22 pairs are non-sex chromosomes (autosomes) [2]. The autosomes are named according to their number as chromosomes 1 to 22, while the two possible sex chromosomes are called X and Y, whereby females typically have two X chromosomes and males typically have one X chromosome and one Y chromosome [2]. Each of the two sets of chromosomes consists of roughly 3.2 billion base pairs (bp) [31, 32].

To form these chromosomes in a space–efficient manner, the double–stranded DNA is condensed into coiled chromatin fibres, which in turn consist of a tightly

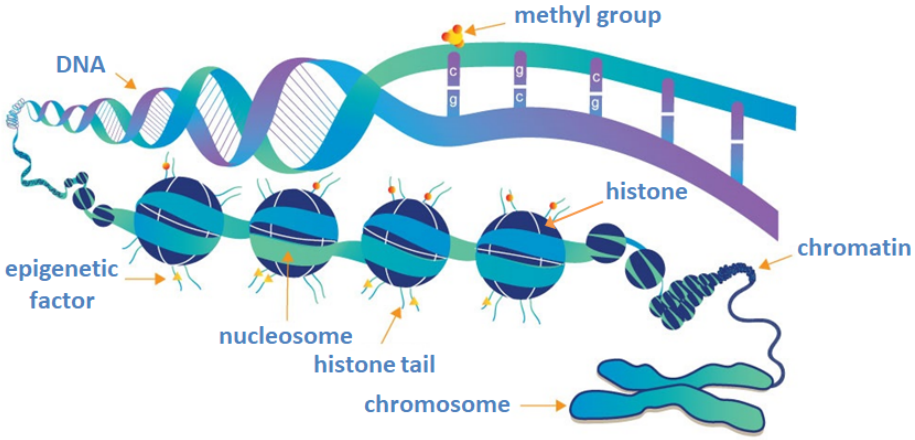

Figure 2.2: *Double–stranded DNA is wrapped around nucleosomes, consisting of a histone octamer, and further condensed into chromatin fibres and finally chromosomes. Further shown are epigenetic DNA–methylation of cytosine and histone tail modifications. Adapted from WhatIsEpigenetics.com [33].*

packed chain of nucleosomes [6]. Each nucleosome consists of circa 146 bp DNA wrapped around a protein octamer, which is comprised of four pairs of histones H2A, H2B, H3 and H4, while the individual nucleosomes are linked by approximately 80 bp long stretches of DNA and the linker histone H1 [6, 21]. A large proportion of amino acids in histones are positively charged, resulting in electrostatic attraction to the negatively charged DNA backbone [27].

These condensed chromatin fibres, called heterochromatin, are difficult to impossible to access for the transcriptional machinery and thus generally transcriptionally inactive, in contrast to the loosely packed euchromatin that is transcriptionally permissive [23].

2.3 Histone Modifications and Variants

The DNA and chromatin can be modified in a heritable manner, and these epigenetic modifications are heavily involved in gene expression regulation [34–36] (Figure 2.2). For instance, DNA methylation of cytosines in cytosine–guanine dinucleotides (CpG) of promoters can silence the expression of the corresponding gene by interfering with the binding ability of DNA–binding proteins involved in transcription [37]. However, in this thesis, the focus was on histone variants and post–translational modifications of histones, specifically histone methylation and acetylation.

The amino acid chains of proteins are polypeptides with a carboxy–group at one end (C–terminus) and an amino group at the other end (N–terminus) [2]. The N–terminal end of histones, the histone tail, sticks out from the nucleosomes and can be epigenetically modified after translation [6]. Typical histone modifications are acetylation of lysine residues (K), mono–, di– or tri–methylation of lysines and mono– or di–methylation of arginines (R), phosphorylation of serines (S) as well as sometimes threonines (T) and tyrosines (Y), and ubiquitination of lysines [22, 34]. These modifications have different effects on gene expression depending on the modified histone, the modification type and the

type and position of the modified amino acid residue [22, 34]. This resulted in a naming convention whereby the histone is given first, followed by the modified residue type and its position in the histone tail and the modification type. For example, H3K27ac is the acetylation of a lysine residue at position 27 counted from the N–terminus of the tail of histone H3, whereas H3K4me2 describes a di–methylated lysine residue at position 4 in histone H3.

In addition to post–translational modifications, there are variants of the histones H2A, H2B, H3 and H4 that can replace the canonical versions and affect chromatin structure and gene expression [24–26]. As key regulatory factors, both histone variants and histone modifications are contributors to many cellular processes and play a crucial role in cell differentiation and maintenance of cell types [38].

The following subsections go into more detail about histone acetylation and methylation, histone variants, and combinatorial effects of modifications and variants.

2.3.1 Histone Acetylation

Histone acetylation is a transient, covalent modification that is added to histone tail residues by histone acetyltransferases (HATs), which transfer singe acetyl–groups from acetyl–coenzyme A (coA) to lysine [6, 27]. Conversely, they are removed by histone deacetylases (HDACs) and transferred back to coA [6, 27]. Acetylation adds a negative charge to the positively charged lysines, neutralising the positive charge of histone and thus loosening its interaction with the negatively charged ribose–phosphate backbone of the DNA, which is associated with increased chromatin accessibility and contributes to transcriptional activation [6, 27, 36]. Moreover, histone acetylation can be recognised by the bromodomain of transcription factors and contribute to their recruitment [6].

There are various families of HATs and HDACs, specific combinations of which maintain acetylation in different regulatory regions, resulting in the association of different acetylation marks with different genomic regions [6, 36]. For example, histone acetylation marks like H3K9ac and H3K27ac are generally found near the TSS, while marks like H3K4ac or H4K15ac are enriched in the promoter and gene bodies of active genes without peaking at the TSS [36]. Some HATs are not just involved in acetylating specific genes or their regulatory elements, but establish global acetylation patterns that contribute to the general activation of the transcriptional pre–initiation–complex (PIC) [6, 27].

Consequently, histone acetylation is involved in many diverse processes, including nucleosome assembly, myogenesis, DNA replication and repair, transcriptional regulation, the cell–cycle, cell differentiation and apoptosis [6, 27, 28].

2.3.2 Histone Methylation

Similar to histone acetylation, histone methylation participates in a large number of processes, such as transcriptional regulation, cell differentiation, DNA repair, stress response and the cell cycle [4, 28]. It is catalysed by various histone methyltransferases (HMTs) that transfer one to three methyl–groups from S–adenosyl–methionine to lysine residues in histone tails, or one to two

methyl–groups to arginine residues, and be reversed by several histone demethylases (HDMs) [4, 28]. In contrast to transient histone acetylation, its turn–over is much slower and turn–over rates can vary between methylation marks, whereby marks involved in epigenetic inheritance are maintained, whereas marks associated with cell differentiation or environmental signals are more dynamic in nature [4].

Histone methylation likely affects chromatin structure and gene expression by serving as recognition sites for various effectors with methyl–binding domains that can recruit chromatin–remodellers or components of the transcriptional machinery [4, 22, 28]. As a consequence, the effect of methylation marks on gene expression is not nearly as uniform as in the case of histone acetylation, and can depend on the type of the methylated residue and its position in the histone tail, the degree of methylation, the location in various genomic regions, as well as other contextual factors [4, 22].

For instance, H3K4me2 typically occurs in transcriptionally active gene bodies, while H3K4me3 is generally associated with TSSs and 5'–end of active genes, however, both marks can be involved in transcriptional repression if other effector molecules bind that for example facilitate histone deacetylation [4, 22]. In contrast, H3K27me3 is commonly linked to repression, but can have an activating effect if combined with other methylation marks such as H3K4me3 [4].

2.3.3 Histone Variants

The amino acid sequence of histone variants differs in certain places from the canonical versions of H2A, H2B, H3 and H4, which affects their properties and consequently nucleosome and chromatin structure and stability [24–26]. Whereas canonical histones are generally deposited during DNA synthesis of the replication phase (S–phase) of the cell cycle, histone variants can be deposited independent of DNA synthesis throughout the cell cycle [39]. In contrast to canonical histones, which are mainly involved in DNA packaging and gene expression regulation, histone variants contribute to additional processes, such as chromosome segregation, DNA repair, X chromosome inactivation or meiotic recombination [24, 25].

The two most widely studied histone variants are H3.3 and H2A.Z, where the former is associated with transcriptional elongation when it is placed in gene bodies at transcription initiation [25]. H2A.Z is a H2A variant distinct from the other members of that family and similar to H3.3 it has been associated with transcriptionally active euchromatin [24, 25]. Specifically, it tends to be enriched in promoters around TSSs where it correlates with RNA polymerase occupancy, and thus seems to play a role in transcription activation [24, 25]. Additionally, it could contribute to transcriptional activation by hindering promoter DNA methylation, as H2A.Z and DNA methylation were shown to be negatively correlated and knockout of a H2A.Z deposition complex lead to hypermethylation in *Arabidopsis thaliana* [40]. Furthermore, nucleosome stability decreases if both H2A.Z and H3.3 are present, which might contribute to chromatin accessibility [41].

However, it has also been associated with heterochromatin and transcriptional repression, and it has been suggested that this seemingly contradictory effect could be in part a result of differences in H2A.Z acetylation, since H2A.Z is hypoacetylated in heterochromatin while it tends to be acetylated in active

promoters [25]. Closer examination in human cells found that the presence of H2A.Z in promoters aids with the recruitment of the transcriptional machinery during transcription initiation, whereas H2A.Z in gene bodies is associated with inactive transcription and disappears during transcription [42].

Regarding cell differentiation, in murine ESCs, H2A.Z is enriched in silent developmental genes, contributing to their repression, while also being essential for ESC differentiation where it plays an activating role in promoters [43].

2.3.4 Histone Code

Even though there are general patterns for individual histone marks and variants, there are seeming contradictions and contextual differences in effect, as discussed in the previous subsections. Therefore, the concept of a histone code has been proposed whereby histone modifications and variants have a combinatorial effect on chromatin structure and the recruitment of protein complexes involved in chromatin–remodelling and the transcriptional machinery, and thus gene expression [4, 36, 44, 45].

Whereas individual histone marks or variants are often not sufficient to accurately predict gene expression, combinatorial approaches have been quite successful. For instance, the Encyclopedia of DNA Elements (ENCODE) Project Consortium (2012) predicted gene expression based on promoter histone modifications and variants by first predicting if genes were expressed (classification) and then to which degree (quantitative regression) [35]. They were able to achieve a strong Pearson correlation of $r = 0.9$ between predicted and actual gene expression, with an area under curve (AUC) of 0.95 for the classification. Furthermore, they found that some marks or variants were much more informative than others, whereby H3K9ac, H3K4me3, H3K4me2, H3K27ac, H3K79me2 and H2A.Z were the most informative in predicting if a gene was expressed at all, and H3K79me2, H3K9ac, H3K4me3, H3K36me3 and H3K27ac for subsequent prediction of the degree of the expression.

Methods have also been developed to predict differential gene expression between two states based on selected differential histone modifications, which proved more difficult than prediction within a single condition [46]. One of the latest approaches is DeepDiff by Sekhon et al. (2018), which outperforms previous methods by achieving Pearson correlations between 0.4 and 0.8 between predicted and actual differential gene expression based on H3K4me1, H3K4me3, H3K36me3, H3K9me3 and H3K27me3, depending on the compared cell types [46].

2.4 Transcription Factors

Transcription is partly controlled by transcription factors (TFs), which are small proteins that can bind to the promoter and to enhancer regions, where they contribute to transcriptional activation or repression in variety of way, as reviewed in detail by Lambert et al. (2018) [3]. For instance, they can recruit chromatin–remodelling proteins that alter the accessibility of the DNA for the transcriptional machinery. Some of them also directly recruit the RNA polymerase or strengthen its ability to bind to the proximal promoter region, thus facilitating transcription initiation. Additionally, transcription factors are part

of the transcription initiation complex together with the RNA polymerase and other co–factors, and the mediator complex in between. On the other hand, some transcription factors interfere with transcription by blocking important binding sites in the promoter, or by recruiting co–repressors.

Transcription factors play important roles in many processes, such as functioning as master regulators in cell differentiation transitions [10], as core components of the circadian pacemaker [47], cell cycle progression [48] or tumour suppression [11], to name just a few. As a result, dysregulation involving transcription factors can lead to problems for individual cells and the entire organism, for example the disruption of tumour suppressors can contribute to the development of cancer [11].

2.5 MicroRNAs

MicroRNAs (miRNAs) are about 22 bp long, non–coding RNAs that are encoded within introns, or sometimes even exons, of host genes or as their own genes [49, 50]. Canonically, transcription initially yields a much longer primary, single–stranded transcript, the primary–microRNA (pri–miRNA), which can contain the precursors for one or more miRNAs [50]. In the nucleus, the pri–miRNA is cleaved by the microprocessor complex Drosha into hairpin–shaped precursor–microRNAs (pre–miRNAs), which are subsequently exported to the cytoplasm, where they are cleaved into double–stranded, mature miRNAs by the Dicer complex [50]. Some miRNAs are encoded by small introns, so called mirtrons, that result in a pre–miRNA after splicing without requiring processing by the *Drosha* complex [51].

In the cytoplasm, the mature miRNA duplex is separated into the active and the passenger strand, the latter of which is degraded [50]. The active strand is incorporated into the RNA–induced silencing complex (RISC) and guides it to complementary target mRNA transcripts, which are then degraded by the Argonaute component if fully complementary [8]. In cases where the miRNA is only partially complementary with the mRNA, RISC can interfere with translation by blocking the translational machinery [8], or by destabilising the mRNA through removal of the 3'–poly–adenosine tail at 3'–end, which marks it for degradation [9].

Consequently, miRNAs silence gene expression, with target sites in more than 60% of human protein–coding genes [13], and are important regulators of cell growth, differentiation and proliferation [14], apoptosis [14] and signalling [12]. Disruption of miRNA synthesis or function has been associated with many diseases, ranging from cardiovascular and neurodegenerative diseases to cancer, a compilation of which can be found in the Human microRNA Disease Database (HMDD) [52].

2.6 TF–miRNA Co–Regulatory Motifs

Initially, the regulatory roles of TFs and miRNAs were studied separately, however, examination of gene regulatory networks revealed that they can co–regulate target genes and form network motifs [7]. In addition to simple co–regulation, where the TF and miRNA do not regulated one another, this can

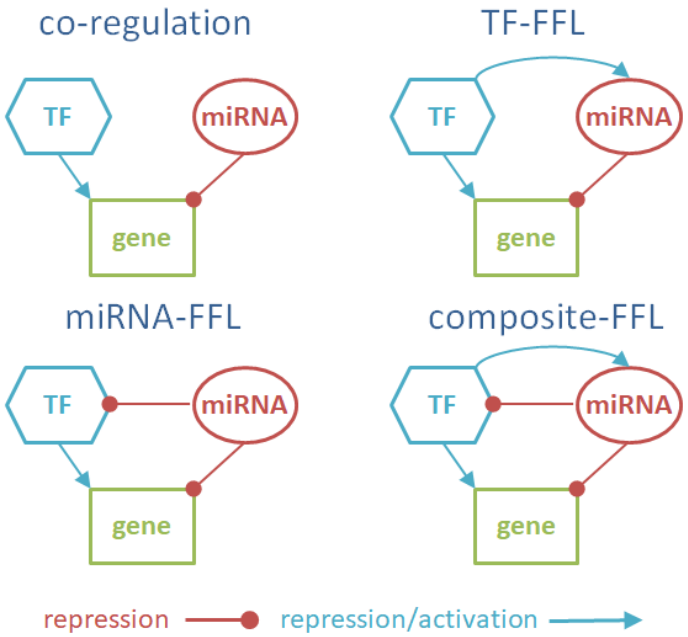

Figure 2.3: *Transcription factor (TF) and miRNA co–regulatory motif types, three of which form feedforward loops (FFLs). Adapted from Hamed et al. (2015) [15].*

take the form of feedforward loop (FFL) in which at least one of them regulates the other (Figure 2.3).

The TF–miRNA–FFLs can be further distinguished into three cases based on the interaction between the TF and the miRNA: (i) miRNA–FFLs in which the miRNA the miRNA represses the TF without being regulated in turn by the TF, (ii) TF–FFLs in which the TF regulates the miRNA and the miRNA does not repress the TF, and (iii) composite–FFLs in which the miRNA represses the TF while being regulated by the TF [7, 15].

In contrast to the repressive function of the miRNA, the regulation of the miRNA or the shared target genes by the TF can be activating or repressing. As a consequence, TF–FFLs can indirectly reinforce the direct effect of the TF, called coherent FFLs, when for example a repressor activates a co–regulatory miRNA or when an activator represses the miRNA. On the other hand, TF–FFLs can be incoherent, where the direct and indirect regulation are contradictory, such as FFLs in which an activator activates a repressive co–regulatory miRNA or a repressor represses the miRNA. Similarly, a miRNA–FFL is coherent when the miRNA represses a co–regulatory activator, and incoherent when it represses a repressor.

If the TF and the miRNA can regulate a shared target gene on their own, similar to a logical “OR” relationship, then coherent co–regulatory FFLs facilitate rapid or strong regulatory effects [7, 16, 18]. If both the TF and the miRNA are required for the regulation of a shared target gene, like a logical “AND”, then the slightly delayed nature of the indirect regulation in coherent FFLs guards against transient signals, only reacting to persistent ones [18]. Incoherent TF–miRNA co–regulatory FFLs are much rarer than their coherent

counterparts, and seem to play a role in noise–buffering against stochastic signalling, fine–tuning of expression or maintaining protein steady states and thus homeostasis [16, 17].

TF–miRNA co–regulatory motifs have been identified as important regulatory mechanisms in many processes, including the cell cycle, development and cell differentiation, and also play roles in disease–specific regulatory networks, such as cancer, schizophrenia or interstitial lung disease [7, 17, 19].

In addition to specific studies, more general services for TF–miRNA co–regulatory motif analyses have recently been developed. One of them is the webserver TFmiR by Hamed et al. (2015), which constructs a disease–specific TF–miRNA co–regulatory network based on de–regulated mRNA and miRNA input, and on that basis identifies network motifs and key players in combination with over representation and functional similarity analyses [15]. Another example is the webtool CMTCN by Li et al. (2018) that allows users to explore cancer–specific TF–miRNA co–regulatory motifs with incorporated cancer gene expression data, and performs network topology and enrichment analyses [20].

Chapter 3

Materials

The previous chapter introduced the basics of gene expression and involvement in cell differentiation and human development. Furthermore, the regulatory roles of histone modifications and histone variants as well as TF–miRNA co–regulatory motifs were explained, which have been studied separately, so far. The goal of this thesis was to explore the relationship of gene expression and histone modifications and histone variants in human TF–miRNA co–regulatory motifs. Due to the involvement of both the motifs and the histone modifications and variants in cell differentiation, it was decided to conduct the exploration in cell differentiation transitions. This chapter lists and explains the different types of data used to accomplish this goal.

First, identification and exploration of TF–miRNA co–regulatory motifs required gene regulatory information. Furthermore, a human reference genome and transcriptome with corresponding gene definitions was required in order to establish the genomic context of gene expression and histone modification or variants measurements. In the same vein, promoter regions were defined to study histone modifications or variants in promoter regions and gene bodies separately. Lastly, gene expression data and histone modification and variant data of different cell types involved in cell differentiation transitions was needed. Table 3.1 gives an overview of the major data sources and types of information used in this thesis, while the following sections provide detailed explanations.

3.1 Gene Regulatory Information

For the purpose of identifying and analysing TF–miRNA co–regulatory motifs, gene regulatory information was required. There are several large databases that offer different types of human gene regulatory data, many of which were combined in the recently updated internal database of TFmiR [15].

TFmiR is a webserver that conducts analyses on networks constructed from sets of de–regulated mRNAs and miRNAs, including co–regulatory TF–miRNA motifs. As such its regulatory database was a well suited basis for the motif analyses conducted in this thesis. An overview of the sources integrated in that database can be found in Table 3.2. Only interactions labelled as experimentally verified were utilised, yielding 226,156 regulatory interactions between 22,308 genes.

Table 3.1: *Data sources and materials used in this thesis.*

Source	Information	References
Ensembl	reference transcriptome and gene information for GRCh38: ID, name, type, chromosome, strand, start and end, TSS	[53]
FANTOM5	miRNA host genes	[54, 55]
MiRIAD	miRNA host genes	[56, 57]
ENCODE	RNA–seq and histone modification or histone variant ChIP–seq data	[35, 58]
TFmiR	gene regulatory information	[15]
NCBI	GRCh38 chromosome lengths	[59]

Table 3.2: *Overview of the databases and other sources comprising the human gene regulatory database obtained from TFmiR [15]. "Gene" in the interaction type column is a general term that can include TFs and miRNAs, as well as other types of genes.*

Interaction Type	Source	Validation	Version	Reference
TF → gene	TransFac	experimental	v11.4	[60]
	ORegAnno	experimental	v3.0, Nov. 2016	[61]
	TRRUST	experimental	v2	[62]
	TRED	predicted	2007	[63]
TF → miRNA	TransmiR	experimental	v2.0, Oct. 2018	[64]
	PMID20584335	experimental	Apr. 2009	[65]
	ChIPBase	predicted	v1.1, Nov. 2012	[66]
miRNA → gene	miRTarBase	experimental	v7.0, Sep. 2017	[67]
	TarBase	experimental	v7.0	[68]
	miRecords	experimental	Apr. 2013	[69]
	starBase	predicted	v3.0, 2018	[70]
miRNA → miRNA	PmmR	predicted	Mar. 2011	[71]
gene → gene	mentha	experimental	June 2019	[72]
	STRING	predicted	v11.0	[73]

3.2 Reference Genome and Transcriptome

The differential gene expression and differential histone modification or variant analyses performed in the course of this thesis required a genomic reference. The Genome Reference Consortium (GRC) currently maintains the two most commonly used human reference genomes, namely GRC Human Build 37 (GRCh37) (released in February 2009) [31] and GRCh38 (released in December 2013) [32], which are also known as human genomes 19 (hg19) and 38 (hg38), respectively. For this thesis GRCh38 was chosen since it is the newest of the two, specifically patch 12 (released in December 2017).

Gene definitions for GRCh38 were obtained from Ensembl (release 96) [53], a genome browser for vertebrates. Due to alternative splicing in humans, a single gene can result in several different transcripts [74], and Ensembl provides these definitions on the transcript– and the gene–level. Even though it would have been possible to conduct the differential gene expression and differential histone modification or variant analyses on the transcript–level, regulatory information was only available on the gene–level (see Section 3.1). Therefore, gene–level definitions were chosen, in particular Ensembl's internal ID, commonly used name, type, description, as well as the start and end position on the respective chromosome and strand, and TSS for each gene.

The gene starts provided were the start position of the 5'–most transcript associated with the respective gene in Ensembl's database, while the gene ends were the 3'–most transcript end. For miRNAs, the available start and end positions corresponded to the processed pre–miRNA, not the longer pri–miRNA. The corresponding TSSs obtained from Ensembl were equivalent to the 5'–gene start on the particular strand.

Additionally, a matching human reference transcriptome was needed for the chosen RNA–sequencing (RNA–seq) quantification method during the differential expression analysis. The GRCh38 reference transcriptome was obtained in FASTA–format from Ensembl as well, and included entries for 226,365 transcripts for both coding and non–coding genes. For each transcript entry, the Ensembl transcript ID, corresponding Ensembl gene ID, genomic location on the respective chromosome, as well as transcript and gene type were given in the header, followed by the nucleotide sequence.

Lastly, the chromosome lengths of GRCh38 path 12 were needed for the histone modification and histone variant analysis, and were obtained from the National Center for Biotechnology Information (NCBI) [59].

3.3 Promoter Definitions

Histone modifications and histone variants can have different effects depending on their location in promoter regions and gene bodies [28]. For this reason, promoters and gene bodies were examined separately in this thesis. Consequently, promoter definitions were needed in addition to the gene definitions discussed in the previous section.

Initially, the plan was to use one of the several databases that provide experimental information on human promoters or TSSs, namely TSSs from the Eukaryotic Promoter Database [75] as well as FANTOM5 [54] promoters, which could

be accessed via Ensembl. However, there were several issues with integrating the promoter or TSS information with the available gene definitions.

The first problem was correctly associating the promoter or TSS definitions with the gene definitions. For instance, the FANTOM5 promoters obtained from Ensembl were only provided with genomic positions, not with associated gene IDs, and consequently could not be matched unambiguously to the gene definitions. Similarly, the TSSs from Eukaryotic Promoter Database (EPD) were provided with general gene names not IDs, which could sometimes be matched to several Ensembl gene IDs and thus multiple gene definitions. Additionally, the 16,455 EPD–TSSs only covered 28% of the 58,788 Ensembl gene definitions.

Examining the position of the EPD–TSSs relative to the corresponding Ensembl–TSSs revealed that 80% of EPD–TSSs were downstream of the matching Ensembl–TSS (Table 3.3). Since the Ensembl–TSSs were defined as the start of the 5'–most transcript, it was suspected that these downstream–positioned EPD–TSSs might correspond to some of the other transcripts. This was confirmed after acquiring the Ensembl transcript–level definitions, and matching the Ensembl transcript TSSs to the EPD–TSSs.

Lastly, there were problematic cases that seemed unlikely to be a correct match due to the large distance between the EPD TSS and the gene body. For example, there were instances where the EPD–TSS was up to 153,388,201 base pairs upstream from the corresponding Ensembl–TSS (Table 3.3), while the human chromosomes range in length between 50,818,468 and 248,956,422 base pairs [59].

On account of the insufficient coverage and integration issues, it was decided to instead make use of the Ensembl gene–level TSSs and to homogeneously define the promoter region of each gene as an interval of fixed length around its TSS.

Commonly used promoter definitions in humans include -900 to +100 base pairs relative to the TSS [76], ± 2,000 base pairs [77], ± 2,500 base pairs [78, 79], and -2000 to +500 base pairs [80]. The latter was chose for this thesis since it encompasses the region in which histone methylation and acetylation marks peak near the TSS [80] while only having a small overlap with the gene body, which aids in distinguishing between histone modifications and variants in promoter regions versus the gene body on gene expression.

Table 3.3: *Position of EPD TSSs in base pairss relative to the corresponding Ensembl TSSs broken down by strand and relative location. For each category, the number of EPD TSSs (n), closest (Min.), median, mean and furthest (Max.) relative position is given, as well as the standard deviation (SD).*

Strand	Location	n	Min.	Median	Mean	Max.	SD
plus	upstream	1,919	0	-1	-86,617	-153,388,201	3,501,606
	downstream	6,113	1	107	10,011	2,072,490	8,795
minus	upstream	792	0	-105	-20,501	-858,189	839,677
	downstream	7,081	1	59	10,011	2,072,490	58,795

However, another problem arose with regard to miRNA promoters, namely that many miRNAs are not intergenic and do not possess their own promoters. Instead, intragenic miRNAs can be located alone or in clusters within introns of their host genes, or more rarely within exons, and are consequently expressed via the promoter of their host gene [49].

Information on miRNA host genes was obtained from the FANTOM5 miRNA atlas [54, 55] and miRIAD [56, 57], a database for intergenic and intragenic miRNAs. Both databases provided information on slightly different sets of miRNAs but were congruent for the intersection consisting of 950 miRNAs. In total, 1,179 of the 1,879 (62.7%) miRNAs contained in the Ensembl gene set were classified as intragenic. The promoter region of those intragenic miRNAs was set to that of their respective host gene.

3.4 Expression and Modification Data

For the purpose of studying the correlation between differential gene expression and differential histone modifications or variants in TF–miRNA co–regulatory motifs, it was decided to use RNA–seq data for gene expression and chromatin immunoprecipitation DNA–sequencing (ChIP–seq) data for histone modifications and histone variants.

RNA–seq is a high–throughput technology for genome–wide transcriptome profiling that offers several advantages over older, commonly used approaches such as RNA microarrays, where transcripts are measured through hybridisation to complementary probes on microarray chips [81, 82]. While the microarray approach is cost–effective and easy to analyse, it struggles with low abundance transcripts, hybridisation issues, and limitation to the specific sequences on the chip [82]. In contrast, RNA–seq consists of first creating a library from the RNA transcripts that is subsequently sequenced with next–generation sequencing (NGS) [83], which offers advantages such as not being limited to specific sequences, larger detection ranges, less background noise, and higher accuracy and reproducibility [82, 83].

