Integrative analysis on the effect of DNA methylation on gene regulation in tumor and healthy tissues

DOCTORAL THESIS
for acquiring the academic degree of
Doctor of Philosophy
(Ph.D.)

Submitted to the Faculty of Life Sciences of Humboldt-Universität zu Berlin

by

Katarzyna Wręczycka
born 19.07.1990 in Warsaw, Poland

President of Humboldt-Universität zu Berlin

Prof. Dr.-Ing. Dr. Sabine Kunst

Dean of the Faculty of Life Sciences of Humboldt-Universität zu Berlin
Prof. Dr. Dr. Christian Ulrichs

Reviewers:

1.

2.

3.

Date of oral examination:

Abstract

DNA methylation is one of the main epigenetic modifications in the human genome. It has been shown to play an essential role in cell-type specific gene regulation, and consequently, in development and maintenance of a cell identity. The influence of DNA methylation on gene expression and on the binding of transcription factors (TFs) to DNA is a complex process that requires an integrative approach. The aim of this thesis is to unveil the mechanisms of transcriptional regulation and the complex dynamics between DNA methylation and TF machinery in two areas: in disease and health.

DNA methylation in cancer plays a variety of roles, transforming the healthy gene regulation to a disease pattern. Neuroblastoma is an early childhood cancer that arises from aberrant differentiation of the neural crest tissues of the sympathetic nervous system. The clinical spectrum of neuroblastoma ranges from spontaneously resolving tumors to a highly aggressive cancer. Known genetic aberrations only partially mirror this diversity, pointing towards epigenetic involvement in neuroblastoma pathogenesis. To address the regulatory role of DNA methylation in an unbiased manner at a single nucleotide resolution and genome-wide scale in primary neuroblastomas, analyzed a cohort of 24 pairs of samples, consisting of samples from whole genome bisulfite-sequencing (Bisulfite-seq) and matching RNA-seq samples. We confirmed previously identified methylation-based clustering, separating high-risk from low-risk tumors and MYCN-amplified from non-MYCN-amplified tumors, and a MYCN-driven deregulation of DNA methylation at regulatory elements in MYCN-amplified samples. Additionally, we applied an integrative approach of combining Bisulfite-seq, RNA-seq, publicly available ChIP-seq for tumor-derived H3K27ac marks, and known DNA motifs. It revealed that, a subset of key TF networks deregulated specifically in high-risk neuroblastomas can be modeled on the basis of DNA methylation alone. Our findings suggest that epigenetic mechanisms are involved in neuroblastoma deregulation programs, and they can be attractive targets for further investigation.

In healthy human cells, CpG-rich clusters, such as CpG islands are mostly maintained in an unmethylated state. In technical terms, detection of TF binding to these regions comes with its flaws. High-occupancy target (HOT) regions are segments of the genome with unusually high number of TF sites, and their role has been discussed in recent years as either a technical bias of ChIP-seq or that they posses a biological function. HOT regions usually occur at CpG islands, coincide with house-keeping gene promoters, and genes that are stably expressed across multiple cell types. Our study uncovered that HOT regions can be at least partially explained

by formation of persistent DNA-RNA hybrids plus the displaced single-stranded DNA. These secondary DNA structures are typically formed during transcription, and similarly to HOT regions are associated with a DNA methylation-free state. Furthermore, we propose strategies how to deal with HOT regions for the future ChIP-seq studies.

Finally, this work provides novel insights into gene regulation and transcription factor binding due to DNA methylation dynamics in cancerous and non-cancerous cells.

Zusammenfassung

Die DNA-Methylierung ist eine der wichtigsten epigenetischen Modifikationen im menschlichen Genom. Sie spielt eine wesentliche Rolle bei der zelltypspezifischen Genregulation und folglich bei der Entwicklung und Aufrechterhaltung der Zellidentität. Der Einfluss der DNA-Methylierung auf die Genexpression und auf die Bindung von Transkriptionsfaktoren (TFs) an die DNA ist ein komplexer Prozess, der einen integrativen Ansatz erfordert. Ziel dieser Arbeit ist es, die Mechanismen der Transkriptionsregulation und die komplexe Dynamik zwischen DNA-Methylierung und TF-Maschinerie in zwei Bereichen aufzuklären: in Krankheit und Gesundheit.

In Krebszellen beeinflusst DNA-Methylierung verschiedene Faktoren und verwandelt dabei die gesunde Genregulation in ein Krankheitsmuster. Das Neuroblastom ist eine frühkindliche Krebserkrankung, die aus einer aberranten Differenzierung des Neuralleistengewebes des sympathischen Nervensystems entsteht. Das klinische Spektrum des Neuroblastoms reicht von spontan abklingenden Tumoren bis hin zu einem hochaggressiven Krebs. Bekannte genetische Aberrationen spiegeln diese Vielfalt nur teilweise wider, was auf eine epigenetische Beteiligung an der Neuroblastom-Pathogenese hindeutet. Um die regulatorische Rolle der DNA-Methylierung in primären Neuroblastomen in einer unvoreingenommenen Art und Weise mit einer Auflösung von einzelnen Nukleotiden und auf genomweiter Ebene zu untersuchen, analysierten wir eine Kohorte von 24 Probenpaaren, bestehend aus Proben aus Ganzgenom-Bisulfit-Sequenzierung (Bisulfit-seq) und dazu passenden RNA-seq-Proben. Wir bestätigten ein zuvor identifiziertes Methylierungs-basiertes Clustering, das Hoch-Risiko- von Niedrig-Risiko-Tumoren und MYCN-amplifizierte von nicht-MYCN-amplifizierten Tumoren trennt, sowie eine MYCN-getriebene Deregulation der DNA-Methylierung an regulatorischen Elementen in MYCN-amplifizierten Proben. Zusätzlich haben wir einen integrativen Ansatz angewandt, der Bisulfit-seq, RNA-seq, öffentlich verfügbare H3K27ac ChIP-seq-Marker aus Tumoren und bekannte DNA-Motive kombiniert. Es zeigte sich, dass eine Untergruppe von Schlüssel-TF-Netzwerken, die speziell in Hochrisiko-Neuroblastomen dereguliert sind, allein auf Basis der DNA-Methylierung modelliert werden kann. Unsere Ergebnisse deuten darauf hin, dass epigenetische Mechanismen in Neuroblastom-Deregulationsprogrammen involviert sind, und sie könnten attraktive Ziele für weitere Untersuchungen sein.

DNA-Methylierung erfolgt meist an Cytosine Nucleotidien innerhalb nebeneinanderliegender Cytosine-Guanine Basenpaare (CpG) und in gesunden menschlichen Zellen werden CpG-reiche Cluster, wie z.B. CpG-Inseln, meist in einem unmethylierten Zustand gehalten. Technisch gesehen hat der Nachweis der TF-Bindung in diesen Regionen seine Tücken.

Abschnitte des Genoms mit einer ungewöhnlich hohen Anzahl von TF-Bindestellen sind als High-Occupancy-Target (HOT)-Regionen bekannt, und ihre Rolle wurde in den letzten Jahren entweder als technische Verzerrung des ChIP-seq-Verfahrens oder als biologische Funktion diskutiert. HOT-Regionen treten in der Regel an CpG-Inseln auf und fallen mit Promotoren von Housekeeping-Genen und Genen zusammen, die über mehrere Zelltypen hinweg stabil exprimiert werden. Unsere Studie deckte auf, dass HOT-Regionen zumindest teilweise durch die Bildung von persistenten DNA-RNA-Hybriden und der verdrängten einzelsträngigen DNA erklärt werden können. Diese sekundären DNA-Strukturen werden typischerweise während der Transkription gebildet und sind, ähnlich wie HOT-Regionen, mit einem unmethylierten Zustand verbunden. Darüber hinaus schlagen wir Strategien für den Umgang mit HOT-Regionen für zukünftige ChIP-seq-Studien vor.

Schließlich liefert diese Arbeit neue Einblicke in die Regulation der Gen-Expression und der Bindung von Transkriptionsfaktoren durch die Dynamik der DNA-Methylierung in gesunden und krebsartigen Zellen.

Author Contributions

Parts of this dissertation have been released before in the following publications:

1. Katarzyna Wreczycka*, Alexander Gosdschan*, Dilmurat Yusuf, Björn Grüning, Yassen Assenov, and Altuna Akalin. "Strategies for analyzing bisulfite sequencing data." Journal of biotechnology (2017)

2. Ricardo Wurmus, Bora Uyar, Brendan Osberg, Vedran Franke, Alexander Gosdschan, Katarzyna Wreczycka, Jonathan Ronen, and Altuna Akalin. "PiGx: reproducible genomics analysis pipelines with GNU Guix." Gigascience (2018)

3. Katarzyna Wreczycka*, Vedran Franke*, Bora Uyar, Ricardo Wurmus, Selman Bulut, Baris Tursun, and Altuna Akalin. "HOT or not: Examining the basis of high-occupancy target regions" Nucleic Acids Research (2019)

* indicates co-first authors

Section 1.3.1 is adapted from Wreczycka, Gosdschan, et al. 2017, with large sections reproduced verbatim. Pre-processing of Bisulfite-seq data in Chapter 2 and 3 was performed using the PiGx BSseq pipeline from Wurmus et al. 2018. Chapter 3 is reproduced with minor edits from Wreczycka, Franke, et al. 2019. Figure captions state when figures are reproduced from any of these publications.

Author contributions for the publications are as follows:

(Wreczycka, Gosdschan, et al. 2017) A.A. designed the study with input from K.W. and A.G. K.W described DNA methylation detection methods, Bisulfite-seq processing, and did differential methylation analysis, reviewed its methods and tools, and wrote a substantial part of the manuscript. A.G. did and described segmentation analysis, and reviewed strategies for dealing with large datasets. A.A. reviewed annotation of DMRs/DMCs and segments, and workflows and tools that do not require programming experience. A.A. supervised the writing, analysis, and ensured its progress.

(Wurmus et al. 2018) R.W, B.U, B.O, V.F, A.G, A.A wrote the manuscript. R.W, B.U, B.O, V.F, A.G implemented RNA-seq, ChIP-seq and scRNA-seq pipelines. K.W.'s initial BS-seq snakemake pipeline was fixed and improved by A.G, B.O, and KW, and incorporated into PiGx PIGx pipelines. A.A., R.W and B.O supervised the project and ensured its progress.

(Wreczycka, Franke, et al. 2019) A.A. and V.F. conceived the idea during discussions on ChIP-seq noise on ENCODE datasets. A.A. designed the study with input from V.F. and K.W.

K.W. downloaded and processed all the ChIP-seq, DRIP-seq, G4-ChIP-seq data from human, mouse, worm and fly. Yeast DRIP-seq data is processed by V.F. HOT region algorithm is designed and implemented by A.A. with contributions from K.W. HOT region predictions for each species is done by K.W. Machine learning approach is implemented by A.A., K.W. and V.F. B.U. and R.W. provided support with data analysis, processing for peak calling and examining HOT region sequence characteristics. K.W., V.F., A.A. and B.U. wrote the manuscript. A.A. supervised the project and ensured its progress.

Declaration

I hereby declare that I completed the doctoral thesis independently based on the stated resources and aids. I have not applied for a doctoral degree elsewhere and do not have a corresponding doctoral degree. I have not submitted the doctoral thesis, or parts of it, to another academic institution and the thesis has not been accepted or rejected. I declare that I have acknowledged the Doctoral Degree Regulations which underlie the procedure of the Faculty of Life Sciences of Humboldt-Universität zu Berlin, as amended on 5th March 2015. Furthermore, I declare that no collaboration with commercial doctoral degree supervisors took place, and that the principles of Humboldt-Universität zu Berlin for ensuring good academic practice were abided by.

Erklärung

Hiermit erkläre ich, die Dissertation selbstständig und nur unter Verwendung der angegebenen Hilfen und Hilfsmittel angefertigt zu haben. Ich habe mich anderwärts nicht um einen Doktorgrad beworben und besitze keinen entsprechenden Doktorgrad. Ich erkläre, dass ich die Dissertation oder Teile davon nicht bereits bei einer anderen wissenschaftlichen Einrichtung eingereicht habe und dass sie dort weder angenommen noch abgelehnt wurde. Ich erkläre die Kenntnisnahme der dem Verfahren zugrunde liegenden Promotionsordnung der Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin vom 5. März 2015. Weiterhin erkläre ich, dass keine Zusammenarbeit mit gewerblichen Promotionsberaterinnen/Promotionsberatern stattgefunden hat und dass die Grundsätze der Humboldt-Universität zu Berlin zur Sicherung guter wissenschaftlicher Praxis eingehalten wurden.

Berlin,

Katarzyna Wręczycka

Contents

1 Introduction

1.1 Thesis outline

In the following sections, I will introduce gene regulation mechanisms by transcription factors and epigenetic regulation in healthy and cancerous cells (section 1.2). In section 1.3, I will describe two genomic assays, namely Bisulfite-seq and ChIP-seq, and their computational processing techniques that are extensively explored in Chapters 2 and 3. The first part of the content of this section is largely reproduced from Wreczycka, Gosdschan, et al. 2017, where I reviewed computational methods and tools for processing Bisulfite-seq data. Approaches for computational integration of DNA methylation with other genomic data will be reviewed in section 1.4. Then, I will introduce, in more detail, few problem areas where I have made contributions. Chapter 2 describes a novel genome-wide DNA methylation at single nucleotide analysis in neuroblastoma, integrated with gene expression, chromatin modifications, and sequence information data that reveals well-known and new regulatory networks in high-risk neuroblastoma subgroups. The Bisulfite-seq raw data from neuroblastoma patient tumors was processed using a software that I co-developed and that was published and described in great detail in Wurmus et al. 2018. Apart from many factors which influence gene regulation and are linked to unmethylated state, such as most of DNA binding proteins, DNA-RNA hybrids have recently been well characterized, not only as by-products of transcription, but are also known to participate in a number of physiological processes, and form across the genomes

of bacteria, yeast, and higher eukaryotes throughout cell cycle. I will describe their impact on transcription factor binding detection issues in ChIP-seq in Chapter 3. These findings were published in Wreczycka, Franke, et al. 2019, from which the content is reproduced, with some editing for clarity in the context of this dissertation. Lastly, I will discuss the overall impact of the work and share some concluding remarks in Chapter 4.

1.2 Elements of gene regulation in healthy and cancerous cells

1.2.1 Interplay between genome organization, gene regulation and epigenetics

The human genome contains 3.3 billion base pairs of DNA, but the genome is more than just a naked stretch of DNA. A double-helix DNA is wrapped around proteins (called histones) forming higher-order structures called nucleosomes that are part of a chromatin fibre and packed into 23 chromosomes (Figure 1.1). The processes of unwinding and winding stretches of DNA are regulated by the epigenome.

The human genome can be classified into coding and non-coding DNA. A DNA fragment that can be transcribed into RNA and then translated into a protein is called a coding sequence. Non-coding DNA is defined as all of the DNA sequences that do not code for proteins. The vast majority of all nucleotides in the human genome are non-coding sequences (circa 98%) and coding sequences are only a small fraction of the genome (< 2%) (Lander 2011). Initially, non-coding DNA was named as "junk DNA" due to its unknown biological function. However, comparative genomics studies uncovered that the majority of mammalian-conserved regions consist of non-coding elements. It indicates that functional elements with regulatory and other functions are situated in non-coding regions of the human genome (Kellis et al. 2014). In every cell of our organism, there occurs coordinated expression of thousands of individual genes, to use the biological information encoded in its genome. It is estimated that the human genome encodes approximately 25,000 genes. Mostly, expression of genes, i.e. the processes of reading genetic information in genes and rewriting it into products (various forms of RNA or proteins), is unique for each cell-type. Every gene is a unit of information that can influence phenotype - observable physical properties of an organism. Many genes require not only coding sequence

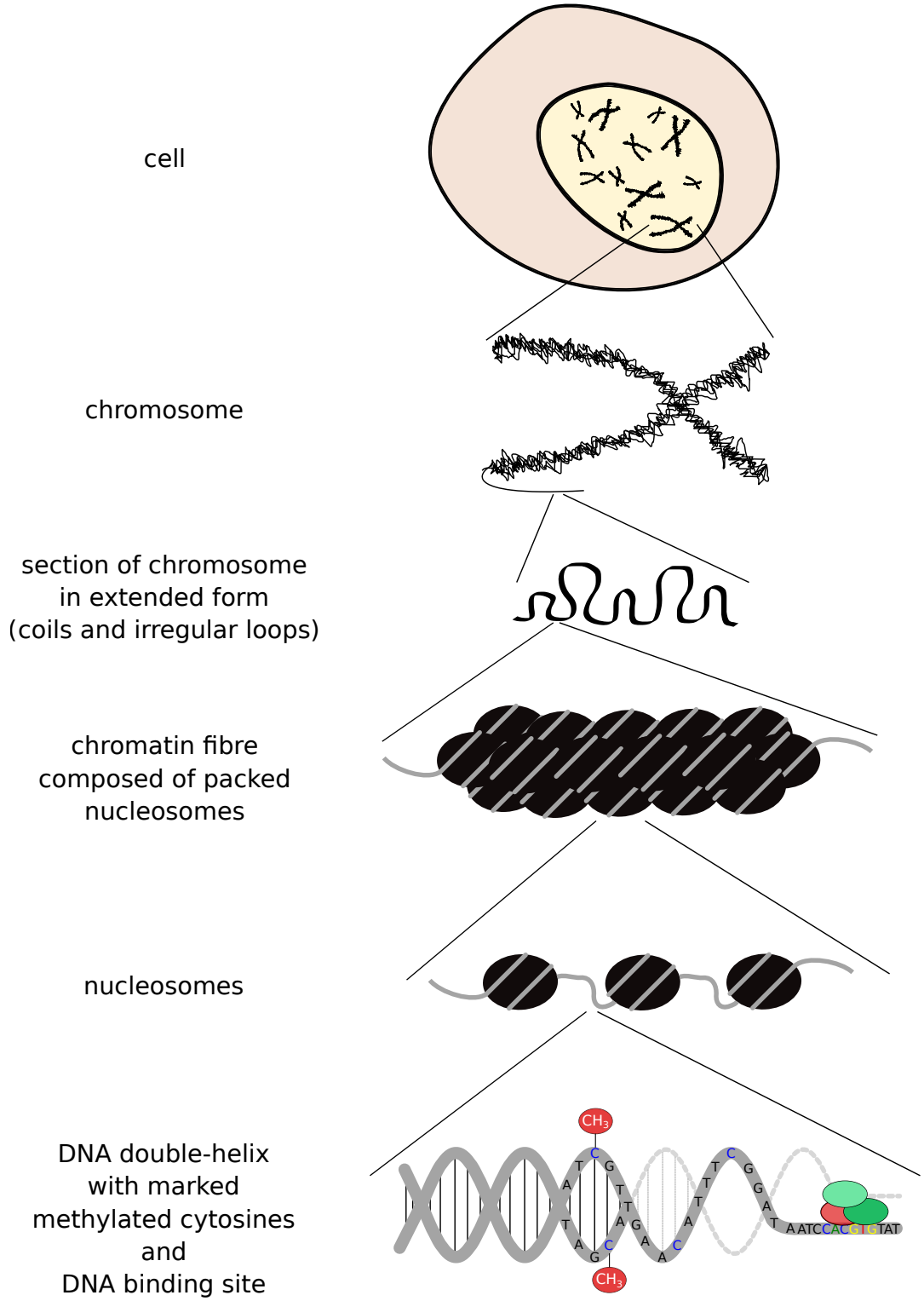


Figure 1.1: A representation of chromosome structure in animals with marked DNA methylation and binding proteins to a DNA sequence.

but also correctly functioning regulatory sequences, that is, multiple cis-acting genomic regions (regions on the same chromosome) for correct spatial, temporal and quantitative expression (Kleinjan 1998). It is accomplished through the sequence-specific binding of proteins (known as transcription factors) which may promote or block the transcription of genetic information from DNA to RNA.

In order for a gene to have a chance to be accessed by transcription factors, the DNA molecule has to be first unwound from histones, nucleosomes, chromatin fibres and a chromosome. DNA unwinding occurs in a cell-type-specific fashion, regulated by stable and heritable modifications that are distinct from DNA sequences and fostered by specialized mechanisms called epigenetic modifications: DNA methylation, chromatin remodeling, histone modifications and non-coding RNA mechanisms.

In the following sections I will provide the reader with some fundamentals on gene regulation mechanisms (regulation of the rate of transcription) and review advances in the understanding of the mechanisms and role of DNA methylation in biological processes in healthy and cancerous cells.

1.2.2 Gene regulation by transcription factors

1.2.2.1 Promoter elements

Before the first steps of gene expression, the two strands of DNA need to be separated in order to expose a gene promoter to a pre-initiation complex (PIC). Gene promoter elements consist of core and proximal promoter DNA fragments in the immediate neighbourhood of the 5' end of a gene of the template strand (the other strand is called the non-template or coding strand), and are usually 100-1000 base pairs long. The site on the DNA within a core promoter from which the first RNA nucleotide is transcribed by RNA polymerase II (RNAP II) is said to be at position +1 and is called the transcription start site (TSS). The PIC is a complex of ~100 proteins that is necessary for the transcription initiation. The minimal PIC includes RNAP II and six helper proteins called general transcription factors that bind to core promoter elements.

Transcription factors (TFs) are proteins that can bind to the DNA at promoter elements and distal elements: enhancers, silencers or insulators (Bulger and Groudine 2011; Shlyueva, Stampfel, and Stark 2014). Both, promoter and distal elements, contain patterns of nucleotides

on a DNA strand recognised by specific transcription factors (Figure 1.2A). For example, for many genes, especially those encoding abundantly expressed proteins, a TATA box located 20-30 base pairs upstream from the cap site directs RNA polymerase II to the start site, and is bound specifically by a general transcription factor II D (TFIID). Other DNA patterns that are part of the core promoter are: downstream promoter elements (DPEs), the B recognition elements (BRE), and CpG islands (see Figures 1.2B and 1.4A). DPEs are thought to have a similar function to a TATA-box, BRE element is recognised by a general transcription factor II B (TFIIB), and lies upstream of the TATA-box. CpG islands are regions with large numbers of repeats of CG dinucleotides (a cytosine nucleotide followed by a guanine nucleotide). In vertebrate genomes, CpG islands extend for 300-3000 base pairs and are located within or in close proximity to sites of about 50-70% of mammalian gene promoters. Additional regulatory complexes (such as the mediator co-activator (Allen and Taatjes 2015), and chromatin remodeling complexes (Clapier et al. 2017) may also be components of the PIC. Proximal promoter elements are relatively short (~15–30 base pairs), and are located within the first ~200 base pairs upstream of TSS.

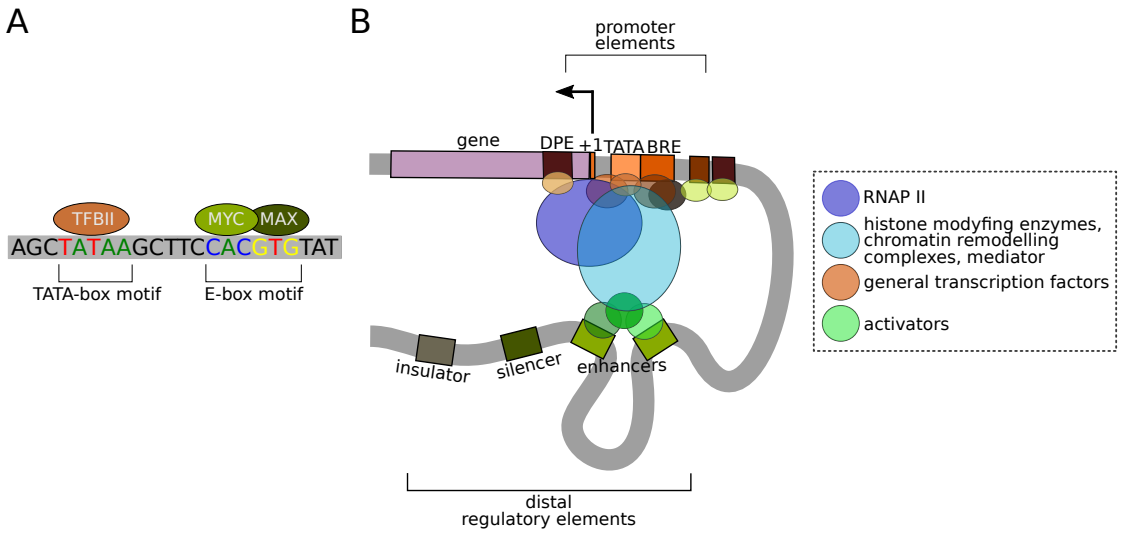


Figure 1.2: Simplified representation of regulatory regions and transcription complexes in animal genomes. (A) Transcription factors bind to specific DNA patterns (DNA motifs). On the left side of the figure TFIIB protein binds to a TATA-box motif, on the right side dimer of MYC and MAX proteins binds to an E-box motif. (B) Initiation of transcription requires pre-initiation complex that consists of RNA polymerase II (RNAPII), general transcription factors, specific transcription factors (activators), mediator, histone modifying enzymes and chromatin remodelling complexes. Figures modified from Akalin 2020.

In terms of protein coding-genes, once the correct transcription factors complex is formed, the RNA polymerase II moves along the template strand, adding more nucleotides to the newly formed mRNA (elongation phase), completes transcription with the breakdown of the DNA and RNA polymerase II, and the release of the newly formed RNA molecule (termination phase). After mRNA is produced it is often spliced, which means that sections of that RNA called introns are removed and sections called exons are left in. Then, based on the final mature mRNA, proteins can be created.

1.2.2.2 Distal regulatory elements

Generally, genes are inactive unless activated by binding transcription factors binding to their promoter elements and regulatory sites. Most transcription factor binding sites in the human genome are found in intergenic regions (a stretch of DNA sequences located between genes) and introns, which indicates the widespread usage of distal regulatory elements in the genome.

Enhancer sequences are regulatory sites found in distal, intergenic regions, and introns that enhance transcription through recruitment of transcription factors that are co-activator proteins (activators). In contrary to enhancers, silencers carry binding sites for repressor proteins (repressors) (Figure 1.2B). As an example of co-activators, enhancer box DNA motif (E-box) with a palindromic canonical sequence of "CACGTG" is found on enhancers that are bound by transcription factors with a basic helix-loop-helix domain (bHLH), such as MAX-MYC heterodimer (Blackwood and Eisenman 1991) (Figure 1.2A).

The activators interact with each other and join a transcription initiation complex through proteins called mediators, enabling the creation and assembly of an effective protein complex at the transcription initiation site. In eukaryotic organisms, the basic mechanism of repressors is based on the principle of inactivation of activators, directly competing for the same binding site or induce a repressive chromatin state in which no activator binding is possible. The activity of activators or repressors is independent of their distance to the promoter they interact with. Many mammalian genes are usually controlled by more than one enhancer region.

A number of studies have shown that enhancers can regulate their target genes even up to a megabase (1,000,000 bp) away (Plank and Dean 2014; Pombo and Dillon 2015; Long,

Prescott, and Wysocka 2016; Furlong and Levine 2018). It's due to the dynamic promoter-enhancer DNA loops, that allow enhancers to get in contact with promoters elements. The formation of the loop-like structure allows activators or repressors to take part in the process of regulating the expression of many genes in a specific region (Figure 1.2B). For the spatial isolation of their regions of activity, chromosomes are divided into regulatory neighborhoods separated by sequences known as insulators. Insulators block enhancer-promoter communication and/or prevent silencing open chromatin regions by spreading heterochromatic DNA. The most studied insulators are bound by CTCF (CCCTC-binding factor) that appear to shape 3D structure of DNA in order to block enhancer activity. CTCF, together with a protein complex cohesin, is associated with topologically associating domains that form insulated neighbourhoods (Dixon et al. 2012; Ciabrelli and Cavalli 2015). Genome-wide studies from different tissues confirm that CTCF binding is largely invariant of a cell type, and CTCF motif locations are conserved, indicating that vertebrates rely heavily on the CTCF activity (Khoury, Achinger-Kawecka, Saul A Bert, et al. 2020a).

1.2.3 Epigenetic regulation

Epigenetic changes to the genome affect how DNA is packaged and expressed without altering the underlying DNA sequence (Allis, David Allis, and Jenuwein 2016). Research into epigenetics has demonstrated that epigenetic regulation of genes expression has a critical role in normal development and cellular functions, including imprinting, X-inactivation and tissue-specific gene expression (A. P. Bird 1986; P A Jones 2001). Chemical modifications of DNA or modifications of chromatin-associated proteins, have a major influence on chromatin structure and gene expression. If a region is not accessible for the transcriptional machinery, for instance, when the chromatin structure is compacted due to DNA methylation of the promoter of a gene or the presence of specific histone modifications, transcription may not start at all. Gene expression might also be controlled post-transcriptionally by non-coding RNAs (Statello et al. 2021), protein modifications and protein-protein interactions (Linhares, Grembecka, and Cierpicki 2020), but they are not subject of this thesis.