To make the RNA transcripts suitable for NGS, longer transcripts are fragmented, all transcripts and fragments are then reverse transcribed to complementary DNA (cDNA), and finally tagged with adaptors of known sequence [83], which serve to anchor the fragments during amplification and sequencing [84]. The sequenced reads can then be aligned or mapped to a reference genome or reference transcriptome in order to estimate transcript abundance [83]. If the fragments are only sequenced from one end to the other, the resulting reads are called single–end, whereas paired–end reads were sequenced from both ends of the fragment [82, 83]. The latter provides additional information and accuracy for studying for instance splice variants or allele specific expression, however, for the gene–level analyses conducted in this thesis, it does not make much of a difference [82].

As reviewed in Park (2009) [85], for histone chromatin immunoprecipitation, the chromatin is fragmented and treated with a specific antibody that binds to the histone modification or histone variant of interest. The antibody tagged histone–DNA complexes are then extracted via precipitation, and the bound DNA fragments isolated. Similar to RNA–seq, ChIP–seq utilises NGS instead of

microarrays to determine the sequence of these DNA fragments, thus conferring similar advantages.

The Encyclopedia of DNA Elements (ENCODE) [35, 58] was chosen as the source for gene expression and histone modification and histone variants data in cell differentiation transitions, specifically, since it offers RNA–seq gene expression data sets for a large number of human cell and tissue types, as well as matching ChIP–seq data sets for various histone modifications and variants. Additionally, ENCODE data sets were produced by uniform and well–documented processing pipelines to ensure reproducibility, quality and consistency [35, 58].

The database was searched to identify cell and tissue types for which RNA–seq samples and at least one histone ChIP–seq sample was available, after excluding data sets that contained errors or were not compliant with ENCODE's quality audit, for example due to insufficient read depth. From the resulting cell and tissue types, only those were selected that were part of a cell differentiation transition with another cell or tissue type in the pruned set, and additionally had ChIP–seq data for at least one histone modification or histone variant in common. Table 3.4 lists the number of samples available for each cell type.

This yielded six cell differentiation transition consisting of nine cell types, an overview of which can be found in Figure 3.1. The first one was the transition from the pluripotent H1–human embryonic stem cell (hESC) cell line to the induced pluripotent stem cell (iPSC) GM23338 cell line. Another transition was from multipotent mesenchymal stem cells (MSCs) to terminally differentiated bone cells (osteoblasts). The third one was a haematopoietic transition from multipotent myeloid progenitor cells (MPCs) to monocytes, a type of white blood cell.

Finally, the remaining three cell types were involved in neural cell differentiation: from multipotent neural stem progenitor cells (NSPCs) to neural progenitor cells (NPCs) to terminally differentiated bipolar neuron cells. In order to examine the differences between direct and more indirect cell differentiation transitions, neural stem progenitor cells and bipolar neuron cells were analysed as well.

The RNA–seq data sets were obtained as paired–end or single–end NGS reads in FASTQ–format, while the ChIP–seq data sets were acquired as already pre–processed and quality controlled alignments to the GRCh38 reference genome in binary sequence alignment/map (BAM)–format.

Detailed information on the individual RNA–seq and ChIP–seq data sets obtained from ENCODE, including experiment and file accessions, can be found in Tables A.1 and A.2 in the appendix, respectively.

Table 3.4: *Number of biological samples for which suitable RNA–seq and histone ChIP–seq data was available in ENCODE.*

		Histone ChIP–seq			
Cell Type	**RNA–seq**	**H2A.Z**	**H3K4me2**	**H3K4me3**	**H3K27ac**
bipolar neuron	6	2	2	2	2
GM23338	6	2	2	2	2
H1–hESC	16	2	2	3	2
mesenchymal stem cell	4	1	–	–	–
monocyte	5	–	–	4	3
myeloid progenitor	1	–	–	5	1
neural progenitor cell	4	2	2	2	2
neural stem progenitor cell	2	1	–	6	6
osteoblast	4	2	–	–	2

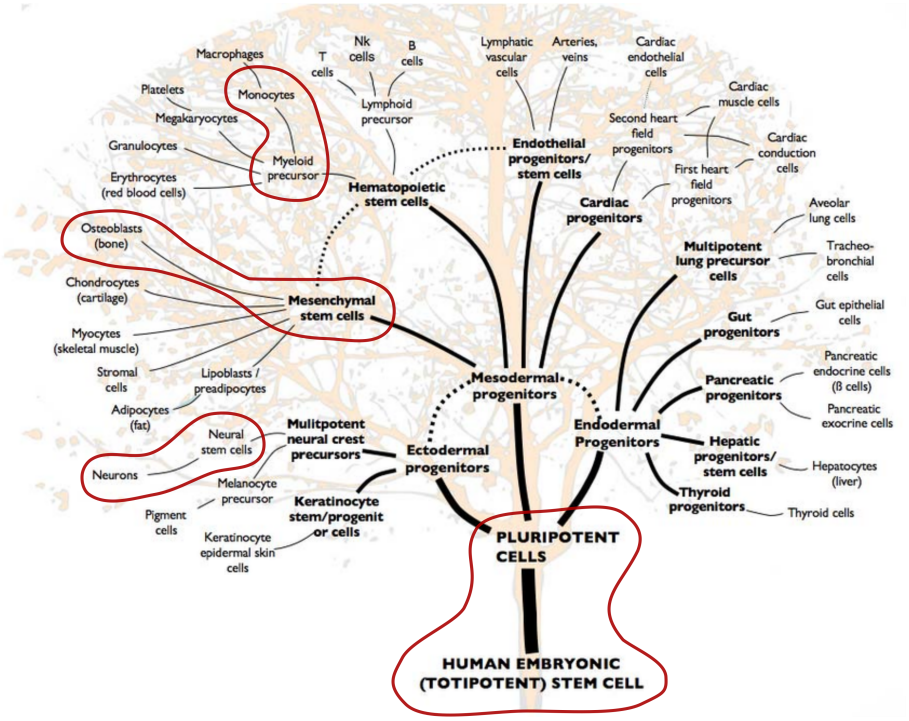

Figure 3.1: *Human cell differentiation tree, starting with embryonic stem cells as the root and branching out to terminally differentiated cell types. Differentiation transitions examined in this thesis are marked in red. Altered from Winickoff et al. (2009) [30].*

Chapter 4

Methods

The purpose of this thesis was to explore the correlation of gene expression and histone modifications or histone variants in human TF–miRNA co–regulatory motifs. Specifically, the differential gene expression of six cell differentiation transitions was studied, including ESC, neural, bone cell and blood cell differentiation, with respect to the three histone modifications H3K4me2, H3K4me3 and H3K27ac, and the histone variant H2A.Z.

Chapter 2 introduced background information on gene expression, as well as the regulatory roles of these co–regulatory motifs and histone modifications and variants. The materials used in this thesis, such as reference genome and transcriptome, gene and promoter definitions, gene regulatory information and expression and modification data sets, were discussed in Chapter 3.

First, differential gene expression analysis and differential histone modification or variant analysis was performed, in order to assess changes in gene expression and histone modifications or variants during cell differentiation transitions. Next, gene regulatory networks were built for the examined transitions and searched for TF–miRNA co–regulatory motifs. Then, the correlation between differential gene expression and differential histone modifications or variants was examined genome–wide as a baseline, as well as in genes involved in different types of TF–miRNA co–regulatory motifs, whereby histone modifications and histone variants in promoters and gene bodies were investigated separately. Furthermore, the correlations in TF–miRNA co–regulatory motifs were compared to those of motifs in randomised networks. Moreover, functional annotation enrichment analysis was conducted. The last section describes how the work conducted in this thesis was implemented.

4.1 Differential Gene Expression Analysis

In this thesis, gene expression was appraised based on RNA–seq data sets from various samples of different cell types (Table 3.4). In a first step, the RNA–seq data was quantified for each sample. For the purpose of assessing the change in gene expression for each cell differentiation transition, the quantified samples were subsequently grouped by cell type and statistically compared between the respective cell types. For each transition, this produced the log_2–transformed

fold change in expression for each gene, as well as an assessment of the statistical significance of that change.

4.1.1 RNA–Seq Quantification

There are generally speaking two major groups of RNA–seq quantification approaches: alignment–based and alignment–free methods [86, 87]. As the name suggests, the first group computes a full alignment of the RNA–seq reads to the reference genome, and then counts reads mapped to genomic features, such as genes, based on the resulting alignment files [88]. There are many tools for both of these tasks, resulting in many possible alignment–based RNA–seq quantification pipelines [88]. A popular pipeline employs the aligner *STAR* [89] coupled with *featureCount* [90], which tends to perform well [88]. However, alignment–based RNA–seq quantification pipelines suffer from long run–times [86, 87], which is problematic if larger numbers of samples need to be quantified.

The second, more recent group of approaches foregoes the time–consuming alignment step, for example by using k–mer counting (*Sailfish* [91]), pseudo–alignments (*Kallisto* [92]), or quasi–alignments (*Salmon* [93]). This results in considerable gains in speed as well as reduction in required memory [91–93], while at the same time maintaining the accuracy of alignment–based approaches [86, 87]. Consequently, an alignment–free RNA–seq quantification method was selected to quantify the large number of RNA–seq samples examined in this thesis.

Specifically, *Salmon* was chosen, since it outperforms the other alignment–free methods in speed due to multiprocessing, and additionally accounts for several biases and input scenarios [93]. Before the actual quantification, *Salmon* builds a lightweight index of the given reference transcriptome for mapping, which can be re–used for subsequent quantification runs. It consists of a suffix array of the transcriptome, as well as a hash table that maps all sub–sequences of length k (k–mer) in the reference transcriptome to the corresponding interval in the suffix array [93, 94].

In this thesis, the index for the GRCh38 reference transcriptome was constructed with the default $k = 31$, which is appropriate for RNA–seq read lengths of roughly 50–75 base pairs or longer [95], thus fitting the reads used in this thesis, most of which were 76 or 101 base pairs long and none shorter than 51 base pairs. Indexing was accomplished with the following command:

```
salmon index -t <transcriptome_file> -i <index_file>
```

Salmon quantifies RNA–seq reads in two steps. During the first step, transcript abundance is estimated using a quasi–mapping approach where the index is used to find the possible locations of all k–mers of a given read in the reference transcriptome, based on which the most likely origin transcripts of that read are returned [93, 94]. Additionally, various bias models and parameters are learned during this phase. On that basis, the second step further refines and corrects the abundance estimates [93].

In this thesis, reads from both single–end and paired–end RNA–seq samples had to be quantified, which required slightly different input specifications for *Salmon*'s quantification command. The RNA–seq read files were grouped for each sample, and matched with corresponding read files in the case of samples

originating from paired–end runs. Given a single–ended sample with n files $f_1, \ldots, f_n$, the quantification was executed as follows:

```
salmon quant
    -i <index_file> -l A -r <f_1 ... f_n>
    -o <output_directory> -p <number_cores>
    --gcBias --seqBias --incompatPrior 0.0
    --validateMapping
```

Similarly, a paired–end sample with n file pairs (a_i, b_i), $i \in \{1, \ldots, n\}$, was quantified with the following command:

```
salmon quant
    -i <index_file> -l A
    -1 <a_1 ... a_n> -2 <b_1 ... b_n>
    -o <output_directory> -p 16
    --gcBias --seqBias --incompatPrior 0.0
    --validateMapping
```

The parameter `-l` controls how *Salmon* interprets the subsequently provided read file(s). In this case, it was set to automatically detect the run–type (paired–end or single–end), as well as the strand–specificity, or lack thereof. The number of threads used to speed up quantification via multithreading could be set with `-p`. To only consider read–transcript mappings that were compatible with the RNA–seq library type, `--incompatPrior` was set to 0.

Transcript abundance estimation can be confounded by biases resulting from non-uniform and non-random fragmentation of transcripts during preparation of the RNA–seq library, as well as during selection for sequencing [96, 97]. In particular, the beginning and end of transcripts tend to show higher fragment coverage than the rest (positional bias) [96]. In addition, the likelihood of fragments being selected for sequencing is affected by the sequence near their 5'– and 3'–end (sequence–specific bias) [96], as well as their GC–content (GC bias) [97].

The `--posBias`, `--seqBias`, and `--gcBias` flags instruct *Salmon* to learn and correct transcript abundance estimates for these biases in the input RNA–seq reads. Since positional bias correction was still experimental as of writing this thesis, only sequence–specific and GC bias correction was enabled. For the purpose of increasing accuracy, *Salmon* was instructed to run additional checks on the read to transcript mappings via the `--validateMapping` flag.

For each RNA–seq sample, this resulted in a main file containing estimated read counts for each transcript in the reference transcriptome, as well as the bootstraps and various statistics collected during the quantification run.

4.1.2 DESeq2

For the purpose of rigorously assessing the change in gene expression between two biological conditions, it is not sufficient to simply compare RNA–seq read counts of each transcript or gene due to biological variance between and within samples of each state and between states [98, 99], as well as technical variability [100]. Additionally, a change in expression in itself is not necessarily significant, given the possibility that it could be the result of naturally occurring random variation [99].

There are several differential gene expression analysis tools that address these problems with a variety of different statistical models and methods, including

negative binomial distributions (e.g. *edgeR* [101], *DESeq2* [102]), linear models (e.g. *limma* [103]), and non–parametric approaches (e.g. NOIseq [104], SAMseq [105]). Of these, *DESeq2* was chosen for this thesis based on its consistently high true positive rate and accuracy [106].

In preparation of the differential gene expression analysis, the transcript–level abundance estimates produced by the previously described RNA–seq quantification were summarised to the gene–level for each RNA–seq sample with *tximport* [107]. The required transcript ID to gene ID mapping was generated from the GRCh38 reference transcriptome that contained Ensembl transcript and corresponding gene IDs for each transcript.

For each cell differentiation transition from cell type A to cell type B, *DESeq2* analysed the gene expression counts of all samples of A in comparison to those of B. For each gene, this produced the log_2–transformed fold change in expression for the transition from A to B, as well as a p–value describing the statistical significance of this change in expression. By default, *DESeq2* uses the Benjamini–Hochberg procedure for multiple hypothesis testing correction and additionally yields the corrected p–value. A more detailed discussion of multiple hypothesis testing and Benjamini–Hochberg can be found in Section 4.4.3. However, for this thesis independent hypothesis weighting (IHW) was used due to its increased statistical power, especially for high–throughput genomic applications such as differential gene expression analysis [108]. The false discovery rate (FDR) of the differential gene expression analysis was controlled at $\alpha = 0.01$ in this thesis, and genes with corrected p–value ≤ 0.01 were thus considered to be differentially expressed.

4.2 Differential Histone Modification Analysis

Similar to differential gene expression analysis, differential histone modification analysis consists of first quantifying the modifications in the compared samples, and then statistically evaluating the modification or variant difference of genomic regions [109]. Approaches include comparing ChIP–signal peaks, locally weighted regression and hidden Markov models (HMMs) [109].

For this thesis, *HistoneHMM* [110], a HMM method, was chosen since it is comparatively fast, outperforms other differential histone modification analysis tools in terms for true positive to false positive rate, was tested on ChIP–seq data from ENCODE [110].

HistoneHMM partitions the reference genome into n equally sized bins b_i of length m, with $i \in \{1, \ldots, n\}$. Given two ChIP–seq files A and B that were aligned to the reference genome, a bivariate HMM is constructed, comprised of four modification states s: a bin is modified in A and B, only modified in A, only modified in B, or modified in neither A nor B.

This HMM allows transitions from each state to all other states, as well as self–transitions. A series of bins consecutively occurring in the genome can be viewed as a series of transitions from one modification state to another (or staying in the same state), each of which has a transition probability. However, it is unknown (hidden) in which state each bin is. The goal is consequently to use the known paired read counts of A and B to identify the most probable modification state of each bin.

To that end, a bivariate distribution of the observed read counts in A and B is computed, based on which the emission probabilities are estimated that describe the probability of observing a given read count pair given an underlying modification state, followed by an estimation of the transition probabilities. Finally, the hidden state probabilities $p_{s,i}$ are inferred for each bin b_i, and b_i is classified according to the state s with the highest $p_{s,i}$.

The acquired ChIP–seq data sets were already quality–controlled and aligned to the GRCh38 reference genome by the ENCODE processing pipeline. However, the samples belonging to each combination of cell type and histone modification or variant had to be indexed and merged into a single ChIP–seq alignment file before conducting the differential histone modification analysis, which was accomplished with *SAMTools* [111].

For each cell differentiation transition from cell type A to B, *histoneHMM* was then applied to each histone modification or variant for which both A and B had available ChIP–seq data. *HistoneHMM* offers the option to train the parameters on corresponding gene expression data, if provided, which leads to improved performance of the classification. However, since the aim of the thesis was to analyse the correlation between differential gene expression and differential histone modifications, using gene expression to estimate histone modifications would have confounded that analysis, and *histoneHMM* was consequently run without providing the available gene expression data. Otherwise the default parameters were selected that had also been used to evaluate *histoneHMM* on ChIP–seq data from ENCODE [110], namely bin size $m = 1,000$ base pairs.

The results for each combination of cell differentiation transition from A to B and histone modification or variant included the genomic coordinates and modification classification of each bin. Additionally, consecutive bins with the same classification as either differentially modified in A or differentially modified in B were combined into a single histone mark, thus capturing marks that were longer than the given bin size. These marks were then mapped onto the reference promoters and gene bodies by computing the intersection of the corresponding genomic ranges with *BEDTools* [112].

4.3 Gene Regulatory Networks

Prior to identifying and analysing TF–miRNA co–regulatory motifs, a gene regulatory network had to be constructed from the available human regulatory data. In that network, individual genes were represented by network nodes, while the directed edges embody regulatory interactions between regulators and targets. Genes that neither regulated another gene nor were regulated by at least one gene were removed from the network.

However, the differential gene expression analysis described in Section 4.1 did not provide gene expression information for all genes in the Ensembl gene definition set, nor for all genes comprising the overall gene regulatory network. Furthermore, the set of genes for which the differential gene expression analysis provided expression information differed between the examined cell differentiation transitions.

Since analysing the correlation between differential gene expression and differential histone modifications required gene expression information, it had to be

ensured that gene expression information was available for all genes involved in TF–miRNA co–regulatory motifs in the respective cell differentiation transition. This could be accomplished in a number of different ways:

One option was to prune the gene regulatory network such that only nodes were included for which gene expression data was available in all six cell differentiation transitions. This approach, while simplifying subsequent analyses, posed the problem of losing information that might be relevant for cell differentiation specific differences.

A second option consisted of first identifying TF–miRNA co–regulatory motifs in the overall network, and subsequently extracting a set of such motifs for each cell differentiation transition that only contained those motifs for which gene expression information was available for both the TF and the miRNA, as well as for at least one shared target genes of the two.

However, removing a large number of TFs, miRNAs and other genes due to lacking gene expression information can be viewed as creating a pruned, cell differentiation transition specific sub–network of the overall regulatory network. As explained in more detail in Section 4.4, TF–miRNA co–regulatory motifs were defined as those TF–miRNA pairs that shared a statistically significant number of target genes. Consequently, TF–miRNA co–regulatory motifs that share a significant number of target genes in the overall network might not do so in cell differentiation transition specific sub–networks, if the respective TF or miRNA occur in the sub–network at all, while other motifs might be significant in the sub–network but not in the overall network.

In order to better capture the cell differentiation transition specific regulatory motifs, it was decided to first construct a sub–network for all cell differentiation transitions, made out of all genes with available expression information in that transition, and to then perform a motif search on each sub–network. Furthermore, an additional sub–network comprised of only significantly differentially expressed genes was built for each differentiation transition.

4.4 TF–miRNA Co-Regulatory Motifs

Identifying TF–miRNA co–regulatory motifs in gene regulatory networks consists of first finding all TF–miRNA pairs that share one or more target genes, and then selecting all such pairs for which the number of shared targets is statistically significant [15, 113].

The following subsections describe in detail how co–regulatory TF–miRNA pairs were identified, assessed for statistical significance while correcting for multiple hypothesis testing, and finally how the selected motifs were classified.

4.4.1 Finding Co–Regulatory Pairs

The gene regulatory networks constructed in Section 4.3 were each internally stored as an edge list, as well as a set of adjacency lists that contained the target gene(s) of each gene node. Additionally, they contained a set of transcription factors and a set of miRNAs.

On that basis, a list of all TF–miRNA pairs with at least one shared target gene was then generated by iterating over all pairs in the Cartesian product of the transcription factor set and the miRNA set. For each pair, the intersection

of the adjacency lists of the respective transcription factor and miRNA was computed. All transcription factor and miRNA pairs with a non–empty set of shared target genes were potential candidates for co–regulatory motifs.

4.4.2 Statistical Significance of Co–Regulatory Pairs

Given T transcription factors and M miRNAs in a gene regulatory network, the approach described above generates up to $T \times M$ TF–miRNA co–regulatory pairs. However, the number of shared target genes for these pairs might not necessarily be noteworthy given the number of genes regulated by the transcription factor, the number of genes regulated by the miRNA as well as the total number of genes in the network that are regulated by at least one transcription factor or miRNA.

Thus, for the purpose of this work, TF–miRNA co–regulatory motifs were considered to be only those TF–miRNA pairs that shared a statistically significant number of target genes. That statistical significance was assessed with the hypergeometric test, following the motif finding approach used by TFmiR [15].

Let N be the number of all genes targeted by at least one transcription factor or miRNA in the network. Given a co–regulatory pair consisting of a transcription factor t with n target genes and a miRNA m with K target genes that share $x \in \{0, \ldots, \min(n, K)\}$ target genes, the probability of t and m sharing x target genes under the hypergeometric distribution is:

$$P(X = x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} \tag{4.1}$$

However, it is not sufficient to just consider the probability to have exactly x shared target genes in order to assess the statistical significance of t and m sharing x target genes. Instead, the probability of sharing at least x target genes needs to be considered to account for the possibility that t and m could share more than x target genes by chance. That probability can be calculated in two equivalent ways:

$$\begin{aligned} P(X \geq x) &= \sum_{k=x}^{\min(n,K)} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \\ &= 1 - \sum_{k=0}^{x-1} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \end{aligned} \tag{4.2}$$

The first approach directly sums up the probabilities $P(X = k)$ for all $k \in \{x, \ldots, \min(n, K)\}$, whereas the latter approach sums up the probabilities $P(X = k)$ for all $k \in \{0, \ldots, x-1\}$, resulting in the probability $P(X < x)$ to have less than x shared target genes, and then subtracting it from 1. Which approach is more efficient depends on whether x is closer to $\min(n, K)$ or 1.

$P(X \geq x)$ is then the p-value that gives the probability that transcription factor t and miRNA m share x target genes or more under the null–hypothesis that the number of shared target genes for this pair was a result of chance.

4.4.3 Multiple Hypothesis Testing Correction

The p–value was computed as described in Section 4.4.2 for all n TF–miRNA pairs with at least one shared target gene in a given network. This can be viewed as testing n null–hypotheses H_i that TF–miRNA pair i has its number of shared target genes due to chance alone, with $i \in \{1, \ldots, n\}$. Given a significance threshold α, the null-hypothesis H_i is rejected if $p_i < \alpha$, where p_i is the p–value of the i–th TF–miRNA pair. Testing multiple hypotheses in this manner increases the chance of accumulating false positives for which the null–hypothesis was incorrectly rejected [114].

For example, given a significance threshold $\alpha = 0.05$ and a network with $T = 2,580$ transcription factors and $M = 583$ miRNAs, there are potentially up to $T \times M = 1,504,140$ TF–miRNA pairs with shared target genes. If 20% of these pairs have 1 or more shared target genes, then the hypergeometric test is performed for $n = 300,828$ pairs. The expectation is that around $n \cdot \alpha = 15,041$ of these TF–miRNA pairs are going to test as statistically significant with $p_i < 0.05$ just by chance alone.

Therefore, multiple hypothesis testing correction was conducted before selecting TF–miRNA co–regulatory pairs as motifs.

There are two major groups of commonly used multiple hypothesis testing correction methods [115]. The first one controls the so called family–wise error rate (FWER), which gives the probability of having one or more false positives among all tested hypotheses. Given the desired significance level α, the goal is to control the FWER such that FWER $\leq \alpha$ [114, 116].

One of the most commonly used ways of controlling the FWER is the Bonferroni correction [116] that divides the desired significance level by the number of hypothesis tests conducted and then only considers p-values as significant that are below this adjusted significance level [114, 116]:

$$p_i \leq \frac{\alpha}{n} \tag{4.3}$$

The second group of methods controls the false discovery rate (FDR), which gives the proportion of false positives among the significant results for which the null–hypothesis was rejected with $p_i < \alpha$. Similar to the FWER, the goal is to control the FDR such that FDR $\leq \alpha$ [114].

The FWER approach is more conservative than the FDR approach and increases the chance of generating false negatives, especially if the number of tests or expected number of significant results is large, which negatively affects its statistical power [114, 115]. For the task of identifying significant co–regulatory TF–miRNA pairs as motifs, it was important to minimise the number of false negatives while keeping the rate of false positives among the significant pairs below the chosen significance level.

Therefore, a FDR approach was chosen to correct for multiple hypothesis testing at a significance level of $\alpha = 0.05$, specifically a two–stage version of the Benjamini–Hochberg procedure [117].

The original, one–stage version of the Benjamini–Hochberg procedure is one of the most widely used methods to control the FDR [115], and has been utilised successfully in previous approaches for identifying co–regulatory TF–miRNA motifs [15].

Given n hypothesis tests and the desired FDR–controlled significance level α, the corresponding p–values are sorted in ascending order such that $p_1 \leq \cdots \leq p_n$. Then, the largest k is identified for which

$$p_k \leq \frac{k}{n}\alpha. \tag{4.4}$$

All p–values with sorted index $i \leq k$ are then accepted as significant, if such a k exists. If the n tests are independent, then this method controls the FDR at

$$\frac{n_0}{n}\alpha \leq \alpha, \tag{4.5}$$

where n_0 is the unknown number of true null–hypotheses [114, 117]. If n_0 is close to n, this controls the FDR close to α as expected. However, the smaller n_0 is compared to n, the lower the level at which the FDR is actually controlled, which, as discussed before, comes with an increased risk of false negatives. To constrain the FDR at exactly α, the Benjamini–Hochberg procedure would have to be run with the following adjusted level [117]:

$$\alpha^* = \frac{n}{n_0}\alpha \tag{4.6}$$

The two–stage method by Benjamini, Krieger and Yekutieli approximates this by estimating n_0 from the given p–values [117]. First, the procedure applies the one–stage Benjamini–Hochberg approach with slightly lowered significance level α':

$$\alpha' = \frac{\alpha}{1+\alpha} \tag{4.7}$$

Let n_s be the resulting number of p–values that are significant at α'. If $n_s = n$, the FDR is already controlled at $\alpha' < \alpha$, and the null–hypothesis is thus rejected for all n tests and the procedure stops.

Otherwise, the number of true null-hypotheses n_0 is estimated from the number of rejected null-hypothesis as $\hat{n}_0 = n - n_s$, followed by another application of the one–stage Benjamini–Hochberg procedure, this time with further adjusted significance level $\hat{\alpha}^*$:

$$\hat{\alpha}^* = \frac{n}{\hat{n}_0}\alpha' \tag{4.8}$$

If $n_s = 0$, then $\hat{\alpha}^* = \alpha'$, which means the second stage amounts to a repeat of the first stage. Consequently, the second stage can be skipped in this specific case, and like in the first stage, none of the n tests are considered to be significant.

If $0 < n_s < n$, then $\hat{\alpha}$ approximates α^* (see Equation 4.6), which increases its statistical power, especially if n_0 is small compared to the total number of tests [115, 117, 118]. Furthermore, even though the two–stage Benjamini–Krieger–Yekutieli approach like many FWER and FDR controlling approaches technically requires the hypothesis tests to be independent, it tends to perform well even if there is (positive) correlation between the tests [115, 117].

For these reasons it was employed in conjunction with the hypergeometric test to assess the statistical significance of all co–regulatory TF–miRNA pairs in a given network at $\alpha = 0.05$, instead of the one–stage Benjamini-Hochberg procedure that had been used in previous TF–miRNA co–regulatory motif search approaches [15].

4.4.4 Classification of Co–Regulatory Motifs

Finally, the identified TF–miRNA co–regulatory motifs identified in the previous step were grouped by their type (see Figure 2.3). Motifs where the transcription factor and the miRNA did not interact with each other were classified as "co–regulatory". Cases where the transcription factor and the miRNA regulated each other were categorised as composite–FFLs. Motifs where the transcription factor regulated the miRNA, but the miRNA not the transcription factor, were classified as TF–FFLs. Lastly, in miRNA–FFLs the miRNA regulated the transcription factor without being regulated by the transcription factor.

Due to missing information on the precise nature of TF $\rightarrow$ miRNA and TF $\rightarrow$ gene interactions in the regulatory database, i.e. activation or repression, the FFLs types were not further distinguished into coherent or incoherent FFLs.

4.5 Expression and Modification Correlation

Based on the results of the differential gene expression and differential histone modification or variant analyses, as well as the TF–miRNA co–regulatory motifs in the cell differentiation transition specific gene regulatory networks, the relationship between histone modifications or variants and gene expression in these motifs could be analysed.

To that end, the correlation between differential expression and differential modification or variant was first computed genome–wide for all cell differentiation transitions to obtain a baseline and a point of comparison with the literature. The same was subsequently done for genes involved in different TF–miRNA co–regulatory motif types for each transition, and the correlation in these motifs was additionally assessed with randomised comparisons. Furthermore, the correlation was compared between promoter and gene body, as well as between histone modifications or variants in the same region for each examined group.

4.5.1 Correlation Measurement

The differential gene expression analysis detailed in Section 4.1 was performed for each cell differentiation transition t from cell type A to B, and produced the log_2–fold change f_g in gene expression for each gene g and whether it was differentially expressed.

Histone modifications and variants h were distinguished by their location l in the promoter region or gene body. For each combination of transition and histone modification, the differential histone modification analysis described in Section 4.2 yielded a set of intervals i for each region of each gene. These intervals were classified as either unmodified in both A and B, only modified in A, only modified in B, or modified in both A and B.

During preliminary testing, a quantitative and an ordinal, categorical approach were considered. Comparative examples are visualised in Figure 4.1.

For the quantitative approach, the change in expression $x_t(g)$ was described by the log_2–fold change for differentially expressed genes, while 0 was used for non–differentially expressed genes:

$$x_t(g) = \begin{cases} f_g, & g \text{ differentially expressed in } t \\ 0, & g \text{ not differentially expressed in } t \end{cases} \tag{4.9}$$

The change in modifications or variants $y_{t,h,l}(g)$ (Equation 4.10) was calculated by summing up the interval values $v_{t,h,l,g}(i)$ (Equation 4.11). Cases where the interval was unmodified in both A and B, or modified in both, were considered to be non–differentially modified and thus assigned 0. When the interval was only modified in A, $v_{t,h,l,g}(i)$ was set to -1, representing a loss in modification or variant in that interval. Conversely, instances where B was modified but not A were considered to be a gain in modification or variant, and assigned 1.

$$y_{t,h,l}(g) = \sum_{i \in l} v_{t,h,l,g}(i) \tag{4.10}$$

$$v_{t,h,l,g}(i) = \begin{cases} -1, & i \text{ only modified in } A \\ 1, & i \text{ only modified in } B \\ 0, & \text{otherwise} \end{cases} \tag{4.11}$$

Thereby, if $y_{t,h,l}(g) = 0$, the region was either not differentially modified at all or differentially modified in A and B to the same extent, whereas $y_{t,h,l}(g) < 0$ indicated that it was either only differentially modified in A or to larger extent in A than B, and vice versa for $y_{t,h,l}(g) > 0$.

For the categorical approach, non–differentially expressed genes were assigned 0, while differentially expressed genes that were up–regulated in B with a log_2–fold change > 0 were set to 1, and differentially down–regulated genes were set to -1 (Equation 4.12).

$$x_t(g) = \begin{cases} 1, & g \text{ differentially up–regulated from } A \text{ to } B \\ -1, & g \text{ differentially down–regulated from } A \text{ to } B \\ 0, & g \text{ not differentially expressed} \end{cases} \tag{4.12}$$

For the categorical change in modification or variants (Equation 4.13), intervals that were modified both in A and B or not modified in either of them were removed from consideration as non–differentially modified, leaving for each region only intervals that were either differentially modified in A or in B. If all intervals in a region were differentially modified in A and not B, then this was considered to be a loss in modification and assigned -1, analogous to down–regulation in expression. Similarly, instances where all intervals were differentially modified in B and not A were a gain in modification and consequently assigned 1. Cases without any differentially modified intervals were categorised as non–differential and assigned 0. The same was done for ambiguous cases where some intervals were differentially modified in A and some in B.

$$y_{t,h,l}(g) = \begin{cases} 1, & l \text{ of } g \text{ only differentially modified in } B \\ -1, & l \text{ of } g \text{ only differentially modified in } A \\ 0, & \text{otherwise} \end{cases} \tag{4.13}$$

In order to gauge which of the two approaches was better suited, the genome–wide correlation was calculated with Pearson's, Spearman's and Kendall's correlation coefficients and compared to the literature. While both approaches

yielded overall similar correlation coefficients in many instances, the categorical approach produced overall stronger correlations (0.034 ± 0.08). Additionally, examining the cases with the largest difference between the two approaches showed that the quantitative identified only a small positive correlation between H3K27ac and gene expression, while the categorical approach was able to capture a moderate to strong positive correlation for the same cases, which is more in line with the activating effect of H3K27ac [35, 119, 120].

Thus, the categorical approach was used for all subsequent analyses. Out of the three correlation coefficients, Pearson's r was chosen since the rank–based nature of Spearman's ρ and Kendall's τ was unnecessary due to the categorical nature of the variables.

4.5.1.1 Pearson's Correlation Coefficient

Let $X_t = \{x_t(g_1), \ldots, x_t(g_n)\}$ be the change in expression for n genes g_i in cell differentiation transition t, with $x_t(g_i)$ defined as in Equation 4.12, and $Y_{t,h,l} = \{y_{t,h,l}(g_1), \ldots, y_{t,h,l}(g_n)\}$ the change in histone modification or variant h by their location l in the promoter or gene body of those genes as given in Equation 4.13. The corresponding Pearson's correlation coefficient $r_{t,h,l}$ is then the covariance (COV) of X_t and $Y_{t,h,l}$ divided by the product of the respective standard deviations (SDs) (Equation 4.14), with $\bar{X}_t$ and $\bar{Y}_{t,h,l}$ the means of X_t and $Y_{t,h,l}$, respectively. An example can be found in Figure 4.2.

$$\begin{aligned} r_{t,h,l} &= \frac{\mathrm{COV}(X_t, Y_{t,h,l})}{\mathrm{SD}(X_t) \cdot \mathrm{SD}(Y_{t,h,l})} \\ &= \frac{\sum_{i=1}^{n} \left(x_t(g_i) - \bar{X}_t\right) \left(y_{t,h,l}(g_i) - \bar{Y}_{t,h,l}\right)}{\sqrt{\sum_{i=0}^{n} \left(x_t(g_i) - \bar{X}_t\right)^2} \sqrt{\sum_{i=0}^{n} \left(y_{t,h,l}(g_i) - \bar{Y}_{t,h,l}\right)^2}} \end{aligned} \tag{4.14}$$

Positive correlation coefficients, with $0 < r_{t,h,l} \leq 1$, describe a linear relationship between differential gene expression and differential histone modifications, where a gain in modification during the cell differentiation transition tended to be accompanied by an up–regulation of gene expression, or a loss of modification by down–regulation. Similarly, for coefficients with $-1 \leq r_{t,h,l} < 0$, a gain in modification tended to coincide with down–regulation, or loss in modification with up–regulation. The closer $r_{t,h,l}$ to 1 or -1, the stronger the relationship, while $r_{t,h,l} = 0$ corresponds to no correlation.

There was a small number of cases where the change in expression X_t or the change in histone modification $Y_{t,h,l}$ was constant, mostly in smaller gene sets. If X_t is constant for a set of genes, then $x_t(g_i) = \bar{X}_t$ for all $i \in \{1, \ldots, n\}$, which results in a standard deviation of 0 and thus a division by 0. The same problem occurs if $Y_{t,h,l}$ is constant. Consequently, $r_{t,h,l}$ is only defined if neither X_t nor $Y_{t,h,l}$ are constant. Instances where the correlation coefficient was not defined were labelled with the respective cause.

4.5.1.2 Statistical Dependence of Expression and Modification

While Pearson's correlation coefficient assesses the effect size of the relationship between differential gene expression and differential histone modifications or

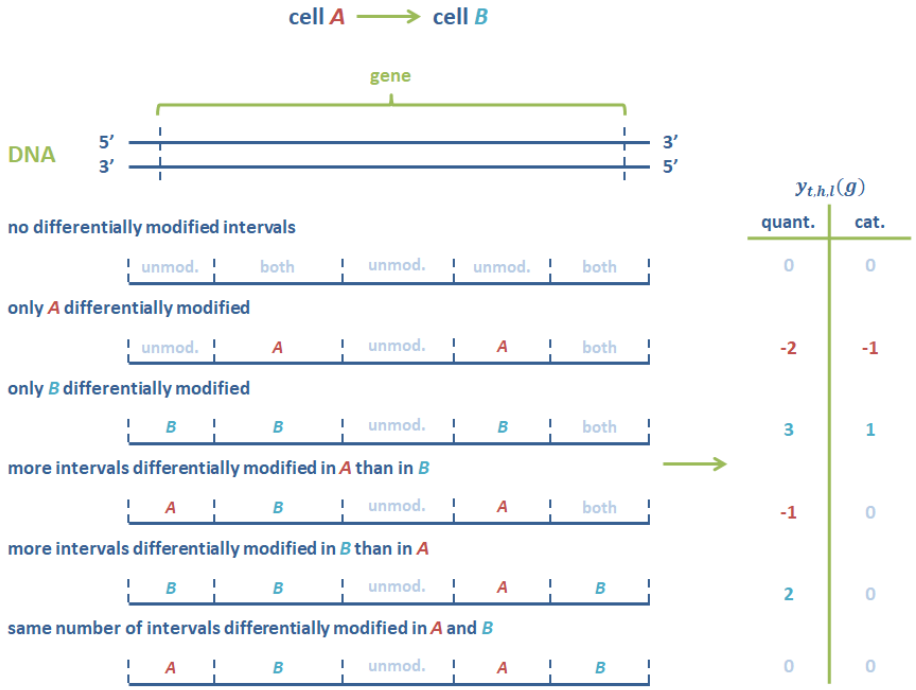

Figure 4.1: *Examples of quantification versus categorisation of differential modifications $y_{t,h,l}(g)$ of histone h in location l (promoter or gene body) of gene g for cell differentiation t from cell type A to B. The differential histone modification analysis classified intervals of g as either unmodified in A and B, modified in both, only differentially modified in A, or only differentially modified in B.*

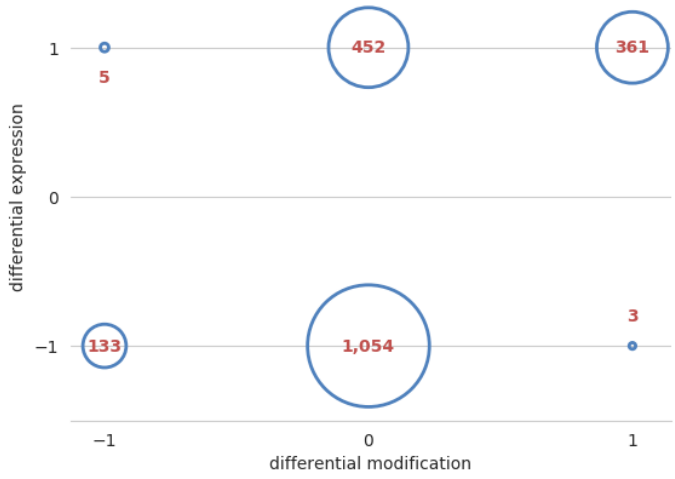

Figure 4.2: *Paired categorical differential gene expression and differential H3K4me3 modifications in the gene body of 2,008 differentially expressed genes in the cell differentiation transition from neural progenitor cell to bipolar neuron. The number of genes in each category are given in red, corresponding to the size of the blue circles. The Pearson correlation between differential expression and differential modification is $r = 0.549$, with $p < 0.05$ for the χ^2–test of independence.*

variants, it does not determine if the relationship is statistically significant. To that end, Pearson's χ^2–test of independence was employed due to its non–parametric nature and suitability for comparing two categorical variables with more than two categories [121].

Given a set of n genes g_i in cell differentiation transition t, let $x_t(g_i)$ be the categorical differential gene expression of g_i in t and $y_{t,h,l}(g_i)$ be the corresponding differential histone modification or variant h in location l, with $i \in \{1, \ldots, n\}$, $x_t(g_i) \in \{-1, 0, 1\}$ and $y_{t,h,l}(g_i) \in \{-1, 0, 1\}$. This results in n observation pairs $(x_t(g_i), y_{t,h,l}(g_i))$ of coinciding differential expression and differential modification.

The occurrence of each of the nine possible combinations of differential expression and differential modification can be counted in a so called contingency table O, where the rows i correspond to the $r = 3$ possible differential expression values and the columns j to the $c = 3$ possible differential modification values. In that observation table, the cell $O_{i,j}$ then contains how many of the n pairs had differential expression i paired with differential modification j. Based on the contingency table, the observed frequency of each possible differential expression value and each possible modification value can be computed as follows:

$$\begin{aligned} f_i &= \frac{1}{n} \sum_{j=1}^{c} O_{i,j} \\ f_j &= \frac{1}{n} \sum_{i=1}^{r} O_{i,j}, \end{aligned} \tag{4.15}$$

whereby f_i is the fraction of observations in row i, whereas f_j is the fraction of observations in column j. The expected occurrence of combination (i, j) under independence is then:

$$E_{i,j} = n f_i f_j \tag{4.16}$$

The χ^2–test statistic is subsequently computed by comparing the observed to the expected occurrences:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \tag{4.17}$$

It can be converted to a p–value representing the probability to obtain the observed differential gene expression and differential histone modification or variant pairs under the null–hypothesis that they are independent. The χ^2–test of independence was conducted for all computed Pearson correlation coefficients, and adjustment for multiple hypothesis testing at $\alpha = 0.05$ was accomplished with the approach by Benjamini, Krieger and Yekutieli.

4.5.2 Genome–Wide Correlation

Genome–wide correlation between gene expression and histone modifications or histone variants has been examined for a large number of different histone marks and variants in various cell types and conditions, including for H2A.Z, H3K4me2, H3K4me3 and H3K27ac, which were studied in this thesis [22, 119].

To evaluate if the categorical approach of this work was able to capture the same trends as those related works and to have a baseline for the corresponding correlations in TF–miRNA co–regulatory motifs, the genome–wide correlation was computed for each cell differentiation transition.

Additionally, it was calculated for only differentially expressed genes in the respective transitions, as well as the three disjoint subsets of TFs, miRNAs and remaining genes, to explore potential correlation differences between the key players involved in the co–regulatory motifs.

For all gene sets, the χ^2–test of independence was performed and the resulting p–values corrected for multiple hypothesis testing at $\alpha = 0.05$ with the procedure by Benjamini, Krieger and Yekutieli.

4.5.3 Number of TF–miRNA Motifs

The occurrence of TF–FFLs, miRNA–FFLs, composite–FFLs and simple co–regulatory motifs was counted, for each complete or differential gene regulatory network of the examined cell differentiation transitions. Following the example set by the TFmiR webserver [15], a randomisation approach was employed to ascertain if the motifs were significantly enriched in these networks.

For each network with m directed edges, each consisting of a source node S and target node t, $k = 5,000$ randomised versions of that network were generated. TFmiR initially utilised a degree–preserving randomisation procedure, whereby two edges $e_1 = s_1 \rightarrow t_1$ and $e_2 = s_2 \rightarrow t_2$ were selected at random and their target nodes were switched, resulting in $e_1 = s_1 \rightarrow t_2$ and $e_2 = s_2 \rightarrow t_1$. This step was repeated $2m$ times.

However, this approach did not consider the interaction type of the randomly chosen edges, thus for example turning a TF–gene into a miRNA–gene interaction, even though TFs and miRNAs work through different regulatory mechanisms. Follow–up work by Sadegh et al. (2017) [122] compared the degree–conserving randomisation approach with a method that additionally conserves the interaction type, and found that the latter was better at capturing biologically meaningful dynamics.

Therefore, the degree– and interaction–type–conserving randomisation approach was chosen in this thesis. To that end, the edges of each network were split into four separate edge lists according to their type: TF $\rightarrow$ miRNA, TF $\rightarrow$ gene, miRNA $\rightarrow$ TF, and miRNA $\rightarrow$ gene. These separate edge lists were randomised k times with $2m$ iterations of the degree–conserving edge–switching method, and subsequently put back together into k randomised networks.

TF–miRNA co–regulatory motifs were identified in the randomised networks as described in Section 4.4, and the occurrence of each motif type was counted. Let n_{motif} be the number of times a motif type occurred in the original network, n_i the number of times it occurred in random network i, with $i \in \{1, \ldots, k\}$, and c the number of times where $n_i \geq n_{\text{motif}}$. The p–value representing the probability of the motif to occur at least as often in the biological network as it did was then calculated as:

$$p = \frac{c}{k} \tag{4.18}$$

Correction for multiple hypothesis testing at $\alpha = 0.05$ was performed for all such generated p–values with the method by Benjamini, Krieger and Yekutieli.

4.5.4 Correlation in TF–miRNA Motifs

For each cell differentiation transition regulatory network, the TF–miRNA co–regulatory motifs were aggregated by the four types examined in this thesis: TF–FFLs, miRNA–FFLs, compite–FFLs and simple co–regulation. A duplicate–free gene set was then constructed from all TFs, miRNAs and co–regulated target genes involved in the respective motif type and transition.

For each combination of transition t, histone mark or variant h, location l in promoter or gene body, and motif type comprised of gene set s, the Pearson correlation coefficient $r_{t,h,s,l}$ between differential gene expression and differential histone modification was computed. The resulting correlations were statistically compared to motif correlations in randomised networks, as well as to the corresponding genome–wide correlations.

4.5.4.1 Randomised Comparison

The correlation $r_{t,h,s,l}$ was computed in the same way for motifs in the randomised networks generated in Section 4.5.3, in order to assess if the correlation in real motifs was stronger than in random motifs. Since the value range of Pearson's correlation coefficient includes negative and positive values, it was possible that actual and random coefficients would differ in sign, and thus a stronger correlation could mean a coefficient that was more extreme with the same sign, or a coefficient that was more extreme regardless of sign.

Let $r_{t,h,s,l}(i)$ be the matching correlation in randomised network i, with $i \in \{1, \ldots, k\}$ and $k = 5,000$. Then, c_{abs} is the number of cases where $|r_{t,h,s,l}| \geq |r_{t,h,s,l}(i)|$, and c_{sign} the number of times where $|r_{t,h,s,l}| \geq |r_{t,h,s,l}(i)|$ and additionally $r_{t,h,s,l}$ and $r_{t,h,s,l}(i)$ have the same sign. The corresponding p–values are calculated as:

$$\begin{aligned} p_{\text{abs}} &= \frac{c_{\text{abs}}}{k} \\ p_{\text{sign}} &= \frac{c_{\text{sign}}}{k} \end{aligned} \tag{4.19}$$

Multiple hypothesis testing correction at $\alpha = 0.05$ was conducted separately for p_{abs} and p_{sign} with the procedure by Benjamini, Krieger and Yekutieli.

4.5.4.2 Genome–Wide Comparison

In addition to the randomised assessment in the previous subsection, the motif correlations were compared to the genome–wide correlation baseline, as well as to the correlation of in only differentially expressed genes. The difference between the correlation coefficients was statistically evaluated based on Fisher's r to z transform, where z unlike r has an approximately normal sampling distribution [123].

Let $r_{t,h,s,l}$ be the correlation between differential gene expression and differential histone modifications in a motif type comprised of gene set s of size $n_{t,h,s,l}$ in cell differentiation transition t for histone modification or variant h in located l in the promoter region or the gene body. Then, the corresponding

z–value and standard error (SE) is computed as follows:

$$\begin{aligned} z_{t,h,s,l} &= \frac{1}{2} \ln \frac{1 + r_{t,h,s,l}}{1 - r_{t,h,s,l}} \\ SE_{t,h,s,l} &= \frac{1}{\sqrt{n_{t,h,s,l} - 3}} \end{aligned} \tag{4.20}$$

Additionally, let $\bar{s}$ be the complement of s, consisting of all $\bar{n}_{t,h,s,l}$ genes for which differential gene expression data was available in t but were not involved in the particular motif type in t, and $\bar{r}_{t,h,s,l}$ the correlation coefficient for $\bar{s}$ and $\bar{z}_{t,h,s,l}$ the z–value. The combined SE of $z_{t,h,s,l}$ and $\bar{z}_{t,h,s,l}$ is then:

$$SE = \sqrt{\frac{1}{n_{t,h,s,l} - 3} + \frac{1}{\bar{n}_{t,h,s,l} - 3}}, \tag{4.21}$$

On that basis, the z–score of the comparison is:

$$z = \frac{z_{t,h,s,l} - \bar{z}_{t,h,s,l}}{SE} \tag{4.22}$$

Furthermore, this was done for the complement consisting only of differentially expressed genes. The z–values were then used to compute the corresponding two–tailed p–values, which were controlled at $\alpha = 0.05$ for multiple hypothesis testing with the Benjamini–Krieger–Yekutieli method.

4.6 Promoter and Gene Body Correlation

The effect of histone modifications or histone variants can differ based on their location in the promoter region or the gene body [28]. Thus, the correlation coefficients $r_{t,h,s,p}$ of differential gene expression and differential histone modifications in the promoter were compared to the corresponding coefficients in the gene body $r_{t,h,s,g}$, for genome–wide correlations and correlations in TF–miRNA co–regulatory motifs for all cell differentiation transitions t and histone modifications or variants h.

Both $r_{t,h,s,p}$ and $r_{t,h,s,g}$ were computed from the same set s of n genes, and had one variable, the differential gene expression, in common. There are several statistical tests for comparing non–independent correlation coefficients with one shared variable, including one based on Fisher's r to z transform (Equation 4.20) by Dunn and Clark (1969) [124] that performs comparatively well in terms of FDR and statistical power [125]. In this thesis, a modified version proposed by Steiger (1980) [126] was used.

Let $r_{p,g}$ be the correlation between the histone modifications in the promoter and respective gene body, and $\bar{r}$ the mean of $r_{t,h,s,p}$ and $r_{t,h,s,g}$. The z–score of the comparison was then computed as follows:

$$z = \frac{(z_{t,h,s,p} - z_{t,h,s,g}) \cdot \sqrt{n - 3}}{\sqrt{2 - 2c}}, \tag{4.23}$$

where $z_{t,h,s,p}$ and $z_{t,h,s,g}$ are the Fisher–transformations (Equation 4.20) of $r_{t,h,s,p}$ and $r_{t,h,s,g}$, respectively, and

$$c = \frac{r_{p,g}\left(1 - 2\bar{r}^2\right) - \frac{1}{2}\bar{r}^2\left(1 - 2\bar{r}^2 - r_{p,g}^2\right)}{\left(1 - \bar{r}\right)^2}. \tag{4.24}$$

The two–tailed p–value was subsequently calculated from the z–score, and the Benjamini–Krieger–Yekutieli procedure was employed to correct for multiple hypothesis testing at $\alpha = 0.05$.

4.7 Annotation Enrichment

In addition to the differential gene expression and differential histone modification or variant analysis, an annotation enrichment analysis was performed to establish a biological function and process context. To that end, ShinyGO (version 0.60) [127] was employed, which offers web–based enrichment analyses for the Gene Ontology (GO) [128, 129] and the Kyoto Encyclopedia of Genes and Chromosomes (KEGG) [130, 131] annotations with Ensembl gene IDs.

Specifically, the top 30 significantly enriched annotations compared to the general human background were computed for each of the four TF–miRNA co–regulatory motif types in each cell differentiation transition, as well as differentially expressed genes. This was done for the GO categories "Biological Process" and "Molecular Function" and for KEGG pathway annotations, with an FDR–controlled p–value threshold of 0.05.

4.8 Implementation

The main pipeline and functionality of the work presented in this chapter, as well as the pre–processing detailed in Chapter 3, was implemented in Python 3 (version 3.6.8) [132]. The packages *numpy* (version 1.16.3) [133] and *pandas* (version 0.24.2) [134] were used for array and data frame handling. For the statistical analyses, the Pearson's correlation coefficient, hypergeometric and χ^2–test implementation of the *scipy* (version 1.2.1) [135] packages was used. Plots were generated with the *matplotlib* (version 3.0.3) [136] and *seaborn* (version 0.9.0) [137] packages.

Due to the large amount of data handled by the pipeline, intermediate pre–processing and analysis results were stored in SQLite databases that were handled with Python's built–in *sqlite3* package. To improve runtime, the built–in *multiprocessing* package was used.

The C++ program *Salmon* (version 0.13.1) [93] was employed to built the reference transcriptome index and to quantify the RNA–seq data for the differential gene expression analysis. The resulting transcript counts were summarised to the gene level and prepared for the subsequent differential analysis with the R (version 3.4.4) [138] package *tximport* (version 1.6.0) [107]. The differential gene expression analysis itself was conducted with the R package *DESeq2* (version 1.18.1) [102] using default parameters.

The ChIP–seq BAM files were checked for duplicates, indexed and merged with the program *SAMtools* (version 1.7) [111] in preparation for quantification and differential analysis. For the differential histone modification and histone variant analysis itself, the R package *histoneHMM* (version 1.7) [110] was used with default parameters. The results were mapped onto the reference gene and

promoter regions with the C++ program *BEDTools* (version 2.26.0) [112, 139]. The gene regulatory networks were randomised with the help of the R package *igraph* (version 1.2.4.1) [140].

These programs and R scripts were automatically executed on the command line with Python's built–in *subprocess* package.

Chapter 5

Results and Discussion

In this thesis, the correlation of differential gene expression and differential histone modifications or variants in human TF–miRNA co–regulatory motifs was analysed for different cell differentiation transitions, and contrasted with genome–wide correlation patterns and motifs in randomised networks.

Background information on TF–miRNA co–regulatory motifs, histone modifications and variants and their respective roles in gene expression regulation, cell differentiation and general human development was introduced in Chapter 2. Subsequently, Chapter 3 detailed the genomic and transcriptomic references, gene regulatory information, and gene expression and histone mark data sets used in this thesis.

In total, six human cell differentiation transitions were selected, which had sufficient RNA–seq expression data and paired histone ChIP–seq data for at least one of the histone modifications H3K4me2, H3K4me3, and H3K27ac, or the histone variant H2A.Z. The transition between pluripotent H1–hESC cell line and the likewise pluripotent iPSC GM23338 cell line had paired data for all four histone marks, as did the transition from multipotent NPC to terminally differentiated bipolar neuron cell. In contrast, the transitions from multipotent NSPC to NPC and to bipolar neuron cells had paired data for H2A.Z, H3K4me3 and H3K27ac, but not H3K4me2. For the haematopoietic transition from multipotent MPC to monocytes, matching H3K4me3 and H3K27ac data was available, whereas for the multipotent MSC to terminally differentiated osteoblast transition that was only the case for H2A.Z.

Chapter 4 outlined the pre–processing and analysis methodology. First, differential gene expression and differential histone mark analysis was conducted. Based on the regulatory data, a general gene regulatory network was built, as well as a two networks for each cell differentiation transition: one that was only comprised of differentially expressed genes and one of all genes that were expressed, differentially or not, in the transition. Furthermore, each network was randomised 5,000 times to enable randomised comparisons. Subsequently, four types of TF–miRNA co–regulatory motifs were identified in each network: TF–FFLs, miRNA–FFLs, composite—-FFLs and simple co–regulation. The correlation between differential gene expression and differential histone marks was computed for each motif type in each network, as well as genome–wide. Lastly, GO and KEGG annotation enrichment analysis was conducted for genes involved in the different motif types, as well as differentially expressed genes.