1.2.3.1 DNA methylation

DNA methylation is the covalent addition of the methyl group to cytosine by an enzyme called DNA methyltransferase 1 (DNMT1), resulting in 5-methylcytosine (5-mC or mC) (Figure 1.3). DNA methylation is generally associated with gene silencing (Zemach et al. 2010; Hashimshony et al. 2003; Venolia and Gartler 1983). There are two mechanisms associated with DNA methylation that lead to silencing. Firstly, gene silencing is based on the induction of spread of repressive chromatin structure. CpG sites can be bound by methylated CpG binding proteins such as MeCP1 and MeCP2 that contain DNA binding and transcriptional repression domain. Secondly, DNA methylation inhibits transcription factor binding. Hence, if the promoter of a gene is methylated, then it is located in condensed form of chromatin, thereby preventing transcription initiation complexes from reaching it, and as a consequence the gene is silenced. In each cell type different sets of genes might be activated or silenced, and since DNA methylation is a heritable mark, this makes it a cell-type specific pattern (Schübeler 2015). DNA methylation is reversible, but mostly remains stable through cell division. The stability of epigenetic marks, such as DNA methylation is important, because according to Waddington's epigenetic landscape theory (Waddington 1956), the more differentiated a cell becomes, the harder it gets to go back to the pluripotent state. Therefore, DNA methylation plays an important role in cell-type specific regulation of a gene expression, and cell-type identity.

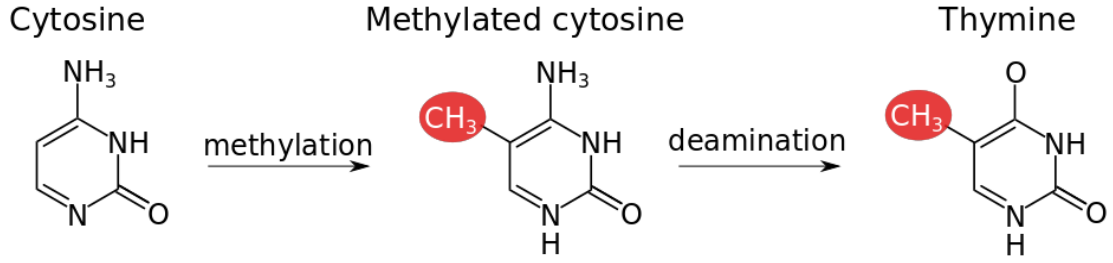


Figure 1.3: Schematic representation of the reaction of cytosine methylation and deamination of methylated cytosine to thymine. Figure reprinted and modified from Day and David Sweatt 2010.

DNA methylation occurs almost exclusively within a symmetric CpG context. The "CpG" is a term describing a cytosine (C) followed by guanine (G) in a DNA nucleotide and linked by a phosphate bond (p). There are roughly 28 million CpGs in the human genome, and most of them (70% to 80%) are methylated. CpGs are not uniformly distributed over the genome, but

rather are associated with CpG-dense sites in proximal promoter elements, upstream of a gene, called CpG islands (Z. D. Smith and Meissner 2013). CpG islands are found in around 60% of promoters of mammalian genes, usually unmethylated, and mostly found in housekeeping genes in vertebrates (housekeeping genes are expressed in all cells of an organism under normal conditions and are required for the maintenance of basic cellular function) (Bibikova 2016). It has been demonstrated that methylation of CpGs as far as 100 kb away from the TSS of a gene can be associated with its expression (Kamalakaran et al. 2011; Fleischer et al. 2014). Recently, it has been shown that methylation differences between tissues, or between healthy and disease samples, can be found not only at CpG islands, but also in a short distance from the CpG islands called CpG island shores (Irizarry et al. 2009) and shelves (Bibikova 2016).

CpG islands are highly CpG rich regions, and it is argued that there is an evolutionary mechanism behind CpG island formation (Cohen, Kenigsberg, and Tanay 2011). However, in general, CG dinucleotides are very uncommon in vertebrate genomes. In fact, in humans and mice, CGs are the least frequent dinucleotide, making up less than 1% of all dinucleotides. There are fewer CpGs than expected by chance in the genome, because DNA methylation is highly mutagenic. Methylated cytosine ammonia compounds (NH_3) can be relatively, easily and spontaneously lost and then methylated cytosine is converted (deaminated) to a thymine residue (Figure 1.3). Therefore, because CpG dinucleotides steadily mutate to TpG dinucleotides, there are more TG residues and fewer CG residues in a genome as expected. Contrastingly, spontaneous deamination of unmethylated cytosine creates uracil residue, a mutation that is quickly recognized and fixed back to cytosine by the cell.

In contrast to CpG islands, intergenic regions and repetitive DNA elements tend to be methylated (Figure 1.4). DNA methylation function at intergenic regions is thought to be important for maintaining genomic integrity. Mouse embryonic stem cells deficient DNMT1, although viable, die when induced to differentiate, and DNMT1 conditional knockout mouse fibroblasts die within a few cell divisions after DNMT1, gene deletion (Jackson-Grusby et al. 2001; Panning and Jaenisch 1996). Their other function is to silence cryptic start sites or cryptic splice sites (Neri et al. 2017). For instance, if a cryptic promoter is not silenced, it can interfere with a closely located promoter of another gene. As a result, RNA polymerases II might be laid down on both of these promoters, consequently bump into each other,

resulting in transcription interference, and inability to generate the whole transcript of a non-cryptic promoter.

Repetitive elements are patterns of DNA that occur in multiple copies throughout the genome. It's desirable by an organism to repress or silence them to avoid their spreading (by "cut-and-paste" or "copy-and-paste" mechanisms) to random parts of the genome that can consequently disrupt its genomic functions. Repetitive elements have such strong promoters, that they can be active in every cell at all times and can additionally read out into neighbouring genes - start transcription of neighbouring genes when they are not supposed to be transcribed. Repeats can be silenced by marking CpG in their promoters with a methylation group and/or by cytosine to thymine mutation in CpG context (CpGs to TpGs mutation).

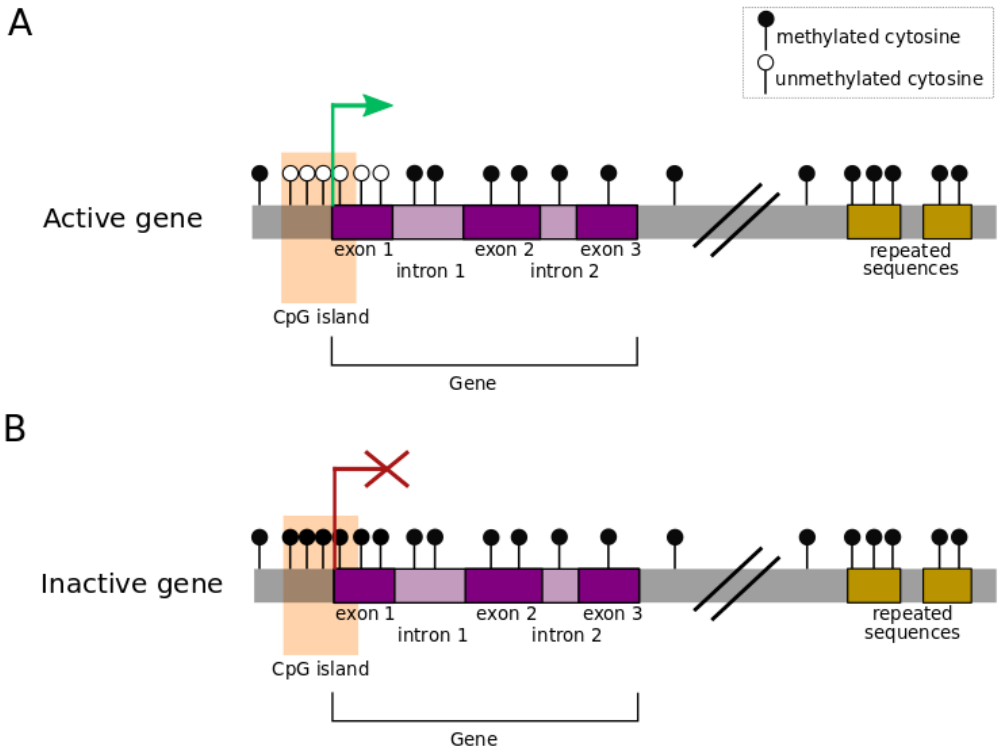


Figure 1.4: Classic paradigm of DNA methylation association with transcription. DNA methylation of the CpG island and the first exon of a gene prevents initiation of gene expression. (A) A representation of an active gene. (B) A representation of a repressed gene. Methylated cytoisnes are depicted as black, and unmethylated cytosines as white lollipops.

1.2.3.2 Histone modifications

Histones are proteins that provide structural support to a chromosome and their modifications make chromatin more compact or more open for proteins to bind to the DNA (Bednar et al. 1998; Fischle, Yanming Wang, and David Allis 2003). They are the chief protein components of chromatin, acting as spools around which DNA winds (Luger et al. 1997). It enables chromatin to be either densely packed (called heterochromatin or closed chromatin), or to be loosely packed (called euchromatin or open chromatin). Highly packed heterochromatin is thought to be inaccessible to transcriptional machinery and therefore not transcribed.

Histones undergo post-translational modifications, by covalent modifications of their long and unstructured N-terminal tails. Modifications of the tails include: acetylation (ac), methylation (me) and phosphorylation (p). Using their tails, histones interact with neighboring nucleosomes and the modifications on tails affect the nucleosomes' affinity to bind DNA, and therefore influence DNA packaging around nucleosomes. The common nomenclature of histone modifications is as follows - the name of the histone (e.g H3), the single-letter amino acid abbreviation (e.g., K for Lysine) and the amino acid position on the tail, the type of modifications (e.g. acetylation) and the number of modifications (methylation is known to occur in more than one copy per residue, such as 1,2,or 3 is mono-, di- or tri-methylation respectively). For example, H3K27ac denotes the acetylation of the 27th residue that is a lysine from the start of the tail (i.e. the N-terminal) of the H3 protein (Bannister and Kouzarides 2011).

Histone modifications are proposed to affect chromosome function through at least two distinct mechanisms. The first mechanism suggests that modifications may alter the electrostatic charge of the histone resulting in a structural change in histones or their binding to DNA. The second mechanism proposes that these modifications are binding sites for protein recognition modules, such as the bromodomains or chromodomains, which recognize acetylated lysines or methylated lysines, respectively (Füllgrabe, Kavanagh, and Joseph 2011).

Different modifications on histones tails are used in different combinations for various purposes, including: active promoters and enhancers are commonly marked by H3K9ac, and H3K4me1/me2/me3; transcribed regions are enriched for H3K14ac; enrichment of H3K4me1/me2, H3K27ac and the histone acetyltransferase p300 is enhancer-specific; and repressed genes may be located in large domains of H3K9me3/me2/me1 or H3K27me3/me2/me1 (selected

histone modifications are depicted in Figure 1.5). Consequently, histone modifications indicate activity of genes and regulatory regions (V. W. Zhou, Goren, and Bernstein 2011; Bannister and Kouzarides 2011).

Some proteins can modify histones (and other proteins), and silence target genes. Polycomb-group proteins (PcG) are a family of protein complexes that consist of two proteins - Polycomb repressive complex 1 (PRC1) and Polycomb repressive complex 2 (PRC2). PRC2 contains the histone methyltransferase EZH2 and other proteins (Margueron and Reinberg 2010) that maintain H3K27me3 and lead to heterochromatinization through the binding of PRC1, which contains the chromodomain protein Pc. PRC1 recognizes H3K27me3 to inhibit transcriptional elongation through H2A monoubiquitylation and to compact the chromatin structure (Wenlai Zhou et al. 2008).

A CTCF protein, already mentioned in section 1.2.2.2, binds to insulator sequences, is associated with boundaries between active and repressive histone marks, and topologically associating domains (Henikoff 2008; Phillips and Corces 2009). Histone modification profiles have proven to be particularly useful for identifying enhancer elements in an unbiased fashion. In addition to specific histone modifications, enhancers are preferentially occupied by sequence-specific DNA-binding proteins and co-activators such as p300 (Visel et al. 2009) (as depicted on Figure 1.5).

1.2.4 Introduction to cancer epigenetics

All cancers have a series of common traits that govern the transformation of normal cells to cancer (Hanahan and Weinberg 2000). These hallmarks of cancer include: evading growth suppressors, sustained proliferative signalling, enabling replicative immortality, inducing angiogenesis, activating invasion and metastasis, and resisting cell death. In addition to the traditional hallmarks of cancer, more recently, new hallmarks have been described (Hanahan and Weinberg 2011) - abnormal metabolic pathways, avoiding immune destruction, and two enabling characteristics: inflammation and genomic instability. Originally, cancer was thought to be an accumulation of genetic mutations in DNA that influence each of these cancer hallmarks (Nowell 1976). However, it is already well established that cancer is not just a genetic disease, but rather genetic mutations partnered with epigenetic abnormalities (Mohammad, Barbash, and Creasy 2019).

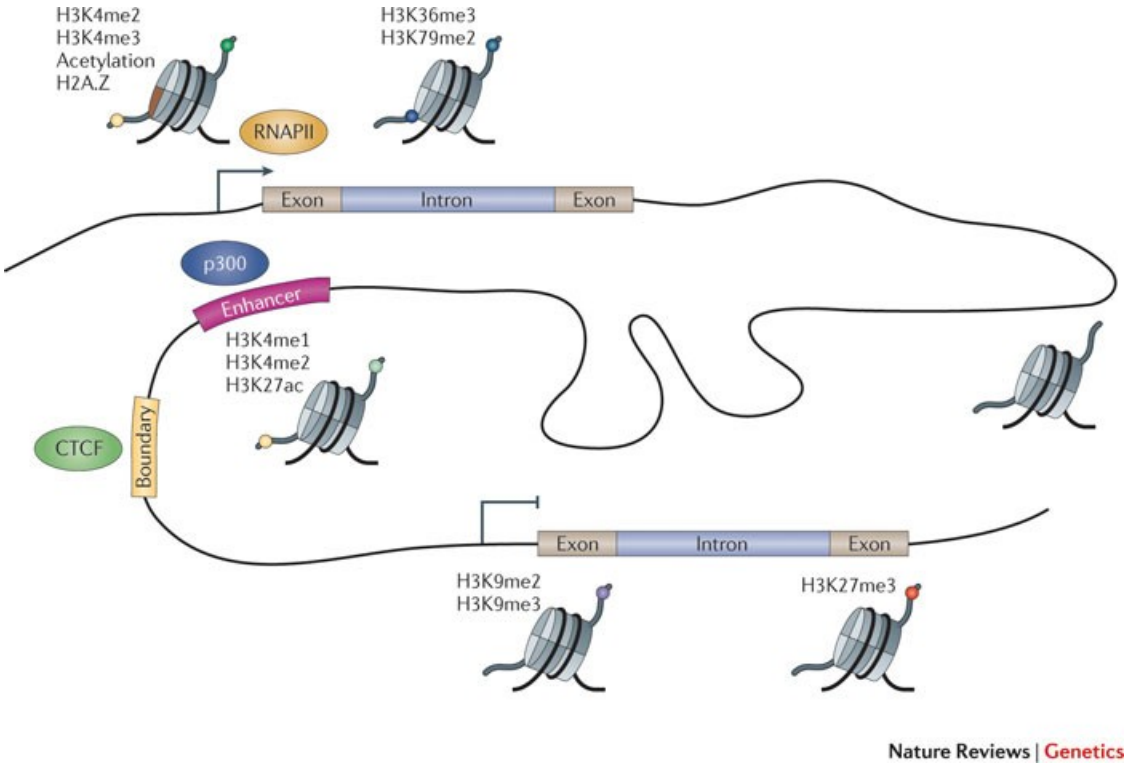


Figure 1.5: Histone modifications associated with functional elements in mammalian genomes. In the figure are depicted histone modifications and selected proteins for active promoters (H3K4me2, H3K4me3, acetylation, and H2A.Z), transcribed regions (H3K36me3 and H3K79me2), repressed genes (H3K9me2 and/or H3K9me3 or H3K27me3), enhancers (H3K4me1, H3K4me2, H3K27ac, and binding of p300), boundary elements, and insulators (binding of CTCF) and histone methyltransferases associated with gene silencing (polycomb-group of proteins, pcG). Abbreviations: RNAPII, RNA polymerase II; p300, histone acetyltransferase p300; CCCTC-binding protein, CTCF. Figure reprinted and modified from V. W. Zhou, Goren, and Bernstein 2011.

1.2.4.1 Genetic aberrations in cancer

Traditionally, cancer has been viewed as a genetic disease that is driven by sequential acquisition of mutations, leading to the constitutive activation of oncogenes and the loss of function of tumor suppressor genes. Oncogenes are genes that are involved in growth promotion to allow dividing rapidly, and immortality of a cell. Tumor suppressors are genes normally able to make the cell die or decrease the proliferation, but their inactivation inhibits them to regulate the cell. However, it has become increasingly evident that combination of activation of oncogenes and inactivation of tumor suppressors can happen both genetically and epigenetically.

At the genetic level, overexpression of oncogenes can happen by oncogene amplification in a genome (by creating many copies of the same gene in different parts of a genome without

a proportional increase in other genes (Albertson 2006)), inactivation of a tumor suppressor by mutating it, or hyper-methylation of its promoter, or entirely deleting tumor suppressor from the genome. In an extreme case, if a tumor suppressor is an E3 ligase, which regulates degradation of an oncogene, then the tumor suppressor is reduced, and the onco-protein is accumulated at the same time (e.g. TRIM33 and SMAD Mészáros et al. 2017). The best known example of proto-oncogenes, genes that can become an oncogene due to mutations or increased expression, is the MYC gene family. In Burkitt's lymphoma, chromosomal translocations cause c-MYC to be placed downstream of the highly active promoter region, leading to overexpression of MYC (D. Liu et al. 2007). Amplification of the n-MYC gene (also called MYCN) is a well established prognostic factor for the high-risk stage of neuroblastoma (J. M. Cheng et al. 1993; G. M. Brodeur et al. 1984; Matthay et al. 2016). p53 is one of the best known tumor suppressor genes that encodes the protein p53, which is involved in many human cancers, and therefore nicknamed "the Guardian of the Genome" (Lane 1992; Danilova 2020). p53 has many different essential functions in a cell including DNA repair, inducing apoptosis, transcription, and regulating the cell cycle (Harris 1996). Mutated p53 is involved in colon cancers, breast cancers, lung cancers, leukemias, lymphomas, sarcomas, and neurogenic tumors (Kandoth et al. 2013). Its popularity among cancers makes it an attractive target for new cancer therapies (Mantovani, Collavin, and Del Sal 2019).

Epigenetic alterations, including DNA methylation, histone modifications and histone variants, nuclear architecture, and noncoding RNAs, are now considered markers to study the cancer states. Mostly, the actual original driver in tumorgenesis remains unknown, that is, what is the cause versus the consequence for these genetic and epigenetic mistakes that are made in cancer. However, one of the reasons that makes epigenetic abnormalities clinically relevant is the fact that they are reversible and more dynamically regulated compared to genomic evolution (Locke et al. 2019).

1.2.4.2 DNA methylation aberrations in cancer

DNA methylation abnormalities have been observed for almost all cancer types (Feinberg and Vogelstein 1983; Koch et al. 2018). Particularly, decrease of global level of methylation (genome-wide hypomethylation) was the first observed epigenetic feature in cancer (Feinberg

and Vogelstein 1983), and later it followed by hypermethylation of CpGs islands at promoters of tumor suppressor genes (Esteller 2002) (Figure 1.6A,B). As cancer progresses over time, these changes take part in transformation of a normal tissue through to hyper-plastic tissue, neoplasia to a metastatic tissue (Figure 1.6C).

In a normal cell, CpG islands generally are not methylated, but intergenic regions, introns and repetitive elements are (Figure 1.6A). In contrast, CpG islands in cancer are more likely to be methylated and the rest of the genome rather unmethylated (Figure 1.6B). It happens in order to silence the promoters of tumor suppressor genes. Although DNA methylation is mitotically heritable, it can lock tumor suppressor genes in an inactive state in daughter cells in terms of epigenetic silencing. Therefore, the cancer methylome is a combination of both somatically acquired DNA methylation changes and characteristics reflecting the cell of origin (Steliarova-Foucher et al. 2017; Fernandez et al. 2012; Hovestadt, D. T. W. Jones, et al. 2014). The latter property allows, for example, tracing of the primary site of highly differentiated metastases of cancers of unknown origin (Moran et al. 2016). It has been shown that DNA methylation profiling is highly robust and reproducible even from small samples and poor quality material (Hovestadt, Remke, et al. 2013).

Hypermethylation of CpG islands and CpG islands shores In the study published in 70s by Knudson 1971, it was proposed that in order for a cancer to occur, it needs to have multiple hits, because just one hit might affect only one copy of two alleles of a gene coding for tumor suppressor or an oncogene. As a consequence, it might not start mechanisms of apoptosis of an altered cell, but it can be sufficient to cause tumorigenesis. Epigenetic alterations can act as one of the "two hits". Hence, hypermethylation events at CpG islands function equivalently to coding-region mutations or deletions, can affect involved signalling pathways, causing tumorigenesis and not starting mechanisms of apoptosis of a cell. In contrast, chromothripsis announced in 2011 similarly involves multiple mutations, but asserts that they may all appear at once (Stephens et al. 2011). Importantly, in comparison to genetic mutations, DNA hypermethylation is reversible, which brings up an option to remove some of these DNA methylation patterns therapeutically, and improve outcome for a patient. In terms of genetic mutations, it's worth mentioning that with CRISPR-base editors (Anzalone, Koblan, and D. R. Liu 2020), it's also

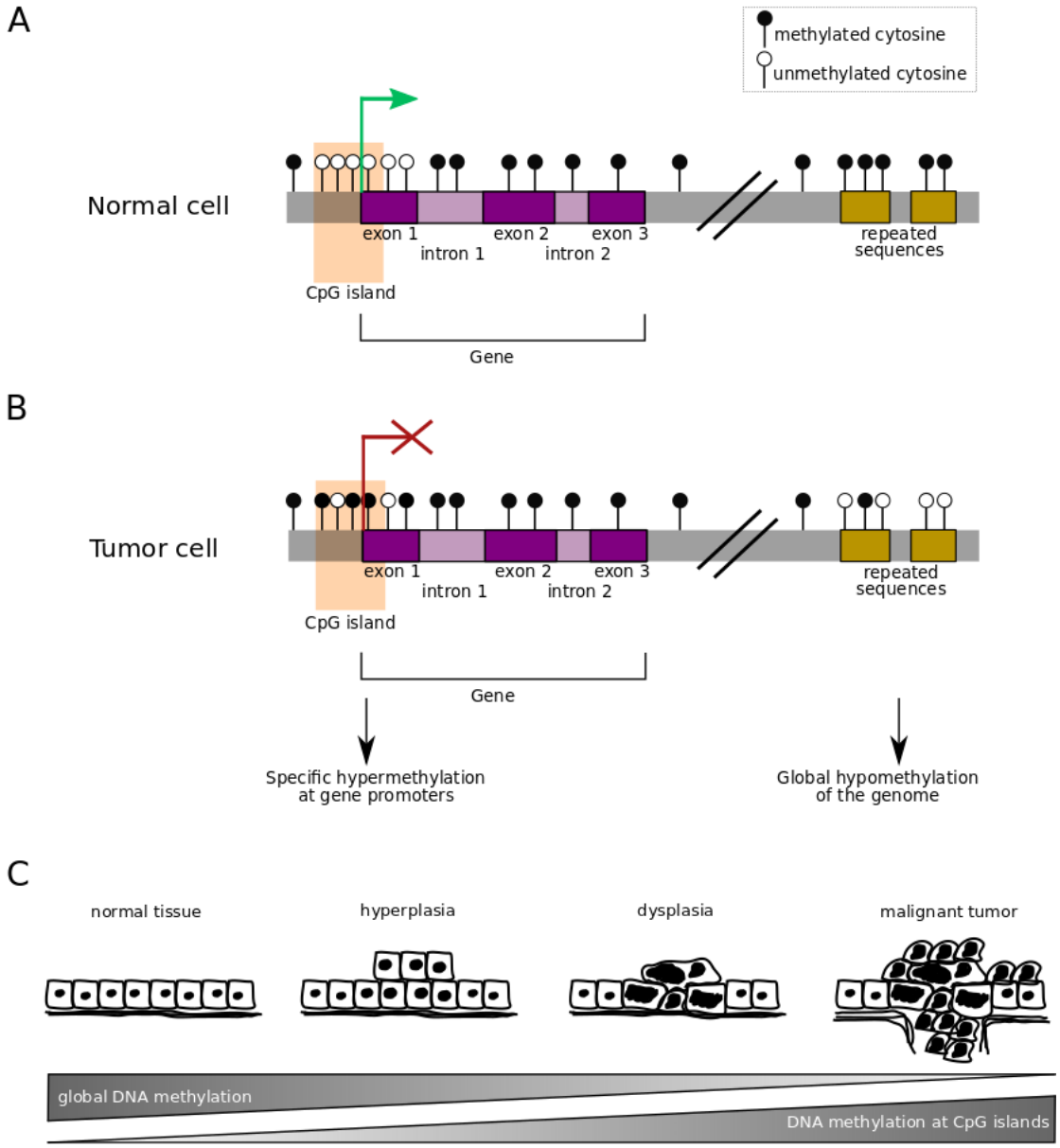


Figure 1.6: Classical DNA methylation characteristics in cancer. DNA methylation patterns in a healthy (A) and cancerous cell (B). (C) Neoplastic progression and changes in DNA methylation patterns. Normal tissue over time might be transformed into hyperplasia, dysplasia and a malignant tumor (invasion fase). Figure reprinted and modified from Barillot et al. 2012.

becoming possible to repair genetic mutations. Therefore, genetic and epigenetic single gene biomarkers can be used prognostically, and diagnostically (Thangue and Kerr 2011).

The most common hypermethylated regions in cancer are located within CpG islands, but recently it has been discovered that hypermethylation of regions that flank CpG islands (up to 2kb distant) with less CG-density play a role in tumorigenesis (Irizarry et al. 2009;

Rao et al. 2013). Although biomarkers found at CpG islands are very good predictors of gene expression, there is evidence that tissue- and cancer-specific differentially methylated regions occur more frequently within CpG islands shores than within CpG islands themselves (Irizarry et al. 2009; Doi et al. 2009).

A particular tumor type where CpG islands of a set of genes are frequently methylated is called CpG island methylator phenotype (CIMP) (J.-P. Issa 2004). CIMP is characterized as sets of CpGs that are hypermethylated and differ by tumor type. This frequent occurrence of hypermethylation has been described for several different types of cancers, first described for colorectal cancer (Toyota et al. 1999), later for glioma (Abe et al. 2005; Noushmehr et al. 2010), neuroblastoma (Abe et al. 2005) and many more (J.-P. Issa 2004). CIMPs can be particularly interesting for diagnosis of cancer subtype and useful for prognosis since panels of CpGs might have stronger prognostic value and more sensitive read out than single CpGs.

Global loss of DNA methylation Global loss of DNA methylation occurs, to some extent, in every tumor. It seems to happen quite early in tumorigenesis, and progresses over time. Regions affected by global hypomethylation in cancer are mostly intergenic regions, repetitive regions, but also to a smaller extent, non-CpG island promoters (or CpG-poor promoters) that can cause oncogene activation (Feinberg and Vogelstein 1983; Miranda and Peter A. Jones 2007; M. A. Kerachian and M. Kerachian 2019). DNA methylation function is to maintain genomic stability, but hypomethylation of intergenic regions and repetitive elements might lead to losing chromosomes or gaining chromosomes, deletions or insertions of parts of chromosomes, and illegitimate recombination between chromosomes (an exchange of genetic material between chromosomes due to alignment between repeats). As an example, in a normal cell, these illegitimate recombinations couldn't happen because the repetitive elements, that cause it, would be heavily methylated and heterchromatonized. Additionally, repetitive elements can be activated because of hypomethylation - they have the ability to "copy and paste" or "cut and paste" their genetic material to any other place in the genome. They may disrupt the coding region of a gene, or for example, activate neighbouring genes.

Long-range DNA methylation alterations Most studies on DNA methylation dynamics have largely focused on global loss of methylation. However, recent genome-wide approaches have revealed additional various types of large genomic blocks with specific DNA methylation patterns associated with gene regulatory activities, such as blocks that partially lose their default hypermethylated state, termed partially methylated domains (PMDs) (Lister et al. 2009), low-methylated regions (Stadler et al. 2011), highly methylated domains (HMDs), unmethylated regions (UMRs) (Burger et al. 2013), and long-range epigenetic activation and suppression regions (LRES) (Coolen et al. 2010; Saul A Bert et al. 2013).

Partially methylated domains (PMDs) are large continuous regions with low methylation, range from a few hundreds kb up to several Mbp in size, and cover a large portion of the genome (50–75%) (Lister et al. 2009; Stadler et al. 2011). They are linked to repressive chromatin (heterochromatic histone modification mark H3K27me3), gene silencing, and are gene-poor (Berman et al. 2011; Hon et al. 2012; Wen et al. 2009). PMDs are present in up to 75% of the genome in human and mouse cells irrespective of their tissue or cell origin (Salhab et al. 2018). PMDs have been described for a variety of cell types: adipocyte tissue (Lister et al. 2009), SH-SY5Y neuronal cells (Schroeder et al. 2011), and human cancers (Brinkman et al. 2019; Hovestadt, D. T. W. Jones, et al. 2014; Kasper Daniel Hansen et al. 2011; Timp et al. 2014). They have been linked to nuclear-lamina-associated domains (Berman et al. 2011), and in combination with broad histone marks are involved in late replication (Wanding Zhou et al. 2018; Salhab et al. 2018).