This chapter first gives a general overview of the number of expressed genes and the composition of the gene regulatory networks. Then, results of the genome–wide correlation analyses are presented and discussed, followed by the results and discussion of correlation in TF–miRNA co–regulatory motifs. Lastly, limitations of this work as well as future work are discussed.

5.1 Data Overview

5.1.1 Number of Expressed Genes

In total, 58,788 human gene IDs were obtained from Ensembl, of which 2,593 (4.4%) belonged to TFs, 1,879 (3.2%) to miRNAs and the remaining 54,316 (92.4%) to other types of genes. However, differential gene expression data was not available for many of them in the examined cell differentiation transitions.

The total number of genes with available differential gene expression data ranged from 34,458 (58.61%) genes in the transition from MSC to osteoblast to 46,943 (79.95%) genes in the H1–hESC to GM23338 transition. Most of the TFs (90.5% – 94.8%) had differential expression information, but only 18.9% to 44.9% of miRNAs (519 ± 183), while the bulk of genes without differential expression data were other types of genes.

Closer examination revealed, that in an average of 1,514 ± 303 cases, the IDs given in the reference transcriptome did not match the Ensembl gene IDs. In most cases, however, the problem was lacking expression in one or both of the cell types of a differentiation transition (17,286 ± 4,805 cases), disproportionately in pseudogenes, and only in 164 ± 44 cases due to extreme outliers or other abnormalities that were rejected by *DESeq2*.

On average, 2,850 ± 1,900 genes were differentially expressed. The number of differentially expressed miRNAs was low in general (20 ± 15), whereas the number of differentially expressed TFs was on average 266 ± 167. An overview can be found in Table 5.1.

In general, the total number of differentially expressed genes in each transition matched what one would expect based on the closeness of the cell types. For instance, in the neural development pathway from NSPC to NPC to bipolar neuron, the number of differentially expressed genes between NSPC and bipolar neuron (5,218) was much higher than between neural NSPC and NPC (1,376) or NPC and bipolar neuron (2,008). Similarly, between H1–hESC and the iPSC GM23338 there were 3,831 differentially expressed genes and 4,343 between MPC and monocyte.

It was thus surprising that only 326 genes were differentially expressed between MSC and osteoblast, of which 25 were TFs and 301 other genes, without any differentially expressed miRNA, even though miRNAs play a role in lineage commitment and osteogenic differentiation of MSCs [141].

5.1.2 Gene Regulatory Networks

The human gene regulatory database obtained from TFmiR contained 22,308 genes that could be matched to the Ensembl gene IDs used in this thesis, as well as 226,156 regulatory interactions between those genes. It included regulatory interactions for 99.5% of TFs in Ensembl, but only for 583 (31%) of miRNAs.

Table 5.1: *Number of genes with available gene expression data in cell differentiation transitions. That number is further broken down into differentially expressed genes and different types of genes. The number of available Ensembl gene IDs is given as well.*

Transition	Expression	Total	TFs	miRNAs	Other
H1–hESC	all	46,943	2,459	845	43,639
→ GM23338	differential	3,831	264	42	3,525
mesenchymal stem cell	all	34,458	2,374	446	31,638
→ osteoblast	differential	326	25	0	301
myeloid progenitor	all	42,200	2,400	622	39,178
→ monocyte	differential	4,343	300	27	4,016
neural progenitor	all	37,827	2,373	440	35,014
→ bipolar neuron	differential	2,008	280	5	1,723
neural stem progenitor	all	38,980	2,386	406	36,188
→ bipolar neuron	differential	5,218	455	19	4,744
neural stem progenitor	all	34,629	2,347	357	31,925
→ neural progenitor	differential	1,376	32	27	1,317
all Ensembl IDs		58,788	2,593	1,879	54,316

Due to the removal of edges involving nodes without Ensembl ID, a small number of TFs and miRNAs did not have associated interactions in which they were the regulator. An overview of the number of nodes and edges can be found in Tables 5.2 and 5.3, respectively.

Furthermore, most regulatory interactions occurred between TFs and other TFs (15.5%) and TFs and other genes (81.8%), whereas the number of 2,723 (1.2%) TF → miRNA interactions was comparatively much lower. Only 356 (0.16%) and 544 (0.2%) interactions were miRNA → TF and miRNA → TF interactions, respectively. There was also a small number (1.2%) of not further specified gene → gene interactions. Almost all target genes were regulated by at least one TF or miRNA.

Due to the different number of genes with available differential expression data in the cell differentiation transitions, a separate gene regulatory network was built for each of them, by removing nodes and corresponding interactions without expression data. Most TFs with differential expression data in the different cell differentiation transitions could be included in the respective gene regulatory network, but on average only 201 ± 54 miRNAs had interaction as well as available differential gene expression information. Of the 22,097 TF or miRNA targets in the complete network, 18,447 ± 832 could be incorporated into the transition networks.

The low number of miRNAs nodes and miRNA → target interactions proved especially problematic when each network was further pruned to only contain differentially expressed genes. In those differential networks, the overall number of nodes (1,451 ± 1,103) was quite low for some transitions, with only 169 and 171 nodes for the transitions from NSPC to NPC and from MSC to osteoblast, respectively, and on average they only contained 3 ± 4 miRNAs.

Table 5.2: *Number of nodes, each representing a gene, in the complete network constructed from the human gene regulatory database obtained from TFmiR, as well as nodes with expression information in cell differentiation transition specific networks. Differential network versions only include genes that were significantly differentially expressed in the respective cell differentiation specific network. Genes that did not interact with at least one other gene in the respective network were excluded.*

Transition	Network	Nodes	TFs	miRNAs	Regulators	Targets	TF/miRNA Targets
H1–hESC	complete	19,924	2,448	298	2,726	19,795	19,789
→ GM23338	differential	1,651	232	11	91	1,633	1,633
mesenchymal stem cell	complete	17,697	2,365	214	2,577	17,614	17,610
→ osteoblast	differential	171	24	0	9	171	170
myeloid progenitor	complete	19,123	2,391	208	2,637	19,030	19,023
→ monocyte	differential	2,272	279	2	132	2,266	2,264
neural progenitor	complete	18,177	2,364	177	2,557	18,099	18,093
→ bipolar neuron	differential	1,547	270	0	98	1,543	1,541
neural stem progenitor	complete	18,512	2,376	156	2,570	18,440	18,434
→ bipolar neuron	differential	2,893	437	6	195	2,883	2,882
neural stem progenitor	complete	17,811	2,338	150	2,521	17,740	17,734
→ neural progenitor	differential	169	17	1	5	165	164
complete network		22,308	2,580	583	3,044	22,105	22,097

Table 5.3: *Number of edges, each representing an interaction between two genes, in the complete network constructed from the human gene regulatory database obtained from TFmiR, as well as in cell differentiation transition specific networks. Differential network versions only include interactions between genes that were both significantly differentially expressed in the respective cell differentiation specific network. None of the networks contained miRNA–miRNA interactions.*

Transition	Network	Edges	TF → miRNA	TF → TF	TF → gene	miRNA → TF	miRNA → gene	gene → gene
H1–hESC	complete	208,296	1,231	33,291	170,897	180	278	2,774
→ GM23338	differential	2,833	13	461	2,353	1	3	45
mesenchymal stem cell	complete	194,939	792	32,483	158,865	149	223	2,780
→ osteoblast	differential	383	0	58	328	0	0	4
myeloid progenitor	complete	202,327	810	32,784	165,922	127	224	2,816
→ monocyte	differential	6,828	1	1,073	5,738	0	0	70
neural progenitor	complete	197,041	599	32,283	161,423	122	216	2,753
→ bipolar neuron	differential	5,383	0	1,036	4,335	0	0	50
neural stem progenitor	complete	199,404	518	32,548	163,604	112	178	2,801
→ bipolar neuron	differential	10,474	9	1,876	8,561	0	4	82
neural stem progenitor	complete	195,126	422	32,184	159,801	111	195	2,765
→ neural progenitor	differential	218	0	20	199	0	0	5
complete network		226,156	2,723	35,063	184,954	356	544	2,878

5.2 Genome–Wide Correlation

The genome–wide correlation was computed for all combinations of cell differentiation transition, histone modification or variant marks, location in promoter or gene body, gene type (all, TF, miRNA, other), and for all genes or only differentially expressed genes. The correlation coefficients can be found in Table 5.4.

Overall, there was a small positive genome–wide correlation ($r = 0.11 \pm 0.06$) between differential gene expression and differential acetylation of H3K27 across all cell differentiation transitions and genomic regions. The exception was a very small negative correlation ($r = -0.008$) in the transition from NSPC to NPC in the gene body. H3K7ac ChIP–seq data was not available for the MSC to osteoblast transition.

Similarly, there was a slightly smaller positive genome–wide correlation ($r = 0.08 \pm 0.04$ for H3K4me2 in the transition from H1–hESC to GM23338 and from NPC to bipolar neuron; for the other four transition there was no paired H3K4me2 ChIP–seq data. The genome–wide correlation for H3K4me3 was also positive ($r = 0.11 \pm 0.07$), except for the transition from MSC to osteoblast, which again lacked ChIP–seq data for H3K4me3.

The genome–wide histone variant H2A.Z pattern was more complicated. While there was a very small positive correlation ($r = 0.03 \pm 0.007$) in the promoter region, the general pattern in the gene body varied depending on the cell differentiation transitions. In the transitions from H1–hESC to GM23338, from MSC to osteoblast and from NSPC to NPC there was also a very small positive correlation ($r = 0.03 \pm 0.016$), whereas in the transitions from NSPC to bipolar neuron and from NPC to bipolar neuron the small correlation was negative ($r = -0.1$ and $r = -0.005$, respectively). There was no H2A.Z ChIP–seq data for MPC to monocyte transition.

The general patterns were quite weak when considering all genes and without regard for genomic regions. However, strong correlations existed for more specific combinations of gene subsets and modification or variant location, which is discussed in more detail in the following subsections.

5.2.1 Promoter versus Gene Body Modifications

Since different histone modifications or variants are associated with different genomic regions, i.e. promoter and gene body as seen above for H2A.Z, the correlation was statistically compared between the two. Table 5.4 contains the correlation coefficients, while Table B.2 in the appendix details the difference and statistical significance of all comparisons.

The correlation difference between genome–wide correlation for H3K27ac in promoter and gene body was very small for all transitions (0.005 ± 0.05), which was also the case for H3K4me2 (0.04 ± 0.001) and H3K4me3 (0.05 ± 0.07). The differences were nevertheless statistically significant with $p < 0.05$, likely due to the large sample sizes of 34,629 to 46,943 genes.

As mentioned above, the genome–wide correlation for H2A.Z was overall mildly positive for the promoter region, positive for the gene body in the transition from H1–hESC to GM23338, MSC to osteoblast and NSPC to NPC, whereas there was a stronger negative pattern for the gene body in the NPC to bipolar neuron and NSPC to bipolar neuron transitions. In the three transitions

Table 5.4: *Genome–wide correlation between differential gene expression (XX) and differential histone modifications (Y) in cell differentiation transitions. For each transition, the correlation is given for all n genes (A) of a certain type, as well as only for differentially expressed genes (D). Histone modifications and variants are distinguished by their location in the promoter or gene body region. A statistically significant dependent relationship between expression and modification or variant is marked in bold, with $p < 0.05$ and after correcting for multiple hypothesis testing. Combinations for which insufficient data points were available are labelled as "–". Negative correlation is highlighted in red, positive correlation in blue, and darker highlighting signifies stronger correlation. X=0 and Y=0 mark cases without differential gene expression or without differential modifications, respectively, while X=1 marks constant up–regulation. Histone modifications without ChIP–seq samples are left blank.*

				H2A.Z		H3K4me2		H3K4me3		H3K27ac	
Transition	**Genes**		n	**prom.**	**gene**	**prom.**	**gene**	**prom.**	**gene**	**prom.**	**gene**
H1–hESC	all	A	46,943	**0.038**	**0.022**	**0.030**	**0.069**	**0.059**	**0.057**	**0.071**	**0.117**
↓		D	3,831	**0.091**	**0.065**	**0.066**	**0.215**	**0.091**	**0.107**	**0.240**	**0.387**
GM23338	TFs	A	2,459	**0.080**	**0.026**	**0.083**	**0.086**	**0.082**	**0.063**	**0.069**	**0.138**
		D	264	**0.167**	**0.056**	0.119	**0.236**	0.104	0.133	**0.204**	**0.415**
	miRNAs	A	845	0.006	-0.006	0.016	0.006	-0.008	-0.008	0.060	**0.106**
		D	42	Y=0	Y=0	Y=0	Y=0	Y=0	Y=0	0.238	**0.299**
	other	A	43,639	**0.036**	**0.022**	**0.028**	**0.068**	**0.059**	**0.058**	**0.071**	**0.117**
		D	3,525	**0.087**	**0.065**	**0.062**	**0.214**	**0.090**	**0.106**	**0.243**	**0.390**
mesenchymal	all	A	34,458	**0.036**	**0.024**						
stem cell		D	326	**0.149**	**0.209**						
↓	TFs	A	2,374	**0.146**	0.055						
ostebolast		D	25	0.363	0.553						
	miRNAs	A	446	X=0	X=0						
		D	0	–	–						
	other	A	31,638	**0.029**	**0.021**						
		D	301	0.126	**0.189**						
myeloid	all	A	42,200					**0.110**	**0.181**	**0.078**	**0.137**
progenitor		D	4,343					**0.195**	**0.321**	**0.249**	**0.362**
↓	TFs	A	2,400					**0.147**	**0.227**	**0.146**	**0.190**
monocyte		D	300					**0.210**	**0.386**	**0.323**	**0.406**
	miRNAs	A	622					0.017	**0.093**	0.041	0.057
		D	27					Y=0	**0.038**	**0.038**	**0.038**
	other	A	39,178					**0.108**	**0.178**	**0.076**	**0.134**
		D	4,016					**0.193**	**0.315**	**0.246**	**0.360**
neural	all	A	37,827	**0.033**	**-0.101**	**0.086**	**0.129**	**0.113**	**0.263**	**0.158**	**0.164**
progenitor		D	2,008	**0.132**	**-0.272**	**0.338**	**0.485**	**0.242**	**0.549**	**0.597**	**0.575**
↓	TFs	A	2,373	**0.032**	**-0.144**	**0.157**	**0.162**	**0.174**	**0.319**	**0.252**	**0.242**
bipolar		D	280	0.095	**-0.305**	**0.370**	**0.446**	**0.269**	**0.640**	**0.634**	**0.614**
neuron	miRNAs	A	440	0.013	0.012	-0.064	0.032	**0.005**	0.003	-0.089	-0.022
		D	5	–	–	–	–	–	–	–	–
	other	A	35,014	**0.034**	**-0.095**	**0.082**	**0.127**	**0.108**	**0.256**	**0.152**	**0.157**
		D	1,723	**0.139**	**-0.267**	**0.336**	**0.491**	**0.239**	**0.536**	**0.594**	**0.570**
neural stem	all	A	38,980	**0.041**	**-0.005**			**0.076**	**0.141**	**0.172**	**0.156**
progenitor		D	5,218	**0.105**	**0.005**			**0.098**	**0.196**	**0.347**	**0.361**
↓	TFs	A	2,386	0.018	**-0.058**			**0.105**	**0.221**	**0.233**	**0.253**
bipolar		D	455	0.038	**-0.112**			**0.158**	**0.425**	**0.375**	**0.495**
neuron	miRNAs	A	406	-0.015	-0.102			**0.214**	-0.016	0.084	-0.014
		D	19	–	–			–	–	–	–
	other	A	36,188	**0.044**	0			**0.073**	**0.136**	**0.168**	**0.152**
		D	4,744	**0.113**	**0.023**			**0.093**	**0.182**	**0.341**	**0.353**
neural stem	all	A	34,629	**0.022**	**0.052**			**0.036**	**0.012**	**0.064**	**-0.008**
progenitor		D	1,376	**0.080**	0.061			**0.194**	**0.161**	**0.126**	**0.150**
↓	TFs	A	2,347	0.036	-0.003			**0.102**	**0.039**	**0.018**	**-0.001**
neural		D	32	0.417	-0.049			**0.600**	**0.521**	0.238	**0.465**
progenitor	miRNAs	A	357	0.029	0.007			Y=0	Y=0	0.049	0.016
		D	27	X=1	X=1			X=1,Y=0	X=1,Y=0	X=1	X=1
	other	A	31,925	**0.021**	**0.054**			**0.033**	**0.012**	**0.066**	**-0.004**
		D	1,317	**0.068**	0.058			**0.138**	**0.116**	**0.115**	**0.127**

with the same H2A.Z correlation sign in promoter and gene body, the difference was almost non–existent (0.001 ± 0.02), whereas in the other two, the gene body correlation was lower by 0.09 ± 0.06, changing sign from positive to negative.

5.2.2 Correlation for Gene Subsets

The correlations of the different gene subsets (TFs, miRNAs, other) are given in Table 5.4, while Table B.1 in the appendix lists the differences between general and differentially expressed correlation for all subsets, genomic regions and differentiation transitions.

5.2.2.1 Differentially Expressed Genes The correlation between differential gene expression and differential histone modifications or variants was on average 0.15 ± 0.11 stronger in differentially expressed genes compared to the genome–wide correlation. In particular, the correlation was small to moderate ($r = 0.34 \pm 0.16$) for H3K27ac in differentially expressed genes across all genomic regions. For H3K4me2 and H3K4me3 it was, while still small, also stronger than genome–wide with $r = 0.28 \pm 0.18$ and H3K4me3 $r = 0.25 \pm 0.13$, respectively. In the promoter region, the differential correlation was $r = 0.11 \pm 0.03$ for H2A.Z. In the gene body, it was moderately negative ($r = 0.272$) in the transition from NPC to bipolar neuron, and otherwise weakly positive ($r = 0.09 \pm 0.08$) with a slightly stronger correlation for the MSC to osteoblast transition ($r = 0.209$).

5.2.2.2 Transcription Factors The correlation was on average a bit stronger by 0.04 ± 0.04 for TFs than for non–TF and non–miRNA genes (other), and by 0.07 ± 0.2 when excluding very weak cases with a genome–wide correlation of $|r| < 0.1$. For H3K27ac, the overall TF correlation was on average positive with $r = 0.15 \pm 0.09$, and $r = 0.42 \pm 0.142$ for differentially expressed TFs. In the case of H3K4me3, it was positive as well with $r = 0.15 \pm 0.09$ ($r = 0.35 \pm 0.19$) for (differentially expressed) TFs, which was a bit stronger than for H3K4me2 where the correlation was $r = 0.12 \pm 0.04$ and $r = 0.29 \pm 0.14$ for all versus differentially expressed TFs, respectively. For promoter region H2A.Z, the average correlation was weakly positive for all TFs ($r = 0.06 \pm 0.05$) and differentially expressed TFs ($r = 0.022 \pm 0.16$), whereas it was only weakly positive for the transition from H1–hESC to GM23338 and from MSC to osteoblast in the gene body, with a strong correlation of $r = 0.55$ for differentially expressed TFs in the latter. For gene body H2A.Z for the other three transitions with available data, the average correlation was weakly negative for all and differentially expressed TFs, with $r = -0.07 \pm 0.07$ and $r = -0.16 \pm 0.13$, respectively.

5.2.2.3 MiRNAs The correlation in miRNAs generally matched the sign of the genome–wide correlation and was generally very weak with $r = 0.017 \pm 0.05$. There were not enough differentially expressed miRNAs to compute Pearson's correlation coefficient ($n < 20$) in three of six cell differentiation transitions. In the case of the transition of MSC to osteoblast, none of the miRNAs were differentially expressed and in the transition from NSPC to NPC all differentially expressed miRNAs were up–regulated, both of which lead to an undefined correlation coefficient. Similarly, none of the miRNA in the H1–hESC to GM23338 transition were differentially modified. In total, the correlation ($r = 0.13 \pm 0.12$) could only be computed for five groups of differentially

expressed miRNAs: H3K27ac in H1–hESC to GM23338 (promoter and gene body), as well as H3K4me2 (gene body) and H3K27ac (promoter and gene body) in MPC to monocyte.

5.2.3 Differentially Expressed Gene Annotations

For differentially expressed genes, GO and KEGG annotation enrichment analysis was performed, a more detailed summary of which can be found in Table B.3 in the appendix.

The differentially expressed genes in the H1–hESC to GM23338 and MPC to monocyte transitions were not enriched in annotations, except for the MPC to monocyte transition that was enriched in molecular function GO annotations with extracellular binding and signalling. In the transition from MSC to osteoblast, the differentially expressed genes were enriched for biological process GO annotations concerning embryonic organ development and morphogenesis, as well as regulation of development, cell proliferation and vasculogenesis, and extracellular signalling KEGG pathways.

The two neural cell differentiation transitions from NSPC and NPC to bipolar neuron were enriched in neural differentiation, neurogenesis, axonogenesis and nervous system development GO annotations (biological process), while the transition between the two progenitor cell types (NSPC to NPC) was enriched in nucleosome assembly and organisation, chromatin silencing and remodelling. Molecular functions were enriched in neurotransmitter, ion channel and transmembrane transport activity, and molecular binding. KEGG pathway annotations were enriched in axon guidance, cytoskeleton regulation, various signalling pathway, and disease processes.

5.2.4 Discussion

The observed positive correlation pattern for H3K27ac was consistent with the literature, which characterises H3K27ac as an activating mark of gene expression [6, 27, 36]. H3K27ac is one of the histone marks that peaks at the TSS [36], which is where the promoter and gene body definitions used in this thesis overlapped and as such could explain the negligible difference between promoter and gene body correlation.

Similarly, even though both H3K4me2 and H4K4me3 have been found to repress transcription in specific circumstances, they are generally described as activating marks, fitting the positive correlation pattern observed here [4, 22]. H3K4me2 marks peak in the middle of the gene body while also extending to the TSS and H3K4me3 peaks near the 5'–end of the gene body and the TSS [4, 22], thus resulting in similar but sightly higher correlation for the gene body compared to the promoter due to slightly overlapping definitions.

H2A.Z in promoter regions around the TSS is generally associated with active promoters [24, 25], which fit the small positive correlation pattern for promoter H2A.Z that was observed here. In gene bodies, H2A.Z can be associated with transcriptional repression [42], which matched the pattern for two of five cell differentiation transitions with available H2A.Z data. Furthermore, a study on murine ESCs found that H2A.Z is enriched in silent developmental genes and contributes to their repression, while being essential for ESC differentiation via promoter activation [43]. This was consistent with the negative correlation

of H2A.Z and gene expression in the gene bodies of the two transitions from NSPC and NPC to the terminally differentiated bipolar neuron cells. Additionally, this correlation was stronger in differentially expressed genes, which were enriched in neuron differentiation and fate commitment, as well as nervous system development processes.

However, the three transitions from H1–hESC to GM23338, MSC to osteoblast and NSPC to NPC had a positive correlation between differential gene expression and differential H2A.Z presence in gene bodies. That correlation was very small for both all and only differentially expressed genes in the first two transitions, but could be strong for differentially expressed genes ($r = 0.21$), and especially differentially expressed TFs ($r = 0.55$), in the latter, even though the differentially expressed genes were enriched for development and morphogenesis. As discussed in Talbert et al. (2010) [25], this might be explained by differences in H2A.Z acetylation or ubiquitylation, or by the presence of other histone variants like H3.3.

While histone modifications and variants play important roles in gene expression regulation, they are by far not the only factors contribute to gene expression, nor are they only involved in gene expression regulation. The stronger correlation in differentially expressed genes compared to all genes, the vast majority of which were not differentially expressed, could thus be explained by other cellular processes generating noise in the latter case, while in the differentially expressed subsets, gene expression processes were more prominent and expression levels more pronounced between the compared cell types. Stronger correlations for TFs could be explained similarly, since their primary function is participation in gene expression.

Most correlation coefficients were statistically significant with $p < 0.05$. Instances where that was not the case generally involved a very small sample size, which is problematic for the χ^2–test of independence [121], or very small correlation coefficients. For large gene sets involving 30,000 or more genes, even these very weak correlations tested as statistically significant, which might say more about the sample size than the differential gene expression and histone modification correlation [142].

Lastly, almost non–existent correlation as well as the many instances of constant histone modifications for miRNAs are discussed in Section 5.4.

5.3 TF–miRNA Co–Regulatory Motifs

5.3.1 Number of Co–Regulatory Motifs

TF–miRNA co–regulatory motifs were identified in the gene regulatory network of each cell differentiation transition, as well as the complete regulatory network constructed from the TFmiR database. In the complete network, a total of 11,927 motifs were found, of which 11,786 (98.8%) were cases of simple co–regulation, 85 (0.7%) were miRNA–FFLs, 52 (0.4%) TF–FFLs and four (0.03%) composite–FFLs. Table 5.5 presents an overview.

Those numbers were much lower in the transition networks, with $3{,}132 \pm 881$ co–regulations, 31 ± 6 miRNA–FFLs, 15 ± 3 TF–FFLs and only one composite–FFL, except for the H1–hESC to GM23338 transition, which had two. There were no TF–miRNA co–regulatory motifs in the differential networks due to the

Table 5.5: *Number of TF–miRNA co–regulatory motifs in the complete network constructed from the human gene regulatory database obtained from TFmiR, as well as in the cell differentiation specific networks. The total motif count is further broken down into miRNA–FFLs, TF–FFLs, and composite–FFLs and simple co–regulatory motifs. Differential network versions only include interactions between genes that were both significantly differentially expressed in the respective cell differentiation specific network. Motif counts that were significantly higher with $p < 0.05$ than in randomised networks are marked in italic, or in bold if they were additionally still statistically significant after correcting for multiple hypothesis testing.*

Transition	Network	Total	Co–Reg.	miRNA	TF	Comp.
H1–hESC	complete	**4,831**	**4,771**	41	17	2
→ GM23338	differential	1	1	0	0	0
mesenchymal stem cell	complete	**3,110**	**3,061**	*34*	14	1
→ osteoblast	differential	0	0	0	0	0
myeloid progenitor	complete	**3,192**	**3,144**	29	18	1
→ monocyte	differential	0	0	0	0	0
neural progenitor	complete	**3,082**	**3,032**	31	18	1
→ bipolar neuron	differential	0	0	0	0	0
neural stem progenitor	complete	**2,221**	**2,185**	26	9	1
→ bipolar neuron	differential	0	0	0	0	0
neural stem progenitor	complete	**2,643**	**2,601**	*25*	*16*	1
→ neural progenitor	differential	0	0	0	0	0
complete network		**11,927**	**11,786**	85	52	4

low number of miRNAs nodes and miRNA → target interactions, apart from a single co–regulation in the transition from H1–hESC to GM23338.

TF–miRNA co–regulatory motifs occurred significantly more often in the examined biological networks compared to randomised networks ($p < 0.05$). Upon closer examination, this was mainly an effect of co–regulation being significantly enriched, whereas the number of TF–FFLs, miRNA–FFLs and composite–FFLs was not statistically significant, apart from TF–FFLs in one and miRNA–FFLs in two cell differentiation transitions. However, those exceptions were no longer significant after correcting for multiple hypothesis testing, and might thus be false positives.

5.3.2 Correlation in Co–Regulatory Motifs

Due to the lack of TF–miRNA co–regulatory motifs in differential networks, only motifs in the cell differentiation transition networks consisting of both differentially and non–differentially expressed genes were considered. An overview of the motif correlations can be found in Table 5.6. Motif correlations were compared to those in randomised networks, a selection of substantial differences can

be found in Figure 5.1, while the full comparisons can be found in Figures B.1, B.2 and B.3 in the appendix.