Smaller localized low-methylated regions (LMRs) are located in distal regulatory regions, and have been suggested to act as regulatory elements that define cellular identity (Lister et al. 2009; Stadler et al. 2011). Their average methylation is 30% and they contain a few hundred to a few thousand base pairs with less than 30 CpGs, and they do not colocalize with CpG islands. LMRs map to active histone modification marks, DNAse I hypersensitive regions, and usually contain transcription factor binding sites (Lister et al. 2009; Stadler et al. 2011). Their DNA methylation is inversely correlated with the activity of distal regulatory regions. A comparison of neuronal and stem-cell methylomes by study of Stadler et al. 2011 confirmed that DNA-binding factors are necessary and sufficient to create LMR, as cell-type-specific LMRs are occupied by cell-type-specific transcription factors. This study revealed a crosstalk between DNA-binding factors and local DNA methylation, where DNA binding factors

locally influence DNA methylation enabling the identification of active regulatory regions in a cell-type-specific manner. Apart from LMRs, the segmentation method of methylome described by Stadler et al. 2011 identified: unmethylated regions (UMRs, mostly CpG islands, more than 30 CpGs) and fully methylated regions (FMRs). UMRs correspond to proximal regulatory elements (unmethylated CpG islands) and most of the genome was classified as FMRs as the majority of CpGs in a genome are methylated. Finally, this study also showed an evidence that CTCF binding within CpG poor regions is not affected by methylation status of its binding site, but rather that binding itself initiates local demethylation (CTCF binding sites are not confined by DNA methylation).

A study by Burger et al. 2013 defined highly methylated domains (HMDs) that show contrasting chromatin signatures to PMDs (more heterochromatic and rather gene-poor regions). They observed cell-type-specific changes from PMD to HMD, and vice versa, occurring in genomic regions that contain genes functionally enriched for cell-type-specific properties. Unmethylated regions (UMRs) share characteristics with LMRs in terms of their length being relatively short (a few hundred to a few thousand basepairs), and low methylation patterns (below 50%), but they are located mostly on CpG islands, and correspond to proximal regulatory elements (Stadler et al. 2011).

Recently, new types of aberrations have been shown to exist - even larger epigenetic alterations covering many megabase regions that may encapsulate tens of genes. Long Range Epigenetic Silencing (LRES) and Long Range Epigenetic Activation (LREA) were both first discovered in prostate cancer in comparison to normal prostate epithelium (Frigola et al. 2006). It has been shown that LRES overlap with PMDs in colon, and overall, the genes within these two sets overlapped significantly (Berman et al. 2011). In normal cells, LRES are large, spanning many Mb, hypomethylated regions marked with H3 and H4 acetylation marks. In contrast, in cancer cells DNA methylation is observed, chromatin is compacted, hypoacetylation of H3 and H4, and H3K9me and H2K27ac. It makes LRES posses similar features to hypermethylated CpG islands, but they span clearer, enlarged blocks, thousands to millions bases. Within these long-range intervals occurs an exchange of repressive chromatin marks, from normal to cancer, such as gain of DNA methylation, and loss of H3K27me. In each case, genes are inactivated, even if for different reasons. Neighbouring to LRESes there are also large, whole megabase

spanning intervals regions of activation. LREAs are characterized by gain of active chromatin marks and loss of repressive marks that go with that activation (Saul A Bert et al. 2013).

Since these large regions that are epigenetically activated or inactivated together, contain a large number of different genes, function of each individual gene can't be assigned to these regions, but rather depends on alteration in the underlying nuclear architecture (Khoury, Achinger-Kawecka, Saul A. Bert, et al. 2020b). It's unlikely that all genes within a megabase or several megabase interval would be all tumor suppressors or oncogenes. It has been shown that nuclear architecture is the major factor - active regions loop into the centre of the nucleus and inactive regions loop out to the nuclear periphery. Studies on nuclear architecture influencing epigenetics are still ongoing, but nuclear disorganisation has been used to diagnose cancer for over a century by pathologists. After staining the cells and their nucleus, pathologist can see under the microscope that normal and cancer cells have different nuclear sizes, nuclear shapes, ploidy (number of chromosome copies), slightly denser staining regions (for heterochromatin) and slightly lighter staining regions (for euchromatin) (Zink, Fischer, and Nickerson 2004). Aberrant nuclear architecture is one of the hallmarks of cancer, is a well established method to diagnose neoplastic tissues, and hopefully, in the future together with epigenetic aberrations will help us in understanding cancer progression.

1.2.4.3 Histone modification aberrations in cancer

Histone modifications, similarly to DNA methylation, go through local and global alterations in cancer. Moreover, there is a strong relationship between histone modifications, and cancer features like genomic stability, and DNA repair. Genomic instability is increased when histone modifications involved in a chromosome structure are altered. Particularly, histone modifications or some histone variants alterations that are found on centromeres might lead to an imbalanced number of chromosomes in each daughter cell, because they are positioned at the point of attachment to the mitotic spindle. Histone modifications have a strong relationship to DNA repair due to the fact that mutations in cancer occur most commonly in regions of heterochromatin (that in normal cells was originally euchromatin) (Zheng et al. 2014). DNA repair mechanisms in heterochromatin are slower and less efficient in comparison to

euchromatin, because they include additional steps of unwinding DNA heterochromatin and re-compacting DNA after the repair.

Cancerous cells exhibit global alterations in histone modifications, mainly changes associated with overall activation - decreased monoacetylated and trimethylated forms of histone H4 (decreased H4ac and H4me3) (Fraga et al. 2005). Additionally, in contrast to normal cells, in which expressed genes are marked with H3K4me3 and heavy H3ac and H4ac marks, cancer cells exhibit an accumulation of epigenetically silencing marks. These genes in cancerous cells are silenced by hypermethylation at their CpG island regions, and histones in their neighbourhood show a different make up: decrease of active marks: H3ac, H4ac, H3k4me and increase of inactive marks: H3K9me and H3K27me.

Histone modification changes, just like DNA methylation, might be tumor-specific. There is growing evidence which suggests that histone-modifying enzymes are found deregulated in human cancers. An extensive analysis of expression patterns of histone-modifying enzymes was able to discriminate between tumor samples and their normal counterparts, and cluster the tumor samples according to cell type (Özdağ et al. 2006). It indicates that changes in the expression of histone-modifying enzymes have important and tumor-specific roles in cancer development.

However, tumor-type specific histone modifications are not as technically stable as DNA methylation marks and as simple to assess in the laboratory. Hence, DNA methyltransferase inhibitors (DNMTi) were the first FDA-approved epi-drugs to reach the market in back in 2004, in comparison to the first histone deacetylase inhibitors (HDACi) in 2006 (Berdasco and Esteller 2019). Combination of DNMTi and HDACi to obtain sustainable chromatin modulation, and stable re-activation of silenced genes, has been widely explored in the treatment of leukemia and refractory advanced non-small cell lung cancer (Prebet et al. 2014; Juergens et al. 2011). So far, clinical results for the combination of DNMTi and HDACi has been controversial, mainly because of the limited sample size of the cohorts and side-effects (Peter A. Jones, J.-P. J. Issa, and Baylin 2016).

1.3 Genomic assays and data types

Next-generation sequencing (NGS), also known as high-throughput sequencing, is the catch-all term used to describe a number of different sequencing technologies to determine the sequence of nucleotides in DNA. These technologies allow for sequencing of DNA and RNA, and are typically characterized by being highly scalable. In contrast to the previously used Sanger sequencing (or first-generation sequencing) (Sanger and Coulson 1975), NGS is much quicker, cheaper, and allows the entire genome to be sequenced at once. It is accomplished by fragmenting the genome into small pieces, randomly sampling for a fragment, and sequencing it using one of the variety of technologies, such as those described in the next sections. Sequencing an entire genome is possible because multiple fragments are sequenced at once (giving it the name "massively parallel" sequencing) in an automated process. NGS revolutionised the study of genomics and molecular biology, and imposed increasing demands on statistical methods and bioinformatic tools for the management of the huge amounts of data generated by these technologies.

Data processing and analysis constitute substantial work behind this thesis. Therefore, in this section I will present a description of the type of data used and an overview on its processing techniques.

1.3.1 Bisulfite sequencing for detection of DNA methylation

To date, the gold standard in methylome mapping has been bisulfite sequencing. In this method, DNA is chemically treated with sodium bisulfite, which results in the conversion of unmethylated cytosines (C) to uracils (U), and after PCR the resulting uracils are ultimately sequenced as thymines (T), whilst methylated cytosines remain unchanged (Frommer et al. 1992)(see Figure 1.3 for schematic representations of cytosine, methylated cytosine and thymine). Bisulfite treatment in combination with NGS is termed Bisulfite-treated DNA sequencing (Bisulfite-seq or WGBS). Bisulfite-seq is considered the gold standard for assaying DNA methylation due to its global coverage at a single-base resolution (for more details see techniques for profiling genome-wide DNA methylation I reviewed in (Wreczycka, Gosdschan, et al. 2017)). To perform Bisulfite-seq, the genomic DNA is first randomly fragmented to the desired size (circa 200 bp), converted into a sequencing library by ligation to adapters, treated with bisulfite and

amplified using PCR. A precise recall of cytosine methylation requires not only sufficient sequencing depth, but also strongly depends on the quality of bisulfite conversion and library amplification. The benefit of this shotgun approach is that it typically reaches coverage of over 90% of the CpGs and non-CpGs.

Despite its clear advantages, Bisulfite-seq remains the most expensive technique, and standard library preparation requires relatively large quantities of DNA (100ng - 5μg); as such, it is usually not applied to large numbers of samples (Stirzaker et al. 2014). Additionally, chemical bisulfite reaction damages and degrades DNA, resulting in fragmentation and loss. Bisulfite libraries demonstrate some GC bias and are enriched for methylated regions. There have been efforts to overcome these limitations (Vaisvila et al. 2019), but nevertheless, to achieve high sensitivity in detecting methylation differences between samples, high sequencing depth is required, which leads to significant increase in sequencing cost.

Bisulfite-seq data processing Bisulfite sequencing applies routine sequencing methods on bisulfite-treated genomic DNA to determine methylation status at CpG dinucleotides. The methodologies to analyze bisulfite-treated DNA can be generally divided into strategies based on methylation-specific PCR (MSP), and strategies employing PCR performed under non-methylation-specific conditions. Microarray-based methods also use PCR based on non-methylation-specific conditions (Laird 2003).

After converting raw data as images from a sequencer, detected and sequenced DNA fragments from a Bisulfite-seq experiment are aligned to a reference genome which enables genome-wide mapping of a whole methylome on the genome at base pair resolution. Bisulfite-seq analysis steps focus on counting the number of C to T conversions and based on that, quantifying the methylation proportion per base. This is simply done by identifying C-to-T conversions in the aligned reads and dividing the number of Cs by the sum of Ts and Cs for each cytosine in the genome (methylation calling) (Figure 1.7).

Reliable quantification of Bisulfite-seq depends on quality control before alignment, the alignment methods and post-alignment quality control. Since the quality of base-calling is inconsistent, it can change between sequencing runs, and within the same read, it is important to inspect the base quality (which represents the level of confidence in the base calls). Miscalled

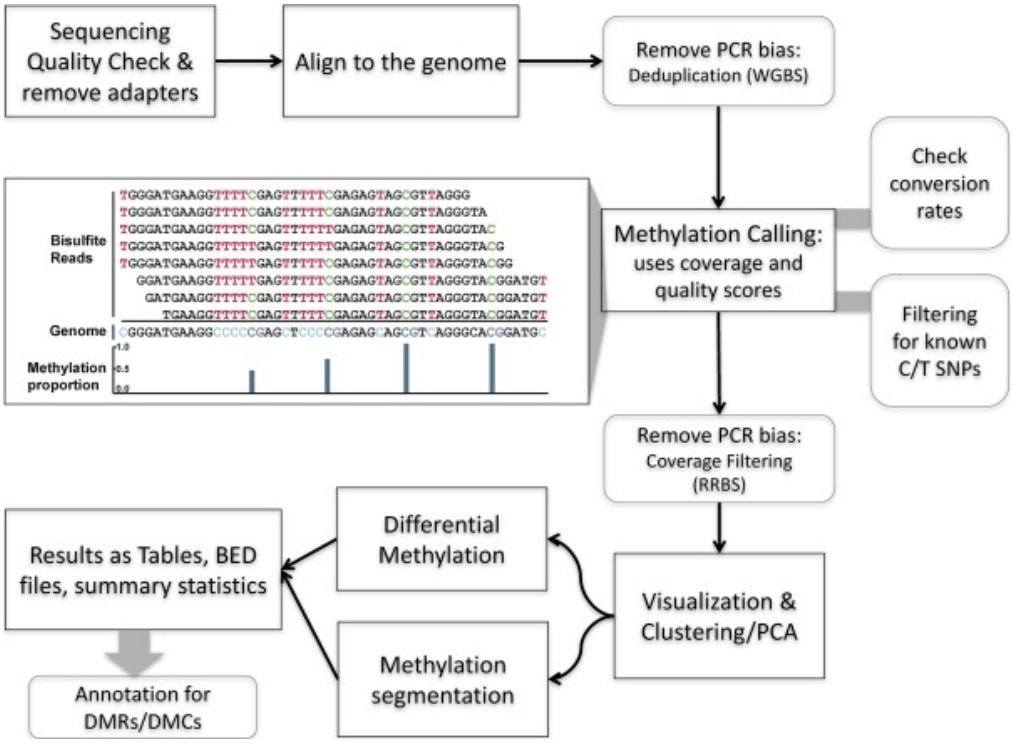


Figure 1.7: Workflow for analysis of DNA methylation using data from bisulfite sequencing experiments. Figure reprinted from Wreczycka, Gosdschan, et al. 2017.

bases can be counted as C-T conversions erroneously, and such errors should be avoided if possible. Furthermore, sometimes adapters can be sequenced, and if not properly removed, they could either lower the alignment rates or cause false C-T conversions. After the alignment and methylation calling, there is still a need for further quality control. During the end repair step following the fragmentation, unmethylated Cs are introduced at the ends of the DNA fragments (Bock 2012). It is recommended to trim low quality bases on sequence ends and removing adapters to minimize issues with false C-T conversions and to increase alignment rates.

Once pre-alignment quality control and processing is done, the next step is the alignment where potential C-T conversions should be handled. The Bisulfite-seq alignment methods mostly rely on modifications of known short-read alignment methods: in silico C-T conversion of reads and genomes (Krueger and Andrews 2011; Xi and W. Li 2009), specific score matrix that can tolerate C-T mismatches (Frith, Mori, and Asai 2012), and masking Ts in the reads and matches them to genomic Cs (Krueger and Andrews 2011; Xi and W. Li 2009).

It has been shown that the majority of CpGs with high inter-population differences contain

common genomic SNPs (minor allele frequency > 0.01) (Daca-Roszak et al. 2015). It is advised, in order to ensure more reliable interpretation of the data, to remove known C/T SNPs which can interfere with methylation calls.

The last post-alignment quality procedure addresses PCR bias. A simple way could be to remove reads that align to the exact same genomic position on the same strand (deduplication). Post-mapping and deduplication analysis steps include tabulation of the fractional methylation of CpG sites, the segmentation of genomic methylation patterns across the genome, the selection of differentially methylated sites between pairs of treatments, and annotating them into genomic regions (such as introns, promoters, enhancers and other regions of interest).

Methods for differentially methylated cytosines detection Once methylation proportions per base are obtained, generally, the dynamics of methylation profiles are considered next. The most common task in omics data is feature selection. In terms of DNA methylation it is a selection of cytosines or groups of neighbouring cytosines which are important with respect to the DNA methylation differences usually between disease and control samples. For example, DNA methylation varying in CpGs marking specific cell types exhibits patterns that correlate with underlying cell fractions and can act as biomarkers, whereas those driven by genetic variants might not (Schübeler 2015).

A popular supervised feature selection strategy is to select CpGs for which there is a significant difference in the average between phenotypes, defining differentially methylated cytosines in a test sample relative to a control (DMCs). In simple comparisons between such pairs of samples (i.e. test and control), methods such as Fisher's Exact Test (implemented in methylKit (Akalin, Kormaksson, et al. 2012) and RnBeads (Assenov et al. 2014)) can be applied when there are no replicates for test and control cases. There are also methods based on hidden Markov models (HMMs) such as ComMet, included in the Bisulfighter methylation analysis suite (Saito and Mituyama 2015) or the MethPipe software package (Song et al. 2013). These tools are sufficient to compare one test and one control sample at a time; if there are replicates, they can be pooled within groups to a single sample per group (Akalin, Kormaksson, et al. 2012). This strategy, however, does not take into account biological variability between replicates.

Regression-based methods are generally used to model methylation levels in relation to the sample groups and variation between replicates. Differences between currently available regression methods result from the choice of distribution to model the data on and the variation associated with it. In the simplest case, linear regression can be used to model methylation per given CpG or loci across sample groups. The model fits regression coefficients to model the expected methylation proportion values for each CpG site across sample groups. Hence, the null hypothesis of the model coefficients being zero could be tested using t-statistics. Such models are available in the limma package (Ritchie et al. 2015). limma was initially developed for the detection of differential gene expression in microarray data, but it is also used for methylation data. It is the default method applied in RnBeads. It uses moderated t-statistics in which standard errors have been moderated across loci, i.e. shrunk towards a common value using the Empirical Bayes method. Another method that relies on linear regression and t-tests is the BSmooth method (Kasper D Hansen, Benjamin Langmead, and Irizarry 2012). The main difference is that BSmooth applies a local-likelihood smoother to smooth DNA methylation across CpGs within genomic windows, assumes that the data follows a binomial distribution and parameters are estimated by fitting a linear model inside windows. It calculates signal-to-noise ratio statistics similar to t-test together with Empirical Bayes approach to test the difference for each CpG.

However, linear regression based methods might produce fitted methylation levels outside the range [0,1] unless the values are transformed before regression. An alternative is logistic regression, which can deal with data strictly bounded between 0 and 1 and with non-constant variance, such as methylation proportion/fraction values.

In the logistic regression, it is assumed that fitted values have variation $np(1-p)$, where p is the fitted methylation proportion for a given sample and n is the read coverage. More specifically, at a given base, the methylation proportion P_i is modelled, for sample $i = 1, ..., n$ (where n is the number of biological samples) through the logistic regression model:

$$log(\frac{P_i}{1-P_i}) = \beta_0 + \beta_1 Treatment_i \tag{1.3.1.1}$$

where $Treatment_i$ denotes the treatment indicator for sample i, $Treatment_i = 1$ if sample is in the treatment group and $Treatment_i = 0$ if sample i is in the control group. The

parameter β_0 denotes the log odds of the control group and β_1 the logodds ratio between the treatment and control group. Therefore, independent tests for all the bases of interest are against the null hypothesis $H0 : \beta_1 = 0$. If the null hypothesis is rejected it implies that the logodds (and hence the methylation proportions) are different between the treatment and the control group and the base would subsequently be classified as a differentially methylated cytosine. However, if the null hypothesis is not rejected it implies no statistically significant difference in methylation between the two groups.

If the observed variance is larger or smaller than assumed by the model, one speaks of under- or overdispersion. This over/under-dispersion can be corrected by calculating a scaling factor and using that factor to adjust the variance estimates as in $np(1-p)s$, where s is the scaling factor. MethylKit can apply logistic regression to test the methylation difference with or without the overdispersion correction. In this case, Chi-square or F-test can be used to compare the difference in the deviances of the null model and the alternative model. The null model assumes there is no relationship between sample groups and methylation, and the alternative model assumes that there is a relationship where sample groups are predictive of methylation values for a given CpG or region for which the model is constructed.

More complex regression models use beta binomial distribution and are particularly useful for better modeling the variance. Similar to logistic regression, their observation follows binomial distribution (number of reads), but methylation proportion itself can vary across samples, according to a beta distribution. It can deal with fitting values in [0,1] range and performs better when there is greater variance than expected by the simple logistic model. In essence, these models have a different way of calculating a scaling factor when there is overdispersion in the model. Further enhancements are made to these models by using the Empirical Bayes methods that can better estimate hyperparameters of beta distribution (variance-related parameters) by borrowing information between loci or regions within the genome to aid with inference about each individual loci or region. Some of the tools that rely on beta-binomial or beta model are as follows: MOABS (Sun et al. 2014) and DSS (Feng, Conneely, and Wu 2014), RADMeth (Dolzhenko and A. D. Smith 2014), BiSeq (Hebestreit, Dugas, and Klein 2013) and methylSig (Y. Park et al. 2014).

Our comparison of differential methylation methods in a review article (Wreczycka, Gosdschan, et al. 2017) revealed that the performance of different methods are comparable. One can choose particular methods based on the overall goal of their research. The methods that are stringent and limit the false positive rates are good for subsequent validation studies (DSS, limma, BSmooth, methylKit with F-test and overdispersion correction), however these methods sacrifice sensitivity (true positive rate) for the sake of reducing false positives. A very relaxed method, such as the default methylKit method, has the best accuracy overall but also highest false positive rate. A good alternative to stringent and relaxed methods is Chi-square test after overdispersion correction implemented in methylKit. This method has high sensitivity without sacrificing too much for specificity.

Methods for differentially methylated region detection Differential methylation can also be called at the regional level and there are a number of reasons why identifying differentially methylated regions (DMRs) is desirable. First, due to the activity of DNA methyltransferases and other enzymes modifying the epigenome, DNA methylation scales up to approximately 500 bp and beyond (Guo et al. 2017). Calling DMRs removes some of the spatial redundancy, helping to reduce the dimensionality of the data. Second, calling differential methylation at the regional level may offer increased robustness, especially in the context of limited coverage of Bisulfite-seq data (Libertini et al. 2016). Finally, DNA methylation alterations that extend to the regional level are thought to be more functionally important than alterations that affect only isolated sites.

Some statistical algorithms for calling DMRs are designed to detect DMRs via aggregating DMCs together within a predefined regions, such as CpG islands or CpG shores. RADmeth (Dolzhenko and A. D. Smith 2014), and eDMR (S. Li et al. 2013) group p-values of adjacent CpGs and produce differentially methylated regions based on distance between differential CpGs, and combination of their p-values using weighted Z-test. DSS sets some thresholds on the p-values, number of CpG sites and length of regions before aggregation. Similarly, BSmooth defines DMRs by taking consecutive CpGs and cutoff based on the marginal empirical distribution of t and DMRs are ranked by sum of t-statistics in each CpG. BiSeq, on the other hand, first agglomerates CpG sites into clusters and smoothes methylation within clusters,

uses beta regression and Wald test to examine a group effect between control and test samples (with maximum likelihood for bias reduction). Apart from the various ways of clustering nearby CpGs or DMCs, many other methods rely on HMMs or other segmentation methods to segment the differential CpGs into hypo- and hyper-methylated regions and combine them to DMRs, such as MOABS, Methpipe, ComMet and methylKit. Other methods define DMRs directly based on pre-defined windows. When input for functions for differential methylation calling are regions, then data is summarized per region. The regions can be either predefined (such as regions with biological meaning like CpG islands) or user-defined with criteria like fixed region length for tiling windows that cover the whole genome, fixed numbers of significant adjacent CpG sites and smoothed estimated effect sizes.

1.3.2 ChIP-seq for detection of DNA binding proteins

Chromatin immunoprecipitation (ChIP) is the most direct way to identify the binding sites of a single DNA-binding protein or the locations of modified histones in the nucleus. It includes precipitation (IP) of a protein antigen out of solution using an antibody that specifically binds to that protein. ChIP has been used since 1998 (Solomon, Larsen, and Varshavsky 1988), later in combination with hybridization arrays (such as ChIP-on-chip), and coupled with NGS (ChIP-seq) in 2007 (Johnson et al. 2007; Barski et al. 2007; Robertson et al. 2007; Mikkelsen et al. 2007). ChIP-seq allows single base-pair resolution, greater coverage, larger dynamic range and, in general, significantly improved data in comparision to previous techniques (P. J. Park 2009).

ChIP-seq is the most common genome-wide assay to determine transcription factor binding sites locations (Johnson et al. 2007; Robertson et al. 2007; Mikkelsen et al. 2007). Briefly, the conventional method is as follows: DNA and its bound proteins are cross-linked by treating cells with formaldehyde and the chromatin is sheared by sonication or endonuclease digestion into small fragments, which are generally in the 250-600 bp range. Then, protein-specific antibodies are used to immunoprecipitate the DNA-protein complex. Finally, the crosslinks are reversed, and purified DNA can be sequenced on any of the next-generation sequencing platforms (Figure 1.8A).

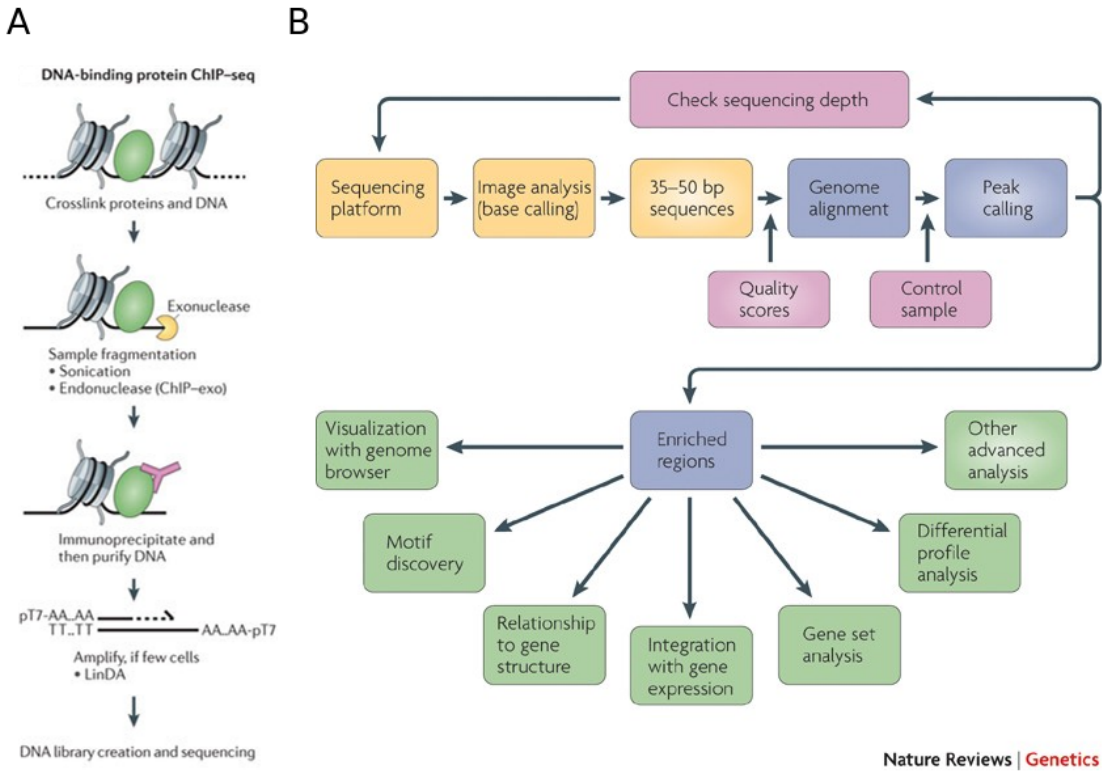


Figure 1.8: (A) Schematic illustration of the steps in a typical ChIP-seq protocol. Figure reprinted with modifications from Furey 2012 (B) Overview of ChIP-seq computational analysis. Figure reprinted from P. J. Park 2009.

ChIP-seq data processing DNA fragments in the library from ChIP-seq are usually 250-600 bp long, but they are sequenced only for the first 30-50 bp from the 5' end of the DNA strands. Therefore, the alignment of these fragments to the genome results in two peaks (one from each strand) that flank the binding location of the protein of interest. This strand-specific pattern can be used for the optimal detection of enriched regions. To create an approximate distribution of all fragments, each tag location can be extended by an estimated fragment size in the appropriate orientation and the number of fragments can be counted at each position. Consequently, the data can be represented as a signal over genomic coordinates flanking proteins of interest, and each base is associated with binding strength (a normalized count of the reads covering the location). Peak calling, using data from the ChIP profile and a control profile, generates a list of enriched genomic regions that are ordered by false discovery rate as a statistical measure. Subsequently, the profiles of enriched regions are viewed with a browser and various advanced analyses are performed (Figure 1.8B). As with many high-throughput techniques, ChIP-seq is

also susceptible to technical and biological biases. Thus, designing control experiments is an important part of ChIP-seq experimental design (P. J. Park 2009; Kidder, Hu, and K. Zhao 2011).

First, biases arise during genome fragmentation step, open chromatin regions are easier to shear than closed chromatin regions, and thus it leads to higher background signals (Leonid Teytelman et al. 2009). Consequently, IP generates more complexes from the open chromatin regions, resulting in more sequencing reads. To correct this fragmentation bias, the fragmented genomes are divided into two portions. One portion goes through the IP step and then the sequencing step, whereas the other portion is sequenced directly to serve as input control. This direct sequencing result contains the shearing bias of fragmentation, and thus can be used to normalize the sequencing results from the IP protocol (Kharchenko, Tolstorukov, and P. J. Park 2008).