5.3.2.1 H3K27ac The motifs showed a positive correlation pattern with $r = 0.25 \pm 0.13$, with the exception of a negative pattern ($r = -0.08 \pm 0.03$) for co–regulation, TF–FFLs and miRNA–FFLs in the bodies of genes involved in the transition from NSPC to NPC. In that transition, the correlation for co–regulation was non–existent ($r = 0$) in the promoter and none of the genes in the single composite–FFLs were differentially expressed, nor differentially acetylated, resulting in an undefined coefficient. The same problem occurred in the transition from MPC to monocyte. The correlation for TF–FFLs in the transition from H1–hESC to GM23338 was also non–existent with $r = 0$, and smaller ($p < 0.05$) than for TF–FFLs in randomised networks. In contrast, in the same transition, the correlation for miRNA–FFLs ($r = 0.17$ and $r = 0.32$ for promoter and gene body, respectively) was higher than in randomised networks ($p < 0.05$), both in the absolute strength as well as with regard to the sign.

5.3.2.2 H3K4me2 The overall pattern was a small positive correlation ($r = 0.19 \pm 0.19$) for the two transitions for which H3K4me2 ChIP–seq data was available, namely H1–hESC to GM23338 and NPC to bipolar neuron. There was no differential modification for the composite–FFL in the latter and the promoter region of TF–FFLs and composite–FFLs in the former. The miRNA–FFLs in the H1–hESC to GM23338 transition had a very small negative correlation with $r = -0.003$. None of the correlations were significantly stronger or weaker than in randomised networks.

5.3.2.3 H3K4me3 Similar to H3K4me2, there was a pattern of small positive correlation ($r = 0.16 \pm 0.12$). All composite–FFLs correlation coefficients were undefined due to constant non–differential gene expression or constant non–differential histone methylation. There were three cases of non–existent correlation ($r = 0$), namely for the gene bodies of miRNA–FFLs in the transition from MPC to monocyte, and for the promoters of co–regulation and TF–FFLs in the H1–hESC to GM23338 transition. The latter case was smaller than in randomised networks ($p < 0.05$), as was the very weak negative correlation ($r = -0.003$) for the miRNA–FFLs in the same network. The promoter correlation of miRNA–FFLs in the transition from NSPC to bipolar neuron was higher than in randomised networks ($p < 0.05$), but not in the absolute sense.

5.3.2.4 H2A.Z In the transition from NSPC to NPC there was a very small positive correlation between expression and H2A.Z presence ($r = 0.05 \pm 0.03$), where the correlation of $r = 0.009$ for gene bodies of miRNA–FFLs was significantly smaller than in randomised networks, while no gene in the composite–FFLs was differentially expressed. Similarly, the correlation was positive ($r = 0.17 \pm 0.07$) in the transition from H1–hESC to GM23338, except for the negative correlation ($r = -0.058$) for co–regulation in gene bodies, and constant non–differential presence in the only two composite–FFLs. The correlation in gene bodies of miRNA–FFLs ($r = 0.28$) was stronger than in randomised networks ($p < 0.05$), regardless of sign, while the correlation in gene bodies of TF–FFLs ($r = 0.013$) and miRNA–FFLs ($r = 0.16$) was also higher than in randomised networks ($p < 0.05$) but only with regard to the positive sign.

Table 5.6: *Correlation between differential gene expression (X) and differential histone modifications or variants (Y) in m TF–miRNA co–regulatory motifs comprised of n genes, for different cell differentiation transitions. For each transition, the correlation is given for TF–FFLs (TF), miRNA–FFLs (miRNA), composite–FFLs (comp.) and simple co–regulatory motifs (co–reg.). Histone modifications and variants are distinguished by their location in the promoter or gene body region. Negative correlation is highlighted in red, positive correlation in blue, and darker highlighting signifies stronger correlation. A statistically significant dependent relationship between expression and modification or variant is marked in bold, with $p < 0.05$ and after correcting for multiple hypothesis testing. X=0 and Y=0 mark cases without differential gene expression or without differential modifications, respectively. Coefficients that were significantly higher than motifs of the same type in randomised networks with $p < 0.05$ are underlined, while cases where it was smaller are italicised. Histone modifications without ChIP–seq samples are left blank.*

				H2A.Z		H3K4me2		H3K4me3		H3K27ac	
Transition	**Motif**	m	n	**prom.**	**gene**	**prom.**	**gene**	**prom.**	**gene**	**prom.**	**gene**
H1–hESC	co–reg.	4,771	1,036	**0.097**	**-0.058**	0	**0.127**	0	0.060	0.077	**0.127**
↓	TF	17	84	0.207	0.133	Y=0	**0.335**	*0*	**0.204**	*0*	0.248
GM23338	miRNA	41	88	**0.276**	0.159	-0.003	**0.246**	*0.003*	**0.196**	0.165	**0.319**
	comp.	2	13	Y=0	Y=0	Y=0	0.677	Y=0	Y=0	0.123	0.135
mesenchymal	co–reg.	3,061	904	**0.128**	0.041						
stem cell	TF	14	64	**-0.081**	**-0.168**						
↓	miRNA	34	64	0.202	0.114						
ostebolast	comp.	1	10	X=0	X=0						
myeloid	co–reg.	3,144	872					**0.115**	**0.244**	**0.188**	**0.250**
progenitor	TF	18	85					**0.320**	**0.311**	**0.276**	**0.288**
↓	miRNA	29	68					0.036	0	**0.314**	0.243
monocyte	comp.	1	10					X=0	X=0	X=Y=0	X=0
neural	co–reg.	3,032	839	-0.060	**-0.252**	**0.081**	**0.122**	0.003	**0.287**	**0.214**	**0.276**
progenitor	TF	18	90	**0.137**	**-0.253**	0.024	0.144	Y=0	**0.198**	**0.339**	**0.306**
↓	miRNA	31	72	0.028	-0.138	0.023	0.147	Y=0	**0.164**	0.180	0.275
bipolar neuron	comp.	1	10	Y=0	0.167	Y=0	Y=0	Y=0	Y=0	0.167	**0.218**
neural stem	co–reg.	2,185	780	0.011	**-0.075**			**-0.043**	**0.210**	**0.230**	**0.249**
progenitor	TF	9	53	-0.158	-0.129			Y=0	**0.385**	**0.390**	0.361
↓	miRNA	26	60	0.014	-0.126			0.024	0.209	0.141	0.268
bipolar neuron	comp.	1	10	Y=0	-0.559			Y=0	Y=0	0.667	0.559
neural stem	co–reg.	2,601	784	0.051	0.063			0.006	0.020	0	**-0.051**
progenitor	TF	16	74	0.041	0.075			Y=0	**-0.014**	**0.020**	-0.068
↓	miRNA	25	54	0.077	*0.009*			Y=0	Y=0	0.048	-0.118
neural progenitor	comp.	1	10	X=0	X=0			X=Y=0	X=Y=0	X=Y=0	X=0

The pattern was positive ($r = 0.12 \pm 0.06$) for co–regulation and miRNA–FFLs in the transition from MSC to osteoblast, and negative for TF–FFLs ($r = -0.08$ and $r = -0.17$ for promoter and gene body, respectively). The latter two motifs had a stronger negative correlation than in randomised networks ($p < 0.05$), which was also the case in the absolute sense for the gene body. None of the genes involved in the single composite–FFL were significantly expressed.

In contrast, the pattern was negative ($r = -0.19 \pm 0.15$) for the two transitions from NSPC and NPC to bipolar neuron, with the exception of very small positive correlations for promoters of co–regulation ($r = 0.011$) and miRNA–FFLs ($r = 0.014$) in the NSPC to bipolar neuron transition, and the promoters of miRNA–FFLs ($r = 0.028$) in the NPC to bipolar neuron transition. The promoter region of composite–FFLs was not differentially enriched in H2A.Z

for both transitions. None of the correlations were significantly stronger than in randomised networks.

The cases with stronger or weaker correlation than in randomised networks with $p < 0.05$ were not statistically significant after multiple hypothesis testing correction, which might mean that those cases were false positives. In addition, due to the low number of composite–FFLs, the number of involved genes was between 10 and 13, while it is generally recommended to have at least 20 data points for Pearson's correlation coefficient.

5.3.3 Comparison with Genome–Wide Correlation

The differential gene expression and differential histone modification or variant correlation was compared between genes in TF–miRNA co–regulatory motifs and all genes not in those motifs, as well as differentially expressed genes not in those motifs. Table B.4 in the appendix presents a detailed overview of all differences, while Tables 5.4 and 5.6 contain the genome–wide correlation coefficients and motif correlations, respectively.

5.3.3.1 All Genes The correlation in TF–miRNA co–regulatory motifs was overall stronger by 0.06 ± 0.12, not counting cases where a sign change occurred. There were several motifs whose correlation was significantly stronger ($p < 0.05$) by an average of 0.14 ± 0.06: co–regulation for H3K27ac correlation ($r = 0.19$ and $r = 0.25$ for promoter and gene body, respectively) in the transition from MPC to monocyte, gene body H3K4me2 correlation ($r = 0.34$) in TF–FFLs in the H1–hESC to GM23338 transition, co–regulation for gene body H3K27ac in the transitions from NSPC to bipolar neuron ($r = 0.25$) and from NPC to bipolar neuron ($r = 0.28$), co–regulation for promoter H2A.Z in the transition from MPC to osteoblast ($r = 0.13$), and finally co–regulation for gene body H2A.Z for the NPC to bipolar neuron transition ($r = -0.25$). Promoter H3K4me3 was significantly weaker ($p < 0.05$) for co–regulation in the NPC to bipolar neuron transition ($r = 0.003$).

Furthermore, in three cases the correlation was significantly lower ($p < 0.05$) by an average of 0.1 ± 0.02 in co–regulation motifs, resulting in a sign change from positive to negative: promoter H2A.Z in the NPC to bipolar neuron transition ($r = -0.06$), gene body H2A.Z in the H1–hESC to GM23338 transition ($r = -0.06$), and for promoter H3K4me3 in the transition from NSPC to bipolar neuron ($r = -0.04$).

5.3.3.2 Differentially Expressed Genes In most cases, the correlation in TF–miRNAs co–regulatory motifs was weaker by 0.17 ± 0.12, of which 44.7% were significantly weaker ($p < 0.05$), not including cases where a sign change took place. In some cases, the correlation in motifs was stronger by an average of 0.1 ± 0.11, of which none were statistically significant. The latter occurred mostly frequently in TF–FFLs and miRNA–FFLs in the transition from H1–hESC to GM23338.

In addition to the three instances of a significant sign change in the comparison with all genes, there were two more such cases here, since the difference became more extreme due to the stronger positive correlation in differentially expressed genes compared to all genes.

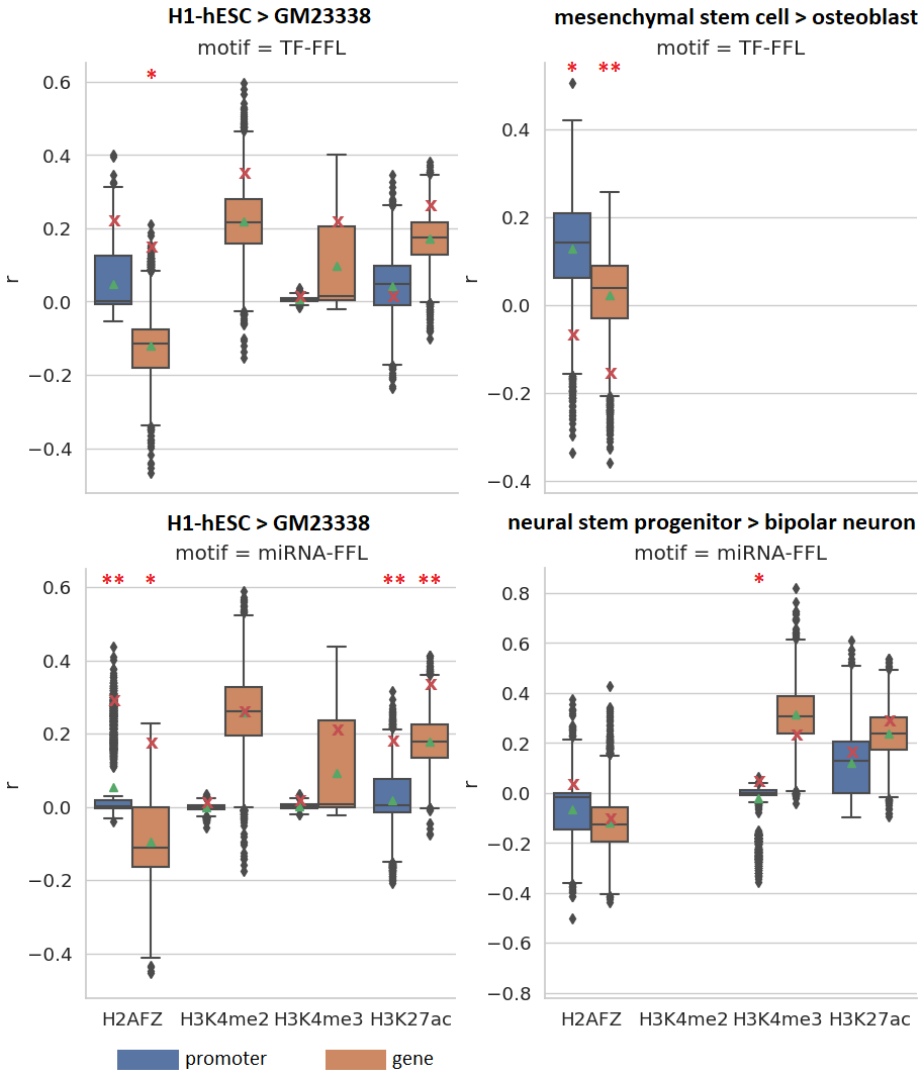

Figure 5.1: *Comparisons of differential gene expression and differential histone modification or variant correlation r in actual TF–miRNA co–regulatory motif types (red X) and motif types in 5,000 randomised regulatory networks (box plots) for two cell differentiation transitions, distinguished by promoter region and gene body. The box plots give the median (middle black bar), mean (green triangle), and outliers (black diamonds). Single stars mark cases with $p_{\text{sign}} < 0.05$ (stronger correlation with the same sign), and double stars cases with $p_{\text{abs}} < 0.05$ (stronger absolute correlation). The p–values were not statistically significant after correcting for multiple hypothesis testing.*

For both types of comparisons, there were additional cases with $p < 0.05$ for the correlation difference that were not statistically significant after multiple hypothesis testing correction and thus excluded from this summary as potential false positives. Furthermore, there were instances of larger effect size differences of > 0.2, usually for TF–FFLs or miRNA–FFLs, that were not statistically significant, potentially due to the small number of these motifs and consequently small number of involved genes.

5.3.4 Promoter versus Gene Body Modifications

Similar to the genome–wide examination, the correlation was statistically compared between promoter and gene body, in order to investigate potential differences. Table 5.6 contains the correlation coefficients in the TF–miRNA co–regulatory motifs, while Table B.5 in the appendix details the difference and statistical significance of all comparisons.

5.3.4.1 H3K27ac Overall, the difference between promoter and gene body correlations was quite small (0.04 ± 0.08) and none of them were statistically significant. However, in the transition from NSPC to NPC, the promoter and gene body differed in sign for TF–FFLs ($r = 0.02$ and $r = -0.07$, respectively) and miRNA–FFLs ($r = 0.05$ and $r = -0.12$, respectively).

5.3.4.2 H3K4me2 The correlation for gene bodies was consistently higher (0.07 ± 0.08) than in the promoter, whereby the difference was statistically significant ($p < 0.05$) for co–regulation in the transtion from H1–hESC to GM23338, with $r = 0$ and $r = 0.13$ in the promoter and gene body, respectively.

5.3.4.3 H3K4me3 Similar to H3K4me2, the correlation was generally higher (0.06 ± 0.1) in the gene body than in the promoter region. There two exceptions in the transition from MPC to monocyte, where the promoter correlation was minimally higher (0.02 ± 0.02) than in the gene body for TF–FFLs and miRNA–FFLs. The difference was statistically significant with $p < 0.05$ for TF–FFLs and miRNA–FFLs in the H1–hESC to GM23338 transition, and co–regulation in the NPC and NSPC to bipolar neuron transitions.

5.3.4.4 H2A.Z In the transition from NSPC to NPC, there was not much difference between the promoter and gene body correlation (-0.005 ± 0.04). The correlation was overall slightly stronger (0.04 ± 0.07) in the gene body that in the promoter. For co–regulation in the transition from NPC to bipolar neuron, the gene body correlation was significantly ($p < 0.05$) stronger than in the promoter region ($r = 0. - 25$ and $r = -0.06$, respectively). In the same transition, a sign change occurred for miRNA–FFLs and TF–FFLs, resulting in a significant difference ($p < 0.05$) for the latter, with $r = 0.14$ and $r = -0.25$ for promoter and gene body, respectively.

In the transition from MSC to osteoblast, the correlation in the promoter was stronger than in the gene body for miRNA–FFLs ($r = 0.2$ and $r = 0.11$, respectively) and co–regulation ($r = 0.13$ and $r = 0.04$, respectively), where the latter difference was statistically significant with $p < 0.05$, whereas the negative correlation was stronger in the gene body than in the promoter ($r = -0.17$ and $r = -0.08$, respectively) for TF–FFLs. Lastly, the correlation in the promoter

region was higher than in the gene body (0.07 ± 0.07) in the H1–hESC to GM23338 transition, whereby a statistically significant ($p < 0.05$) sign change occurred for co–regulation ($r = 0.1$ and $r = -0.06$ in promoter and gene body, respectively).

5.3.5 Motif Annotations

GO and KEGG annotation enrichment analysis was performed for each TF–miRNA co–regulatory motif for each cell differentiation transition with FDR–controlled p–value threshold of 0.05. A detailed summary of the top 30 enriched annotations for each category can be found in Table B.6 in the appendix.

Overall, KEGG pathway annotations were generally enriched for cancer, viral infections, cell senescence and cell cycle. Enriched molecular function GO annotations mainly included binding to HDACs, HATs, TFs, enzymes, ubiquitin–ligases, kinases, cyclins, DNA and phosphatases. Biological process GO annotations for the FFL motifs were overall enriched for regulation of transcription, development, cell differentiation, apoptosis and cell proliferation, except for simple co–regulation which was only enriched for transcription regulation in all cell differentiation transitions.

The genes involved in the single composite–FFL were the same for all cell differentiation transitions, except the transition from H1–hESC, and thus shared the same annotation enrichment. The FFL was enriched in ossification, neural crest migration, hypoxia response, in addition to transcription and apoptosis regulation.

5.3.5.1 H1–hESC to GM23338 Co–regulation was additionally enriched for pluripotency in KEGG. Composite– and TF–FFLs showed enrichment for hypoxia response, and the latter in addition for structure morphogenesis.

5.3.5.2 MSC to Osteoblast Similar to the previous transition, co–regulation was enriched for pluripotency, as well as osteoclast differentiation. TF–FFLs were additionally enriched in hypoxia response and structure morphogenesis, and miRNA–FFLs in organ morphogenesis.

5.3.5.3 MPC to Monocyte In this transition, TF–FFLs showed further enrichment for structure morphogenesis and cytokine response, and miRNA–FFLs for haematopoiesis and immune differentiation, while co–regulation was enriched for pluripotency and immune differentiation.

5.3.5.4 NPC to Bipolar Neuron The annotations matched the general case.

5.3.5.5 NSPC to Bipolar Neuron MiRNA–FFLs were further enriched in neuron generation and G1/S–phase transition, and the TF–FFLs in hypoxia response and structure morphogenesis.

5.3.5.6 NSPC to NPC Both TF– and miRNA–FFLs were additionally enriched in hypoxia response.

5.3.6 Discussion

5.3.6.1 General Annotation Patterns Unsurprisingly, TF–miRNA co–regulatory motifs were generally enriched for regulatory process annotations for general transcription as well as development, differentiation and cell proliferation, which matched their previously established roles in these processes as well as the cell differentiation context [7, 19]. Enrichment in KEGG disease pathways such as cancer was similarly consistent with previous findings [7, 15, 20]. Matching the regulatory role of these motifs, molecular function was enriched in regulatory binding to DNA as well as effectors such as HATs, HDACs, TFs, ubiquitin–ligases or phosphatases, suggesting that these motifs might in part regulate gene expression by contributing to the establishment, removal or maintenance of epigenetic modifications.

While differentially expressed genes were enriched in processes and functions that were more specific to the respective cell differentiation transition, the enrichment of TF–miRNA co–regulatory motifs in them was more general. This might be a result of motifs including both differentially and not differentially expressed genes, as well as motifs in which none of the genes were differentially expressed in a given cell differentiation transition.

5.3.6.2 TF–miRNA Co–Regulatory Motif Patterns Overall, TF–miRNA co–regulatory motifs exhibited a stronger correlation than was observed genome–wide, significantly so in some instances. In a few other instances, the motif correlation was negative while the corresponding genome–wide correlation was positive, resulting in a significant sign change. However, the correlation in motifs was generally lower than for differentially expressed genes.

The simple TF–miRNA co–regulation motifs were the most abundant, likely due to their simplicity, and also exhibited only very general regulatory process enrichment, apart from pluripotency enrichment for cell differentiation transitions including one or two pluripotent cell types. While they occurred significantly more often than in randomised networks, the there was no instance where the correlation was significantly stronger compared to those randomised networks. However, there were instances where the correlation differed significantly from the genome–wide correlation.

MiRNA–FFLs are more complex, and as such were much less abundant than simple co–regulation. TF–FFLs were even less abundant than miRNA–FFL, even though there were more TF $\rightarrow$ miRNA than miRNA $\rightarrow$ TF interactions in the gene regulatory networks.

In some transitions, such as from H1–hESC to GM23338, the correlations in TF–FFLs and miRNA–FFLs, was both stronger compared to the genome–wide correlation and additionally stronger than in the vast majority of randomised networks. Furthermore, both exhibited no correlation in some cases with positive genome–wide correlation, which was also very unusual compared to randomised networks. TF–FFLs in the transition from MSC to osteoblast had a negative correlation for H2A.Z, while the genome–wide correlation was positive and quite strong for differentially expressed genes, which was also highly unusual compared to randomised networks. However, even though the p–value was < 0.05 in these cases, it was not statistically significant after multiple hypothesis testing correction, so that it is not clear if these instances were false positives.

Furthermore, miRNA–FFLs were sometimes additionally enriched for cell differentiation specific development and differentiation processes, while TF–FFLs were further enriched in hypoxia response in transitions involving stem cells, as well as structure morphogenesis in some transitions. Oxygen concentrations and conversely hypoxia play a role in stem cell maintenance and differentiation [143].

There were overall two composite–FFLs, one involving ten genes that occurred in all six cell differentiation transitions, and one comprised of three genes that only occurred in the H1–hESC to GM23338 transition. Composite–FFLs were enriched in hypoxia response, as well as ossification and neural crest migration, which matched some of the cell differentiation transitions. However, the co–regulatory TFs and miRNAs were not differentially expressed in any of the six transitions, and it is thus questionable to which degree the motif would play a role in these specific differentiation transitions.

The correlation for these FFLs were either quite strong or not defined due to constant differential gene expression or lack of differential histone marks. Since it is generally recommended to have 20 or more data points for Pearson's correlation coefficient, the observed correlations might not be reliable. Additionally, randomised correlations comparisons could not be performed reliably, since many randomised networks did not contain any composite–FFLs.

Overall, the four different motif types exhibited different patterns and deviated from the genome–wide correlation pattern in some instances.

5.3.6.3 Histone Mark–Centric Patterns All four TF–miRNA co–regulatory motif types showed positive correlation with H3K27ac for both promoter and gene body, with little difference between the two, which matches both the observed genome–wide pattern and the activating effect of H3K27 acetylation in the literature [6, 27, 36]. As discussed for the genome–wide correlation, the small difference between promoter and gene body could be explained by overlapping promoter and gene body definitions around the TSS, which is where H3K27ac marks tend to peak [36]. The exception to this pattern occurred in the transition from NSPC to NPC, where differential expression and H3K27ac were negatively correlated for the gene body, which was not the case genome–wide.

Consistent with the observed genome–wide correlation pattern, as well as the generally activating effect of H3K4me2 and H3K4me3 [4, 22], the TF–miRNA co–regulatory motifs exhibited a positive correlation between differential gene expression and H3K4 di– and tri–methylation. In comparison with the genome–wide case, the difference between promoter and gene body methylation was more pronounced, whereby the correlation was higher in the gene body than in the promoter region. Specifically, many motif types were not differentially methylated in the promoter region, resulting in an undefined correlation coefficient, and others had a very small or even zero correlation. Both H3K4me2 and H3K4me3 are generally more strongly associated with the gene body, the former peaking in the middle while the latter typically peaks at the 5'–end near the TSS [4, 22]. In the discussion of the genome–wide case, the similarity between promoter and gene body correlation was ascribed to overlapping promoter and gene body definitions near the TSS, if that is true, then the lack of differential methylation in promoters seems to suggest that the methylation marks were located more towards the middle or 3'–end of the gene body.

The correlation was overall positive for promoter H2A.Z, fitting the genome–wide pattern and the general association of H2A.Z with active promoters [24, 25]. However, in contrast to the genome–wide case, TF–FFLs in NSPC to bipolar neuron and MSC to osteoblast transitions and co–regulation in the NPC to bipolar transition exhibited negative correlation in for the promoter region.

H2A.Z can be associated with transcriptional repression in gene bodies [42], which was the case for inactive developmental genes in a murine ESCs study [43], but also with transcriptional activation, which has been explained with differently acetylated and ubiquitinated H2A.Z or the presence of other histone variants [25]. For H2A.Z presence in gene bodies of TF–miRNA co–regulatory motifs, a similar pattern emerged to the genome–wide case: The correlation was generally positive for the transitions from H1–hESC to GM23338, MSC to osteoblast and NSPC to NPC, while the transitions to bipolar neuron from NSPC and NPC showed a negative correlation. However, while the genome–wide correlation was positive for gene body H2A.Z in the transitions from H1–hESC to GM23338 and from MSC to osteoblast, it was negative for co–regulation in the former and TF–FFLs in the latter, whereby both differences were statistically significant with $p < 0.05$.

5.3.6.4 Cell Differentiation Transition Patterns While the correlations with different histone modifications or variants generally followed a similar pattern across cell differentiation transitions, as discussed above, there were differences between them regarding the seeming importance of the different histone marks as well as some pronounced differences, such as the correlation sign for H2A.Z in gene bodies.

For example, the correlation for the transition from NSPC to NPC was very weak for all four histone marks, both genome–wide and for TF–miRNA co–regulatory motifs, except for differentially expressed TFs. In contrast, the transition from H1–hESC to GM23338 had similar correlations for most marks and locations but stronger correlations for H3K27ac and gene body H3K4me3, while in the transition from NPC to bipolar neuron had stronger overall correlations with weaker correlation for promoter H2A.Z and promoter H3K4me3, relatively speaking.

This was consistent with the idea of the histone code, whereby histone marks have combinatorial effects on chromatin structure and gene expressions, as well as other contextual factors, which can differ from cell type to cell type, and thus between cell differentiation transition [4, 36, 44, 45]. In the same vein, benchmarking results from *DeepDiff* showed that prediction of differential gene expression based on H3K4me1, H3K4me3, H3K36me3, H3K9me3 and H3K27ac worked better for some combinations of cell types than others [46].

5.4 Limitations and Future Work

While the methodology employed in this thesis was able to identify consist patterns, there were several limitations and problematic aspects.

First, the Ensembl gene definitions only included pre–miRNA not pri–miRNA definitions. However, the pri–miRNA is the relevant transcriptional unit for miRNAs with regard to promoter definitions and chromatin modifications. As a consequence, the gene body of miRNAs was only 180 bp long on average, and

the promoter definition not only contained the entire gene body, but likely also most of the pri–miRNA gene body. This could explain the overall barely existing genome–wide correlation for miRNAs, as well as the comparatively frequent instances of lacking differential histone marks.

Furthermore, only 18.9%to 44.9% of the total 1,879 miRNAs had differential expression information, which could partly be a result of using standard RNA–seq, instead of accompanying small RNA–seq that is more suitable for measuring miRNA expression. Another problem involving miRNA were the comparatively small numbers of miRNA $\rightarrow$ TF and miRNA $\rightarrow$ gene regulatory interactions. This likely contributed to the even smaller portion of miRNAs with differential expression data (8–16%) in the transition specific regulatory networks, and thus motif identification.