Second, uneven regulatory binding in the genome may result in bias during the IP step due to antibody cross-reactivity. Although the antibody in IP binds specifically to its antigens, i.e. the target TFs, it can also have competing high affinity toward a different antigen, and bind nonspecifically to other proteins, especially closely related family members. To control for this bias, a mock IP can be generated using the IP protocol, with the mock IP lacking specific antibody-antigen interactions. The mock IP either uses an antibody that can't recognize the TF of interest, e.g., IgG, or the TF is not tagged with the epitope for the antibody used in the IP, e.g., GFP. Therefore, the mock IP control mimics only the nonspecific interactions in the IP. In addition to the nonspecific interactions, the mock IP also controls for sonication bias (P. J. Park 2009; Kidder, Hu, and K. Zhao 2011; Landt et al. 2012). However, most IgG antibodies are not obtained from true preimmune serum from the same animal in which the specific antibody was raised; and IgG antibodies usually immunoprecipitate much less DNA than specific antibodies do.

Another point is that the bulk sequencing methods disregard cell-type specific binding patterns, which are important for cell identity but this information is averaged out in bulk datasets, such as bulk ChIP-seq.

In Chapter 3, I will describe evidence that suggests that false positive peaks in ChIP-seq data may be substantial, discuss technical and biological reasons behind them, and I will cover suggestions how to deal with these undesirable spurious sites.

1.4 Computational integration of DNA methylation with other types of genomic data

There are many factors that limit the interpretability of the DNA methylation data generated in a typical epigenome-wide association study. Cell type heterogeneity, genetic variation, alteration to DNA methylation levels caused by phenotype itself can be puzzling (Wahl et al. 2016). DNA methylation is a predictor of gene expression, but it can be also outperformed by chromatin state information encoded in histone modification marks (Karlic et al. 2010; Ernst et al. 2011). Thus, when multiple NGS platforms are used on the same biological samples, such as including gene expression matched to the same samples for which DNA methylation is available, the resulting multi-modal datasets allow to probe biological processes from multiple aspects. This is referred to as multi-platform or multi-omics data integration, and is common in the field of cancer research (Hoadley et al. 2014).

As mentioned in sections 1.2.3.1 and 1.2.4.2, the relationship between DNA methylation and gene expression is complex. It is a challenging task, because even if DNA methylation profile of the gene itself is known, one should include DNA methylation levels at distal regulatory elements, and most of enhancers loop over their nearest genes to target genes much further away, causing uncertainty as to which genes an enhancer may regulate. For example, in the context of cancer, DNA methylation patterns at distal regulatory regions account for more of the intra-tumor expression variation than DNA methylation at promoters (Aran, Sabato, and Hellman 2013) and enhancers are among the most cell-type specific regions (Ziller et al. 2013). Some studies have reported that gene-body methylation levels are more predictive than the more classical TSS region (X. Yang et al. 2014; Peter A. Jones 2012). These meta-analysis are consistent with other studies demonstrating that it is the TSS, first exon and 3′ end that show the strongest monotonic associations (Libertini et al. 2016; Jiao, Widschwendter, and Teschendorff 2014; Brenet et al. 2011). Additionally, methylated promoters are generally associated with gene silencing, whereas unmethylated promoters are associates with both transcribed and untranscribed states (Walsh and Bestor 1999).

Other approaches of integration of DNA methylation and gene expression is not to assign a unique DNA methylation value to a gene, but instead, to use information about a DNA methylation signature over a gene (and beyond) as a predictor of gene expression. One of such

methods uses differential methylation profile in the vicinity of promoters of genes (typically centered on a 10–30 kb window around the TSS of genes), and quantify the similarity of these gene-based profiles using specific distance metrics. Then, it uses an unsupervised clustering technique to arrange the signatures according to their shapes, and identify which clusters of genes exhibit statistically significant changes in expression (VanderKraats et al. 2013). A supervised version of this approach uses a random-forest classifier, which has been shown to improve the prediction of gene expression, highlighting the importance of the TSS and 3′ end as the most predictive gene regions (Schlosberg, VanderKraats, and Edwards 2017).

1.4.1 System-level integration of DNA methylation

A powerful system-level integrative approach is to exploit association of DNA methylation at regulatory regions with transcription factor binding to infer patterns of regulatory activity. Although DNA methylation at regulatory sites has been traditionally viewed as dictating transcription factor affinity, the opposite has also been observed (Stadler et al. 2011). Moreover, there are classes of TFs (for example POU and NFAT families) that prefer binding to methylated sequences (Yin et al. 2017).

First generation of methods to predict transcription factors involved in particular transcriptional response utilize Position Probability Matrices (PWMs) of TFs motifs to predict binding sites in enhancers in the vicinity of regulated genes (Dhaeseleer 2006). These methods range from simple PWM matching (Kel et al. 2003; G. Tan and Lenhard 2016; Grant, Bailey, and Noble 2011; Gama-Castro et al. 2016; Heinz et al. 2010) to modeling-based approaches (Pique-Regi et al. 2011; Zhong, He, and Bar-Joseph 2013; Jankowski, Tiuryn, and Prabhakar 2016; X. Chen et al. 2017). Another approach is transcription factor enrichment analysis (TFEA)(Rubin et al. 2020) that draws inspiration from Gene Set Enrichment Analysis (GSEA)(Subramanian et al. 2005), and detects positional motif enrichment within a list of ranked regions of interest, such as enhancers. Still, the identification of key transcription factors in regulatory networks by only motif enrichment methods is generally associated with a high false-positive rate with too many candidate factors to investigate.

It gave rise to novel system epigenomics methods for inferring TF binding activity. Some of them exploit inverse correlation between DNA methylation and regulatory element activity

to infer disrupted regulatory networks associated with a disease risk (Teschendorff et al. 2015; Yuan et al. 2015). For example, enhancer linking by methylation and expression relationships (ELMER) algorithm first identifies enhancers, whose DNA methylation levels are altered in cancer. Then, it uses the matched mRNA expression of putative gene targets to construct cancer-specific enhancer-gene networks (Silva et al. 2018). ELMER uses TF-binding motif enrichment analysis for correlated enhancers and mRNA expression of enriched TFs to identify cancer-specific activated TFs. Other similar approaches include tracing enhancer networks using epigenetic traits (TENET) (Rhie et al. 2016) and RegNetDriver (Dhingra Priyanka et al. 2017). TENET refines ELMER's method by identifying tissue-specific enhancer-gene links, and RegNetDriver constructs tissue regulatory networks by intergrating cell-type-specific open chromatin data from publicly available databases such as ENCODE (T. E. P. Consortium 2012) and RMEC (Kundaje et al. 2015). Mapping disease-associated molecular alterations in that tissue onto the corresponding tissue-specific network can reveal which TFs are deregulated in disease. ELMER identified RUNX1 as a key TF determining clinical outcome in kidney cancer (Silva et al. 2018), and RegNetDriver revealed that most of the functional alterations of TFs in prostate cancer were associated with DNA methylation changes but that TF hubs were preferentially altered at the copy-number level (Dhingra Priyanka et al. 2017). By integrating a gene function network such as a protein-protein interaction (PPI) network, the functional epigenetic module (FEM) tool can identify differentially expressed gene modules (Jiao, Widschwendter, and Teschendorff 2014). A SMITE algorithm extends that approach by calculating significance based modules, based on transcriptome and epigenome (Wijetunga et al. 2017). Although these tools use correlations between enhancer DNA methylation and mRNA target expression to find the more likely targets, these correlations are themselves subject to potential confounders such as cell-type heterogeneity.

In several studies, linear regression-based approach have been applied, in which the sequence information is used to model gene expression or chromatin marks, and to learn the TFs that play a major role in gene regulation (T. F. Consortium and Center 2009; Piotr J Balwierz et al. 2014; Osmanbeyoglu et al. 2014). The multivariate linear model represents the signal expression of data as a linear combination of motif scores (motif occurrences) and their influential weights ("motif activities") (Figure 1.9A-C). Due to the large number of genes the weights are generally

highly significant and they can be interpreted as a proxy for TFs activities. The error or noise term in the model represents all signal that cannot be explained by the model, i.e. the linear combination of the motif activity scores. More recent approaches use linear regression with L2-regularization (Ridge Regression) (Madsen et al. 2018), Bayesian Ridge Regression (Piotr J Balwierz et al. 2014) or Bayesian Linear Mixed Models (Lederer et al. 2020). Bayesian Linear Mixed Models approach relaxes the rigid independence between samples, however on real biological data, it does not perform better than the Bayesian Ridge Regression method due to the fact that large fraction of the gene expression signal ends up in the noise term of the model (Lederer et al. 2020). These linear models can be extended to any signal, and in context of DNA methylation since it's a signal restricted to the $[0, 1]$ range, it requires log2 transformation, such as: $log_2((Y + 1)/(100 - Y + 1))$. As a consequence of using DNA methylation as response variable, opposite to the case of gene expression or open chromatin marks, the positive weights in the model, or motif activities, are associated with silencing (Figure 1.9D). Next steps in the analysis might include filtering TFs that are expressed in desirable conditions, and TFs that expression is correlated with changes of motif activities (Figure 1.9E), providing lists of enriched Gene Ontology categories of TFs, and creating PPI networks of TFs (Figure 1.9F).

Another set of integrative algorithms are tailored for integrating DNA methylation data that are generated in conjunction with other data types from the same samples: for instance, this may include sequence information, mutations, copy-number variants, mRNA, microRNAs and protein expression. Performing simultaneous inference using all data types together offers more powerful and unbiased framework to reveal system-level associations and extract novel biological insights. These categories of methods are particularly viable and flexible because any data can be ingested as long as they can be represented as a generic matrix of values. As an example, a joint NMF algorithm was applied to the matched DNA methylation, mRNA and miRNA expression data sets for ovarian cancer from The Cancer Genome Atlas (TCGA), revealing novel perturbed pathways (S. Zhang et al. 2012). Other matrix factorization methods include iCluster (R. Shen, Olshen, and Ladanyi 2009) that performed integrative DNA methylation and mRNA analysis of oestrogen receptor (ER)+ breast cancer, and MOFA (Argelaguet, Velten, et al. 2018; Argelaguet, Arnol, et al. 2020). Another powerful method based on deep learning is Janggu (Kopp, Monti, et al. 2020) and Multi-omics Autoencoder Integration (maui) (Ronen,

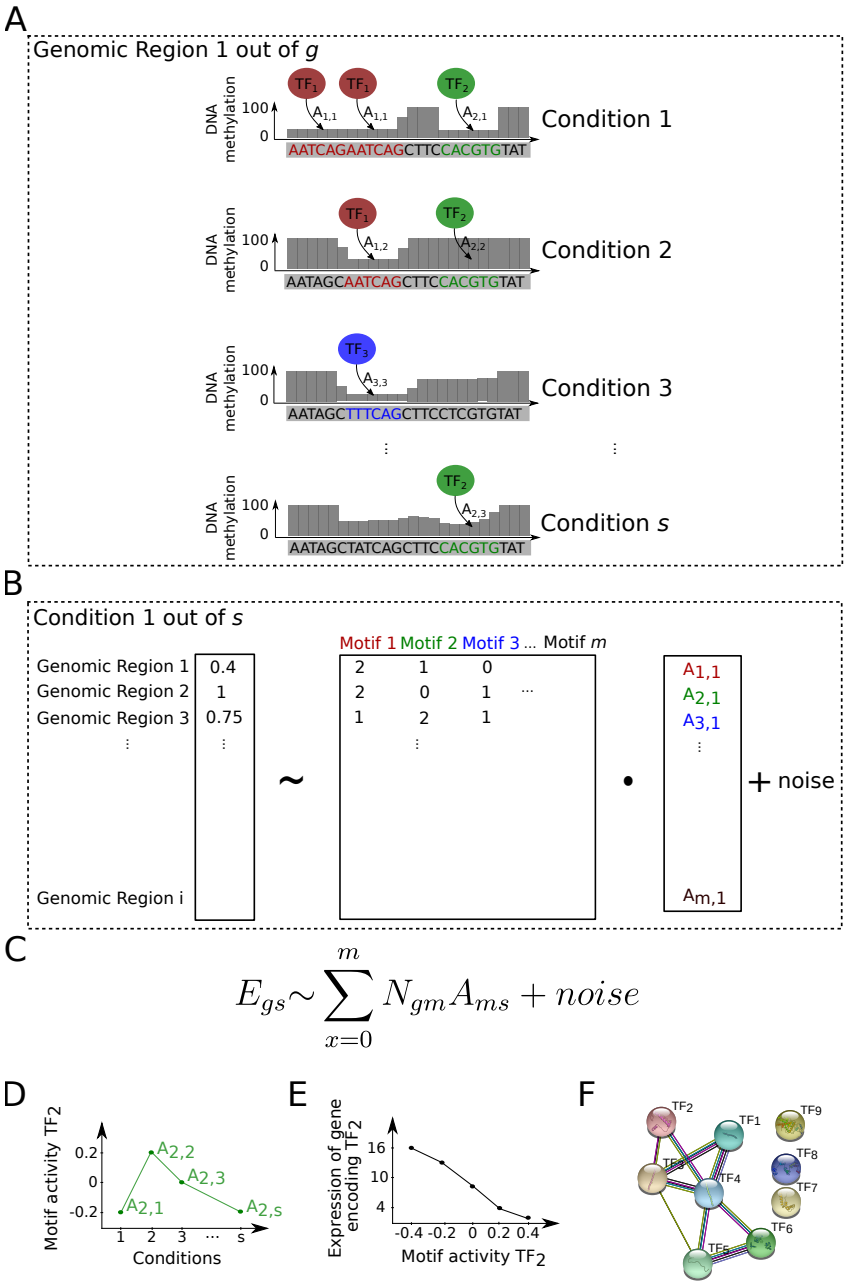


Figure 1.9: (A) Schematic representation of transcription factors binding to their DNA motifs in a genomic region 1, and in conditions 1 to s. On Y axis is depicted DNA methylation percentage. (B) Schematic representation of the model for condition 1 used in (Piotr J Balwierz et al. 2014; Madsen et al. 2018; Lederer et al. 2020). (C) The model is used to explain the signal levels E_{gs} in terms of bindings sites N_{gm} and unknown motif activities A_{ms}, which are inferred by the model. Results might include (D) creating motif activity profiles, (E) correlation of motif activity profiles with gene expression, (F) based on result TFs creating PPI networks using STRING database (Szklarczyk et al. 2018).

Hayat, and Akalin 2019) that gives deep learning's applicability in deconvoluting non-linear relationships in large datasets using a flexible model architecture.

The remainder of this thesis In the chapters that follow, I will describe in depth new insights in gene regulation by DNA methylation in neuroblastoma tumor cells and healthy cells. In Chapter 2, I will introduce novel genome-wide DNA methylation analysis using Bisulfite-seq in 24 neuroblastoma patient-derived samples. First, I will use differential methylation correlated with gene expression. Next, I will focus on a system-level integration approach based on Bayesian Ridge Regression (Piotr J Balwierz et al. 2014) of neuroblastoma DNA methylation patterns with sequence information using DNA motif occurrences, H3K27ac chromatin modifications, and matched gene expression in order to reveal the impact of DNA methylation disturbances in high-risk neuroblastoma to transcription factor networks. In Chapter 3, I will expand the knowledge about previously found false positive signals in ChIP-seq experiments, and show an evidence that they are associated with DNA-RNA hybrids, so called R-loops, in unmethylated functional genomic regions.

2

Genome-wide DNA methylation analysis in neuroblastoma

Methylation of cytosines has increasingly attracted attention as a potential biomarker, and is now acknowledged as a universal feature of tumorigenesis (Moss et al. 2018). It has emerged as a hallmark of many cancer types including glioblastoma (Klughammer, Kiesel, Roetzer, Fortelny, Nemc, Nenning, Furtner, Nathan C. Sheffield, et al. 2018b), medulloblastoma (Hovestadt, D. T. W. Jones, et al. 2014), myelomonocytic leukemia (Stieglitz et al. 2017), colorectal cancer (M. A. Kerachian, Javadmanesh, et al. 2020) and nervous system tumors (Capper et al. 2018).

In this chapter, I describe computational analysis of the methylome of a paediatric cancer neuroblastoma. It's a clinically heterogeneous cancer, and the diversity of its high-risk subgroups is not yet fully understood. On the molecular level, since the last four decades, the best established high-risk neuroblastoma biomarker has been a MYCN oncogene amplification (G. Brodeur et al. 1984), and with the bulk next-generation sequencing era, handful of other genes (Peifer et al. 2015; Ackermann et al. 2014; Mossé et al. 2008; Zeineldin et al. 2020), structural variant changes (Depuydt et al. 2018; Richard P Koche et al. 2020), and lineage specific transcription factors increasingly attracted attention in the field (Boeva et al. 2017; Groningen et al. 2017; Durbin et al. 2018; Decaesteker et al. 2018). Yet, they only partially mirror the heterogeneity of neuroblastoma. Neuroblastoma displays very few somatic mutations (Gröbner et al. 2018; Lawrence et al. 2013), which pinpoints to a possible involvement of

epigenetic changes in its pathogenesis (Jubierre et al. 2018).

Inline with these observations, in this chapter, I describe a novel approach to explain neuroblastoma diversity - a genome-wide and single-nucleotide neuroblastoma methylome analysis that revealed tumor-specific DNA methylation alterations which affect transcription factor regulatory networks, and consequently might influence tumor formation and progression.

2.1 Introduction

Neuroblastoma (NB) is a tumor of early childhood and is the most common malignancy diagnosed in the first year of life, with 25–50 cases per million individuals (Stiller and Parkin 1992). This tumor type commonly emerge near the adrenal gland and are characterized by high heterogeneity ranging from spontaneous regression to rapid progression (Matthay et al. 2016). The outcome for high-risk NB cases is poor despite intensive multi-modal treatment, with long-term survival less than 50% (Molenaar et al. 2012; Pugh et al. 2013). Genome-wide studies revealed several recurrent molecular abnormalities in primary NB cases, the best characterized is amplification of MYCN oncogene (Schwab et al. 1983; Richard P. Koche et al. 2019), and other include upstream TERT rearrangements (Peifer et al. 2015; Valentijn et al. 2015; Brady et al. 2020), mutations of ALK and ATRX (Zeineldin et al. 2020), and deletions of chromosomal regions in 1p, 3p and 11q, and gain of 17q (Depuydt et al. 2018; Richard P Koche et al. 2020). MYCN amplification is a defining feature of high-risk NB (Weiss 1997) and has been used as a risk factor that is associated with a poor prognosis (Matthay et al. 2016).

NB is an embryonal tumor that arises from aberrant differentiation of the neural crests derived from sympathoadrenal progenitors cells and adrenal chromaffin cells, that later obtain migratory mesenchymal-like phenotype during development. Transition between differentiation states in NB development occurs quickly but the mechanisms leading to epigenetic and transcriptional reprogramming are still under investigation. This complex lineage specification and differentiation process is regulated by a handful of master transcription factors, such as ISL1, HAND2, PHOX2B, GATA3, and TBX2, which are a part of a core regulatory circuit (CRC) (Saint-André et al. 2016; Banerjee et al. 2020). Their expression is elevated in all NB cell states compared to cell lines derived from other tumor types, indicating regulation of cell growth and survival in NB (Durbin et al. 2018). NB can be classified into at least two states based on gene

expression, enhancer and super-enhancer profiles - one with a committed adrenergic (ADRN) signature and the other with mesenchymal, migratory neural crest properties (MES). Members of the noradrenergic CRC consists of MYCN, ISL1, GATA3, HAND2, PHOX2A, PHOX2B, TBX2, ASCL1 (Durbin et al. 2018; Decaesteker et al. 2018; Boeva et al. 2017; Groningen et al. 2017; L. Wang et al. 2019), and a co-regulator LMO1 (L. Wang et al. 2019), and mesenchymal signature is marked by transcription factors such as FOS, JUN, AP-1 (Boeva et al. 2017), MEOX1, MEOX2, SIX1, SIX4, SOX9, NAI2, VIM, FN1, SMAD3 and WWTR1 (Groningen et al. 2017).

MYCN, a part of the ADRN CRC, is also a part of the gene family encoding DNA binding basic helix-loop-helix (bHLH) proteins that recognise E-box CANNTG DNA motifs, and have well-established role in promoting tumor genesis in several human cancers (Entz-Werlé et al. 2005; Kwok et al. 2005). MYCN is associated with E-box binding motifs in an affinity-dependent manner, binding to strong canonical E-boxes at promoters of genes involved in growth and proliferation and evading abundant weaker non-canonical E-boxes clustered at enhancers in NB. bHLH transcription factors TWIST1 and HAND2 together with other transcription factors such as KL7, KL13, TFAP2B, NR2F2 and ESRRG co-occupy circa 80% of MYCN regulatory regions. Loss of MYCN leads to a global reduction in transcription, which is most marked at MYCN target genes with the greatest enhancer occupancy. At the deregulated level, highly abundant MYC proteins, which can only bind to open chromatin, invades the cell's cis-regulatory regions binding to promoter-distal enhancers regions enriched in non-canonical E-box motifs (Zeid et al. 2018; Murphy et al. 2009; Guccione et al. 2006). MYCN binding sites have been reported to co-localize with methylation changes in NB, and suggest a dual role for MYCN - a classical transcription factor affecting the activity of individual genes, and a mediator of global chromatin structure (Murphy et al. 2009).

NB, like other pediatric cancers, is characterized by a low mutational burden which suggests that epigenetic mechanisms may drive its development and progression (Pugh et al. 2013). Epigenetic changes have been shown to play an important role in NB, with well-characterised examples of silencing of tumor suppressor genes by DNA methylation (Q. Yang et al. 2007; Q. Yang 2004; Lázcoz et al. 2006; Hoebeeck et al. 2009; Dreidax et al. 2013; Charlet et al. 2017), and by repressive histone modifications (Charlet et al. 2017; C. Wang et al. 2012; Dreidax et al. 2013). Recent methylation-array based study on patient derived data by Henrich et al. 2016,

showed distinct patterns for MYCN-amplified versus non-amplified tumors together with large segmental DNA copy number alterations (1p deletion, 11q deletion, 17q gain), MYCN and MYC expression and other key clinical parameters suggesting that deregulated methylation is a fundamental feature of high-risk NB. Charlet et al. 2017 compared the methylation pattern of NB cell lines to that of human neural crest precursor cells using methylation arrays and discovered hypermethylated genes, including MEGF10, a cell engulfment and adhesion factor gene, was epigenetically repressed in the NB cell lines.

In this study, we aim to gain a deeper understanding of DNA methylation landscape in a single nucleotide and genome-wide manner of MYCN-amplified and non-MYCN-amplified high-risk neuroblastomas and the association of DNA methylation aberrations with neuroblastoma-specific transcription factor regulatory networks.

2.2 Methods and Data

Whole-genome bisulfite sequencing analysis To preprocess the whole genome bisulfite sequencing data, we used a customized version of the PiGx BSseq pipeline (Wurmus et al. 2018). Its steps included quality control using FastQC (Babraham 2018b) (version 0.11.8), trimming using Trim Galore! (Babraham 2018a) (version 0.6.4, cutadapt version 2.6), alignment to the reference human genome version GRCh38 using bwa-meth (Pedersen et al. 2014) (version 0.2.2), deduplication using Picard MarkDuplicates (*Picard* n.d.) (version 2.20.4), and methylDackel R package (*MethylDackel* n.d.) (version 0.5.1, with arguments –minDepth 1 -q 5 -p 5) for methylation calling.

Differential methylation was performed using methylKit (Akalin, Kormaksson, et al. 2012) that applies logistic regression to test the methylation differences. Covariates such as batch effects, age and gender, and correction for overdispersion were added to the model. Chi-square-test was used to compare the difference in the deviances of the null model and the alternative model. The null model assumes there is no relationship between sample groups and methylation, and the alternative model assumes that there is a relationship where sample groups are predictive of methylation values for a given CpG for which the model is constructed. methCP R package (Boying Gong 2019) was used to merge differentially methylated CpGs to regions. For differentially methylated regions we used q-value cutoff 0.01 and minimal methylation difference of 25%.

Differentially methylated regions were annotated using genomation R package (Akalin, Franke, et al. 2015). HR_nMNA high versus low TERT DMRs were calculated between samples of TERT expression calculated as DESeq2's median of ratios with values below 4.6, and above 8.

Methylation microarrays analysis DNA methylation and gene expression of 105 patients neuroblastoma samples from Infinium 450k methylation arrays (Illumina), and 4 x 44k oligonucleotide microarrays (Agilent Technologies) are published by Henrich et al. 2016. R package missMethyl was used for preprocesing of methylation data, and subset-quantile within array normalisation was used. Gene expression data was downloaded from the GEO database with GSE73517 (GSE73517_quantile.csv.gz file provided by the authors). DMRs were called similarly to WGBS DMRs.

Genomic annotation Gene coordinates were obtained from RefSeq GRCh38 assembly, CpG islands from the cpgIslandExt table from UCSC's goldenpath service, CpG shores were defined as 2000 bp upstream of CpG islands and CpG shelves as 4000 bp upstream of CpG islands. Super-enhancers were downloaded from (Boeva et al. 2017) and converted to hg38 assembly using liftOver R package. Enhancers were defined as H3K27ac peaks from Boeva et al. 2017. Neuroblastoma cell line specific enhancers were defined as union of peaks from and filtered to peaks present in at least 3 neuroblastoma cell lines (CLBBER, CLBGA, CLBMA, CLBPE, Gimen, IMR32, LAN1, N206, SHEP, SHSY5Y, SJNB12, SJNB1, SJNB6, SJNB68, SKNAS, SKNBE2C, SKNDZ, SKNFI, TR14, CHP212, GICAN, NB69, SKNSH, NBEB), pdx-specific enhancers and NCC-specific were defined accordingly as provided by Boeva et al. 2017. Enhancers from neuroblastoma tumors were defined as H3K27ac ChIP-seq peaks downloaded from Gartlgruber et al. 2020, and present in at least 3 samples.

To associate DMRs with genes we used R package RGreat (McLean et al. 2010) (the function submitGreatJob() with default parameters, i.e. rule="basalPlusExt", adv_upstream=5.0, adv_downstream=1.0), and then Spearman correlation was used to filter DMRs correlated with expression of their associated genes (<-0.5 or >0.5). Gene ontology of correlated genes was performed using gProfiler2 R package with default parameters (*gprofiler2* n.d.).

RNA-seq analysis Gene expression was estimated using DESeq2 R package (Love, Huber, and Anders 2014) that uses counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene.

System-level integration of DNA methylation, gene expression and sequence information We exploit the ISMARA webserver expert mode (`https://ismara.unibas.ch/mara/`) to model DNA methylation in terms of computationally predicted regulatory sites for transcription factors (Piotr J Balwierz et al. 2014). ISMARA uses bayesian ridge regression, which assumes that a response variable is a vector of gene expression or ChIP-seq signal on regions of interest, and explanatory variable is a matrix of occurrences of DNA motifs within regions of interest. For our purposes, in order to identify the key transcription factors driving the methylation changes, the response variable was $log2((Y+1)/(100-Y+1))$, where Y is the average DNA methylation within MNA or HR_nMNA DMRs where methylation was correlated with nearby genes as described in the "Whole-genome bisulfite sequencing analysis" paragraph above. Estimated beta coefficients were called motif activities, and they indicate contributions of each DNA motif to the model. The model was run on each neuroblastoma sample separately. Motifs with the highest positive z-score are the most significant predictors of consistently high methylation across samples. To predict individual target regions for each motif ISMARA calculates the difference of the log-likelihood of the model in which only the binding sites for the motif in the target region have been removed. For each motif, a list of all target regions / DMRs is provided. Calculating DNA motif occurrences was performed using numMotifHits function from the motifcounter R package with arguments order=3, alpha=0.005, gran=0.05 (Kopp and Vingron 2017). PWM matrices of DNA motifs were downloaded from the JASPAR 2020 database (Fornes et al. 2019). Functions to prepare input data, and visualize results are part of the motifActivity R package (`https://github.com/BIMSBbioinfo/motifActivity/`). PPI networks were build using the STRING database (Szklarczyk et al. 2018). Each DNA motif was assigned one or more transcription factors that bind to them, therefore many motifs can share their transcription factors. PPI networks show TFs whose expression was correlated with motif activity > 0.5 or < -0.5, and in case of microarray data >0.4 or <-0.4.

Additional analyses Gender of the samples were estimated based on number of Y chromosomes reads, percentage of reads mapping to X and Y chromosome and XIST expression.

2.3 Results

2.3.1 DNA methylation stratifies neuroblastoma risk groups

In this study, we analysed 24 whole genome bisulfite sequencing samples from primary neuroblastomas risk groups: ST4S stage, low risk (LR), intermediate-risk (IMR), high-risk with MYCN amplification (MNA), high-risk without MYCN amplification (HR_nMNA). In order to get insights of global DNA methylation patterns across risk-groups, we performed hierarchical clustering of DNA methylation percentage based on the top 5000 most variable CpGs across all samples. We identified three groups of patients: 1) high-risk MYCN amplified, 2) a mixture of high-risk without MYCN amplification, intermediate- and low-risk and 3) low-risk patient groups (see Figure 2.1 and Supplementary Figure A.1A). We investigated potential variables that could influence DNA methylation levels, such as sequencing batches, age and gender, and included them in further analysis as covariates (see Supplementary Figures A.1B-D). MYCN-amplified sub-type is well separated from the rest of the samples, and the HR_nMNA samples from the low-risk groups, but the HR_nMNA group shows more heterogeneity than MYCN-amplified group based on DNA methylation levels. Together, global DNA methylation patterns can identify risk-groups of neuroblastoma patients, and in particular discriminate between MNA versus other risk-groups, and high-risk versus ST4S and low-risk groups.