In the same vein, the differential networks that only contained differentially expressed genes were too strict for the given expression and interaction data, containing no TF–miRNA co–regulatory motifs. It might have been better to instead, or in addition, build intermediate gene regulatory networks from differentially expressed genes as well as genes directly interacting with differentially expressed genes, thus obtaining more cell differentiation transition specific networks that contain enough genes and regulatory interactions to contain TF–miRNA co–regulatory motifs.

Moreover, this work only considered histone modifications or variants in isolation even though these histone marks seem to generally function in combination, therefore identifying relatively small effects. Additionally, while the categorisation approach for differential gene expression and differential histone marks worked reasonably well, it was rather simplistic and was likely not able to capture many nuances. In future, it would probably be better to employ an approach that considers different modifications or variants in combination with a more sophisticated model.

Lastly, the employed approach for motif correlations merged all TFs, miRNAs and shared target genes of a motif type into a single gene set, which allowed exploration of very general differences between the genome–wide correlation baseline and correlation in these motif types. However, it did not examine if there is a tight relationship between marks on individual components of these motifs and gene expression of target genes.

For instance, in a coherent TF–FFL, where the TF represses the shared target gene and activates the miRNA that also represses the shared target gene, histone marks might amplify the co–regulatory effect of the TF and miRNA, if for example TF and miRNA carried activating marks, while the repressed target gene is additionally modified with repressive marks. Since the correlation in the different motif types differed amongst each other as well as from the genome–wide baseline, it might be worthwhile to study this relationship in more detail.

Chapter 6

Conclusion

This work explored the correlation of differential gene expression and differential histone modification or variants in TF–miRNA co–regulatory motifs in human cell differentiation transitions.

Differential gene expression and differential histone modification and variant analyses were conducted. Subsequently, TF–miRNA co–regulatory motifs were identified in cell differentiation specific gene regulatory networks, as well as randomised networks. The correlation between differential gene expression and differential histone marks was computed for each motif type and contrasted with the genome–wide correlation as a baseline, as well as motifs in randomised networks. Additionally, correlation with histone marks in promoter regions was compared to histone marks in gene bodies. Lastly, an annotation enrichment analysis was performed for motif types and differentially expressed genes.

Overall, genome–wide correlation patterns matched the general patterns of individual histone marks or variants given in the literature, and the correlation was stronger for differentially expressed genes, which were enriched in cell differentiation specific process annotations. Differences between promoter and gene body were largely histone mark–specific, however, cell differentiation transition specific differences were observed as well.

TF–miRNA co–regulatory motifs were generally significantly enriched in biological cell differentiation networks compared to randomised networks. The correlation patterns in these were also consistent with previously reported patterns of the respective histone modification or variation. However, they were overall stronger compared to corresponding genome–wide correlations, and there were instances in which the motif and genome–wide correlation differed in sign.

Additionally, motif types showed correlation differences amongst each other, whereby simpler and more common motifs had more consistent trends across cell differentiation transitions and different histone marks, whereas more complex and less abundant motif types such as TF– and miRNA–FFLs exhibited stronger fluctuations. Correlation in the latter was in some cases unusually strong compared to motifs of the same type in randomised networks.

Furthermore, TF–miRNA co–regulatory motifs were enriched in molecular function annotations that seem to suggest contribution to the maintenance, establishment and removal of epigenetic chromatin and DNA modifications. While the more complex motif types were generally enriched in process annotations such as regulation of transcription, development, cell differentiation and pro-

liferation, they differed in some specifics across the various cell differentiation transitions.

In conclusion, the results of this general exploration identified differences between TF–miRNA co–regulatory motifs and genome–wide correlation patterns as well as between motif types that warrant a closer examination of the role of histone marks in these motifs.

Appendix A

Extended Materials

Table A.1: *Experiment and file accessions for all RNA–seq data sets used in the thesis, grouped by cell type, laboratory, and experiment. Some experiments produced suitable RNA–seq data sets for multiple biological samples. These samples (S) were assigned a numerical ID for each experiment. Additionally, the sequencing run–type is given as either single–read or paired–end. All listed files were acquired in FASTQ–format from ENCODE. Location abbreviations: University of Massachusetts (UMass), Cold Spring Harbor Laboratory (CSHL), California Institute of Technology (Caltech), University of California San Francisco (UCSF).*

Cell Type	Laboratory	Experiment	S	Run	Accession(s)
bipolar neuron	Barbara Wold, Caltech	ENCSR023VVO	1	single	ENCFF861BZO ENCFF277ETM
			2	single	ENCFF162SWD ENCFF216DGZ
bipolar neuron	Thomas Gingeras, CSHL	ENCSR603RPC	1	single	ENCFF912BHI
			2	single	ENCFF257FZI
bipolar neuron	Thomas Gingeras, CSHL	ENCSR968WKR	1	paired	ENCFF644CJJ ENCFF774ZGB
			2	paired	ENCFF743MBB ENCFF040TFC
common myeloid progenitor (CD34+)	Zhiping Weng, UMass	ENCSR830HIN	1	paired	ENCFF644NDH ENCFF068EOC
				paired	ENCFF919ARA ENCFF760DAJ
				paired	ENCFF968XRL ENCFF762RCL
				paired	ENCFF790PEK ENCFF598CWC
				paired	ENCFF210EBO ENCFF105BMM

Continued on next page...

Table A.1 – *Continued.*

Cell Type	Laboratory	Experiment	S	Run	Accession(s)
				paired	ENCFF447UNC ENCFF850FEW
				paired	ENCFF359UBE ENCFF574WPO
				paired	ENCFF065OFV ENCFF291VSA
				paired	ENCFF151WCV ENCFF167JUK
				paired	ENCFF196LEM ENCFF102ATN
				paired	ENCFF730BGR ENCFF069XUY
				paired	ENCFF565PAJ ENCFF199NLJ
				paired	ENCFF339YYU ENCFF861PHK
				paired	ENCFF959MBA ENCFF739LBD
				paired	ENCFF412IIT ENCFF501PAR
				paired	ENCFF510WNH ENCFF935WFU
				paired	ENCFF396DUV ENCFF649TLF
GM23338	Barbara Wold, Caltech	ENCSR748GVH	1	single	ENCFF736FIO ENCFF267VJG
			2	single	ENCFF563KDS ENCFF065YUF
GM23338	Thomas Gingeras, CSHL	ENCSR722POQ	1	single	ENCFF004BLT
			2	single	ENCFF450POH
GM23338	Thomas Gingeras, CSHL	ENCSR938LSP	1	paired	ENCFF002DHX ENCFF002DHY
			2	paired	ENCFF002DJY ENCFF002DKC
H1–hESC	Barbara Wold, Caltech	ENCSR962TBJ	1	paired	ENCFF000DJL ENCFF000DJO
				paired	ENCFF000DJM ENCFF000DJP
				paired	ENCFF000DJN ENCFF000DJQ
H1–hESC	Barbara Wold, Caltech	ENCSR000EYP	1	paired	ENCFF000DGR ENCFF000DGZ
				paired	ENCFF000DHA ENCFF000DGS

Continued on next page...

Table A.1 – *Continued.*

Cell Type	Laboratory	Experiment	S	Run	Accession(s)
			2	paired	ENCFF000DGT ENCFF000DHB
				paired	ENCFF000DHC ENCFF000DGU
			3	paired	ENCFF000DGV ENCFF000DHD
				paired	ENCFF000DGW ENCFF000DHE
			4	paired	ENCFF000DGX ENCFF000DHF
				paired	ENCFF000DHG ENCFF000DGY
H1–hESC	Zhiping Weng, UMass	ENCSR043RSE	1	paired	ENCFF953ZDW ENCFF346WNU
H1–hESC	Zhiping Weng, UMass	ENCSR670WQY	1	paired	ENCFF565ZQD ENCFF039JYG
H1–hESC	Joseph Costello, UCSF	ENCSR950PSB	1	paired	ENCFF589VNC ENCFF402HNS
				paired	ENCFF199CUO ENCFF772ZUR
				paired	ENCFF608OLY ENCFF238OAC
				paired	ENCFF687WLZ ENCFF092SYX
				paired	ENCFF350HDB ENCFF806OYD
				paired	ENCFF247WDK ENCFF688IPL
				paired	ENCFF567PCA ENCFF674AED
H1–hESC	Thomas Gingeras, CSHL	ENCSR000CQR	2	paired	ENCFF000FIM ENCFF000FIL
H1–hESC	Thomas Gingeras, CSHL	ENCSR000CQQ	2	paired	ENCFF000FGL ENCFF000FHE
				paired	ENCFF000FIM ENCFF000FIL
H1–hESC	Thomas Gingeras, CSHL	ENCSR000COW	2	paired	ENCFF000FJF ENCFF000FJG
H1–hESC	Thomas Gingeras, CSHL	ENCSR000COV	2	paired	ENCFF000FHG ENCFF000FHH
H1–hESC	Thomas Gingeras, CSHL	ENCSR000COU	1	paired	ENCFF000FEU ENCFF000FFF
			2	paired	ENCFF000FET ENCFF000FFG
H1–hESC	Thomas Gingeras, CSHL	ENCSR000COT	1	paired	ENCFF000FDA ENCFF000FDZ

Continued on next page...

Table A.1 – *Continued.*

Cell Type	Laboratory	Experiment	S	Run	Accession(s)
			2	paired	ENCFF000FDT ENCFF000FEE
mesenchymal stem cell	Zhiping Weng, UMass	ENCSR275ZLF	1	paired	ENCFF378MAM ENCFF663RCH
mesenchymal stem cell	Zhiping Weng, UMass	ENCSR586TYR	1	single	ENCFF276PUU
			2	single	ENCFF225QQQ
mesenchymal stem cell	Zhiping Weng, UMass	ENCSR663WGC	1	paired	ENCFF608QEG ENCFF094XOG
monocyte (CD14+)	Thomas Gingeras, CSHL	ENCSR000CTY	1	paired	ENCFF000HTK ENCFF000HTQ
				paired	ENCFF000HTL ENCFF000HTR
				paired	ENCFF000HTM ENCFF000HTS
			2	paired	ENCFF000HTN ENCFF000HTT
				paired	ENCFF000HTO ENCFF000HTU
				paired	ENCFF000HTP ENCFF000HTV
monocyte (CD14+)	Thomas Gingeras, CSHL	ENCSR000CUC	1	paired	ENCFF000HUU ENCFF000HVA
				paired	ENCFF000HUV ENCFF000HVB
				paired	ENCFF000HUW ENCFF000HVC
			2	paired	ENCFF000HUX ENCFF000HVD
				paired	ENCFF000HUY ENCFF000HVE
				paired	ENCFF000HUZ ENCFF000HVF
monocyte (CD14+)	Zhiping Weng, UMass	ENCSR905LVO	1	paired	ENCFF709TXQ ENCFF246VSY
neural progenitor cell	Thomas Gingeras, CSHL	ENCSR244ISQ	1	paired	ENCFF939FVE ENCFF201WLO
			2	paired	ENCFF726PAY ENCFF996KMK
neural progenitor cell	Thomas Gingeras, CSHL	ENCSR828LSC	1	single	ENCFF193XIP
			2	single	ENCFF994BMY
neural stem progenitor cell	Zhiping Weng, UMass	ENCSR572EET	1	paired	ENCFF834VLU ENCFF652ZQR
neural stem progenitor cell	Zhiping Weng, UMass	ENCSR977XUX	1	paired	ENCFF402BWO ENCFF450JES

Continued on next page...

Table A.1 – *Continued.*

Cell Type	Laboratory	Experiment	S	Run	Accession(s)
osteoblast	Thomas Gingeras, CSHL	ENCSR000CUF	1	paired	ENCFF000GKC ENCFF000GKF
			2	paired	ENCFF000GJC ENCFF000GJE
osteoblast	Thomas Gingeras, CSHL	ENCSR000CUW	1	single	ENCFF000JLG
			2	single	ENCFF000JKU

Table A.2: *Experiment and file accessions for all histone ChIP–seq data sets used in the thesis, grouped by cell type, histone modification or variant, laboratory, and experiment. Some experiments produced suitable ChIP–seq data sets for multiple biological samples (S). These samples were assigned a numerical ID or each experiment. All listed files were acquired as alignments to the GRCh38 reference genome in BAM–format from ENCODE. Location abbreviations: University of Massachusetts (UMass), University of Washington (UW), Eli and Edythe L. Broad Institute of MIT and Harvard (Broad).*

Cell Type	Histone	Laboratory	Experiment	S	Accession(s)
bipolar neuron	H2A.Z	Bradley Bernstein, Broad	ENCSR983CSB	1	ENCFF047IEH
				2	ENCFF017KFY
bipolar neuron	H3K4me2	Bradley Bernstein, Broad	ENCSR821IAK	1	ENCFF148CTR
				2	ENCFF856JAI
bipolar neuron	H3K4me3	Bradley Bernstein, Broad	ENCSR849YFO	1	ENCFF950QWN
				2	ENCFF096QTT
bipolar neuron	H3K27ac	Bradley Bernstein, Broad	ENCSR905TYC	1	ENCFF017WUP
				2	ENCFF751YAL
common myeloid progenitor (CD34+)	H3K4me3	Zhiping Weng, UMass	ENCSR045QJH	1	ENCFF053WVU
				2	ENCFF218BJW
common myeloid progenitor (CD34+)	H3K4me3	Zhiping Weng, UMass	ENCSR192GUR	1	ENCFF218HFN
common myeloid progenitor (CD34+)	H3K4me3	Zhiping Weng, UMass	ENCSR681HMF	1	ENCFF270WBW
				2	ENCFF098CRS
common myeloid progenitor (CD34+)	H3K27ac	Zhiping Weng, UMass	ENCSR620AZM	1	ENCFF150ZJR

Continued on next page...

Table A.2 – *Continued.*

Cell Type	Histone	Laboratory	Experiment	S	Accession(s)
GM23338	H2A.Z	Bradley Bernstein, Broad	ENCSR908FQA	1	ENCFF345LJW
				2	ENCFF499QNE
GM23338	H3K4me2	Bradley Bernstein, Broad	ENCSR740FYN	1	ENCFF913IFO
				2	ENCFF384SDB
GM23338	H3K4me3	Bradley Bernstein, Broad	ENCSR657DYL	1	ENCFF346EPG
				2	ENCFF850PUL
GM23338	H3K27ac	Bradley Bernstein, Broad	ENCSR729ENO	1	ENCFF403VXK
				2	ENCFF259MDS
H1–hESC	H2A.Z	Bradley Bernstein, Broad	ENCSR000APX	1	ENCFF673SFE
				2	ENCFF156BHJ
H1–hESC	H3K4me2	Bradley Bernstein, Broad	ENCSR000ANC	1	ENCFF694EXK
				2	ENCFF115ZGZ
H1–hESC	H3K4me3	Bradley Bernstein, Broad	ENCSR814XPE	1	ENCFF775QSF
				2	ENCFF262WSA
				3	ENCFF272YBU
H1–hESC	H3K27ac	Bradley Bernstein, Broad	ENCSR000ANP	1	ENCFF238SQN
				2	ENCFF242PAC
mesenchymal stem cell	H2A.Z	Zhiping Weng, UMass	ENCSR526NKM	1	ENCFF519YBX
monocyte (CD14+)	H3K4me3	Bradley Bernstein, Broad	ENCSR000ASN	1	ENCFF521ESS
				2	ENCFF393WAI
monocyte (CD14+)	H3K4me3	John Stamatoyannopoulos, UW	ENCSR000DWL	1	ENCFF764OBU
monocyte (CD14+)	H3K4me3	Zhiping Weng, UMass	ENCSR796FCS	1	ENCFF993LIK
monocyte (CD14+)	H3K27ac	Bradley Bernstein, Broad	ENCSR000ASJ	1	ENCFF835ZFV
				2	ENCFF329AXT
monocyte (CD14+)	H3K27ac	Zhiping Weng, UMass	ENCSR012PII	1	ENCFF680VNR
neural progenitor cell	H2A.Z	Bradley Bernstein, Broad	ENCSR677XPJ	1	ENCFF603XTR
				2	ENCFF263UKH

Continued on next page...

Table A.2 – *Continued.*

Cell Type	Histone	Laboratory	Experiment	S	Accession(s)
neural progenitor cell	H3K4me2	Bradley Bernstein, Broad	ENCSR645BCH	1	ENCFF976CGT
				2	ENCFF926HEW
neural progenitor cell	H3K4me3	Bradley Bernstein, Broad	ENCSR661MUS	1	ENCFF224OZC
				2	ENCFF509PFO
neural progenitor cell	H3K27ac	Bradley Bernstein, Broad	ENCSR449AXO	1	ENCFF802FBW
				2	ENCFF826AZF
neural stem progenitor cell	H2A.Z	Zhiping Weng, UMass	ENCSR021HLX	1	ENCFF196URM
neural stem progenitor cell	H3K4me3	Zhiping Weng, UMass	ENCSR354XWM	1	ENCFF396QPN
				2	ENCFF490QCW
neural stem progenitor cell	H3K4me3	Zhiping Weng, UMass	ENCSR416CJP	1	ENCFF934SXI
				2	ENCFF894XCJ
neural stem progenitor cell	H3K4me3	Zhiping Weng, UMass	ENCSR956CTX	1	ENCFF943BCG
				2	ENCFF102CQQ
neural stem progenitor cell	H3K27ac	Zhiping Weng, UMass	ENCSR331CCW	1	ENCFF402KBQ
neural stem progenitor cell	H3K27ac	Zhiping Weng, UMass	ENCSR799SRL	1	ENCFF063LDJ
				2	ENCFF402MPW
				3	ENCFF263UOB
neural stem progenitor cell	H3K27ac	Zhiping Weng, UMass	ENCSR852DMC	1	ENCFF172ICV
				2	ENCFF834FHV
osteoblast	H2A.Z	Bradley Bernstein, Broad	ENCSR000APG	1	ENCFF900VNN
				2	ENCFF264FWL
osteoblast	H3K27ac	Bradley Bernstein, Broad	ENCSR000APH	1	ENCFF835QFJ
				2	ENCFF155WZC

Appendix B

Extended Results

Table B.1: *Difference in genome–wide correlation of differential gene expression and differential histone modifications between all genes of a certain type (r_{all}) to only differentially expressed genes of that type (r_{diff}) in cell differentiation transitions. Histone modifications and variants are distinguished by their location in the promoter or gene body region. Instances where r_{all} and r_{diff} had a different sign are underlined. Cases for which r_{all} or r_{diff} were not available are labelled as "–". Cases where the correlation was lower in differentially expressed genes than all genes are highlighted in red, higher in blue, and darker highlighting signifies a more extreme difference. Histone modifications without ChIP–seq samples are left blank.*

		$r_{\text{diff}} - r_{\text{all}}$							
		H2A.Z		**H3K4me2**		**H3K4me3**		**H3K27ac**	
Transitions	**Genes**	**prom.**	**gene**	**prom.**	**gene**	**prom.**	**gene**	**prom.**	**gene**
H1–hESC	all	0.053	0.043	0.036	0.146	0.032	0.049	0.169	0.277
↓	TFs	0.087	0.031	0.037	0.151	0.022	0.069	0.135	0.277
GM23338	miRNAs	–	–	–	–	–	–	0.178	0.192
	other	0.051	0.043	0.035	0.146	0.031	0.048	0.172	0.273
mesenchymal	all	0.113	0.185						
stem cell	TFs	0.217	0.498						
↓	miRNAs	–	–						
ostebolast	other	0.097	0.168						
myeloid	all					0.085	0.139	0.170	0.226
progenitor	TFs					0.063	0.159	0.177	0.215
↓	miRNAs					–	-0.055	-0.003	-0.018
monocyte	other					0.085	0.138	0.170	0.226
neural	all	0.099	-0.171	0.252	0.356	0.130	0.286	0.439	0.411
progenitor	TF	0.062	-0.160	0.212	0.283	0.095	0.321	0.382	0.372
↓	miRNAs	–	–	–	–	–	–	–	–
bipolar neuron	other	0.105	-0.172	0.253	0.364	0.131	0.280	0.442	0.412
neural stem	all	0.063	0.010			0.022	0.055	0.175	0.205
progenitor	TFs	0.020	-0.055			0.053	0.204	0.141	0.242
↓	miRNAs	–	–	–	–	–	–	–	–
bipolar neuron	other	0.069	0.023			0.020	0.046	0.173	0.200
neural stem	all	0.058	0.008			0.159	0.149	0.061	0.158
progenitor	TF	0.381	-0.047			0.498	0.482	0.220	0.465
↓	miRNAs	–	–			–	–	–	–
neural progenitor	other	0.047	0.004			0.104	0.104	0.049	0.132

Table B.2: *Difference in genome–wide correlation of differential gene expression and differential histone modifications between promoter (r_{prom}) to gene body (r_{gene}) in cell differentiation transitions, for all genes of a type and differentially expressed genes only. A statistically significant difference in correlation with $p < 0.05$ is marked in italic, and in bold if it was still significant after correcting for multiple hypothesis testing. Instances where r_{prom} and r_{gene} had a different sign are underlined. Cases for which r_{prom} or r_{gene} were not available are labelled as "–". Cases where the correlation was lower in the gene body than the promoter are highlighted in red, higher in blue, and darker highlighting signifies a more extreme difference. Histone modifications without ChIP–seq samples are left blank.*

		$r_{\text{gene}} - r_{\text{prom}}$							
		H2A.Z		**H3K4me2**		**H3K4me3**		**H3K27ac**	
Transition	**Genes**	**all**	**diff.**	**all**	**diff.**	**all**	**diff.**	**all**	**diff.**
H1–hESC	all	**-0.015**	-0.026	**0.040**	**0.150**	-0.001	0.016	**0.047**	**0.148**
↓	TFs	**-0.055**	-0.111	0.003	0.117	-0.019	0.028	**0.069**	**0.211**
GM23338	miRNAs	-0.012	–	-0.010	–	0	–	0.046	0.060
	other	**-0.014**	-0.022	**0.040**	**0.151**	-0.001	0.016	**0.045**	**0.147**
mesenchymal	all	*-0.012*	0.060						
stem cell	TFs	**-0.091**	0.190						
↓	miRNAs	–	–						
ostebolast	other	-0.008	0.063						
myeloid	all					**0.072**	**0.126**	**0.058**	**0.114**
progenitor	TFs					**0.080**	**0.176**	0.044	0.083
↓	miRNAs					**0.076**	–	0.016	0
monocyte	other					**0.070**	**0.122**	**0.057**	**0.114**
neural	all	**-0.134**	**-0.404**	**0.042**	**0.147**	**0.151**	**0.307**	0.006	-0.021
progenitor	TF	**-0.177**	**-0.399**	0.005	0.076	**0.145**	**0.371**	-0.010	-0.020
↓	miRNAs	-0.001	–	0.096	–	-0.002	–	0.067	–
bipolar	other	**-0.129**	**-0.406**	**0.045**	**0.155**	**0.148**	**0.297**	0.005	-0.024
neuron									
neural stem	all	**-0.047**	**-0.010**			**0.066**	**0.098**	**-0.015**	0.015
progenitor	TFs	**-0.076**	**-0.150**			**0.116**	**0.267**	0.020	**0.120**
↓	miRNAs	-0.086	–			**-0.230**	–	-0.098	–
bipolar	other	**-0.043**	**-0.090**			**0.064**	**0.089**	**-0.016**	0.011
neuron									
neural stem	all	**0.030**	-0.019			**-0.023**	-0.033	**-0.072**	0.024
progenitor	TF	-0.039	**-0.467**			**-0.063**	-0.079	-0.019	0.226
↓	miRNAs	-0.022	–			–	–	-0.033	–
neural	other	**0.033**	-0.010			**-0.021**	-0.021	**-0.071**	0.012
progenitor									

Table B.3: *Summary of the top 30 significantly enriched annotations for differentially expressed genes in cell differentiation transitions, with FDR controlled p–value < 0.05. Biological process and molecular function annotations are from the Gene Ontology (GO). The motif type abbreviations stand for TF–FFLs, miRNA–FFLs, composite–FFLs and simple co–regulation. (+) marks positive regulation, (-) negative regulation, and (*) general regulation. Cases without enriched annotations are labelled as "n.e.".*

Transition	**Biological Process**	**Molecular Function**	**KEGG**
H1–hESC ↓ GM23338	n.e.	n.e.	n.e.
mesenchymal stem cell ↓ ostebolast	embryonic organ development and morphogenesis, development (+), vasculogenesis (+), proliferation (*)	n.e.	ECM– and neuroactive ligand–receptor interaction
myeloid progenitor ↓ monocyte	n.e.	lipopetide–, Ig2– and extracellular matrix binding, cytokine receptor activity	n.e.
neural progenitor ↓ bipolar neuron	neurogenesis, neuron differentiation, nervous system development, axonogenesis	ion channel and transmembrane transport activity, actin–, cytoskeletal–, kinase–, extrallelular matrix–, growth factor–, DNA–binding	cancer, viral infections, cytoskeleton regulation, axon guidance, amyotrophic lateral sclerosis, various signaling pathways
neural stem progenitor ↓ bipolar neuron	nervous system development (*), neuron fate commitment, neurogenesis, axonogenesis, nucleosome assembly	ligand–gated ion channel activity, neurotransmitter receptor activity	alcoholism, systemic lupus erythematosus
neural stem progenitor ↓ neural progenitor	nucleosome assembly and organisation, chromatin silencing, chromatin remodelling	n.e.	alcoholism, systemic lupus erythematosus, viral carcinogenesis

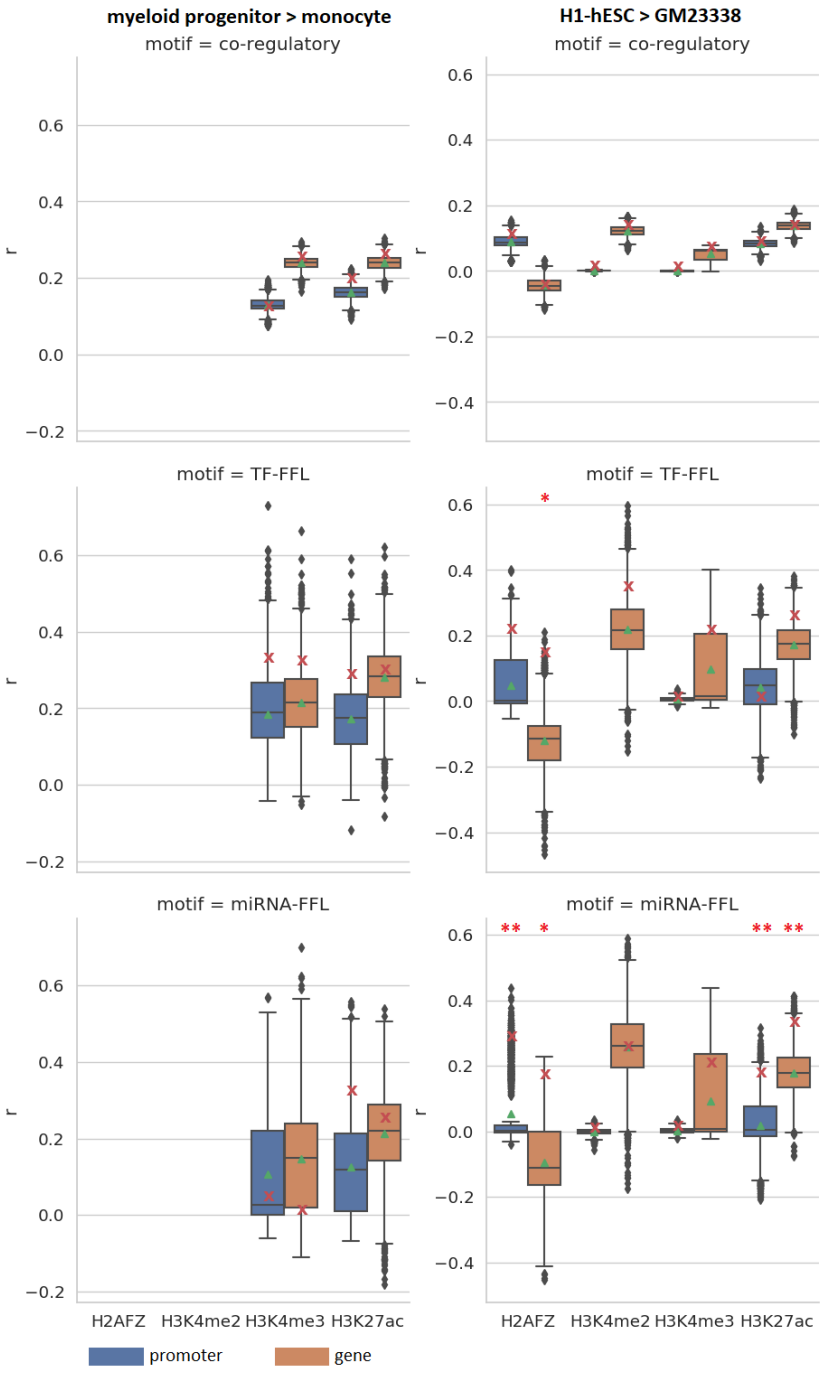

Figure B.1: *Comparison of differential gene expression and differential histone modification or variant correlation r in actual TF–miRNA co–regulatory motif types (red X) and motif types in 5,000 randomised regulatory networks (box plots) for two cell differentiation transitions, distinguished by promoter region and gene body. The box plots give the median (middle black bar), mean (green triangle), and outliers (black diamonds). Single stars mark cases with $p_{\text{sign}} < 0.05$ (stronger correlation with the same sign), and double stars cases with $p_{\text{abs}} < 0.05$ (stronger absolute correlation).*