2.3.2 Differentially methylated regions are enriched at introns and intragenic enhancers

In order to investigate DNA methylation changes associated with high-risk neuroblastomas, we performed differential methylation analysis between high-risk and lower-risk patients subgroups. We calculated differential methylation regions using Chi-square-test with over-dispersion correction and included covariates: batch effects, age, and gender. We called two sets of differentially methylated regions (DMRs): 1) between MYCN amplified samples versus low risk and intermediate risk samples (MNA DMRs) and 2) high-risk MYCN-nonamplified versus low risk and intermediate risk samples (HR_nMNA DMRs). It resulted in 21608 MNA DMRs and

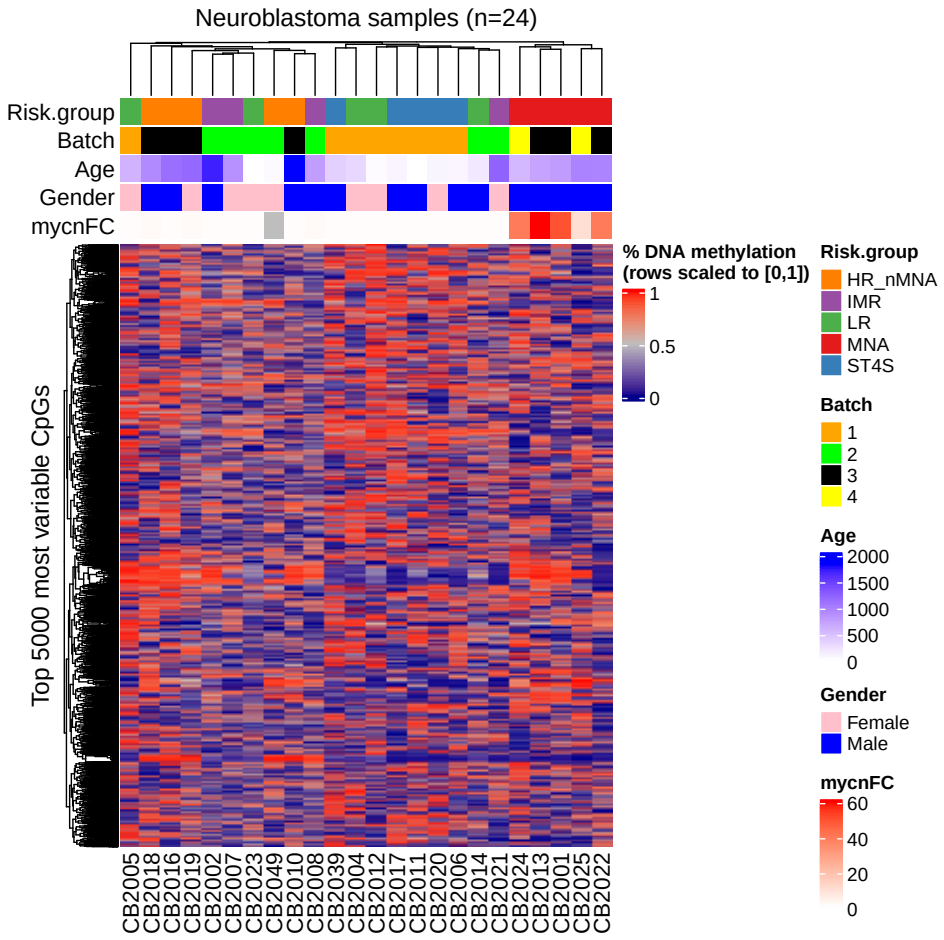


Figure 2.1: Hierarchical clustering of DNA methylation percentage of top 5000 most variable CpGs of 24 neuroblastoma patients. Each column represents a patient and each row represents a CpG. The percentage of DNA methylation is normalized to [0,1] range. Each patient has depicted risk group, sequencing batch, age (in days), and estimated gender (for details see 2.2 Methods and Data).

27859 HR_nMNA DMRs (see Supplementary Figure A.2 for chromosome-wise distribution of DMRs over the genome). MNA DMRs were located in close proximity to transcription start sites (TSSes), as well as hyper-methylated HR_nMNA DMRs, but hypo-methylated HR_nMNA DMRs were located at a notable distance from TSSes (median distance 24kb), suggesting that perturbations of DNA methylation patterns in HR_nMNA neuroblastomas might be occurring at distal genomic regions (Figure 2.2A).

DMRs were located outside of CpG islands and CpG island shores (73% MNA DMRs, 83% HR_nMNA DMRs), and less than half of the DMRs were located at introns (42% MNA DMRs, 40% HR_nMNA DMRs). ~25% of MNA DMRs and 17% of HR_nMNA

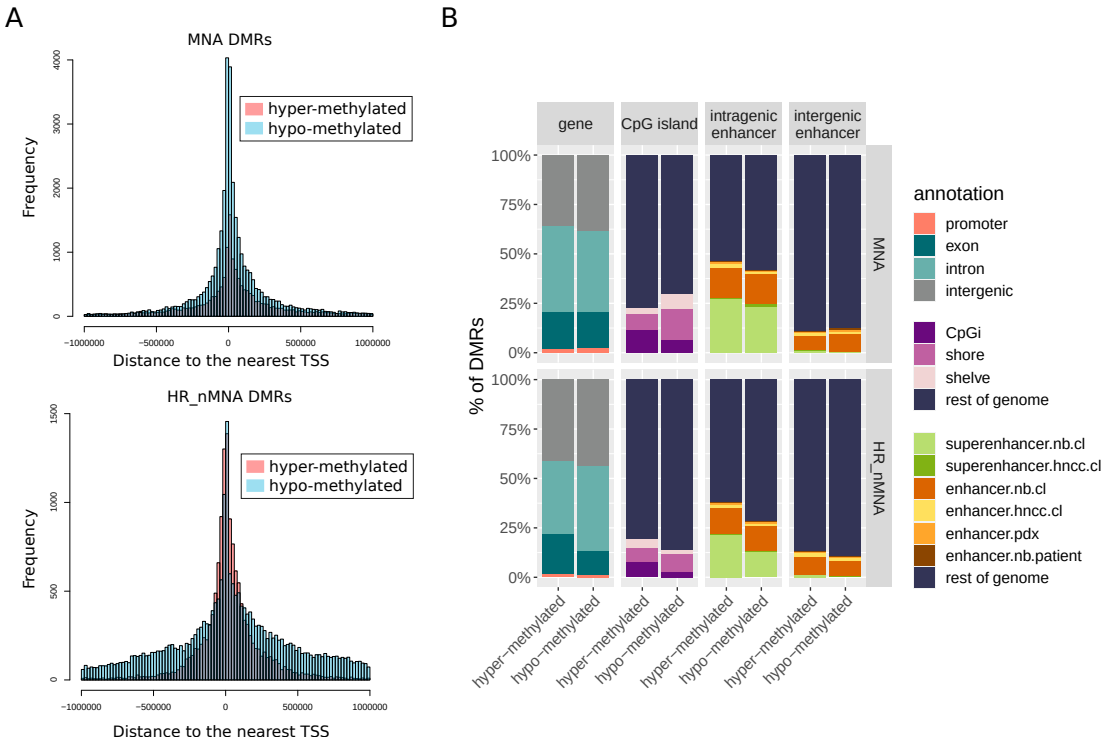


Figure 2.2: (A) Histograms of distances of MNA and HR_nMNA DMRs to the nearest gene TSSes (in bp). Hyper-methylated DMRs are depicted in red hyper-methylated, and hypo-methylated in blue. (B) Annotation of differentially methylated regions (hyper- and hypo-methylated) with genomic regions, such as promoters, eons, introns, intergenic regions, CpG islands, CpG shores, CpG shelves, and enhancer and super-enhancers regions from in neuroblastoma cell lines (terms with a suffix .nb.cl), patient derived xenografts (.pdx), human neural crest cell lines (.hncc.cl) (Boeva et al. 2017), and enhancers from patient-derived neuroblastoma tumors (.patient) (Gartlgruber et al. 2020).

were at intragenic super-enhancers and 15% MNA DMRs and 16% HR_nMNA at intragenic enhancers and super-enhancers regions defined by H3K27ac marks in neuroblastoma cell lines, and ~1% of MNA and HR_nMNA DMRs at enhancers were defined by H3K27ac mark in neuroblastoma tumors (see Methods and Data 2.2) (Figure 2.2B). To summarize, aberrant DNA methylation changes localized mostly at intragenic super-enhancers and enhancers, and in case of high-risk MYCN-nonamplified neuroblastomas, distal, hypo-methylated regions, and global hypomethylation might play a role in tumorigenesis of this particular risk-group.

2.3.3 DNA methylation changes correlate with genes associated with neuronal activity

To estimate the effect of differential DNA methylation on gene expression, we integrated into the analysis, gene expression from matched RNA-seq samples. First, we associated each DMR with genes within 5kb upstream and 1kb downstream of TSSes and then, filtered DMR-gene pairs based on Spearman correlation between their DNA methylation and expression across samples with cut off >0.5 or <-0.5 (see Methods and Data 2.2).

HR_nMNA DMRs overlapped with MNA DMRs only in ~10%, but genes correlated with HR_nMNA DMRs overlapped with MNA genes in 57% (Figure 2.3). It might implicate, that aberrant DNA methylation changes are associated with similar gene expression programs, but are regulated through different regulatory mechanisms in the two high-risk groups.

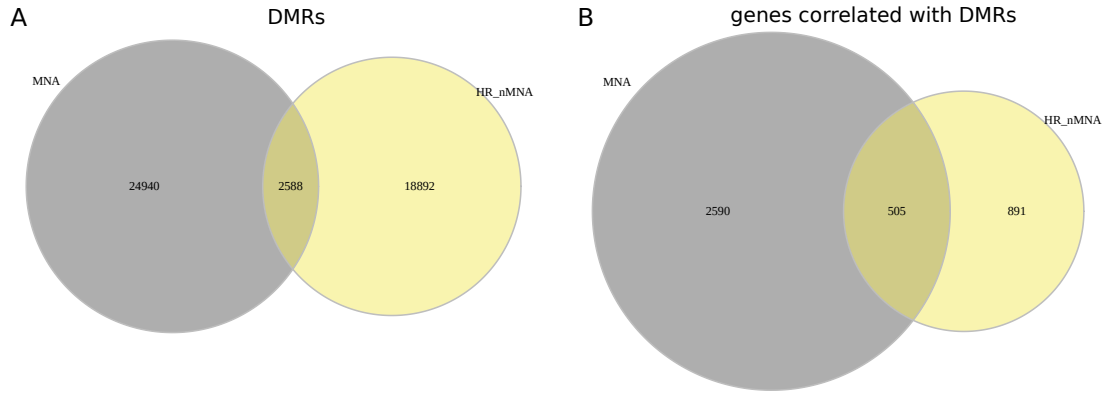


Figure 2.3: Venn diagram showing the overlap of MNA and HR_nMNA (A) and genes correlated with DMRs (B).

Gene ontology analysis of genes correlated with MNA DMRs demonstrated terms related to Rap1 signalling pathway, NCAM (neural cell adhesion molecule) signaling for neurite outgrowth, neuronal system and activated NTRK2 signals through FYN, and sequence specific DNA binding (Supplementary Figure A.3). It has been shown that ALK activation of Rap1 may contribute to cell proliferation and oncogenesis of neuroblastoma, driven by gain-of-function mutant ALK receptors (Schönherr et al. 2010). ALK is a direct transcriptional target of MYCN in neuroblastoma clinical samples with amplified MYCN and in developing tumors of MYCN-transgenic mice (Hasan et al. 2013). NCAM based signaling complexes are involved in a variety of cellular processes of importance for the formation and maintenance of the

nervous system, and can initiate downstream intracellular signals by at least two mechanisms: activation of FGFR and formation of intracellular signaling complexes by direct interaction with cytoplasmic interaction partners such as tyrosine kinases Fyn and FAK. Fyn and FAK interact with NCAM and undergo phosphorylation and this transiently activates the MAPK, ERK 1 and 2, cAMP response element binding protein (CREB) and transcription factors ELK and NFkB (Ditlevsen et al. 2008). In mouse brain, Fyn activation downstream of Bdnf-induced Ntrk2 (TrkB) signaling increases phosphorylation of voltage gated sodium channels by FYN, resulting in decrease of sodium currents (Ahn et al. 2007). Genes which correlated only with HR_nMNA DMRs were associated with generations of neurons, the calcium signalling pathway, and a neuroactive ligand-receptor interaction. Genes that correlated with both MNA and HR_nMNA DMRs were linked to neuronal precursor cell proliferation, nervous system development, and RAP1 signalling pathway.

Next, we examined a relationship between aberrations in DNA methylation with differentially expressed genes. Hypo-methylated MNA DMRs at CpG islands, gene bodies and enhancers were correlated with up-regulated (CpG island=912, gene bodies and enhancers=1970) and down-regulated genes (CpG island=2245, gene bodies and enhancers=4653) (see Figure 2.4). Hypo-methylated HR_nMNA DMRs at gene bodies and enhancers were correlated mostly with only down-regulated genes, at CpG islands with both up- and down-regulated genes, and hyper-methylated regions with up-regulated genes. Hypo-methylated MNA and HR_nMNA DMRs correlating with down-regulated genes were located on introns and enhancers, and hyper-methylated DMRs correlating with up-regulated genes were mostly on introns (Supplementary Figure A.4).

Hyper- and hypo-methylated MNA DMRs correlated with down-regulated genes were associated with the "neuronal system" gene ontology (GO) terms. Hyper-methylated MNA DMRs linked to up-regulated genes were linked to AP2, SP1, E2F, SALL2 DNA motifs, and hypo-methylated MNA DMRs linked to up-regulated genes were linked to multiple c-MYC DNA motifs, and MYCN, MAX, AP2, E2F, MAX, USF, ROX, EGR1, and H1alpha DNA motifs. GO terms for HR_MNA DMRs were obtained, and for hyper-methylated DMRs with up-regulated genes (NHERF1-PRKCZ complex, SCL9A3R1-ACTN4 complex) and

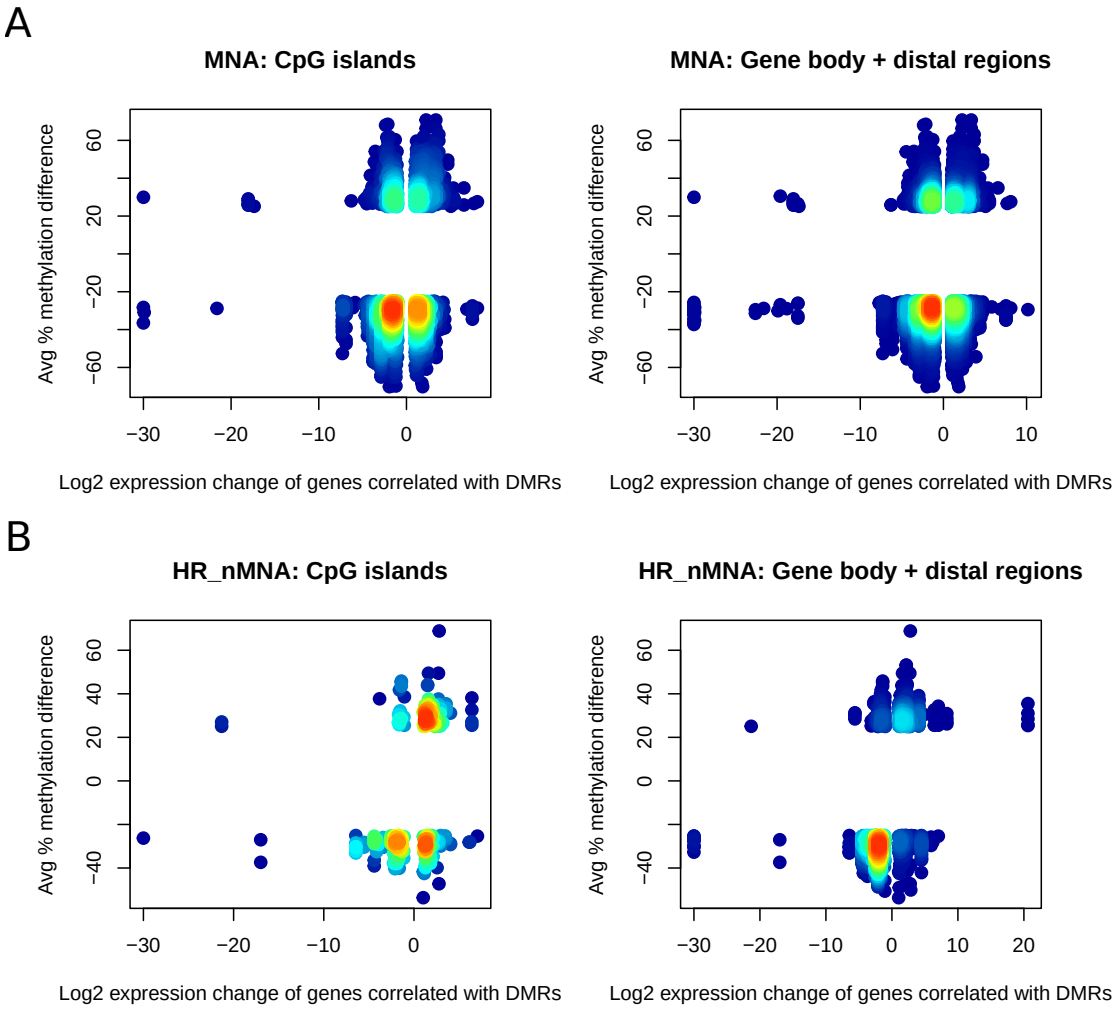


Figure 2.4: Density plot of differential expression of genes and average percentage of DNA methylation within MNA DMRs (A) and HR_nMNA DMRs (B) located either on CpGs islands, or introns and enhancers.

hypo-methylated with up-regulated genes associated with neuronal system terms and GABA receptor activity (see Figure 2.5).

We detected known tumor suppressors, oncogenes, and other functional gene sets such as cytokines and growth factors, protein kinases, translocated cancer genes, and transcription factors linked to the DMRs associated with these genes along with differential expression status of the genes. As an example, TERT, NFIB, PAX5, PDGFRA, TRIP13 oncogenes, SMAD3, PAX5 tumor suppressors, and TWIST1, ASCL1 transcription factors were up-regulated and correlated with hypo-methylated MNA DMRs (Supplementary Figure A.5). BCL6, IKBKE, JAZF1, PRDM16 oncogenes, PTPRT, SFRP1 tumor suppressors, and ARNTL, NEUROD6,

A. MNA DMRs

hyper-methylated DMRs corr. with down-regulated genes

	term.name	term.id	domain	p.value
1	regulation of neuron projection development	GO:0010975	BP	9.47e−06
2	regulation of synapse organization	GO:0050807	BP	1.51e−02
3	cell adhesion	GO:0007155	BP	1.76e−02
4	cell morphogenesis involved in differentiation	GO:0000904	BP	2.51e−06
5	cell morphogenesis involved in neuron differentiation	GO:0048667	BP	7.10e−06
6	synapse assembly	GO:0007416	BP	8.50e−04
7	cell projection	GO:0042995	CC	1.42e−02
8	plasma membrane	GO:0005886	CC	6.92e−03
9	neuron part	GO:0097458	CC	4.08e−04
10	synapse	GO:0045202	CC	3.12e−03
11	Axon guidance	KEGG:04360	keg	4.60e−03

hyper-methylated DMRs corr. with up-regulated genes

	term.name	term.id	domain	p.value
1	negative regulation of signal transduction	GO:0009968	BP	3.31e−02
2	embryonic skeletal system morphogenesis	GO:0048704	BP	2.75e−07
3	stem cell differentiation	GO:0048863	BP	2.68e−02
4	negative regulation of cell communication	GO:0010648	BP	3.30e−02
5	drug transmembrane transporter activity	GO:0015238	MF	2.24e−03
6	Factor: AP−2; motif: MKCCCSCNGGCG; match class: 1	TF:M00189_1	tf	1.03e−02
7	Factor: Sp1; motif: GGNDGGRGGCGGGG; match class: 0	TF:M04953_0	tf	4.80e−02
8	Factor: E2F−1; motif: NGGGCGGGARV; match class: 1	TF:M07206_1	tf	1.56e−02
9	Factor: SALL2; motif: GGGTGGG	TF:M04595	tf	3.44e−02

hypo-methylated DMRs corr. with up-regulated genes

	term.name	term.id	domain	p.value
1	regulation of ion transport	GO:0043269	BP	2.26e−02
2	cell adhesion	GO:0007155	BP	2.85e−02
3	positive regulation of calcium ion import	GO:0090280	BP	4.09e−02
4	signaling	GO:0023052	BP	1.41e−02
5	nervous system development	GO:0007399	BP	8.65e−05
6	plasma membrane part	GO:0044459	CC	1.89e−04
7	neuron part	GO:0097458	CC	2.51e−04
8	neuron projection	GO:0043005	CC	5.78e−04
9	Pancreatic secretion	KEGG:04972	keg	5.18e−03
10	Neuronal System	REAC:R−HSA−112316	rea	3.79e−02

hypo-methylated DMRs corr. with up-regulated genes

	term.name	term.id	domain	p.value
1	developmental process	GO:0032502	BP	4.58e−02
2	animal organ development	GO:0048513	BP	3.63e−02
3	skin 1; melanocytes	HPA:043040	hpa	4.23e−02
4	hsa−miR−6870−5p	MIRNA:hsa−miR−6870−5p	mir	3.49e−02
5	hsa−miR−193b−3p	MIRNA:hsa−miR−193b−3p	mir	6.28e−05
6	O−linked glycosylation	REAC:R−HSA−5173105	rea	9.22e−03
7	Factor: Egr−1; motif: GCGCATGCG	TF:M04869	tf	3.40e−03
8	Factor: Kaiso; motif: GCMGGGRGCRGS; match class: 1	TF:M03876_1	tf	1.25e−02
9	Factor: MAX; motif: CACGTG; match class: 0	TF:M08950_0	tf	6.63e−07
10	Factor: AP−2; motif: MKCCCSCNGGCG; match class: 1	TF:M00189_1	tf	1.44e−02
11	Factor: c−Myc; motif: KACCACGTGSYY; match class: 1	TF:M01154_1	tf	3.36e−08
12	Factor: ROX; motif: GNNNCASGTGGS; match class: 1	TF:M09764_1	tf	3.35e−06
13	Factor: c−Myc:Max; motif: GCCAYGYGSN; match class: 1	TF:M00322_1	tf	1.14e−07
14	Factor: USF; motif: NCACGTGN; match class: 0	TF:M00217_0	tf	1.38e−07
15	Factor: USF2; motif: CASGYG; match class: 1	TF:M00726_1	tf	6.63e−07
16	Factor: Arnt; motif: NNNNNRTCACGTGAYNNNNN; match class: 1	TF:M00539_1	tf	8.69e−05
17	Factor: USF; motif: GYCACGTGNC	TF:M00187	tf	4.57e−02
18	Factor: E2F−1; motif: TTGGCGCGRAANNGNM	TF:M00938	tf	8.52e−04
19	Factor: c−Myc:Max; motif: NNNNNNNCACGTGNNNNNNN; match class: 1	TF:M00615_1	tf	2.78e−06
20	Factor: E2F−4; motif: NGGCGGGAARN	TF:M07084	tf	2.57e−02
21	Factor: Myc; motif: CACGTGS; match class: 0	TF:M00799_0	tf	2.25e−08
22	Factor: E2F−1; motif: NGGGCGGGARV; match class: 1	TF:M07206_1	tf	2.57e−03
23	Factor: N−Myc; motif: NNCCACGTGNNN	TF:M00055	tf	5.03e−09
24	Factor: c−Myc:Max; motif: NNACCACGTGGTNN	TF:M00118	tf	7.85e−05
25	Factor: E2F−4; motif: NNTTCCCGCCNN	TF:M04823	tf	7.42e−03
26	Factor: HIF−1alpha; motif: NCACGT; match class: 0	TF:M02012_0	tf	6.29e−03
27	Factor: c−Myc; motif: NSCACGTGGN	TF:M04743	tf	2.83e−09
28	Factor: HIF−1alpha; motif: NCACGTNN	TF:M07384	tf	3.36e−03
29	Factor: CTCF; motif: NCCRSTAGGGGGCGC	TF:M08911	tf	4.96e−02
30	Factor: CLOCKBMAL; motif: NNNWCACGTGGN	TF:M08869	tf	2.14e−02
31	Factor: CLOCK:BMAL; motif: MCACGTGR; match class: 0	TF:M01116_0	tf	5.42e−03
32	Factor: c−Myc; motif: CACGTGGC; match class: 0	TF:M03867_0	tf	2.73e−06
33	Factor: Max; motif: NNANCACGTGNTNN; match class: 0	TF:M00119_0	tf	3.32e−05
34	Factor: MXI1; motif: NNNCCACGTGSNN; match class: 0	TF:M09775_0	tf	1.62e−04
35	Factor: E2F−1; motif: NNNSSCGCSAANN	TF:M07250	tf	4.82e−02
36	Factor: AP−2; motif: SNNNCCNCAGGCN; match class: 1	TF:M00915_1	tf	8.13e−03
37	Factor: HES−1; motif: GNCACGTGNC; match class: 1	TF:M08767_1	tf	3.96e−07
38	Factor: HAIRYLIKE; motif: NNNNCANGTG; match class: 0	TF:M08885_0	tf	2.65e−03
39	Factor: C−Myc; motif: NGCCACGTGNN	TF:M07601	tf	8.95e−07
40	Factor: E2F−4; motif: GCGGGAAANA	TF:M02090	tf	6.11e−05
41	Factor: AP−2; motif: GSCCSCRGGCNRNRNN	TF:M00800	tf	1.52e−02
42	Factor: MYC; motif: NNCCACGTGCNN; match class: 0	TF:M09812_0	tf	4.77e−08
43	Factor: MYCN; motif: NNCCACGTGGNN; match class: 0	TF:M09814_0	tf	1.89e−06
44	Factor: WT1; motif: NGCGGGGGGGTSMMCYN	TF:M05327	tf	1.12e−02
45	Factor: ARNTLIKE; motif: NNNSCACGTG; match class: 0	TF:M08868_0	tf	8.67e−09
46	Factor: c−Myc; motif: RACCACGTGCTC	TF:M01145	tf	8.95e−04

B. HR_nMNA DMRs

hyper-methylated DMRs correlated with up-regulated genes

	term.name	term.id	domain	p.value
1	NHERF1−PRKCZ complex	CORUM:6790	cor	4.99e−02
2	SLC9A3R1−ACTN4 complex	CORUM:6788	cor	4.99e−02

hypo-methylated DMRs correlated with down-regulated genes

	term.name	term.id	domain	p.value
1	gamma−aminobutyric acid signaling pathway	GO:0007214	BP	2.25e−02
2	postsynaptic membrane	GO:0045211	CC	2.40e−04
3	ion channel complex	GO:0034702	CC	1.68e−02
4	extracellular ligand−gated ion channel activity	GO:0005230	MF	1.76e−02
5	gated channel activity	GO:0022836	MF	3.82e−02
6	GABA−A receptor activity	GO:0004890	MF	9.56e−03
7	Vocal cord dysfunction	HP:0031801	hp	4.66e−02
8	Nicotine addiction	KEGG:05033	keg	3.00e−05
9	Morphine addiction	KEGG:05032	keg	3.00e−02
10	GABAergic synapse	KEGG:04727	keg	2.47e−02
11	Neuronal System	REAC:R−HSA−112316	rea	7.82e−03

Figure 2.5: Gene ontology terms of up- or down-regulated genes correlated with MNA DMRs (A) and HR_nMNA DMRs (B) by gProfiler2 (Raudvere et al. 2019). Abbreviations for data sources (domains): Molecular Functions (MF), Biological Process (BP), Cellular Component (CC), Reactome (rea), KEGG (keg), mirTarBase (mir), Transfac (tf), and CORUM (cor).

SOX1, SOX8, TFAP2A, TFAP2C transcription factors were down-regulated and correlated with hyper-methylated MNA DMRs (Supplementary Figure A.6 and A.7). We detected fewer genes in case of differentially expressed genes and correlated with differentially methylated HR_nMNA DMRs in comparison to described above MNA DMRs. The list of these genes include: RPS6KA4 oncogene, KLF6 tumor suppressor, VIP, CLEC11A, DKK1, FGF10 cytokines

and growth factors (Supplementary Figure A.8).

To summarize, we identified DMRs where DNA methylation is associated with expression of corresponding genes in high-risk neuroblastomas and associated with neuronal activity. MNA DMRs linked with differentially expressed genes contain DNA motifs from the MYCN regulatory network (such as AP2, SP1, E2F, SALL2, MAX, AP2, E2F, MAX, USF, ROX, EGR1, and H1alpha (Durbin et al. 2018; Decaesteker et al. 2018; Boeva et al. 2017; Groningen et al. 2017)). Known oncogenes and tumor suppressors that were shown to play an important role in neuroblastoma are correlated with MNA DMRs such as TERT, SMAD3, NEUROD6, SOX1, SOX8, TFAP2A, and TFAP2C. We also detected TWIST1, ASCL1 transcription factors, dozens of cell differentiation markers, cytokines and growth factors, protein kinases, and translocated cancer genes. HR_nMNA DMRs were correlated with a dozen of genes from functional gene sets, such as cytokines and growth factors, and RPS6KA4 oncogene, and KLF6 tumor suppressor.

2.3.4 Detection of regulatory networks affected by DNA methylation changes specific to high-risk neuroblastomas

We were also interested in decoding regulatory networks influenced by DNA methylation in neuroblastoma. Firstly, we wanted to confirm MYCN dependent methylation landscape in MYCN-amplified samples. Secondly, we hypothesized that MYCN and non-MYCN high-risk groups might 1) share similar regulatory programs and methylation patterns (if not at promoter level then at least at enhancer level) or 2) they have different regulatory mechanisms, but similar expression programs.

A study by Zeid et al. 2018 investigated the consequences of MYCN induction by engineering a tet-on MYCN model in a parental SHEP neuroblastoma cell line grown for multiple passages in a low-MYCN state. As a result, at 0h, 2h and 6h time points they obtained MYCN induction and increase of MYCN binding over time. We explored this data to observe whether MYCN amplified DMRs are enriched in MYCN binding sites in the engineered SHEP cell lines (Figure 2.6). MYCN-amplified DMRs have the highest MYCN ChIP-seq occupancy score after 6h, in the time point when the MYCN occupancy is the highest in the tet-on

MYCN model. It indicates that MYCN-amplified DMRs might act as MYCN regulatory regions, and influence MYCN activity.

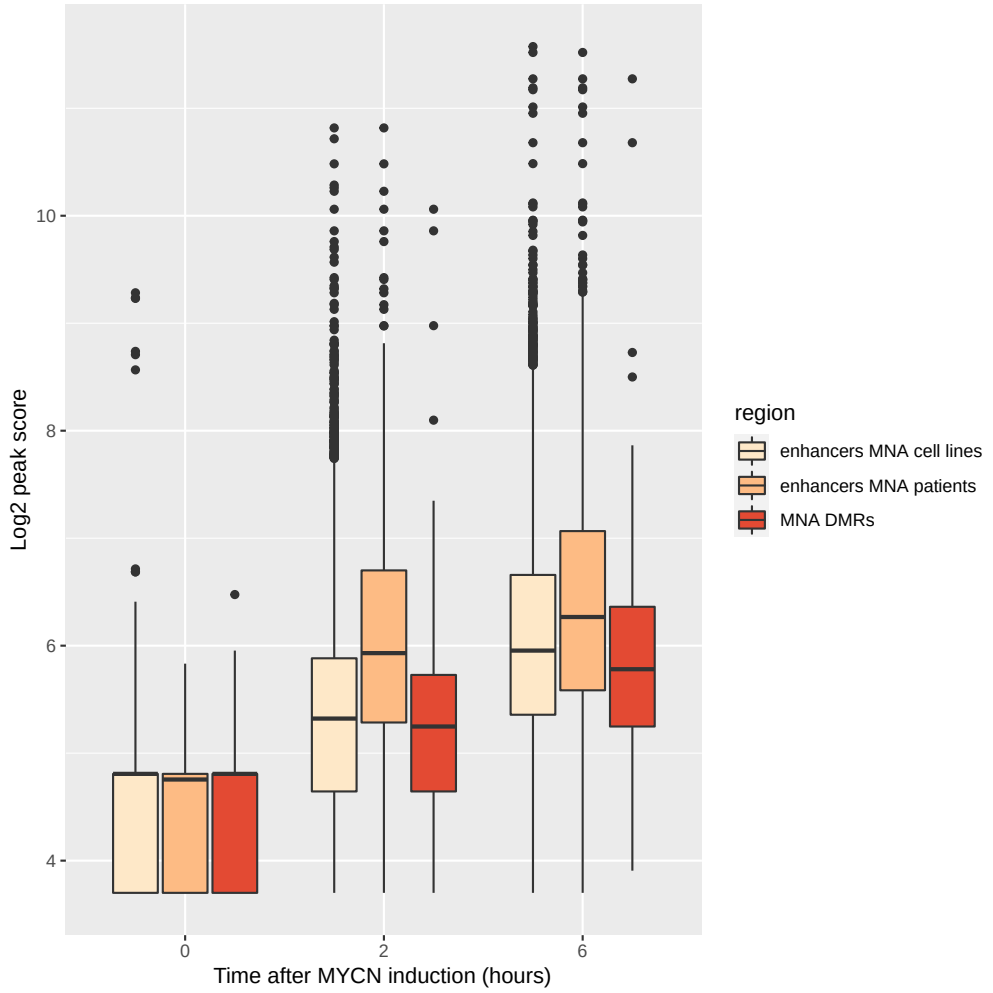


Figure 2.6: MYCN occupancy on MNA DMRs and enhancers. Box plots show MYCN ChIP-seq occupancy from SHEP cell line designed to not have any MYCN activity, and to restore it after 2 and 6 hours (Zeid et al. 2018). Enhancers MNA cell lines and patient-derived are H3K27ac ChIP-seq peaks as described in 2.2 Methods and Data.

To reconstruct transcription factors networks influenced by DNA methylation changes, we applied a computational modelling approach. We used ISMARA webserver (P. J. Balwierz et al. 2014) that performs bayesian ridge regression to investigate whether a number of known DNA motifs on DMRs can explain changes in DNA methylation levels in the DMRs across samples from different risk-groups, and consequently which TFs that bind to these DNA motifs might be associated with the DNA methylation dynamics (for more details see 2.2 Methods and

Data). Our aims were: 1) to identify the key TFs that can explain DNA methylation changes between high-risk and lower-risk groups and their activities (beta coefficients in the model, or so called "motif activities"), 2) identify direct interactions between the TFs in a protein-protein interaction (PPI) network, and which TF pathways are affected, 3) to get target DMRs that are the most influential in the model. We performed the modelling on DMRs that were correlated with nearby genes as described in the section 2.3.3, and occurrences of known DNA motifs were downloaded from the JASPAR database (Fornes et al. 2019). Modelling was performed on each sample separately. To obtain not only significant, but also biologically meaningful DNA motifs, we put a threshold on their motif activities - for each DNA motif we calculated Pearson correlation between its activity and expression of TFs that bind to them >0.5 and <-0.5 across samples (more than one TF can bind to a DNA motif). Then, we created PPI networks of the resulting TFs using the STRING database (Szklarczyk et al. 2018). As a result, we detected two PPI networks: a MNA PPI network that used MNA DMRs as an input, and a HR_nMNA PPI network that used HR_nMNA DMRs.

We detected transcription factors in MNA and HR_nMNA networks that have been reported to play crucial role in high-risk neuroblastomas (Figure 2.7A). Common TFs in MNA and HR_nMNA PPI networks are TFAP2B, ARNT, and JUN (depict in green in Figure 2.7A). MNA network include MYCN, and ISL2 that are part of the NB ADRN CRC components, and also JUN, and FOSL1 TFs that are part of the NB MES CRC. Other TFs included TFAP2B, GATA4, MYOD1, SRY, DLX5, STAT3, FOXP3, SRF, HOD12, STAT3, BACH2, NEUROG2, RORA, ESRRG, and E2F1. TFAP2B together with ISL2 contributes to noradrenergic neuronal differentiation in neuroblastoma (Ikram et al. 2015). It has been shown that FOXP3 is expressed in mesenchymal stem cell, HOD12 is expressed in mesoderm, STAT3 is involved in reprogramming and epithelial-mesenchymal transition (Strobl-Mazzulla and Bronner 2012), and GATA4 appears to be a common feature of neuroblastoma, with highest expression levels in MYCN-amplified tumors (Hoene et al. 2009). BACH2 is involved in neuronal differentiation in neuroblastoma (Shim et al. 2006; Waxman 2019), NEUROG2 and MyoD1 is expressed in neural lineages (Aydin et al. 2019; Q. Y. Lee et al. 2020), RORA is significantly downregulated in MYCN-amplified versus non-MYCN-amplified neuroblastomas (Ribeiro et al. 2016), and ESRRG co-occupies circa 80% of MYCN binding sites (Zeid et al. 2018).

E2F1 regulates MYCN expression in neuroblastomas (Strieder and Lutz 2003). GO enrichment showed enrichment in both types of these TF networks in terms including transcription factor binding to DNA, E-box binding, embryonic morphogenesis, central nervous system neuron development, and exclusively MNA network terms include a signaling pathway pertinent to development (WP2023 Cell Differentiation), and adipogenesis (Figure 2.7B).

Finally, we wanted to get more insights on DNA motifs that have more favourable high-risk motif activities compared to lower-risk sub-types. Motif activity in our context refers to beta coefficients of the model; for example, if a beta coefficient is negative it means that that for every one unit increase in the prediction variable (number of DNA motifs occurrences) the outcome variable (% DNA methylation on DMRs) decreases by the coefficient value. Therefore, in case of transcription factors that bind to open chromatin and unmethylated or lowly methylated regions, we are interested in negative values of motif activity and lower values compared to lower-risk groups, and in case of, for example the MNA network, also lower values compared to HR_nMNA samples. In MNA network, such DNA motifs (cluster 1 in the Supplementary Figure A.9A) include previously described MYCN, E2F1, MYOD1, BACH2, and RORA, and other TFs with important functions in neuroblastoma development: MYBL2 (Raschella et al. 1999), NR2F1 (Ang et al. 2019; Zeid et al. 2018), and SRY (Nagai 2001). DNA motifs with negative motif activities in both high-risk groups (cluster 4) include TFs: ISL2 (a part of ADRN NB CRC), TFAP2B involved in ADRN neuronal differentiation, TLX2 (neural crest homeobox protein), TLX3 and DLX5 involved in developing neural crest-derived tissues (sympathetic neuron progenitors) (Borghini et al. 2006; Hatano et al. 1997; Logan et al. 1998).

In our models, DNA motifs with the highest z-scores are the most significant predictors of consistently high methylation across samples. TFs that bind to such motifs should be then involved in gene silencing and/or bind to methylated sequences. DNA motifs with top z-scores in MNA network are depicted in cluster 3 (e.g. TCF3,GATA4,AHR,ARNT,STAT3), and together with motifs in cluster 4 (e.g. FOS,JUN,ESRRG,SP1,SOX8,NEUROG2,POU2F2,RARG), have the most positive motif activity in MNA samples in comparison to all other risk-groups. Some of these TFs are part of POU (POU2F2) and are associated with NFAT protein families (FOS, JUN, GATA4) that prefer binding to methylated sequences (Yin et al. 2017). NFAT and heterodimer FOS–JUN cooperatively bind to DNA sites and synergistically activate the

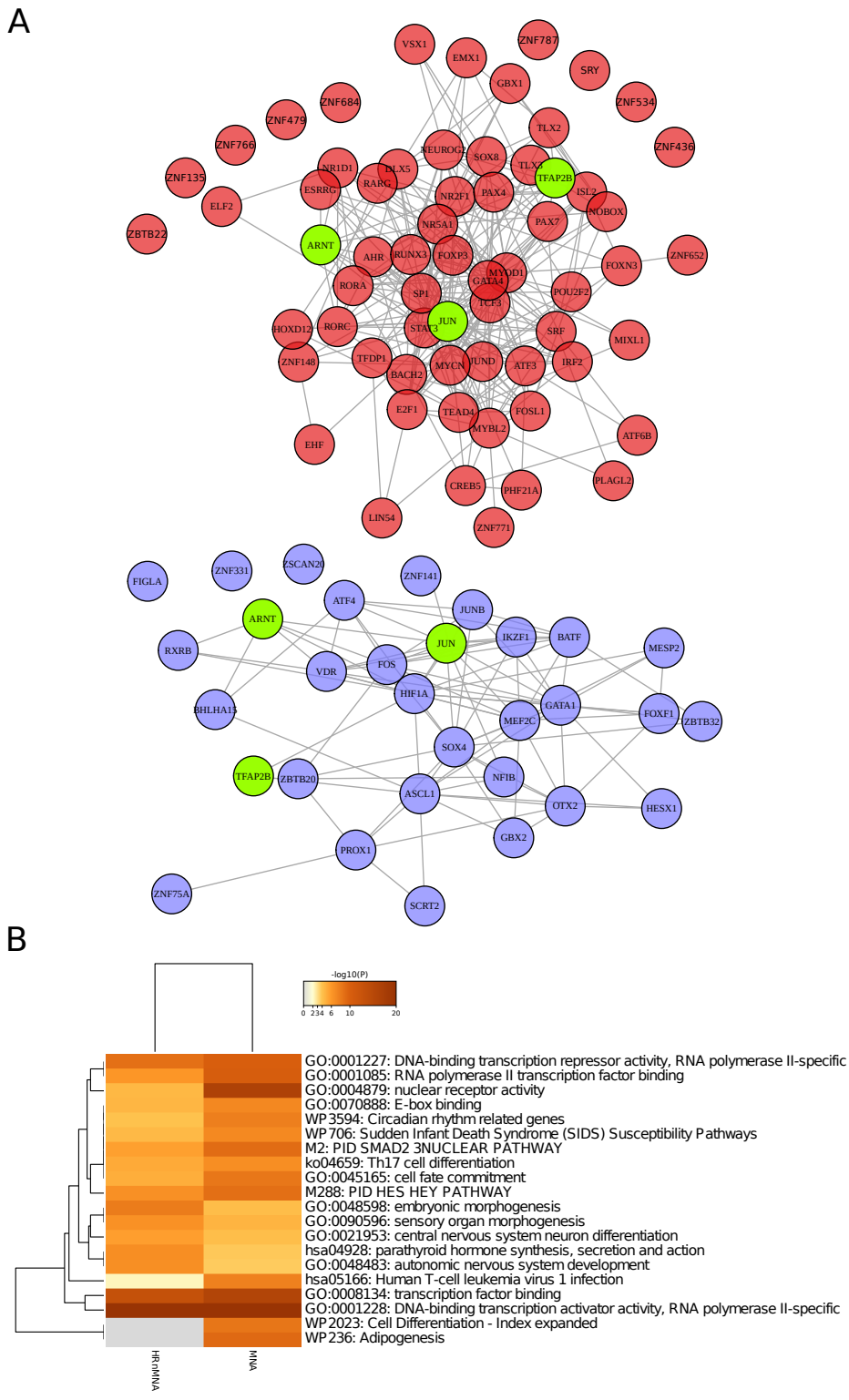


Figure 2.7: (A) Regulatory PPI network based on motif activity results using MNA (in red), and HR_nMNA DMRs (in blue). TFs common in both networks are depict in green. (B) Top 20 Gene Ontology terms (rows) using metascape (Y. Zhou et al. 2019) based on TFs from MNA and HR_nMNA networks (columns).

expression of many immune-response genes (J. Jain et al. 1992). TFs from the cluster 3, such as STAT3, GATA4, TCF3, ARNT are part of the HES/HEY pathway that contain bHLH protein domains. Top target MNA DMRs of these DNA motifs are mostly hyper-methylated (Supplementary Figure A.9A and Supplementary Figure A.22). Most of the target MNA DMRs in MNA network are correlated with MMEL1 (9 DNA motifs), TSPAN9 (5) and REV1 (4) genes. MMEL1 is associated with 1p loss in high-risk neuroblastomas (White et al. 1995; Torkov 2019; Bonvouloir et al. 2001), TSPAN9 takes part in regulating tumor microenvironment (Detchokul et al. 2014), and REV1 is involved in DNA double-strand break repair (Kolas and Durocher 2006). MYCN DNA motif is associated with a MNA DMR that DNA methylation levels across NB samples is correlated with DDX18 gene. DDX18 gene encodes a MYC-related DEAD-box 18 that is a direct target of MYC-MAX heterodimers (Grandori et al. 1996).

HR_nMNA PPI network consist of TFs that have been reported in neuroblastoma, such as TFAP2B (with the highest z-score), JUN and FOS, and also MEF2C that was reported in neurogenesis (Sekiyama, Suzuki, and Tsukahara 2011), SOX4 that is increased in late stages of sympathetic nervous system development (Potzner et al. 2010), and found to have higher expression as NB cells differentiate which is associated with better prognosis (Banerjee et al. 2020) (Figure 2.7A). The HR_nMNA network is composed of TFs with molecular functions such as noradrenergic neuron differentiation (SOX4,ASCL1), sympathetic nervous system development (SOX4,ASCL1,TFAP2B), and biological processes such as transcription factor DNA binding, HMG box domain binding (MEF2C,JUN), R-SMAD (FOS,JUN), E-box binding (ASCL1,HIF1A), p53 binding (GATA1,HIF1A) and nuclear receptor activity (RXRB,VDR). MEF2C binds to a DNA motif whose activity is negative in HR_nMNA and positive in MNA samples (cluster 1 in the Supplementary Figure A.9B), SOX4 has a strong negative activity in HR_nMNA in comparison to all other samples, and as well as FOS, JUN, JUNB, ARNT, TFAP2B and NFIB which are strongly differentially expressed in neuroblastoma versus neuroblasts (Preter et al. 2007) (cluster 2). DNA motifs from cluster 1 and 2 are associated with hypo-methylated top target HR_nMNA DMRs (Supplementary Figure A.23). DNA motifs with positive activity in HR_nMNA samples (cluster 3) include TFs from bHLH family MYC-type such as BHLHA15, ASCL1, MESP2, and GATA1, OTX2, and also ATF4 that coordinates glutamine metabolism in MYCN-amplified neuroblastomas (Ren et al. 2014).

In both types of TF networks, clustering of gene expression of the resulting TFs distinguishes well MYCN-amplified samples from others, however in case of HR_nMNA network, HR_nMNA samples didn't cluster together well (Supplementary Figures A.10 and A.11). Even though clustering based on motif activities don't reflect well gene expression patterns, clustering based on gene expression reveals clear groups of TFs clustering together in MNA and HR_nMNA networks, and still remain clusters of DNA motifs with negative motif activity in MNA and HR_nMNA samples respectively (Supplementary Figures A.12 and A.13). It revealed that in MNA network, MYCN is co-expressed with TLX2 (that interacts with PHOX2B and PHOX2A transcriptional activators of immature sympathetic neurons in neuroblastoma (Reiff et al. 2010), and that are part of the ADRN CRC), DLX2, NR2F1, DLX5, and MYBL2. Whereas, in HR_nMNA network there is a distinct cluster of co-expressed genes in HR_nMNA samples and negative motif activity values are assigned to motifs bound by JUN, FOS (cluster 1), and TFAP2B and SOX4 (cluster 2).

Next, we confirmed our results on an external data by analysing in a similar fashion, 105 DNA methylation and gene expression microarray samples derived from neuroblastoma primary tumors from Henrich et al. 2016. Differentially methylated regions were called between high-risk versus intermediate- and low-risk samples. The results support our finding on our Bisulfite-seq data. MNA network consist of TFs crucial in MYCN-driven tumors, such as previously described MYCN, MYC, MAX, TWIST1, and also SMAD3 that with SOX9 (Decaesteker et al. 2018), and TWIST1 (Gartlgruber et al. 2020) belong to MES CRC (Gartlgruber et al. 2020) (Supplementary Figure A.14A). Additionally, we detected multiple bHLH binding TFs (HAND1,TWIST1,SOX9,TCF3,SMAD3,ARNTL) and TFs binding to E-box motif (TFAP4,TWIST1,NR1D1,TCF3,MAX,ARNTL,MYC) (Supplementary Figure A.14B). TFAP4, MYCN, SMAD3, RORA are associated with negative motif activity values in MNA samples, and hypomethylated regulatory regions, and HAND1,MAX, MYC,SOX9,ARNTL,TWIST1 are associated with positive motif activity and hyper-methylated regions (Supplementary Figure A.14C). HR_nMNA network include TWIST1 and TFAP2B (both are involved in E-box based enhancer invasion in MYCN-dependent networks (Zeid et al. 2018)), and TFs involved in ectoderm development, VAX2 and ZBTB17 (Supplementary Figure A.15).

In order to investigate the differences between high-risk groups and within HR_nMNA group, we performed differential methylation and motif activity analysis on MNA versus HR_nMNA samples and HR_nMNA with high versus low TERT expression (for details see 2.2 Methods and Data). The most significant TFs in MNA vs HR_nMNA network and with the most extreme motif activity values are: MYCN, HAND1, TWIST1, TWIST2, ASCL1, ISL2, STAT1, STAT2, TCF3, FOS, JUN, and NEUROG1, and TCF3 that drives MYC activation in neuroblastoma (Wei et al. 2020), FOXP1 that down-regulation in expression is a common event in high-risk MNA and HR_nMNA neuroblastomas (Ackermann et al. 2014), and up-regulation of FOXN3 leads to increased inhibition of the MYC gene (Karanth et al. 2016; S.-H. Lee et al. 2018) (Supplementary Figure A.17 and Supplementary Figures A.18, A.24). HR_nMNA samples in TFs in HR_nMNA with high vs low TERT expression network don't cluster together based on motif activities (Supplementary Figure A.24 and Supplementary Figure A.21). However, based on co-expression patterns of the resulting TFs, we distinguished two clusters. First cluster contains TFs from the NB CRC such as ASCL1, and HAND1, and TWIST1 expressed in MNA, and the second cluster has TFs such as ARNT, and also NFATC1, MAT, RUNX3 that are involved in Th1 and Th2 cell differentiation stimulating cellular immune response.

Together, firstly, we confirmed that MYCN amplification comes with MYCN-dependent DNA methylation changes in MYCN-amplificated neuroblastomas. Secondly, we built TF networks based on MNA and HR_nMNA DMRs and using known DNA motifs to discover which TFs explain changes in DNA methylation levels in NB high-risk groups. We discovered that MNA TF network consist of DNA motifs that interact with MYCN in neuroblastoma, and components contributing to noradrenergic neuronal differentiation such as ISL2, and TFAP2B, as well as FOS and JUN as part of NB MES CRC, and other TFs that either regulate MYCN expression or take part in neuronal differentiation. We found several TFs that belong to the bHLH TF family (TCF3,SP1,AHR,NR1D1,MYOD1,TCF3,NEUROG2). TFs that were associated with DNA hypomethylation included MYCN, ISL2, TFAP2B, TLX2, MYBL2, NR2F1, SRY, DLX5, E2F1, MYOD1, BACH2, and RORA, and with DNA hypermethylation: POU2F2, FOS, JUN, STAT3, GATA4, TCF3, and ARNT. Top MNA DMRs target of MYCN DNA motif were linked with MMEL1 which is associated with chromosome 1p loss in high-risk

neuroblastomas. Highest-scoring TFs in HR_nMNA network were TFAP2B, JUN, FOS, and MEF2C, and other TFs including SOX4, ASCL1, GATA2, HIF1A. FOS, JUN, JUNB, ARNT, TFAP2B and NFIB, and had the strongest association with hypomethylated HR_nMNA DMRs. JUN, FOS, TFAP2B, SOX4 are co-expressed exclusively in HR_nMNA samples. TFs from bHLH family MYC-type such as BHLHA15, ASCL1, MESP2, GATA1 and OTX2 were associated with hyper-methylated HR_nMNA DMRs. We performed within HR_nMNA samples, comparison to shed more light into HR_nMNA heterogeneity. Gene expression of TFs from the HR_nMNA with high versus low TERT expression network, revealed two TF groups: one that includes ASCL1, HAND1, and TWIST1 with expression programs similar to that of MNA samples, and the other ARNT, NFATC1, MAT, and RUNX3 with similar patterns to lower-risk groups.

2.4 Discussion

The principal finding of our study was to uncover alterations in DNA methylation landscape at genome-wide and single nucleotide level that affects gene regulatory networks associated with high-risk neuroblastomas. We integrated whole genome bisulfite sequencing of 24 primary neuroblastomas from ST4S, low, intermediate MYCN-amplified, and non-MYCN-amplified high-risk groups with gene expression from matched RNA-seq tumors, H3K27ac ChIP-seq samples marking enhancer regions from published high-risk neuroblastoma tumors (Gartlgruber et al. 2020) and cell lines (Boeva et al. 2017), and sequence information.

Recently, several studies based on methylation microarray analysis investigated the potential usage of DNA methylation target regions as biomarkers in neuroblastoma (Henrich et al. 2016; Gómez et al. 2015). Inline with these findings we observed a similar pattern of global, unsupervised discrimination between high-risk and low risk groups based on only DNA methylation, and MYCN-amplified samples forming a distinct cluster. Secondly, differentially methylated regions between high-risk versus low and intermediate-risk groups were mostly located on introns, enhancers and super-enhancers regions detected in neuroblastoma cell lines, and associated with neuronal activity and nervous system Gene Ontology terms, indicating their regulatory functions in neuroblastoma. Interestingly, most of the DMRs correlated to expression of their nearby genes were hypo-methylated and associated with down-regulated

genes. It might indicate recruitment of transcriptional repressors (Gherardi et al. 2013), large chromosomal rearrangements, such as chromosomal 1p and 11q deletions (Depuydt et al. 2018), post-translational modifications (Gu et al. 2011; Otto n.d.), and it can depend on cellular localization, and the presence of specific co-factors. However, remaining hypo-methylated MNA DMRs correlated with up-regulated genes showed strong enrichment in DNA motifs of transcription factors that have been shown to belong to neuroblastoma regulatory networks, such as MYCN, c-MYC, MAX, and AP-2 that are expressed in the emerging neural-crest cells. Nonetheless, since MNA and HR_nMNA DMRs don't overlap much, but genes that are associated with them overlap to a greater extent, we hypothesized that perhaps these two different high-risk groups might share similar gene expression programs but through different regulatory, and potentially epigenetic mechanisms.

Amplification of transcription factor MYCN is a defining feature of high-risk neuroblastoma, and it has been shown by Zeid et al. 2018 that MYCN associates with E-box binding motifs in an affinity-dependent manner. Our hypothesis was that since MYCN, with other bHLH proteins that bind to E-box motifs at promoters and enhancers drive MYCN-amplified neuroblastomas, and it has been shown to be associated with regions of DNA hyper-methylation (Murphy et al. 2009), we should observe some DNA methylation dynamics associated with MYCN amplification in the MNA subtype, and perhaps in the HR_nMNA subtype (or at least some HR_nMNA samples) as well. For that reason, we investigated whether DNA methylation changes in our DMRs can be linked to MYCN binding sites in three SHEP cell lines using MYCN ChIP-seq from (Zeid et al. 2018), each with increasing MYCN activity. We showed that with increasing MYCN activity in a cell, MYCN signal increases on MNA DMRs and enhancers as control regions. Next, we focused on whether high-risk subtypes share similarities in regulation of core regulatory genes, and/or on regulation enhancer level.

It led us to a system-level approach of combining DNA methylation, gene expression, chromatin marks and DNA motif occurrences in order to find aberrant transcription factor regulatory networks in both neuroblastoma sub-types. Our aim was to identify the key TFs driving methylation changes and make detailed predictions regarding their regulatory roles. Therefore, the novelty of our approach included finding TF networks essential in high-risk neuroblastoma development and progression based only on DNA methylation patters in

genome-wide manner together with sequence information. The involvement of the resulting TFs is crucial in neuroblastoma, and supported by the published findings that 1) high expression of many of them is associated with poor prognoses in primary NB tumors, 2) several key TFs are linked to the NB CRC, and 3) most of them function as NB dependencies or are known to regulate NB growth and differentiation. Our analysis pinpoints to TFs that are driven by MYCN and exhibit committed adrenergic signature (MYCN, ISL2, TFAP2B) and are regulated by DNA methylation aberrations in MNA subgroup, and TFs with mesenchymal signature in HR_nMNA subgroup (JUN and FOS). TFAP2B, ARNT, and JUN are part of MNA and HR_nMNA TF networks. TFAP2B is a part of core regulatory circuit based on TF enhancer binding with MYCN and other bHLH TFs (Zeid et al. 2018). Our findings confirm results from previous studies on two neuroblastoma signatures based on enhancer and super-enhancer neuroblastoma profiles (Gartlgruber et al. 2020; Durbin et al. 2018; Decaesteker et al. 2018; Boeva et al. 2017; Groningen et al. 2017; L. Wang et al. 2019), and our idea that both high-risk group might share regulatory programs by involvement of bHLH TFs as described in MNA-amplified genomes (Zeid et al. 2018). Novel regulatory targets that may be important for the formation and development of these tumors, and for therapeutic intervention were also identified.

Neuroblastoma is mostly driven by segmental aberrations (Depuydt et al. 2018), and in case of MYCN amplified samples also other rearrangements, such as formation of extrachromosomal circular DNA containing amplified MYCN and its enhancers (Helmsauer et al. 2020; Richard P. Koche et al. 2019). More than 75% of our DMRs (data not shown) overlap with copy number variation gain and losses found by Depuydt et al. 2018. Additionally our study was limited to DNA motifs from the JASPAR database, and could be extended by incorporating motifs from other databases as well. With more WGBS samples available, a bigger picture of DNA methylation driven regulatory signatures could be revealed by extending our regression-based analysis to multi-omics approach. It could be done by adding to the response variable copy number variation information, open chromatin, and/or gene expression (e.g. regularized multi-task learning (Evgeniou and Pontil 2004)), or by applying increasingly popular and more powerful deep learning approaches (Kopp, Monti, et al. 2020; Ronen, Hayat, and Akalin 2019).

To date, there is still no universal, effective epigenetic biomarker allowing for the diagnosis and prognosis of NB (Jubierre et al. 2018). The distinct sub-type specific methylation patterns,

together with activity of crucial transcription factors in neuroblastoma regulatory networks driven only by DNA methylation suggest that epigenetic mechanisms are involved in NB deregulation programs, and they can be attractive targets for the therapeutic intervention in the future.