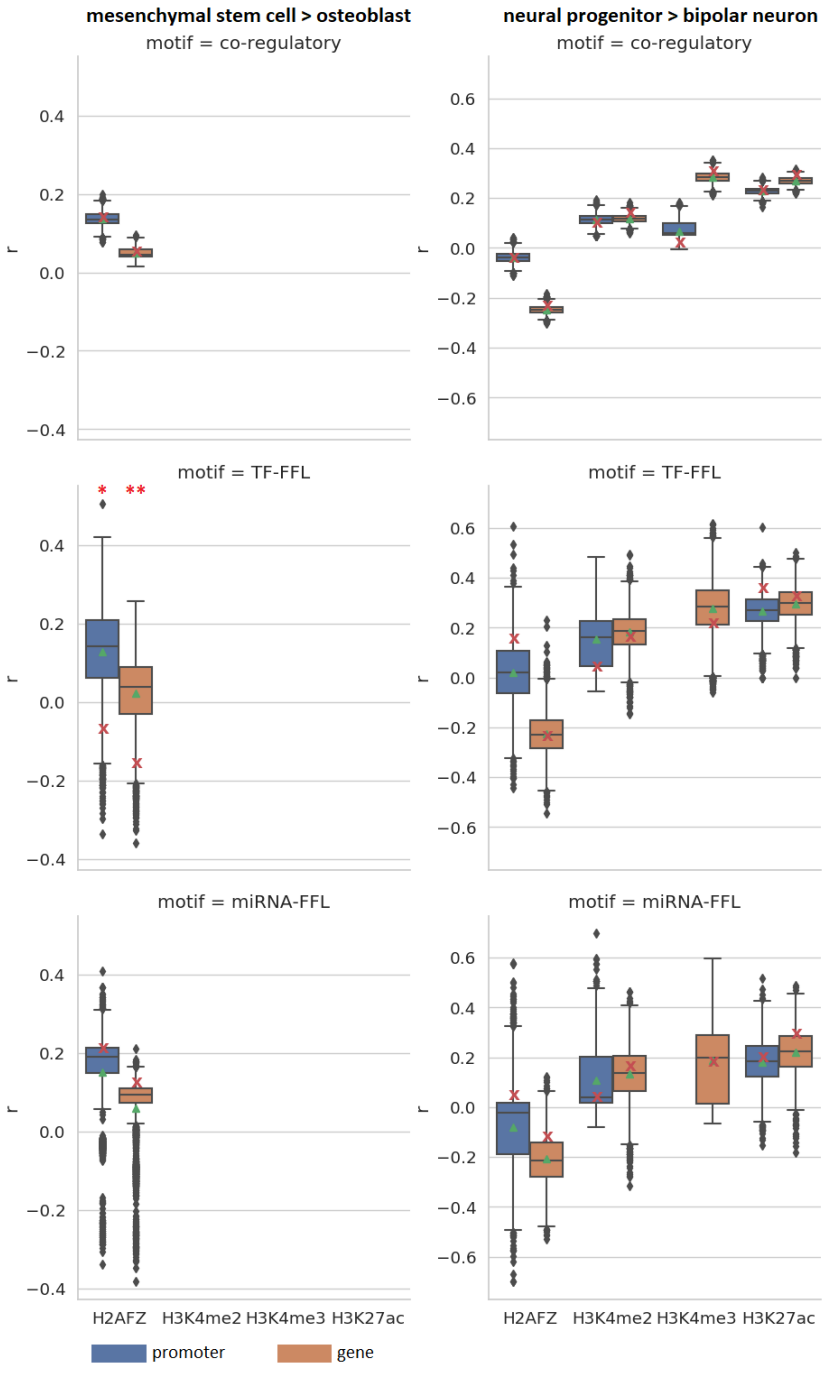

Figure B.2: *Comparison of differential gene expression and differential histone modification or variant correlation r in actual TF–miRNA co–regulatory motif types (red X) and motif types in 5,000 randomised regulatory networks (box plots) for two cell differentiation transitions, distinguished by promoter region and gene body. The box plots give the median (middle black bar), mean (green triangle), and outliers (black diamonds). Single stars mark cases with $p_{\text{sign}} < 0.05$ (stronger correlation with the same sign), and double stars cases with $p_{\text{abs}} < 0.05$ (stronger absolute correlation).*

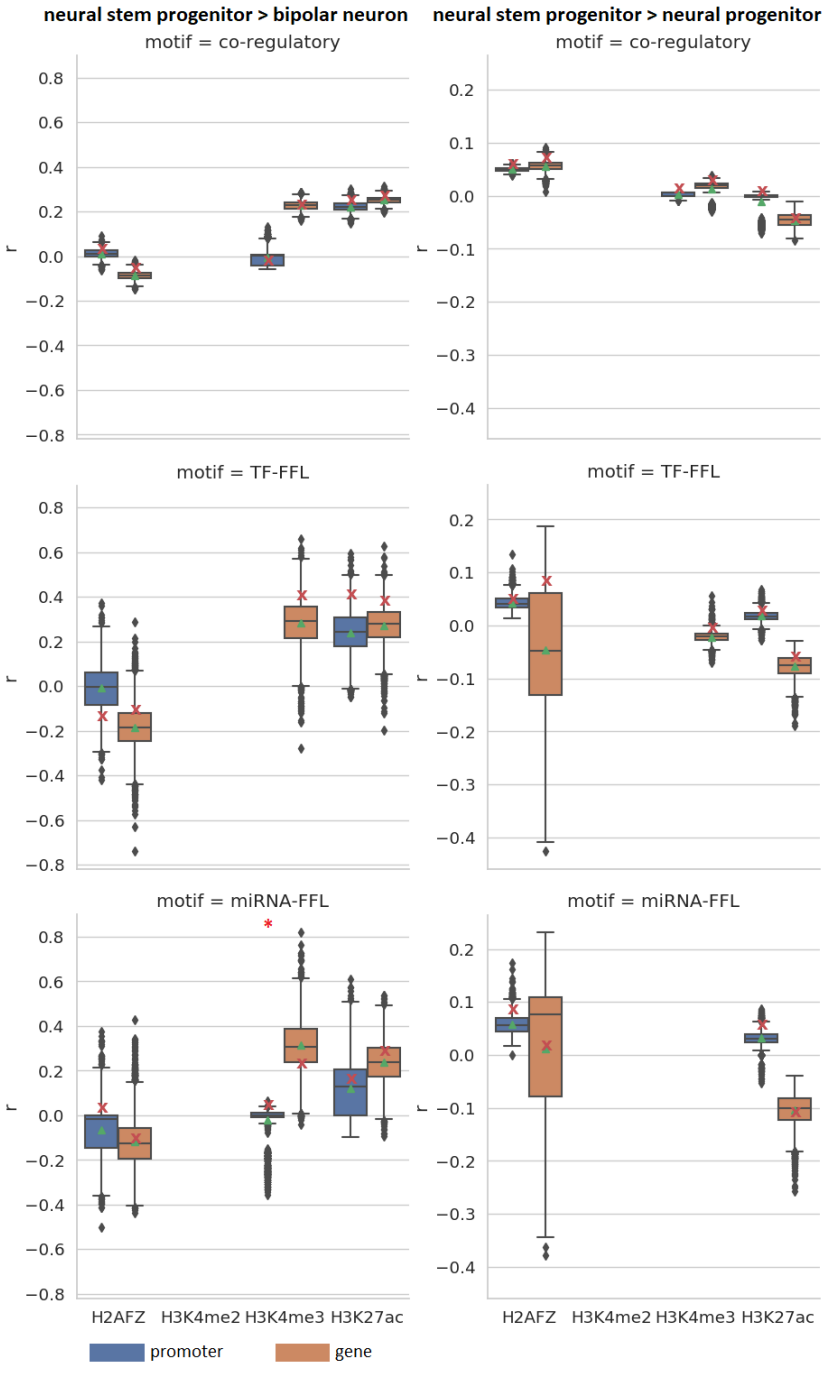

Figure B.3: *Comparison of differential gene expression and differential histone modification or variant correlation r in actual TF–miRNA co–regulatory motif types (red X) and motif types in 5,000 randomised regulatory networks (box plots) for two cell differentiation transitions, distinguished by promoter region and gene body. The box plots give the median (middle black bar), mean (green triangle), and outliers (black diamonds). Single stars mark cases with $p_{\text{sign}} < 0.05$ (stronger correlation with the same sign), and double stars cases with $p_{\text{abs}} < 0.05$ (stronger absolute correlation).*

Table B.4: *Differential gene expression and differential histone modification correlation difference between TF–miRNA co–regulatory motifs (r_{motif}), consisting of m genes, and genome–wide correlation (r_{genome}) of n genes for cell differentiation transitions. The examined motif types were TF–FFLs, miRNA–FFLs, composite–FFLs and simple co–regulatory motifs. For each transition, the difference is computed from the genome–wide (G) correlation for all genes (A), as well as only for differentially expressed genes (D) that were not part of the respective motif. Histone modifications and variants are distinguished by their location in the promoter or gene body region. A statistically significant difference in correlation with $p < 0.05$ is marked in italic, and in bold if it was still significant after correcting for multiple hypothesis testing. Instances where r_{motif} and r_{genome} had a different sign are underlined. Cases for which r_{motif} or r_{genome} were not available are labelled as "–". Cases where the correlation was lower in the motif than genome–wide are highlighted in red, higher in blue, and darker highlighting signifies a more extreme difference. Histone modifications without ChIP–seq samples are left blank.*

					$r_{\text{motif}} - r_{\text{genome}}$							
					H2A.Z		H3K4me2		H3K4me3		H3K27ac	
Transition	**Motif**	m	**G**	n	**prom.**	**gene**	**prom.**	**gene**	**prom.**	**gene**	**prom.**	**gene**
H1–hESC	co–reg.	1,036	A	45,907	0.061	**-0.082**	-0.030	0.059	-0.060	0.003	0.006	0.009
↓			D	3,710	0.009	**-0.129**	-0.066	*-0.084*	**-0.092**	-0.045	**-0.164**	**-0.261**
GM23338	TF	84	A	46,859	0.170	0.111	–	**0.266**	-0.059	0.147	-0.071	0.131
			D	3,819	0.117	0.069	–	0.121	-0.091	0.099	*-0.240*	-0.139
	miRNA	88	A	46,855	*0.239*	0.137	-0.033	0.177	0.239	-0.139	0.094	*0.203*
			D	3,818	0.187	0.095	-0.069	0.032	-0.087	0.091	-0.074	-0.066
	comp.	13	A	46,930	–	–	–	*0.608*	–	–	0.052	0.018
			D	3,829	–	–	–	0.462	–	–	-0.117	-0.252
mesenchymal	co–reg.	904	A	33,554	**0.095**	0.018						
stem cell			D	311	-0.013	**-0.163**						
↓	TF	64	A	34,394	-0.118	-0.193						
ostebolast			D	324	-0.234	**-0.382**						
	miRNA	64	A	34,394	0.166	0.090						
			D	325	0.054	-0.094						
	comp.	10	A	34,448	–	–						
			D	326	–	–						
myeloid	co–reg.	872	A	41,328					0.005	*0.065*	**0.111**	**0.116**
progenitor			D	4,218					*-0.081*	*-0.073*	-0.058	**-0.109**
↓	TFs	85	A	42,115					*0.212*	0.130	0.198	0.152
monocyte			D	4,330					**0.126**	-0.009	0.027	-0.074
	miRNA	68	A	42,132					-0.074	-0.182	*0.236*	0.107
			D	4,334					-0.159	**-0.321**	-0.066	-0.118
	comp.	10	A	42,190					–	–	–	–
			D	4,343					–	–	–	–
neural	co–reg.	839	A	36,988	**-0.097**	**-0.159**	-0.006	-0.008	**-0.113**	0.025	0.057	**0.115**
progenitor			D	1,902	**-0.205**	0.005	**-0.260**	**-0.373**	**-0.244**	**-0.259**	**-0.382**	**-0.297**
↓	TF	90	A	37,737	0.105	-0.153	-0.063	0.015	–	-0.066	0.183	0.142
bipolar			D	1,992	0.006	0.018	**-0.315**	**-0.342**	–	**-0.351**	**-0.256**	**-0.269**
neuron	miRNA	72	A	37,755	-0.005	-0.038	-0.064	0.018	–	-0.100	0.023	0.112
			D	2,001	-0.104	0.133	**-0.316**	**-0.338**	–	**-0.385**	**-0.416**	**-0.299**
	comp.	10	A	37,817	–	0.268	–	–	–	–	0.009	0.054
			D	2,007	–	0.439	–	–	–	–	-0.430	-0.357
neural stem	co–reg.	780	A	38,200	-0.031	*-0.072*			**-0.121**	*0.070*	0.060	**0.094**
progenitor			D	5,052	**-0.097**	*-0.087*			**-0.142**	0.019	**-0.114**	**-0.107**
↓	TF	53	A	38,927	-0.200	-0.124			–	0.244	0.218	0.205
bipolar			D	5,208	-0.264	-0.136			–	0.189	0.044	0.001
neuron	miRNA	60	A	38,920	-0.028	-0.121			-0.052	0.067	-0.031	0.112
			D	5,205	-0.091	-0.132			-0.074	0.013	-0.205	-0.092
	comp.	10	A	38,970	–	-0.554			–	–	0.495	0.403
			D	5,216	–	-0.564			–	–	0.321	0.198
neural stem	co–reg.	784	A	33,845	0.029	0.011			-0.030	0.008	-0.065	-0.044
progenitor			D	1,354	-0.031	0			**-0.191**	**-0.145**	**-0.128**	**-0.189**
↓	TF	74	A	34,555	0.019	0.023			–	-0.026	-0.045	-0.060
neural			D	1,375	-0.039	0.014			–	-0.175	-0.106	-0.218
progenitor	miRNA	54	A	34,575	0.056	-0.043			–	–	-0.017	-0.110
			D	1,373	-0.002	-0.052			–	–	-0.078	-0.268
	comp.	10	A	34,619	–	–			–	–	–	–
			D	1,376	–	–			–	–	–	–

Table B.5: *Differential gene expression and differential histone modification correlation difference between promoter (r_{prom}) and gene body (r_{gene}) in m TF–miRNA co–regulatory motifs, consisting of n genes, in cell differentiation transitions. The examined motif types were TF–FFLs, miRNA–FFLs, composite–FFLs and simple co–regulatory motifs. A statistically significant difference in correlation with $p < 0.05$ is marked in italic, and in bold if it was still significant after correcting for multiple hypothesis testing. Instances where r_{prom} and r_{gene} had a different sign are underlined. Cases for which r_{prom} or r_{gene} were not available are labelled as "–". Cases where the correlation was lower in the gene body than the promoter are highlighted in red, higher in blue, and darker highlighting signifies a more extreme difference. Histone modifications without ChIP–seq samples are left blank.*

				$r_{\text{gene}} - r_{\text{prom}}$			
Transition	**Motif**	m	n	**H2A.Z**	**H3K4me2**	**H3K4me3.**	**H3K27ac**
H1–hESC	co–reg.	4,771	1,036	**-0.155**	**0.126**	0.061	0.050
↓	TF	17	84	-0.073	–	**0.204**	0.248
GM23338	miRNA	41	88	-0.117	0.250	**0.193**	0.155
	comp.	2	13	–	–	–	0.012
mesenchymal	co–reg.	3,061	904	**-0.087**			
stem cell	TF	14	64	-0.087			
↓	miRNA	34	64	-0.087			
ostebolast	comp.	1	10	–			
myeloid	co–reg.	3,144	872			**0.130**	0.062
progenitor	TF	18	85			-0.009	0.012
↓	miRNA	29	68			-0.036	-0.071
monocyte	comp.	1	10			–	–
neural	co–reg	3,032	839	**-0.192**	0.040	**0.284**	0.062
progenitor	TF	17	90	**-0.390**	0.120	–	-0.034
↓	miRNA	31	72	-0.166	0.124	–	0.095
bipolar neuron	comp.	1	10	–	–	–	0.052
neural stem	co–reg.	2,185	780	*-0.086*		**0.252**	0.018
progenitor	TF	9	53	0.029		–	-0.029
↓	miRNA	26	60	-0.140	–	0.185	0.127
bipolar neuron	comp.	1	10	–		–	-0.108
neural stem	co–reg.	2,601	784	0.013		0.015	-0.051
progenitor	TF	16	74	0.034		–	-0.087
↓	miRNA	25	54	-0.068		–	-0.165
neural progenitor	comp.	1	10	–		–	–

Table B.6: *Summary of the top 30 significantly enriched annotations for TF–miRNA co–regulatory motif types in cell differentiation transitions, with FDR controlled p–value < 0.05. Biological process and molecular function annotations are from the Gene Ontology (GO). The motif type abbreviations stand for TF–FFLs, miRNA–FFLs, composite–FFLs and simple co–regulation. (+) marks positive regulation, (-) negative regulation, and (*) general regulation.*

Transition	Motif	Biological Process	Molecular Function	KEGG
H1–hESC ↓ GM23338	co–reg.	transcription (*)	HDAC–, enzyme–, TF–, DNA–binding	cell cycle, cancer, pluripotency (*)
	TF	hypoxia response, differentiation (*), structure morphogenesis, development (*), apoptosis (*), proliferation, transcription (*)	HAT–, HDAC–, enzyme–, kinase–, ubiquitin–ligase–, TF–, DNA–, phosphatase–binding	cancer, viral infections, senescence, cell cycle
	miRNA	development (-), G1/S–phase transition, apoptosis (-), proliferation (*), differentiation (-), transcription (*)	HDAC–, enzyme–, ubiquitin–ligase–, TF–, DNA–binding	cell cycle, senescence, viral infections, cancer
	comp.	apoptosis (*), hypoxia response, mitophagy (*), development (+), transcription (*)	HAT–, HDAC–, enzyme–, ubiquitin–ligase–, TF–, DNA–binding	cell cycle, cancer, senescence, mitophagy, viral infections, endocytosis
mesenchymal stem cell ↓ ostebolast	co–reg	transcription (*)	HDAC–, enzyme–, TF–, DNA–binding	cancer, viral infections, osteoclast differentiation, cell cycle, pluripotency
	TF	hypoxia response, structure morphogenesis, proliferation (*), development (+), differentiation, transcription (*)	HAT–, HDAC–, enzyme–, kinase–, TF–, DNA–binding, DNA–methylation	cell cycle, senescence, cancer, viral infections
	miRNA	organ morphogenesis, apoptosis (-), proliferation (*), transcription (*)	TF–, DNA–binding	cell cycle, senescence, cancer, viral infections

Continued on next page...

Table B.6 – *Continued.*

Transition	Motif	Biological Process	Molecular Function	KEGG
	comp.	ossification, neural crest migration, hypoxia response, apoptosis (*), mitophagy (+), transcription (*)	DNA–binding	cancer, mitophagy, senescence, endocytosis
myeloid progenitor ↓ monocyte	co–reg.	transcription (*)	HDAC–, enzyme–, TF–, DNA–binding	cancer, viral infections, pluripotency, immune differentiation, senescence
	TF	structure morphogenesis, development (+), differentiation (+), proliferation (*), apoptosis (-), cytokine response, transcription (*)	HAT–, enzyme–, kinase–, TF–, DNA–, phosphatase–binding	cancer, viral infections, senescence
	miRNA	haematopoiesis, apoptosis (-), differentiation (+), development (-), proliferation (*), transcription (*)	ubiquitin–ligase–, TF–, cyclin–, DNA– binding	apoptosis, immune differentiation, cancer, viral infections, senescence, cell cycle
	comp.	ossification, neural crest migration, hypoxia response, apoptosis (*), mitophagy (+), transcription (*)	DNA–binding	cancer, mitophagy, senescence, endocytosis
neural progenitor ↓ bipolar neuron	co–reg.	transcription (*)	HDAC–, enzyme–, TF–, DNA–binding	pluripotency, cell cycle, senescence, cancer, viral infections
	TF	differentiation (*), development (+), proliferation (-), apoptosis (*), transcription (*)	HAT–, HDAC–, enzyme–, kinase–, ubiquitin–ligase–, TF–, DNA–, phosphatase–binding	cancer, viral infections, senescence
	miRNA	differentiation (-), development (-), proliferation (*), transcription (*)	enzyme–, kinase–, ubiquitin–ligase–, TF–, DNA–, phosphatase–binding	cancer, viral infection, senescence

Continued on next page...

Table B.6 – *Continued.*

Transition	Motif	Biological Process	Molecular Function	KEGG
	comp.	ossification, neural crest migration, hypoxia response, apoptosis (*), mitophagy (+), transcription (*)	DNA–binding	cancer, mitophagy, senescence, endocytosis
neural stem progenitor ↓ bipolar neuron	co–reg.	transcription (*)	HDAC–, enzyme–, TF–, DNA–binding	cell cycle, cancer, senescence, viral infections
	TF	hypoxia response, structure morphogenesis, differentiation (+), development (+), proliferation (-), transcription (*)	HAT–, HDAC–, enzyme–, kinase–, ubiquitin–ligase–, TF–, DNA–binding, DNA–methylation	cell cycle, senescence, cancer, viral infections
	miRNA	G1/S–phase transition, development (*), proliferation (*), neuron generation, transcription (*)	HDAC–, kinase–, ubiquitin–ligase–, cyclin–, TF–, DNA– binding	cancer, viral infections, cell cycle, senescence
	comp.	ossification, neural crest migration, hypoxia response, apoptosis (*), mitophagy (+), transcription (*)	DNA–binding	cancer, mitophagy, senescence, endocytosis
neural stem progenitor ↓ neural progenitor	co–reg.	transcription (*)	HDAC–, enzyme–, TF–, DNA–binding	cancer, viral infections, cell cycle, senescence
	TF	hypoxia response, proliferation (-), apoptosis (-), development (+), differentiation (*), transcription (*)	HAT–, HDAC–, enzyme–, kinase–, TF–, DNA–binding, DNA–methylation	senescence, cancer, viral infections
	miRNA	hypoxia response, apoptosis (-), differentiation (-), proliferation (+), development (-), transcription (*)	enzyme–, kinase–, cyclin–, ubiquitin–ligase–, TF–, DNA–binding	senescence, cancer, viral infections
	comp.	ossification, neural crest migration, hypoxia response, apoptosis (*), mitophagy (+), transcription (*)	DNA–binding	cancer, mitophagy, senescence, endocytosis

Bibliography

[1] E. Bianconi et al. 'An estimation of the number of cells in the human body'. In: *Ann. Hum. Biol.* 40.6 (2013), pp. 463–471.

[2] David E. Sadava et al. *Life: The Science of Biology.* 8th ed. Sinauer Associates Inc., 23 Plumtree Road, Sunderland, MA 01375 U.S.A.: Sinauer Associates, Inc., 2008. ISBN: 978-0716776710.

[3] Samuel A. Lambert et al. 'The Human Transcription Factors'. In: *Cell* 172.4 (2018), pp. 650–665. ISSN: 0092-8674. DOI: 10.1016/j.cell.2018.01.029.

[4] Eric L. Greer and Yang Shi. 'Histone methylation: a dynamic mark in health, disease and inheritance'. In: *Nature Reviews Genetics* 13 (Apr. 2012), p. 343.

[5] David P. Gavin and Schahram Akbarian. 'Epigenetic and post-transcriptional dysregulation of gene expression in schizophrenia and related disease'. In: *Neurobiology of Disease* 46.2 (2012), pp. 255–262. ISSN: 0969-9961. DOI: 10.1016/j.nbd.2011.12.008.

[6] Loredana Verdone et al. 'Histone acetylation in gene regulation.' In: *Briefings in functional genomics & proteomics* 5.3 (Sept. 2006), pp. 209–21. ISSN: 1473-9550. DOI: 10.1093/bfgp/ell028.

[7] Hong-Mei Zhang et al. 'Transcription factor and microRNA co-regulatory loops: important regulatory motifs in biological processes and diseases'. In: *Briefings in Bioinformatics* 16.1 (Dec. 2013), pp. 45–58. ISSN: 1477-4054. DOI: 10.1093/bib/bbt085.

[8] Arjen van den Berg, Johann Mols and Jiahuai Han. 'RISC-target interaction: cleavage and translational suppression'. In: *Biochimica et biophysica acta* 1779.11 (Nov. 2008), pp. 668–677. ISSN: 0006-3002. DOI: 10.1016/j.bbagrm.2008.07.005.

[9] Ana Eulalio et al. 'Deadenylation is a widespread effect of miRNA regulation'. In: *RNA (New York, N.Y.)* 15.1 (Jan. 2009), pp. 21–32. ISSN: 1469-9001. DOI: 10.1261/rna.1399509.

[10] Pierre-Etienne Cholley et al. 'Modeling gene-regulatory networks to describe cell fate transitions and predict master regulators'. In: *npj Systems Biology and Applications* 4.1 (2018), p. 29. ISSN: 2056-7189. DOI: 10.1038/s41540-018-0066-z.

[11] Marie Classon and Ed Harlow. 'The retinoblastoma tumour suppressor in development and cancer'. In: *Nature Reviews Cancer* 2.12 (2002), pp. 910–917. ISSN: 1474-1768. DOI: 10.1038/nrc950.

[12] Q. Cui et al. 'Principles of microRNA regulation of a human cellular signaling network'. In: *Mol. Syst. Biol.* 2 (2006), p. 46.

[13] Robin C. Friedman et al. 'Most mammalian mRNAs are conserved targets of microRNAs'. In: *Genome research* 19.1 (Jan. 2009), pp. 92–105. ISSN: 1088-9051. DOI: 10.1101/gr.082701.108.

[14] Aurora Esquela-Kerscher and Frank J. Slack. 'Oncomirs – microRNAs with a role in cancer'. In: *Nature Reviews Cancer* 6.4 (2006), pp. 259–269. ISSN: 1474-1768. DOI: 10.1038/nrc1840.

[15] M. Hamed et al. 'TFmiR: a web server for constructing and analyzing disease-specific transcription factor and miRNA co-regulatory networks'. In: *Nucleic Acids Res.* 43.W1 (July 2015), W283–288.

[16] Matteo Osella et al. 'The Role of Incoherent MicroRNA-Mediated Feedforward Loops in Noise Buffering'. In: *PLOS Computational Biology* 7.3 (Mar. 2011), pp. 1–16. DOI: 10.1371/journal.pcbi.1001101.

[17] John Tsang, Jun Zhu and Alexander van Oudenaarden. 'MicroRNA-Mediated Feedback and Feedforward Loops Are Recurrent Network Motifs in Mammals'. In: *Molecular Cell* 26.5 (2007), pp. 753–767. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2007.05.018.

[18] Shai S. Shen-Orr et al. 'Network motifs in the transcriptional regulation network of Escherichia coli'. In: *Nature Genetics* 31.1 (2002), pp. 64–68. ISSN: 1546-1718. DOI: 10.1038/ng881.

[19] Ying Lin et al. 'Transcription factor and miRNA co-regulatory network reveals shared and specific regulators in the development of B cell and T cell'. In: *Scientific Reports* 5 (Oct. 2015), p. 15215.

[20] Ruijiang Li et al. 'CMTCN: a web tool for investigating cancer-specific microRNA and transcription factor co-regulatory networks'. In: *PeerJ* 6 (Nov. 2018), e5951–e5951. ISSN: 2167-8359. DOI: 10.7717/peerj.5951.

[21] Karolin Luger et al. 'Crystal structure of the nucleosome core particle at 2.8 Å resolution'. In: *Nature* 389.6648 (1997), pp. 251–260. ISSN: 1476-4687. DOI: 10.1038/38444.

[22] Bing Li, Michael Carey and Jerry L. Workman. 'The Role of Chromatin during Transcription'. In: *Cell* 128.4 (2007), pp. 707–719. ISSN: 0092-8674. DOI: 10.1016/j.cell.2007.01.015.

[23] Thomas Jenuwein and C. David Allis. 'Translating the Histone Code'. In: *Science* 293.5532 (2001), pp. 1074–1080. ISSN: 0036-8075. DOI: 10.1126/science.1063127.

[24] Kavitha Sarma and Danny Reinberg. 'Histone variants meet their match'. In: *Nature Reviews Molecular Cell Biology* 6.2 (2005), pp. 139–149. ISSN: 1471-0080. DOI: 10.1038/nrm1567.

[25] Paul B. Talbert and Steven Henikoff. 'Histone variants – ancient wrap artists of the epigenome'. In: *Nature Reviews Molecular Cell Biology* 11 (Mar. 2010), p. 264.

[26] Rohinton T. Kamakaka and Sue Biggins. 'Histone variants: deviants?' In: *Genes & Development* 19.3 (2005), pp. 295–316. DOI: 10.1101/gad.1272805.

[27] Mona D. Shahbazian and Michael Grunstein. 'Functions of Site-Specific Histone Acetylation and Deacetylation'. In: *Annual Review of Biochemistry* 76.1 (2007), pp. 75–100. DOI: 10.1146/annurev.biochem.76.052705.162114.

[28] Tony Kouzarides. 'Chromatin Modifications and Their Function'. In: *Cell* 128.4 (2007), pp. 693–705. ISSN: 0092-8674. DOI: 10.1016/j.cell.2007.02.005.

[29] Nick J. Proudfoot. 'Ending the message: poly(A) signals then and now'. In: *Genes & development* 25.17 (Sept. 2011), pp. 1770–1782. ISSN: 1549-5477. DOI: 10.1101/gad.17268411.

[30] D. E. Winickoff, K. Saha and G. D. Graff. 'Opening stem cell research and development: a policy proposal for the management of data, intellectual property, and ethics'. In: *Yale J Health Policy Law Ethics* 9.1 (2009), pp. 52–127.

[31] Deanna M. Church et al. 'Modernizing reference genome assemblies'. In: *PLoS biology* 9.7 (July 2011), e1001091–e1001091. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001091.

[32] V. A. Schneider et al. 'Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly'. In: *Genome Res.* 27.5 (May 2017), pp. 849–864.

[33] WhatIsEpigenetics.com. *Epigenetics: Fundamentals*. URL: https://www.whatisepigenetics.com/fundamentals/ (visited on 1st Aug. 2019).

[34] Shelley L Berger. 'Histone modifications in transcriptional regulation'. In: *Current Opinion in Genetics & Development* 12.2 (2002), pp. 142–148. ISSN: 0959-437X. DOI: 10.1016/S0959-437X(02)00279-4.

[35] The ENCODE Project Consortium et al. 'An integrated encyclopedia of DNA elements in the human genome'. In: *Nature* 489 (Sept. 2012), p. 57. DOI: 10.1038/nature11247.

[36] Zhibin Wang et al. 'Combinatorial patterns of histone acetylations and methylations in the human genome'. In: *Nature Genetics* 40 (June 2008), p. 897.

[37] Peter A. Jones. 'Functions of DNA methylation: islands, start sites, gene bodies and beyond'. In: *Nature Reviews Genetics* 13 (May 2012), p. 484.

[38] Vicky W. Zhou, Alon Goren and Bradley E. Bernstein. 'Charting histone modifications and the functional organization of mammalian genomes'. In: *Nature Reviews Genetics* 12 (Nov. 2010), p. 7.

[39] Hideaki Tagami et al. 'Histone H3.1 and H3.3 Complexes Mediate Nucleosome Assembly Pathways Dependent or Independent of DNA Synthesis'. In: *Cell* 116.1 (2004), pp. 51–61. ISSN: 0092-8674. DOI: 10.1016/S0092-8674(03)01064-X.

[40] Daniel Zilberman et al. 'Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks'. In: *Nature* 456 (Sept. 2008), p. 125.

[41] Chunyuan Jin and Gary Felsenfeld. 'Nucleosome stability mediated by histone variants H3.3 and H2A.Z'. In: *Genes & Development* 21.12 (2007), pp. 1519–1529. DOI: 10.1101/gad.1547707.

[42] Sara Hardy et al. 'The Euchromatic and Heterochromatic Landscapes Are Shaped by Antagonizing Effects of Transcription on H2A.Z Deposition'. In: *PLOS Genetics* 5.10 (Oct. 2009), pp. 1–12. DOI: 10.1371/journal.pgen.1000687.

[43] Menno P. Creyghton et al. 'H2AZ Is Enriched at Polycomb Complex Target Genes in ES Cells and Is Necessary for Lineage Commitment'. In: *Cell* 135.4 (2008), pp. 649–661. ISSN: 0092-8674. DOI: 10.1016/j.cell.2008.09.056.

[44] Brian D. Strahl and C. David Allis. 'The language of covalent histone modifications'. In: *Nature* 403.6765 (2000), pp. 41–45. ISSN: 1476-4687.

[45] Thomas Jenuwein and C. David Allis. 'Translating the Histone Code'. In: *Science* 293.5532 (2001), pp. 1074–1080. ISSN: 0036-8075. DOI: 10.1126/science.1063127.

[46] Arshdeep Sekhon, Ritambhara Singh and Yanjun Qi. 'DeepDiff: DEEP-learning for predicting DIFFerential gene expression from histone modifications'. In: *Bioinformatics* 34.17 (Sept. 2018), pp. i891–i900. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty612.

[47] Steven M. Reppert and David R. Weaver. 'Coordination of circadian timing in mammals'. In: *Nature* 418.6901 (2002), pp. 935–941. ISSN: 1476-4687. DOI: 10.1038/nature00965.

[48] Bing Ren et al. 'E2F integrates cell cycle progression with DNA repair, replication, and G2/M checkpoints'. In: *Genes & Development* 16.2 (2002), pp. 245–256. DOI: 10.1101/gad.949802.

[49] Ludwig Christian G. Hinske et al. 'A potential role for intragenic miRNAs on their hosts' interactome'. In: *BMC genomics* 11 (Oct. 2010), pp. 533–533. ISSN: 1471-2164. DOI: 10.1186/1471-2164-11-533.

[50] Julia Winter et al. 'Many roads to maturity: microRNA biogenesis pathways and their regulation'. In: *Nature Cell Biology* 11 (Mar. 2009), p. 228.

[51] Eugene Berezikov et al. 'Mammalian Mirtron Genes'. In: *Molecular Cell* 28.2 (2007), pp. 328–336. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2007.09.028.

[52] Zhou Huang et al. 'HMDD v3.0: a database for experimentally supported human microRNA–disease associations'. In: *Nucleic Acids Research* 47.D1 (Oct. 2018), pp. 1013–1017. ISSN: 0305-1048. DOI: 10.1093/nar/gky1010.

[53] Daniel R Zerbino et al. 'Ensembl 2018'. In: *Nucleic Acids Research* 46.D1 (Nov. 2017), pp. 754–761. ISSN: 0305-1048. DOI: 10.1093/nar/gkx1098.

[54] Marina Lizio et al. 'Gateways to the FANTOM5 promoter level mammalian expression atlas'. In: *Genome Biology* 16.1 (2015), p. 22. ISSN: 1465-6906. DOI: 10.1186/s13059-014-0560-6.

[55] Marina Lizio et al. 'Update of the FANTOM web resource: expansion to provide additional transcriptome atlases'. In: *Nucleic Acids Research* 47.D1 (Nov. 2018), pp. 752–758. ISSN: 0305-1048. DOI: 10.1093/nar/gky1099.

[56] Ludwig Christian Hinske et al. 'miRIAD-integrating microRNA inter- and intragenic data'. In: *Database : the journal of biological databases and curation* 2014 (Oct. 2014), bau099. ISSN: 1758-0463. DOI: 10.1093/database/bau099.

[57] Ludwig C. Hinske et al. 'MiRIAD update: using alternative polyadenylation, protein interaction network analysis and additional species to enhance exploration of the role of intragenic miRNAs and their host genes'. In: *Database* 2017 (Aug. 2017). ISSN: 1758-0463. DOI: 10.1093/database/bax053.

[58] Carrie A Davis et al. 'The Encyclopedia of DNA elements (ENCODE): data portal update'. In: *Nucleic Acids Research* 46.D1 (Nov. 2017), pp. 794–801. ISSN: 0305-1048. DOI: 10.1093/nar/gkx1081.

[59] NCBI Resource Coordinators. *Human Genome Assembly GRCh38.p12 Chromosome Lengths*. URL: https://www.ncbi.nlm.nih.gov/grc/human/data?asm=GRCh38.p12 (visited on 9th Mar. 2019).

[60] V. Matys et al. 'TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes'. In: *Nucleic Acids Res.* 34.Database issue (Jan. 2006), pp. D108–110.

[61] Robert Lesurf et al. 'ORegAnno 3.0: a community-driven resource for curated regulatory annotation'. In: *Nucleic Acids Research* 44.D1 (Nov. 2015), pp. 126–132. ISSN: 0305-1048. DOI: 10.1093/nar/gkv1203.

[62] H. Han et al. 'TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions'. In: *Nucleic Acids Res.* 46.D1 (Jan. 2018), pp. D380–D386.

[63] C. Jiang et al. 'TRED: a transcriptional regulatory element database, new entries and other development'. In: *Nucleic acids research* 35.Database issue (Jan. 2007), pp. D137–D140. ISSN: 1362-4962. DOI: 10.1093/nar/gkl1041.

[64] Z. Tong et al. 'TransmiR v2.0: an updated transcription factor-microRNA regulation database'. In: *Nucleic Acids Res.* 47.D1 (Jan. 2019), pp. D253–D258.

[65] Chengxiang Qiu et al. 'microRNA evolution in a human transcription factor and microRNA regulatory network'. In: *BMC systems biology* 4 (June 2010), pp. 90–90. ISSN: 1752-0509. DOI: 10.1186/1752-0509-4-90.

[66] J. H. Yang et al. 'ChIPBase: a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data'. In: *Nucleic Acids Res.* 41.Database issue (Jan. 2013), pp. D177–187.

[67] Chih-Hung Chou et al. 'miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions'. In: *Nucleic acids research* 46.D1 (Jan. 2018), pp. D296–D302. ISSN: 1362-4962. DOI: 10.1093/nar/gkx1067.

[68] Ioannis S. Vlachos et al. 'DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions'. In: *Nucleic Acids Research* 43.D1 (Nov. 2014), pp. 153–159. ISSN: 0305-1048. DOI: 10.1093/nar/gku1215.

[69] Feifei Xiao et al. 'miRecords: an integrated resource for microRNA–target interactions'. In: *Nucleic Acids Research* 37.suppl_1 (Nov. 2008), pp. 105–110. ISSN: 0305-1048. DOI: 10.1093/nar/gkn851.

[70] Jun-Hao Li et al. 'starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data'. In: *Nucleic Acids Research* 42.D1 (Nov. 2013), pp. 92–97. ISSN: 0305-1048. DOI: 10.1093/nar/gkt1248.

[71] Debarka Sengupta and Sanghamitra Bandyopadhyay. 'Participation of microRNAs in human interactome: extraction of microRNA–microRNA regulations'. In: *Mol. BioSyst.* 7 (6 2011), pp. 1966–1973. DOI: 10.1039/C0MB00347F.

[72] Alberto Calderone, Luisa Castagnoli and Gianni Cesareni. 'mentha: a resource for browsing integrated protein-interaction networks'. In: *Nature Methods* 10 (July 2013), p. 690. DOI: 10.1038/nmeth.2561.

[73] Damian Szklarczyk et al. 'STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets'. In: *Nucleic Acids Research* 47.D1 (Nov. 2018), pp. 607–613. ISSN: 0305-1048. DOI: 10.1093/nar/gky1131.

[74] Yan Wang et al. 'Mechanism of alternative splicing and its regulation'. In: *Biomedical reports* 3.2 (Mar. 2015), pp. 152–158. ISSN: 2049-9434. DOI: 10.3892/br.2014.407.

[75] René Dreos et al. 'The Eukaryotic Promoter Database: expansion of EPDnew and new promoter analysis tools'. In: *Nucleic Acids Research* 43.D1 (Nov. 2014), pp. 92–96. ISSN: 0305-1048. DOI: 10.1093/nar/gku1111.

[76] Jane M. Landolin et al. 'Sequence features that drive human promoter function and tissue specificity'. In: *Genome research* 20.7 (July 2010), pp. 890–898. ISSN: 1549-5469. DOI: 10.1101/gr.100370.109.

[77] Dan Benveniste et al. 'Transcription factor binding predicts histone modifications in human cell lines'. In: *Proceedings of the National Academy of Sciences of the United States of America* 111.37 (Sept. 2014), pp. 13367–13372. ISSN: 1091-6490. DOI: 10.1073/pnas.1412081111.

[78] Mark B. Gerstein et al. 'Architecture of the human regulatory network derived from ENCODE data'. In: *Nature* 489 (Sept. 2012), p. 91. DOI: 10.1038/nature11245.

[79] Yubo Zhang et al. 'Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations'. In: *Nature* 504.7479 (Dec. 2013), pp. 306–310. ISSN: 1476-4687. DOI: 10.1038/nature12716.

[80] Liina Tserel et al. 'Genome-wide promoter analysis of histone modifications in human monocyte-derived antigen presenting cells'. In: *BMC genomics* 11 (Nov. 2010), pp. 642–642. ISSN: 1471-2164. DOI: 10.1186/1471-2164-11-642.

[81] Almut Schulze and Julian Downward. 'Navigating gene expression using microarrays – a technology review'. In: *Nature Cell Biology* 3.8 (2001), E190–E195. ISSN: 1476-4679. DOI: 10.1038/35087138.

[82] Muhammad Farooq Rai et al. 'Advantages of RNA-seq compared to RNA microarrays for transcriptome profiling of anterior cruciate ligament tears'. In: *Journal of orthopaedic research : official publication of the Orthopaedic Research Society* 36.1 (Jan. 2018), pp. 484–497. ISSN: 1554-527X. DOI: 10.1002/jor.23661.

[83] Zhong Wang, Mark Gerstein and Michael Snyder. 'RNA-Seq: a revolutionary tool for transcriptomics'. In: *Nature Reviews Genetics* 10 (Jan. 2009), p. 57.

[84] Karl V. Voelkerding, Shale A. Dames and Jacob D. Durtschi. 'Next-Generation Sequencing: From Basic Research to Diagnostics'. In: *Clinical Chemistry* 55.4 (2009), pp. 641–658. ISSN: 0009-9147. DOI: 10.1373/clinchem.2008.112789.

[85] Peter J. Park. 'ChIP-seq: advantages and challenges of a maturing technology'. In: *Nature Reviews Genetics* 10 (Sept. 2009), p. 669.

[86] Chi Zhang et al. 'Evaluation and comparison of computational tools for RNA-seq isoform quantification'. In: *BMC Genomics* 18.1 (2017), p. 583. ISSN: 1471-2164. DOI: 10.1186/s12864-017-4002-1.

[87] Celine Everaert et al. 'Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data'. In: *Scientific Reports* 7.1 (2017), p. 1559. ISSN: 2045-2322. DOI: 10.1038/s41598-017-01617-3.

[88] Giacomo Baruzzo et al. 'Simulation-based comprehensive benchmarking of RNA-seq aligners'. In: *Nature Methods* 14 (Dec. 2016), p. 135. DOI: 10.1038/nmeth.4106.

[89] Alexander Dobin et al. 'STAR: ultrafast universal RNA-seq aligner'. In: *Bioinformatics (Oxford, England)* 29.1 (Jan. 2013), pp. 15–21. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bts635.

[90] Yang Liao, Gordon K. Smyth and Wei Shi. 'featureCounts: an efficient general purpose program for assigning sequence reads to genomic features'. In: *Bioinformatics* 30.7 (Nov. 2013), pp. 923–930. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt656.

[91] Rob Patro, Stephen M. Mount and Carl Kingsford. 'Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms'. In: *Nature biotechnology* 32.5 (May 2014), pp. 462–464. ISSN: 1546-1696. DOI: 10.1038/nbt.2862.

[92] Nicolas L. Bray et al. 'Near-optimal probabilistic RNA-seq quantification'. In: *Nature Biotechnology* 34 (Apr. 2016), p. 525. DOI: 10.1038/nbt.3519.

[93] Rob Patro et al. 'Salmon provides fast and bias-aware quantification of transcript expression'. In: *Nature Methods* 14 (Mar. 2017), p. 417. DOI: 10.1038/nmeth.4197.

[94] Avi Srivastava et al. 'RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes'. In: *Bioinformatics (Oxford, England)* 32.12 (June 2016), pp. i192–i200. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btw277.

[95] Rob Patro et al. *Salmon Docs*. URL: https://salmon.readthedocs.io/en/latest/salmon.html (visited on 12th Dec. 2018).

[96] Adam Roberts et al. 'Improving RNA-Seq expression estimates by correcting for fragment bias'. In: *Genome biology* 12.3 (2011), R22–R22. ISSN: 1474-760X. DOI: 10.1186/gb-2011-12-3-r22.

[97] Michael I. Love, John B. Hogenesch and Rafael A. Irizarry. 'Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation'. In: *Nature biotechnology* 34.12 (Dec. 2016), pp. 1287–1291. ISSN: 1546-1696. DOI: 10.1038/nbt.3682.

[98] Davis J. McCarthy, Yunshun Chen and Gordon K. Smyth. 'Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation'. In: *Nucleic acids research* 40.10 (May 2012), pp. 4288–4297. ISSN: 1362-4962. DOI: 10.1093/nar/gks042.

[99] Simon Anders and Wolfgang Huber. 'Differential expression analysis for sequence count data'. In: *Genome Biology* 11.10 (2010), R106. ISSN: 1474-760X. DOI: 10.1186/gb-2010-11-10-r106.

[100] Lauren M. McIntyre et al. 'RNA-seq: technical variability and sampling'. In: *BMC Genomics* 12.1 (2011), p. 293. ISSN: 1471-2164. DOI: 10.1186/1471-2164-12-293.

[101] Mark D. Robinson, Davis J. McCarthy and Gordon K. Smyth. 'edgeR: a Bioconductor package for differential expression analysis of digital gene expression data'. In: *Bioinformatics (Oxford, England)* 26.1 (Jan. 2010), pp. 139–140. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp616.

[102] Michael I. Love, Wolfgang Huber and Simon Anders. 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2'. In: *Genome Biology* 15.12 (2014), p. 550. ISSN: 1474-760X. DOI: 10.1186/s13059-014-0550-8.

[103] Matthew E. Ritchie et al. 'limma powers differential expression analyses for RNA-sequencing and microarray studies'. In: *Nucleic Acids Research* 43.7 (Jan. 2015), pp. 47–47. ISSN: 0305-1048. DOI: 10.1093/nar/gkv007.

[104] Sonia Tarazona et al. 'Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package'. In: *Nucleic acids research* 43.21 (Dec. 2015), e140–e140. ISSN: 1362-4962. DOI: 10.1093/nar/gkv711.

[105] Jun Li and Robert Tibshirani. 'Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data'. In: *Statistical methods in medical research* 22.5 (Oct. 2013), pp. 519–536. ISSN: 1477-0334. DOI: 10.1177/0962280211428386.

[106] Juliana Costa-Silva, Douglas Domingues and Fabricio Martins Lopes. 'RNA-Seq differential expression analysis: An extended review and a software tool'. In: *PloS one* 12.12 (Dec. 2017), e0190152–e0190152. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0190152.

[107] Charlotte Soneson, Michael I. Love and Mark D. Robinson. 'Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences'. In: *F1000Research* 4 (1521 2015). DOI: 10.12688/f1000research.7563.1.

[108] Nikolaos Ignatiadis et al. 'Data-driven hypothesis weighting increases detection power in genome-scale multiple testing'. In: *Nature Methods* 13 (May 2016), p. 577. DOI: 10.1038/nmeth.3885.

[109] Zhen Shao et al. 'MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets'. In: *Genome biology* 13.3 (Mar. 2012), R16–R16. ISSN: 1474-760X. DOI: 10.1186/gb-2012-13-3-r16.

[110] M. Heinig et al. 'histoneHMM: Differential analysis of histone modifications with broad genomic footprints'. In: *BMC Bioinformatics* 16 (Feb. 2015), p. 60.

[111] H. Li et al. 'The Sequence Alignment/Map format and SAMtools'. In: *Bioinformatics* 25.16 (Aug. 2009), pp. 2078–2079.

[112] A. R. Quinlan. 'BEDTools: The Swiss-Army Tool for Genome Feature Analysis'. In: *Curr Protoc Bioinformatics* 47 (Sept. 2014), pp. 1–34.

[113] H. M. Zhang et al. 'Transcription factor and microRNA co-regulatory loops: important regulatory motifs in biological processes and diseases'. In: *Brief. Bioinformatics* 16.1 (Jan. 2015), pp. 45–58.

[114] Yoav Benjamini and Yosef Hochberg. 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing'. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1 (1995), pp. 289–300. ISSN: 00359246.

[115] John R. Stevens, Abdullah Al Masud and Anvar Suyundikov. 'A comparison of multiple testing adjustment methods with block-correlation positively-dependent tests'. In: *PloS one* 12.4 (Apr. 2017), pp. 1–12. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0176124.

[116] Carlo Bonferroni. 'Teoria statistica delle classi e calcolo delle probabilita'. In: *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze* 8 (1936), pp. 3–62.

[117] Yoav Benjamini, Abba M. Krieger and Daniel Yekutieli. 'Adaptive linear step-up procedures that control the false discovery rate'. In: *Biometrika* 93.3 (Sept. 2006), pp. 491–507. ISSN: 0006-3444. DOI: 10.1093/biomet/93.3.491.

[118] Christopher Genovese and Larry Wasserman. 'Operating Characteristics and Extensions of the False Discovery Rate Procedure'. In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 64.3 (2002), pp. 499–517. ISSN: 13697412, 14679868.

[119] H. Kimura. 'Histone modifications for human epigenome analysis'. In: *J. Hum. Genet.* 58.7 (July 2013), pp. 439–445.

[120] Menno P. Creyghton et al. 'Histone H3K27ac separates active from poised enhancers and predicts developmental state'. In: *Proceedings of the National Academy of Sciences* 107.50 (2010), pp. 21931–21936. ISSN: 0027-8424. DOI: 10.1073/pnas.1016071107.

[121] Mary L. McHugh. 'The chi-square test of independence'. In: *Biochemia medica* 23.2 (2013), pp. 143–149. ISSN: 1330-0962.

[122] Sepideh Sadegh et al. 'Randomization Strategies Affect Motif Significance Analysis in TF-miRNA-Gene Regulatory Networks'. In: *Journal of integrative bioinformatics* 14.2 (July 2017), p. 20170017. ISSN: 1613-4516. DOI: 10.1515/jib-2017-0017.

[123] Jacob Cohen et al. *Applied multiple regression/correlation analysis for the behavioral sciences, 3rd ed.* Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers, 2003, pp. xxviii, 703–xxviii, 703. ISBN: 0-8058-2223-2.

[124] Olive Jean Dunn and Virginia Clark. 'Correlation Coefficients Measured on the Same Individuals'. In: *Journal of the American Statistical Association* 64.325 (1969), pp. 366–377. DOI: 10.1080/01621459.1969.10500981.

[125] James B. Hittner, Kim May and N. C. L. A. Y. T. O. N. Silver. 'A Monte Carlo Evaluation of Tests for Comparing Dependent Correlations'. In: *The Journal of General Psychology* 130.2 (Apr. 2003), pp. 149–168. ISSN: 0022-1309. DOI: 10.1080/00221300309601282.

[126] James H. Steiger. 'Tests for comparing elements of a correlation matrix.' In: *Psychological Bulletin* 87.2 (1980), pp. 245–251. DOI: 10.1037/0033-2909.87.2.245.

[127] Steven Xijin Ge and Dongmin Jung. 'ShinyGO: a graphical enrichment tool for ani-mals and plants'. In: *bioRxiv* (2018). DOI: 10.1101/315150.

[128] M. Ashburner et al. 'Gene ontology: tool for the unification of biology. The Gene Ontology Consortium'. In: *Nature genetics* 25.1 (May 2000), pp. 25–29. ISSN: 1061-4036. DOI: 10.1038/75556.

[129] The Gene Ontology Consortium. 'The Gene Ontology Resource: 20 years and still GOing strong'. In: *Nucleic Acids Research* 47.1 (Nov. 2018), pp. 330–338. ISSN: 0305-1048. DOI: 10.1093/nar/gky1055.

[130] Minoru Kanehisa and Susumu Goto. 'KEGG: Kyoto Encyclopedia of Genes and Genomes'. In: *Nucleic Acids Research* 28.1 (Jan. 2000), pp. 27–30. ISSN: 0305-1048. DOI: 10.1093/nar/28.1.27.

[131] Minoru Kanehisa et al. 'KEGG as a reference resource for gene and protein annotation'. In: *Nucleic Acids Research* 44.1 (Oct. 2015), pp. 457–462. ISSN: 0305-1048. DOI: 10.1093/nar/gkv1070.

[132] Python Software Foundation. *Python*. Version 3.6.8. 24th Dec. 2018. URL: https://www.python.org/.

[133] Travis Oliphant. *NumPy: A guide to NumPy*. USA: Trelgol Publishing. 2006–. URL: http://www.numpy.org/.

[134] Wes McKinney. 'Data Structures for Statistical Computing in Python'. In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman. 2010, pp. 51–56.

[135] Eric Jones, Travis Oliphant, Pearu Peterson et al. *SciPy: Open source scientific tools for Python*. 2001–. URL: http://www.scipy.org/.

[136] J. D. Hunter. 'Matplotlib: A 2D graphics environment'. In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[137] Michael Waskom et al. *mwaskom/seaborn: v0.9.0 (July 2018)*. July 2018. DOI: 10.5281/zenodo.1313201.

[138] R Core Team. *R: A Language and Environment for Statistical Computing*. Version 3.4.4. R Foundation for Statistical Computing. Vienna, Austria, 2014. URL: http://www.R-project.org/.

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

[140] Gabor Csardi and Tamas Nepusz. 'The igraph software package for complex network research'. In: *InterJournal* Complex Systems (2006), p. 1695. URL: http://igraph.org.

[141] Lifang Hu et al. 'Mesenchymal Stem Cells: Cell Fate Decision to Osteoblast or Adipocyte and Application in Osteoporosis Treatment'. In: *International Journal of Molecular Sciences* 19.2 (2018). ISSN: 1422-0067. DOI: 10.3390/ijms19020360.

[142] Gail M. Sullivan and Richard Feinn. 'Using Effect Size—or Why the P Value Is Not Enough'. In: *Journal of Graduate Medical Education* 4.3 (2012), pp. 279–282. DOI: 10.4300/JGME-D-12-00156.1.

[143] Hamid Abdollahi et al. 'The role of hypoxia in stem cell differentiation and therapeutics'. In: *The Journal of surgical research* 165.1 (Jan. 2011), pp. 112–117. ISSN: 1095-8673. DOI: 10.1016/j.jss.2009.09.057.