3

Other modes of transcription regulation that affect DNA methylation

DNA methylation at CpG dinucleotides is an essential epigenetic mark and it is assumed that there is an evolutionary reason behind its existence. Methylated CpG residues undergo spontaneous C-to-T deamination that results in a G:T mismatch, which over evolutionary time has resulted in the depletion of CpG residues throughout much of the genome. An exception occurs at the 5′UTR end of many genes, which is known to be enriched in CpG-rich regions, such as CpG islands and regulatory DNA. These CpG-rich clusters are typically maintained in an unmethylated state; even in cells in which the linked genes are not expressed. This phenomenon has been linked to many mechanisms, including presence or absence of specific sequence motifs, local histone modifications, RNA polymerase occupancy, and local nucleosome architecture (Deaton and A. Bird 2011).

Another potential factor that might explain the retention of promoter CpG islands in a DNA-methylation-free state is the formation of persistent RNA-DNA hybrids (Paul A Ginno et al. 2012). RNA-DNA hybrids or so called R-loops, were already discovered in early 90s, and shown to form during transcription when nascent RNA is in close proximity to its DNA template at the transcription site (Bhattacharyya, Murchie, and Lilley 1990; Sanz et al. 2016). However, just recently it has been shown that R-loops are enriched at loci with decreased DNA methylation, increased DNase hypersensitivity (open chromatin regions) (Sanz et al.

2016), and binding sites of RNA/DNA-binding proteins, helicases, and factors implicated in DNA damage (Cristini et al. 2018; Nadel et al. 2015). The formation of R-loops at CpG island promoters is favored by local sequence features (the G-rich nascent RNA base pairing with a C-rich template strand of DNA; called GC skew), which manifest as poor substrates for DNA methyltransferases (Grunseich et al. 2018; Paul A Ginno et al. 2012). R-loops might also form at the 3′UTR end of a number of human genes (P. A. Ginno et al. 2013) and play a key role in transcription termination (Konstantina Skourti-Stathaki, Nicholas J. Proudfoot, and Gromak 2011), which together, suggests that they may posses more general physiological roles (K. Skourti-Stathaki and N. J. Proudfoot 2014).

In the following paragraphs, I will present our study that explores genomic regions that have been reported to be targets of an extreme abundance of transcription factors, called Highly Occupied by Transcription factors regions (HOT regions). Previous studies on HOT regions focused mainly on partially unsuccessful pull-downs of transcription factors by ChIP-seq protocols (D. Jain et al. 2015; Xie et al. 2013; R. A. J. Chen et al. 2014; D. Park et al. 2013; M. B. Gerstein et al. 2010), however, our results support the view that the unspecific enrichment in multiple ChIP-seq experiments can be at least partially explained by specific biological properties and co-occurrence with secondary DNA structures, such as R-loops and G-quadruplexes. Finally, HOT regions, similarly to R-loops, are shown to exist in DNA-methylation free state. Hence, the understanding the genome-wide DNA methylation dynamics is crucial in the study of transcription factor binding sites on the DNA.

The content of the next paragraphs is adapted from Wreczycka, Franke, et al. 2019. Most text is reproduced verbatim, but has been rearranged and edited to fit this format.

3.1 Introduction

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is now a standard method to quantitatively assay the binding sites of a DNA binding protein in the genome. First large scale projects such as ENCODE (ENCODE Project Consortium 2012) and modENCODE (Celniker et al. 2009) used ChIP-seq technology to find binding sites of hundreds of proteins binding to DNA. With more ChIP-seq data available to profile transcription factor binding events, defined as total number of transcription factors bound to genomic regions, it has

become apparent that certain parts of the genome harbour an unusually high degree of overlap between the binding sites of different transcription factors. These regions are called high-occupancy target (HOT) regions and have been observed in multiple species (Mark B Gerstein et al. 2010; Xie et al. 2013; Boyle et al. 2014). Importantly, HOT regions are are enriched in binding events without canonical motifs (Yip et al. 2012). They are thought to have biological importance due to high number of binding sites observed, and are linked to developmental enhancers and distinct regulatory signatures in *Drosophila melanogaster* (Kvon et al. 2012), but previous reports failed to assign a clearly distinctive, conserved function that would explain the requirement for the exuberant number of bound transcription factors.

In this study, we aim to gain a deeper understanding of the nature of HOT regions and the genomic features associated to them. First, we investigate conserved features of HOT regions across different species. To date, there has been no cross-species comparison of HOT regions in terms of sequence features. The sequence features that are shared across species can provide a mechanistic insight into HOT region formation, and enable prediction of HOT regions in other species. With the sequence analysis and subsequent integrative analysis, we primarily aim to uncover the rationale behind the propensity of HOT regions to have unusual number of binding events, many of which are motifless binding events (transcription factors binding to a region without a known motif) (Yip et al. 2012). For us, the plausible explanations for motifless binding are a combination of 1) interaction of transcriptions factors (TFs) where only a handful of them are actually binding to DNA 2) existence of weak binding sites where TFs bind to non-canonical motifs in a weak manner 3) regions with high-affinity for chromatin immunoprecipitation called 'hyper-ChIPable' regions (L. Teytelman et al. 2013). Many of the HOT regions are shown to bind hundreds of proteins based on ChIP-seq experiments (Boyle et al. 2014). Detection of hundreds of proteins occupying an individual HOT region could be explained by extensive protein interaction networks between transcription factors and cofactors, where only a few factors directly bind to DNA. However, only a handful of such interactions were experimentally validated (Xie et al. 2013). Therefore, we seek additional explanations for existence of HOT regions in the genome and their association with motifless binding.

3.2 Methods and Data

ChIP-seq data used to define HOT regions Analyses of human TF binding sites were performed using the UCSC *Homo sapiens* hg19 reference genome, *Mus musculus* mm9, *Drosophila melanogaster* dm3 and *Caenorhabditis elegans* ce10. ChIP-seq files in narrowPeak format were downloaded from the ENCODE (`www.encodeproject.org`) and modENCODE (`data.modencode.org`) portals. Human TF binding sites in narrowPeak format were downloaded from the UCSC Uniform track (`http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/`). In total, 166 *Homo sapiens* TFs, 42 *Mus musculus*, 42 *D. melanogaster* and 83 *C. elegans* were obtained.

Calling HOT regions For a given set of ChIP-seq peaks per species, we determined the summits of the peaks. Following that, we calculated the density of the summits over the genome using 500 bp sliding windows. We calculated the local maxima of the density vector for each chromosome. We made sure that the local maxima of the density vector are the only maxima in 2000 bp surrounding the maxima for human and 1000 bp for other species. This is necessary to remove sub-optimal maxima around the real maxima. 2000 bp threshold was specifically applied for human datasets due to high number of experiments creating multiple local maxima around the real maxima. We then ranked these maxima based on the density scores, which is effectively the number of overlapping ChIP-seq peaks and represents the TF occupancy. These density scores are referred to as TF occupancy throughout the text. We used 99th percentile threshold to define the HOT regions. This is in line with previous methods (M. B. Gerstein et al. 2010). HOT regions were called using only the regulatory peak sets (no RNA polymerase datasets were included). The regions that are not selected as HOT regions are binned according to their TF occupancy percentiles (number of ChIP-seq peak counts) and used as controls in follow-up analyses. Scripts for all analysis are publicly available at `https://github.com/BIMSBbioinfo/HOT-or-not-examining-the-basis-of-high-occupancy-target-regions`.

Assigning HOT regions to genes, expression and open chromatin analysis Distance from HOT regions to the nearest transcription start sites was analysed using GREAT (McLean et al. 2010). Expression values of genes associated with HOT regions across tissues were

obtained from the Expression Atlas EBI database (`www.ebi.ac.uk/gxa`) and fantom5 CAGE expression ("A promoter-level mammalian expression atlas" 2014) (Supplementary Figure B.2A) and from the RoadmapEpigenomics (`https://egg2.wustl.edu/roadmap/data/byDataType/rna/expression/57epigenomes.RPKM.pc.gz`, Figure 3.2B). Accessible sites were obtained from the ENCODE DNAse-seq K562 cell line (ENCFF248FIZ).

Sequence analyses of HOT regions Extraction of 2-, 3- and 4-mers, CpG frequencies (sum of observed G and C divided by length of genomic region equal to 2000), GC skew, the observed/expected ratios for CpG on HOT regions were computed using scripts written in R version 3.3.1 and BSgenome package. We used the following genome assemblies: mm9, hg19, ce10, dm3. The observed/expected ratios for CpG were calculated according to the formula: $[f(CG)/f(C)f(G)] \cdot width(genomic\_region)$, where f denotes the observed frequency of the given mono- or di-nucleotides. De novo motifs were found using R package motifRG (Yao 2017).

MotifRG is a discriminative motif analysis tool which searches for overrepresented motifs in a positive set when compared to a negative set. Sequences from HOT regions were used as the positive set, while the negative set was constructed by sampling equal number of sequences from the HOT and non-HOT regions. Motif analysis was performed separately on HOT regions from *Homo sapiens*, *Mus musculus*, *D. melanogaster* and *C. elegans*. MotifRG was run with the following parameters: start.width = 6, both.strand = TRUE, mask = TRUE, enriched.only = TRUE.

Elastic net construction and PCA for discrimination of HOT regions For training, we used HOT regions defined as regions with TF occupancy percentiles higher than 0.995 for hg19 and 0.99 for other organisms. In order to use as a control, regions with TF occupancy lower than 85th percentile, were sampled matching the number of selected HOT regions. HOT and control regions for mm9 and hg19 were CpG sampled, in order to ensure that the ratio of HOT and control regions that overlap CpG islands were the same. The genomic coordinates of the CpG islands were downloaded from the UCSC Table browser (the cpgIslandsExt table). All models had the following set of features: CpG frequencies, ratio of observed versus expected CpGs, GC skew, and 2-, 3-, 4-mers. Feature matrix was standardized prior to training. Models

trained and tested for the same species were trained using 10-fold cross-validation. Variable importance scores were calculated for each species-specific model as an absolute value of the model coefficients, which were then normalized to a scale from 0 to 100. Average relative importance was calculated as an average of variable importance scores of all models. Area under the ROC curve (AUC) was computed to measure the accuracy of the models. We used elastic net function from the glmnet R package (Friedman, Hastie, and Tibshirani 2010; Simon et al. 2011). For the PCA, we used top 10 features ranked by the average relative importance. Using these same features for all species, we calculated PCA and plotted the color coded scatter plot on principal components for each species. For illustration purposes, we sampled the same number of 'COLD' regions as the number of 'HOT' regions.

Data processing and visualization of KO ChIP-seq, DRIP/RDIP-seq and G4 ChIP–seq samples Fastq files of KO ChIP-seq experiments (see Supplementary Table B.1) were downloaded from the European Nucleotide Archive database (ENA). All fastq files have single-end reads and were uniquely mapped to murine genome version mm9 using Bowtie 1.1.12 (Ben Langmead et al. 2009) with parameters: -p 3 -S -k 1 -m 1 –tryhard -I 50 -X 650 –best –strata –chunkmbs 1000. The bbduk program from the BBMap software 35.14 (B. n.d.) was used for adapter, quality trimming and filtering with parameters: minlength = 20 qtrim = r trimq = 20 ktrim = r k = 25 mink = 11 ref = 'bbmap/resources/truseq.fa.gz' hdist = 1. Two of KO ChIP-seq samples NFAT1_P+I and NFAT1_None did not pass bbduk tests and were excluded from the further analysis. The FastQC 0.11.3 program was used for quality control. Conversion from SAM to BAM file format, sorting and indexing BAM files was done using samtools 0.1.19 (H. Li et al. 2009; H. Li 2011), conversion from BAM to BED file formats and then BED to BedGraph file formats using Bedtools-2.17.0, from BedGraph to BigWig file format using BedGraphToBigWig v4 (Kent et al. 2010). The same pipeline was used for DRIP-seq and RDIP-seq samples, and G4 ChIP-seq (due to lack of detected adapters, bbduk argument 'ref' that indicates a path to adapters was omitted).

The R package genomation (Akalin, Franke, et al. 2015) was used for calculating fold enrichment of KO ChIP-seq samples and plotting heatmaps. Fold enrichment of KO ChIP-seq samples was defined as log2 of IP signal divided by control per base pair. Out of the total 25

KO ChIP-seq samples, 15 are positively and 9 are negatively associated with TF occupancy scores. Heatmaps were binned on x-axis into 50 bins, average for each bin was taken, and winsorized to limit extreme values <0.5 and >0.99 percentile. Some of KO ChIP-seq samples are conditional knockouts using Cre-lox recombination system (See Supplementary Table B.1). Enrichment presented as boxplots of KO ChIP-seq, DRIP/RDIP-seq, G4-ChIP-seq samples (Figure 3.5B and 3.6A, B, D) was calculated as the logarithm base 2 of the number of reads from IP sample overlapping HOT regions (normalized for library size and multiplied by counts per million) divided by the number of the reads from the control sample overlapping HOT regions normalized in the same way. If control was not available, then IP with RNaseH treatment was treated as control. For visualisation purposes, windows on heatmaps were sampled: 3000 windows from 0 to 75th TF occupancy percentile, 3000 from 75th to 99th, and 3000 from 99th to 100th percentile.

IgG samples corresponding to the following antibody ENCAB000AOJ were downloaded from ENCODE. Samples marked with 'extremely low read depth' were removed from the analysis. Samples which belong to the same biosample term id were pooled together. Signal was visualized in a region of ±1 kb around HOT (>99th percentile), MILD (between 99th and 75th percentile), and COLD regions (below 75th percentile). Prior to visualization, the reads were extended to 200 bp in a stranded fashion, and the signal was normalized to per million reads.

Methylation dynamics on HOT regions Methylation over the regions of interest was extracted from the Roadmap Epigenomics Consortium Whole-genome Bisulfite sequencing data sets (Kundaje et al. 2015). Our regions of interest consist of HOT regions, non-HOT regions (regions with lower TF occupancy), and CpG islands not associated with HOT regions (non-HOT CGI). For each region of interest, we extracted overlapping methylation value for each cell type and calculated the mean methylation value per region. We plotted the distribution of mean methylation values for each set of regions: HOT regions, non-HOT regions (binned into different TF occupancy levels), and non-HOT CGI. For each cell type, we calculated the interquartile range and median methylation values for HOT regions and non-HOT CGI. Next, we plotted the distributions of medians and interquartile ranges across all cell types as boxplots to compare the methylation dynamics for HOT regions to non-HOT CGI.

Methylation dynamics on HOT regions with and without R-loops Percent of methylation on HOT regions in the Figure 3.5 was calculated using Bisulfite-seq from NT2 cell line from Sanz et al. 2016. The PiGx BS-seq pipeline (Wurmus et al. 2018) was used to trim, align, and calculate average methylation levels on HOT regions. R-lops were defined as DRIP-seq peaks from NT2 cell line and downloaded from Sanz et al. 2016.

PFAM domains and human protein–protein interactions Reviewed UniProt ("UniProt: a worldwide hub of protein knowledge" 2018) human protein sequences (as of 31 August 2015) were scanned for occurrence of PFAM HMM (both PFAM-A and PFAM-B) models using HMMER3 (Sean R. Eddy n.d.). The HMM scanning detected 9511 types of PFAM domains in 19 275 proteins.

PFAM (El-Gebali et al. 2018) entries for single-stranded DNA-binding domains were collected from the PFAM database by combining all members of the following PFAM clans: OB (for OB-fold domains), KH (for K-homology domains), RRM (for RRM-like domains) and sPC4-like (for the Whirly domain). The collection of these four clans contains 90 different types of PFAM domains.

Human protein–protein interaction data was downloaded from the iRefWeb database (Turner et al. 2010). In order to dissect which protein–protein interactions of TFs are direct (or relatively direct physical interactions of proteins), the interactions were filtered for the following criteria: (i) interactor A (uidA) is from taxa:9606 and interactor B (uidB) is from taxa: 9606; (ii) interaction type between uidA and uidB is one of 'MI:0915 (physical association)', 'MI:0407 (direct interaction)', 'MI:0403 (colocalization)', 'MI:0914 (association)', or 'MI:0191 (aggregation)'; (iii) both uidA and uidB have an ID mapped to UniProt accessions.

3.3 Results

3.3.1 HOT regions cover transcription start sites of stably expressed genes across cell types

HOT regions have been observed in multiple species, such as *H. Sapiens* (human) (Xie et al. 2013; R. A. J. Chen et al. 2014), *D. melanogaster* (fly) (D. Jain et al. 2015), *Saccharomyces* (yeast) (D. Park et al. 2013) and *C. elegans* (worm) (R. A. J. Chen et al. 2014; M. B. Gerstein et al.

2010). In order to find HOT regions, we used available ChIP-seq data from the ENCODE and modENCODE databases (for details see Methods and Data). Based on density of ChIP-seq peaks used as a measure of TF occupancy, we detected 4324 HOT regions in human, 2638 mouse, 422 worm and 408 fly out of 428498 regions with at least one transcription factor peak on a HOT region in human, 245250 in mouse, 40921 in worm, and 37853 in fly, respectively (see Figure 3.1). HOT regions are accessible online via UCSC track hub at `https://bimsbstatic.mdc-berlin.de/hubs/akalin/HOTRegions/hub.txt`.

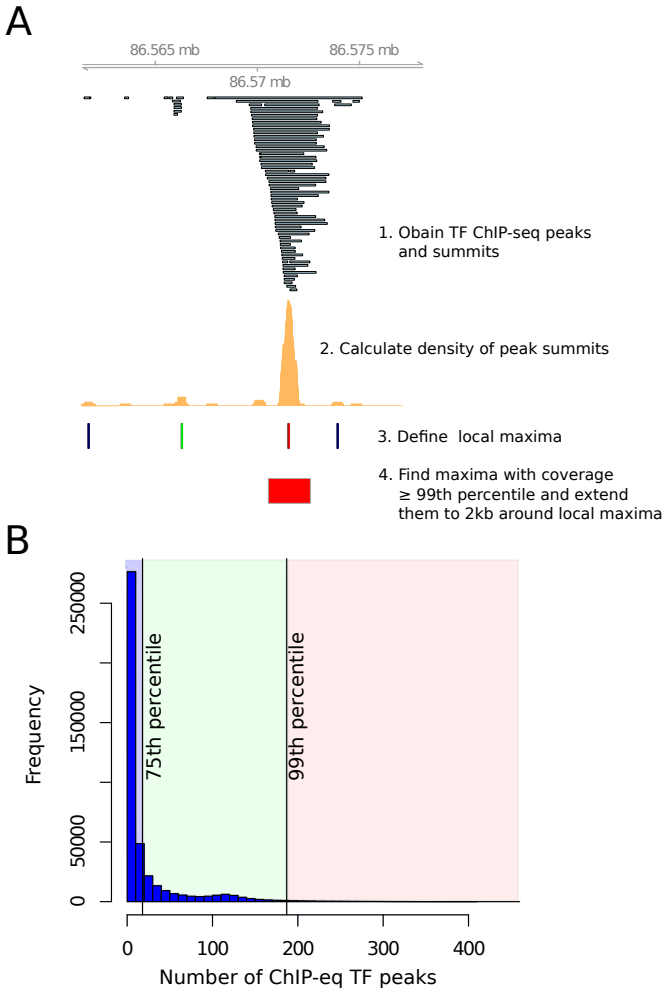


Figure 3.1: (A) Schematic workflow of HOT region definition. (B) The barplot indicates number of ChIP-seq peaks in HOT (red), MILD (green) and COLD (blue) regions. This figure is reproduced from Wreczycka, Franke, et al. 2019.

HOT regions are typically located at promoters. The majority of HOT regions (80%) are in close proximity to TSS (within 5 kb) (Figure 3.2A). In human and mouse, they are

mostly associated with CpG islands (Figure 3.3A), and have higher CpG content and CG frequency than COLD and MILD regions (Figure 3.3B). Genes associated with HOT regions are stably expressed across cell types and tissues, with variability similar to housekeeping genes (Figure 3.4B). Gene expression levels for these genes are generally above the median level of expression for all genes in the respective cell types. Gene Ontology (GO) (McLean et al. 2010) analysis revealed a variety of biological processes highly represented in HOT region-associated genes such as RNA processing, ncRNA processing, ncRNA metabolic process and ribosome biogenesis (Supplementary Figure B.2), which is in line with the findings reported by (Xie et al. 2013). Additionally, although we observe a marginal association between HOT regions and chromatin accessibility, chromatin accessibility alone is not sufficient to explain HOT region formation (Figure 3.2C). Therefore, having knowledge that a region is highly accessible provides no information on whether the region is HOT, likewise, having information that a region is HOT provides no information on how accessible the region is. Consistent with the published features of HOT regions (Xie et al. 2013; Boyle et al. 2014), our findings confirm that genes associated with HOT regions are mostly housekeeping genes - they are required for the maintenance of basic cellular functions and are constitutively expressed.

We hypothesized that sequence characteristics of HOT regions may be shared across species which could explain the existence of HOT regions in multiple species. Therefore, we built machine-learning models that can discriminate HOT regions from non-HOT or so called 'COLD' regions using sequence features in human, mouse, worm and fly. These machine-learning models are primarily used for identifying sequence features that are predictive of HOT regions. As input we used used 2, 3 and 4 bp long k-mer frequencies of HOT and COLD regions, GC content, and CpG observed/expected ratio (O/E ratio). CpG islands are a frequent feature of HOT regions in human and mouse. Although, worm and fly do not have CpG islands, CpG enrichment could be important at least for worm, for which HOT regions are enriched for CpG dinucleotides (R. A. J. Chen et al. 2014). We built a predictive model of "hotness" of genomic regions using a penalized multivariate regression method (Zou and Hastie 2005). For each organism, we built species-specific models using normalized feature matrices as inputs. We achived high accuracy for all models: cross-validation AUC between 0.82 and 0.94 for all the models. The top 10 feature importance averaged across species shows that

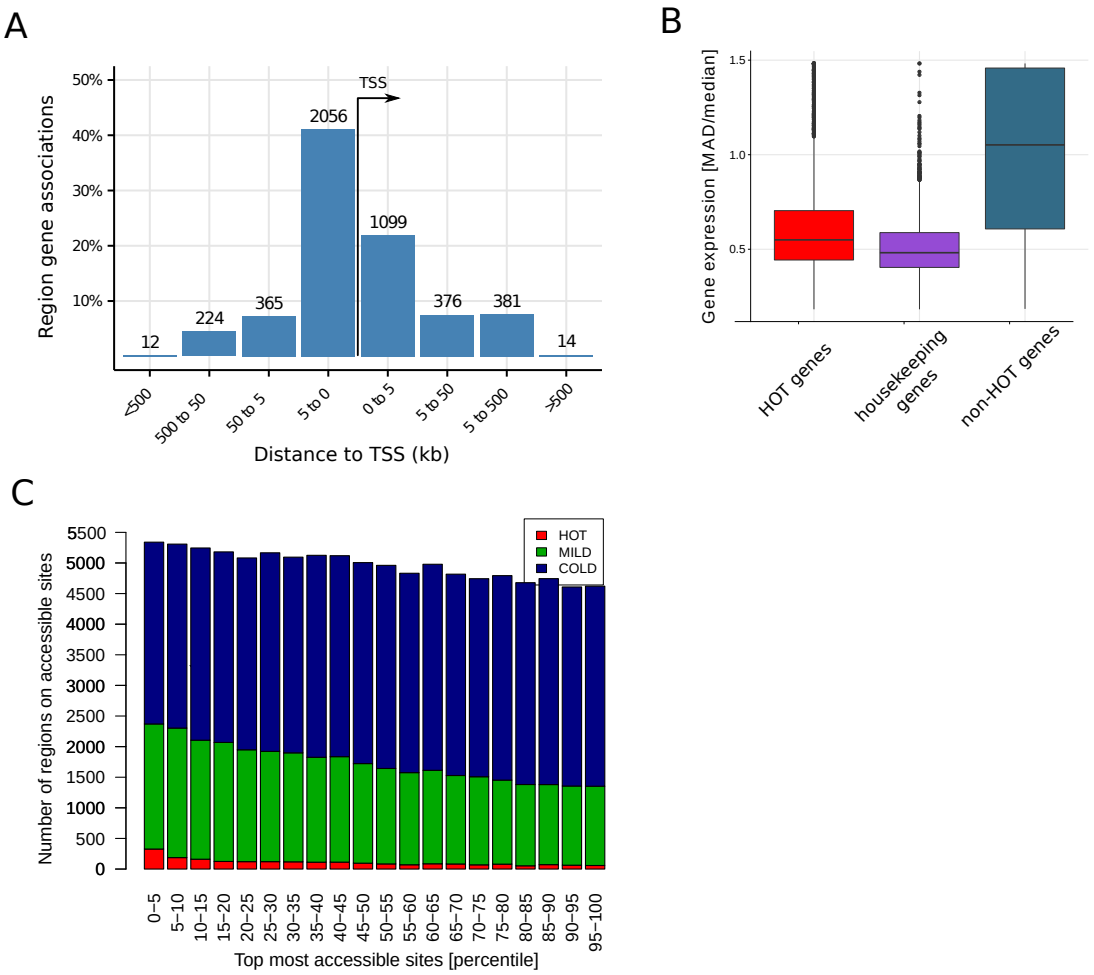


Figure 3.2: (A) HOT regions are located mostly close to transcription start sites and are promoter associated. The figure shows binned orientation and distance between HOT regions and the nearest genes. Associations precisely at 0 refers to the transcription start site of the nearest gene. (B) Gene expression variation on HOT regions. Variation of expression of genes associated with HOT regions is as low as housekeeping genes, and expression is less variable than non-HOT genes. Median absolute deviation (MAD) and median was calculated for each gene across 57 human cell lines and tissues from the Roadmap Epigenomics database. (C) Chromatin accessibility on HOT regions. DNase-seq peak set from K562 cell line was ranked according to their signal value. On X axis are percentiles of DNA-seq peaks according to their ranks, on Y axis percent of HOT regions that overlap DNA-seq peaks. This figure is reproduced from Wreczycka, Franke, et al. 2019.

CpG and GC rich k-mers together with CpG O/E ratio are the most important predictors for all the models (see Figure 3.4A for feature importance across all models, and Figure B.3 for individual models). The most predictive features averaged between all species are sufficient to discriminate HOT from COLD regions. Although localized CpG and GC spikes across genomes of worm and in fly they are not common, we could discriminate HOT and COLD regions across all four species using the same GC/CpG rich top features. We applied principal

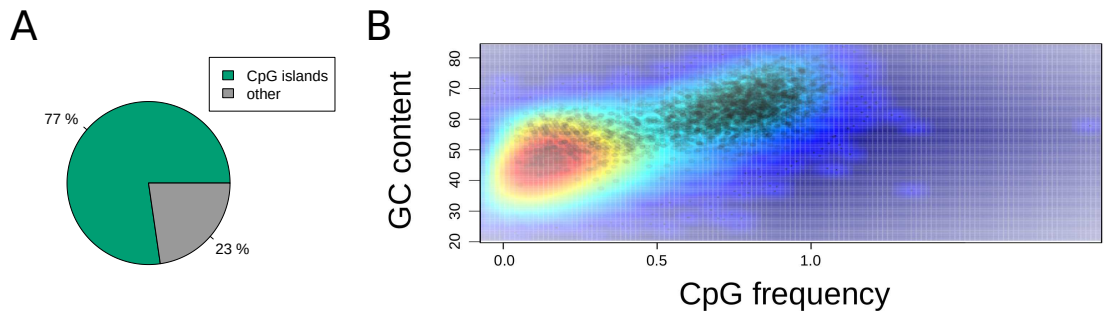


Figure 3.3: HOT regions and CpG-dense regions. (A) Most HOT regions overlap with CpG islands. This figure is reproduced from Wreczycka, Franke, et al. 2019. (B) HOT regions (depicted in black dots) have high CpG frequency and high GC content in comparison to control regions ("MILD" and "COLD" regions are marked depending on their density; from the highest density to the lowest: red, yellow and blue)).

component analysis to visualize the discrimination between HOT and COLD regions using the top features (Figure 3.4B). To determine whether there are higher order sequences that differentiate between HOT and non-HOT regions, we performed discriminative de novo motif analysis on HOT regions from all four species. The resulting motifs were short (5-6 bp with high information content), GC and CpG dominant (Supplementary Figure B.3B). The motifs partially matched binding sites of known transcription factors which bind GC rich sequences, such as SP1 known to bind G-skewed DNA motifs and binds preferentially to the intra-strand G-quadruplex structure in vitro (Raiber et al. 2011).

3.3.2 Enrichment of ChIP-seq signal for knock-out transcription factors in HOT regions

Upon observing common low-level sequence features of HOT regions across species, we investigated whether potential technical biases in ChIP-seq could at least partially explain false positive signals on HOT regions. Previous studies suggest that even if the ChIP-ed protein does not exist in the analysed sample, highly expressed loci might give rise to false-positive peaks in yeast (L. Teytelman et al. 2013) and fly (D. Jain et al. 2015). In order to address this question with a more comprehensive collection of datasets, we downloaded all publicly available experiments where the ChIP-ed transcription factor was not physically present in the cell, as the gene encoding the transcription factor was 'knocked-out'. This set consists of 43 ChIP-seq experiments for knock-out (KO) transcription factors (KO ChIP-seq), only 24 experiments have

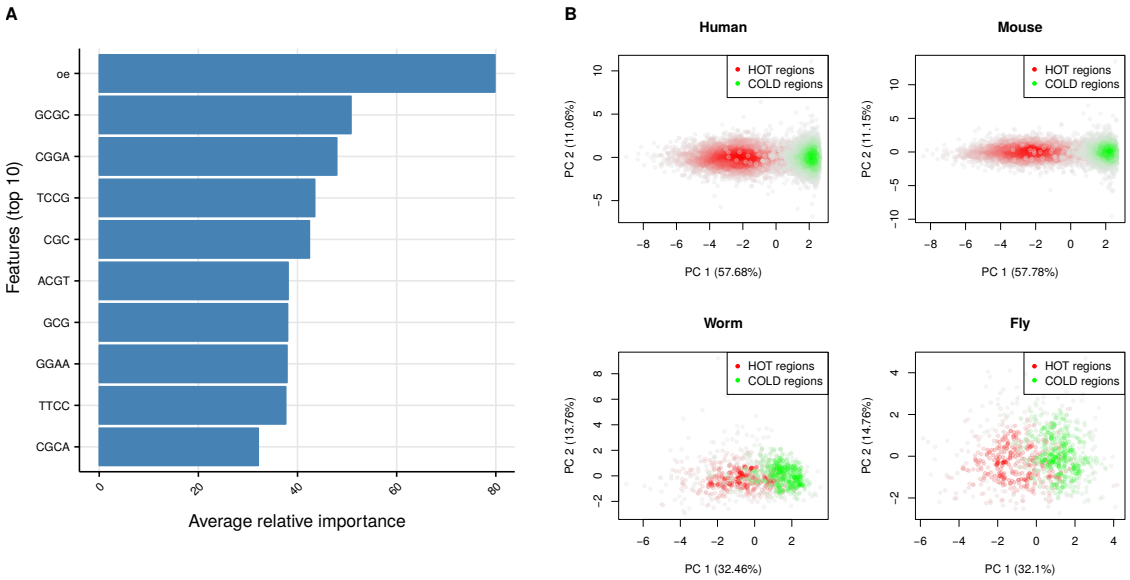


Figure 3.4: k-mer properties of HOT regions. (A) Top 10 features ordered in relative importance averaged across species. Importance scores are scaled to 0–100 scale for each species then averaged. Top feature [oe] means observed/expected ratio for CpG dinucleotides. (B) Principal component analysis using top 10 features shown in A. PCA is carried out for human, mouse, worm and fly separately. Scatter plots using first two principal components are shown, each dot represent HOT and COLD regions. The scatter plot is colored based on density of points, the more dense the points the darker the color. This figure is reproduced from Wreczycka, Franke, et al. 2019.

a control experiment in the form of input DNA or mock-IP (See Supplementary Table B.1 for accession numbers and details). These experiments were carried out by different labs, therefore it reduces the lab-specific bias for KO generation and ChIP-seq experiments. More than half of the KO ChIP-seq experiments show a clear signal enrichment (measured as IP/control) over HOT regions. KO ChIP-seq experiments with strong enrichment on HOT regions are presented in Figure 3.5A and experiments without signal enrichment are shown in Supplementary Figure B.5A. The signal is absent from regions that do not have extreme enrichment of TF binding events. Pooling all of the available signal enrichment for the KO ChIP-seq experiments with strong enrichment on HOT regions also shows the trend where signal enrichment, on average is higher for HOT regions (see Figure 3.5B for signal enrichment of HOT regions and other control regions binned based on their TF occupancy percentiles). We compared ChIP-seq peaks in the wild-type experiments and we observed that KO and WT ChIP-seq scores have a strong correlation on HOT regions. The magnitude of the correlation between WT and KO signal strength indicates that in most cases all WT peaks overlapping HOT regions represent a potential bias in ChIP experiments (Supplementary Figure B.4). We also investigated the

possibility that the signal in the KO experiments might originate from pulldown of highly related proteins—from paralogous transcription factors containing similar epitopes. Out of 24 proteins used in the KO experiments, only seven have known paralogues; eliminating the possibility of this confounding variable (Supplementary Table B.3).

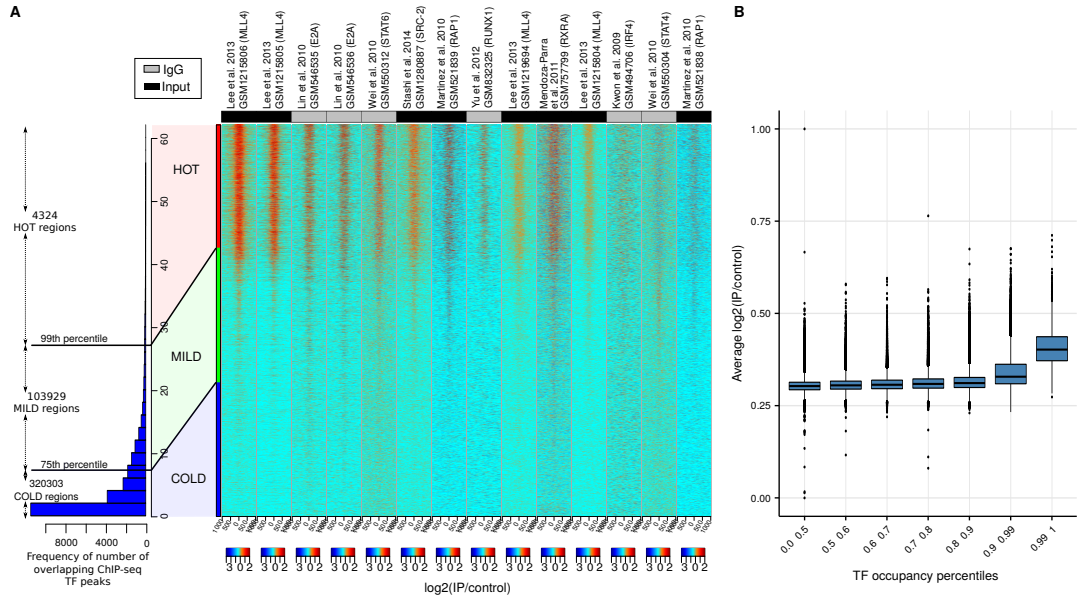


Figure 3.5: (A) Heatmaps show KO ChIP-seq experiments with signal on HOT regions. The barplot shows the distribution of the number of ChIP-seq peaks over all bound genomic regions. Regions overlapping >26 peaks sites (top 1 percentile) are labeled as HOT (4324 regions). The values in the heatmaps are log2(KO – IP/control). The status of the control sample input DNA or IgG is color coded on the top of the heatmaps. (B) Boxplots show distribution of average log2(IP/control) values for different sets of TF occupancy percentile bins. The log2(IP/control) values for each region in A is averaged across KO ChIP-seq experiments. The rightmost boxplot represents HOT regions, TF occupancy >99th percentile, and the rest represent control region sets with different TF occupancy percentiles. This figure is reproduced from Wreczycka, Franke, et al. 2019.

In addition, we noticed that some KO ChIP-seq experiments used IgG 'mock' ChIP-seq as control. The IgG ChIP-seq experiments should ideally control for unspecific binding that could potentially cause a false positive signal, and yet more than half of KO ChIP-seq experiments that use IgG ChIP-seq as control show signal enrichment on HOT regions (see Figure 3.5A). Following up on this, we wanted to see whether the HOT regions show an enrichment of signal in IgG control experiments. We downloaded available IgG control experiments from ENCODE, where antibodies from the same vendor was used in multiple cell types (results shown in Supplementary Figure B.5B). HOT regions showed a consistent enrichment in multiple IgG

experiments, however, the enrichment was weak and showed variability, which was dependent on the cell type (Supplementary Figure B.5C).

3.3.3 Association of HOT regions with R-loops and G-quadruplex DNA

We next investigated the associations of HOT regions with other GC rich features of the genome. One of such features that shares the same type of annotation with HOT regions such as CpG islands are R-loops. An R-loop is a nucleic acid structure that is composed of an RNA–DNA hybrid and a displaced single-stranded DNA (Santos-Pereira and Aguilera 2015). Their formation and stabilization are associated with GC content, CpG islands (P. A. Ginno et al. 2013) and G-quadruplexes (K. Skourti-Stathaki and N. J. Proudfoot 2014). R-loops exist across a broad spectrum of species from bacteria to high eukaryotes and are shared across mammals (Sanz et al. 2016; X. Li 2006). R-loop accumulation is a source of replication stress, genome instability, chromatin alterations, or gene silencing. They are associated with cancer formation and a number of genetic diseases (Santos-Pereira and Aguilera 2015).

R-loops can be detected genome-wide using a method called RNA–DNA immunoprecipitation followed by sequencing (DRIP-seq). It involves immunoprecipitation and sequencing of DNA fragments using the RNA–DNA hybrid specific S9.6 antibody (Paul A Ginno et al. 2012), which was developed by extensively testing for specificity to RNA–DNA hybrids (Lima et al. 2016). We analysed publicly available DRIP-seq datasets to investigate R-loop enrichment on HOT regions (Sanz et al. 2016; Lim et al. 2015; Zeller et al. 2016) (See Supplementary Table B.2 for accession numbers). We observed R-loop enrichment on HOT regions in every analyzed cell line, compared to non-HOT region sets, binned based on their TF occupancy percentiles (Figure 3.6A). We observed this enrichment even when the DRIP-seq experiments with RNAseH treatment were used as controls. The RNAseH treatment removes R-loops and subsequent DRIP-seq experiment results in depleted signal for R-loops. It indicates that the S9.6 antibody most likely binds specifically to R-loops and does not show additional interactions with other forms of DNA and DNA-binding proteins. In addition, we observed DRIP-seq enrichment on HOT regions in *C. elegans* (Figure 3.6B). Together, these results suggest that R-loops, across different species, overlap with HOT regions.

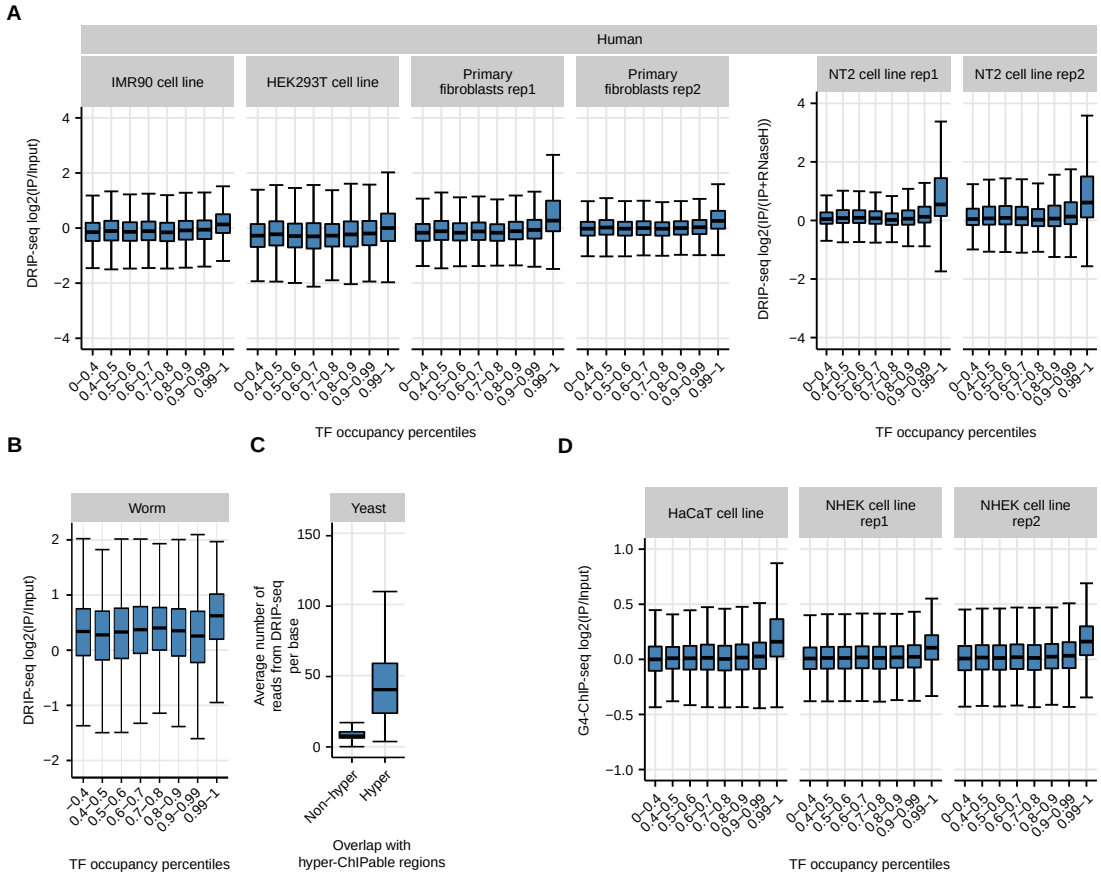


Figure 3.6: R-loops are associated with HOT regions. The boxplots show DRIP-seq log2(IP/control) for HOT regions and control regions binned based on their TF occupancy percentile in (A) Various human cell lines and in (B) worm. Boxplots show DRIP-seq read count per base-pair for hyper-ChIPable regions and all other genes as controls. (C) HyperChIP-able regions in yeast are enriched in R-loops. (D) HOT regions are enriched with G-quadruplex DNA (G4-ChIP-seq). Boxplots show log2(IP/control) for HOT regions and control regions binned based on their TF occupancy percentile. This figure is reproduced from Wreczycka, Franke, et al. 2019.

R-loops colocalize with G-quadruplex DNA (G4) which is a tertiary structure of single-stranded DNA (Q. Y. Lee et al. 2020), and G4-stabilized R-loops leads to increased transcription by a mechanism involving successive rounds of R-loops formation (Q. Y. Lee et al. 2020). G4 forms on the displaced single-stranded G-rich DNA on the opposite side of the R-loop. Therefore, in addition to R-loop enrichment analysis, we calculated the enrichment of G4 on HOT regions by analysing G4-ChIP-seq experiments (Hänsel-Hertsch et al. 2016). To summarize, we observed enrichment of G4 signal on HOT regions, which is consistent with R-loops localization on HOT regions (Figure 3.6D).

In addition, we would also expect to see R-loops in hyper-ChIPable regions in yeast,

originally defined by (L. Teytelman et al. 2013). Indeed, we see enrichment of DRIP-seq signal (Wahba et al. 2016) on published hyper-ChIPable regions (Figure 3.6C).

Since R-loops are associated with HOT regions, occupying TFs must be able to bind RNA–DNA hybrids or single-stranded DNA (ssDNA). Therefore, we investigated whether the TFs assayed by ENCODE have ssDNA binding or RNA–DNA hybrid binding domains, or such GO term annotations (see Supplementary Table B.3). Out of 165 studied TFs in human, only two of them contain at least one of the ssDNA binding domains: BACH1 contains a 'DUF1866' domain (from the RRM clan) and E2F6 contains a BRCA-2_OB3 domain (from the OB clan). Furthermore, none of the 165 TFs have an annotation of the GO term 'single-stranded DNA binding' (GO:0003697). When considering the direct interaction partners of these TFs, 31 out of 165 TFs (18.8%) have at least one direct interaction partner with an ssDNA-binding domain and 11 out of 165 TFs (6.7%) have at least one direct interaction partner with the GO term annotation for ssDNA binding. On the other hand, Cauli_VI domain that mediates the binding of RNASEH1 to RNA/DNA hybrids, is annotated only for two proteins in the whole proteome (RNASEH1 and Ankyrin repeat and LEM domain-containing protein 2 (ANKLE2)) and none of the human proteins have the associated GO term 'DNA/RNA hybrid binding (GO:0071667)' (according to the reviewed UniProt sequence annotations). Therefore, we could not detect any association of TFs or TFs' interaction partners with RNA/DNA hybrid binding function.

3.3.4 Stable hypo-methylation of HOT regions across cell types

We investigated the CpG methylation dynamics on HOT regions, using base-pair resolution methylation data across multiple human cell types. Most of HOT regions are associated with CpG islands and genes with above average expression levels and hence we would expect low methylation on HOT regions (Deaton and A. Bird 2011). Consistent with this observation, we observed hypo-methylation on HOT regions, compared to "MILD" and "COLD" regions. The median methylation levels of HOT regions were similar to the median methylation levels of CpG islands not associated with HOT regions (non-HOT CGI) (see Figure 3.7A for an example cell line, see Supplementary Figure B.6 for all analyzed cell lines). Interestingly, non-HOT CGI had higher variation of DNA methylation than HOT regions despite the median

methylation for both sets being low. This trend was evident in all examined cell types. Across the cell types, non-HOT CGI had 3–4 times higher methylation variation than HOT regions. It indicates that although HOT regions are associated with CpG islands, they are different from non-HOT CGI in their methylation dynamics and they maintain low DNA methylation levels across different cell types (see Figure 3.7B).

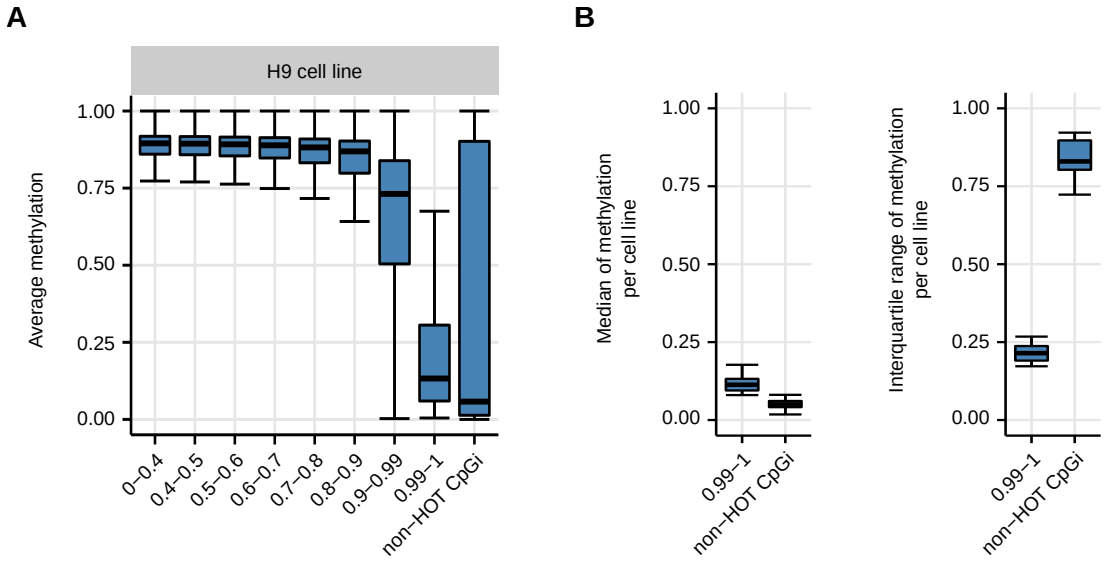


Figure 3.7: (A) HOT regions are hypo-methylated in comparison to controls in H9 cell line. Boxplots show distributions of methylation for HOT regions (rightmost boxplot) and control regions binned based on their TF occupancy percentile. (B) Left boxplot shows distributions of methylation medians across cell types for HOT regions and CpG islands that are not associated with HOT regions (non-HOT CpGi). Right boxplot shows distributions of methylation IQRs (interquartile ranges) across cell types for HOT regions and non-HOT CpGi.

In addition, hypo-methylated CpGs are prevalent in R-loops, and the R-loop associated loci are protected from de novo DNA methylation by preventing binding DNA (cytosine-5)-methyltransferases (DNMTs) (Paul A Ginno et al. 2012; Grunseich et al. 2018). Another mechanism that favours DNA hypomethylation on R-loops is incorporating ten–eleven translocation (TET) DNA demethylases which induces local DNA demethylation (Arab et al. 2019). Hence, both HOT regions and R-loops prefer hypo-methylated regions. We hypothesized, that R-loops might be a factor inhibiting DNA methylation on HOT regions. Our results suggest that it's probably not the case on CpG islands, because HOT regions were in unmethylated state no matter if they overlap with R-loops. HOT regions that were not on CpG islands and overlapped with R-loops were highly methylated and located mainly on introns and terminators.

They tended to be closer to their nearest 3' neighbor compared to HOT regions not overlapping with CpG islands and R-loops (median distances 0.2kb versus 8.8kb, respectively, p-value = 1.855e-09), which is inline with one of the features of terminal R-loops that is to form in close proximity of colliding genes (Sanz et al. 2016). To summarize, from our analysis it is unclear whether R-loops prevent DNA methylation on HOT regions, neither on CpG islands or other regulatory regions (Supplementary Figure B.7).

3.4 Discussion

HOT regions are locations in the genome with remarkably high occupancy of transcription factors. They are formed by the combination of topmost ranking peaks from hundreds of ChIP experiments. HOT regions are mostly associated with promoters of stably expressed genes. They are located in open chromatin regions, however, DNA accessibility does not explain their formation. We showed that the low-level sequence features, such as GC rich and CpG containing k-mers, are shared across HOT regions of different species. Most interestingly, we demonstrated that HOT regions are specifically enriched with false positive signals, using KO transcription factor ChIP-seq. These false positive signals are antibody dependent since KO ChIP-seq experiments show variable intensity of signals on HOT regions. The traditionally suggested controls, such as IgG ChIP-seq, can not reliably control for these artifacts. We showed that HOT regions associate with R-loops, in multiple organisms, as well as G-quadruplex DNA structures. Our results support the view that the peaks observed on HOT regions might be produced by the unspecific enrichment in multiple ChIP-seq experiments, rather than by the pull-down of specific transcription factors.

There might be many causes for the persistent false positive signal on HOT regions. The ChIP-seq signal consists of the signal from actual binding events and the noise. The noise is usually attributed to sequencing depth, library preparation, but most importantly to antibody specificity (Kidder, Hu, and K. Zhao 2011). The observed false positive signal could be obtained through pull-down of non-target proteins; this would however require that all experimentally used antibodies cross-react with a small set of proteins which constitutively bind GC rich promoters in multiple cell lines—a scenario which is highly improbable. The degree of overlap of HOT regions with R-loops suggests another hypothesis—that the antibodies cross-react

directly with polynucleotide epitopes present in the HOT regions (Weitzmann and Savage 1994). R-loops are formed during transcription of GC rich, hypomethylated regions, where the nascent RNA strand displaces one of the DNA strands, forming an RNA:DNA Watson-Crick base pairing with the complementary strand. Such displacement causes R-loop prone regions to contain multiple polynucleotide structures: double stranded DNA, single stranded DNA, RNA:DNA hybrids, single stranded RNA (reviewed in Santos-Pereira and Aguilera 2015; Niehrs and Luke 2020), G quadruplex complexes (Kalsi et al. 1996; Duquette 2004), etc., all of which can be bound by antibodies with a range of affinities (Duquette 2004; Yiqiang Wang, Mi, and Cao 2000; Jin, Sepúlveda, and Burrone 2009; Barbas et al. 1995; Braun and J. S. Lee 1986). Anti-DNA antibodies are abundant in the serum of normal animals immunized with protein fragments (CERUTTI et al. 2005; Ugo Moens et al. 2002; U. Moens et al. 1995; Voynova et al. 2005; Deocharan et al. 2002; Sciascia et al. 2007; Marchini et al. 1995; Desai et al. 1993; Tran et al. 2003; Petrakova et al. 2009), and are frequently polyspecific (Deocharan et al. 2002; Lakamp and Ouellette 2011; Wun et al. 2001; W. Zhang et al. 2010; Reichlin et al. 1994; CAPONI et al. 2002; Yasuda et al. 2009; Kumar et al. 2011; Gaynor et al. 1997)—they can bind both polynucleotide and non-polynucleotide (e.g. peptide, phospholipid) epitopes (Barbas et al. 1995). A recent study by Lentini et al. 2018 has shown that anti-5methylcytosine antibodies nonspecifically enrich short tandem repeat sequences (Lentini et al. 2018). Abundance of epitopes in constrained genomic regions, along with the fact that the HOT regions are associated with CpG islands of housekeeping genes (which are ubiquitously expressed and form R-loops in many cellular systems), and the promiscuity of antibodies, provide a simple explanation for the ubiquity of enrichments observed on HOT regions in various ChIP-seq experiments. Serum of non-immunized, healthy animals usually contains a low percentage of anti-DNA binding antibodies. This could explain why the IgG samples, when used as controls, show a signal on HOT regions, but the intensity of the signal is much lower than from antibodies produced by deliberate immunization. The recommended experimental methods for ascertaining antibody specificity (Wardle and H. Tan 2015) control almost exclusively for binding of antibodies to non-target proteins, so the direct interaction of antibodies with polynucleotide epitopes might be an underappreciated source of false positives in ChIP-seq experiments (see our model summarized in Figure 3.8). The signal on HOT regions could additionally arise by direct binding of TFs to

single-stranded DNA (ssDNA) or RNA–DNA hybrids. Based on the current protein domain annotations, few to none of the TFs have such capabilities.

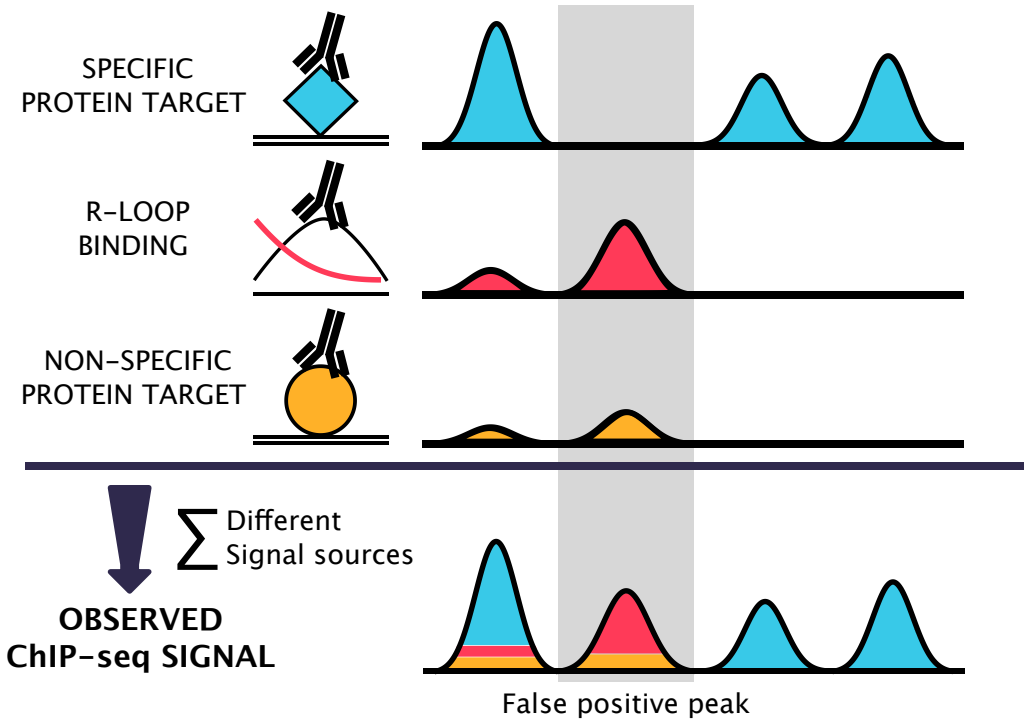


Figure 3.8: The observed ChIP signal arises from a combination of different signal sources. The signal in a ChIP experiment originates from an antibody binding to the intended target protein (blue), and nonspecific antibody binding—either to the non-target proteins (orange) or directly to polynucleotide structures, such as R-loops (red). The error (orange + red) is not proportional to the signal from the targeted protein, rather, it depends on sequence properties, antibody properties and expression characteristics of individual genomic regions. The combination of different noise profiles result in a subset of ChIP-seq peaks being false positives. This figure is reproduced from Wreczycka, Franke, et al. 2019.

In this work, we have focused on regions that show high enrichment in multiple ChIP-seq experiments. Although we provide evidence that HOT regions do not contain several dozens of bound transcription factors, the real extent of detected false positive interactions is probably not limited to HOT regions. With the currently available data, it is not possible to estimate the proportion of an antibody specific error resulting from the enrichment due to the pull-down of non-target proteins vs. the direct binding to polynucleotide epitopes. Examination of the DNA binding properties of monoclonal antibodies, for example with protein binding arrays (Stormo and Y. Zhao 2010; Bulyk n.d.), might provide the required

data for constructing more precise error models.

Lack of a strong signal over HOT regions in a subset of KO ChIP-seq samples shows that by using stringent antibody validation methods, it is possible to perform highly specific ChIP experiments. A level of prudence is needed though—a lack of signal in a KO ChIP-seq experiment might also be caused by technical conditions such as low number of reads, low library complexity or unsuccessful IP.

Our results, consistent with other recommendations (Weitzmann and Savage 1994; Wardle and H. Tan 2015; Parseghian 2013; Uhlen et al. 2016), emphasize the need for critical examination and extensive testing of antibodies prior to their experimental usage. Whenever possible, controls in ChIP-seq experiments should be performed by ChIP-ing of protein in a system where the protein is not physically present, as implemented in Knockout Implemented Normalization method (KOIN) (Krebs et al. 2014). If such controls are unfeasible, we provide lists of HOT regions and the ChIP-seq peaks overlapping with those regions for careful examination. We would like to encourage a careful, and methodical approach, where the existence of HOT regions is taken into account when performing functional association (i.e. colocalization analysis, functional enrichment) with the binding data—it is important to check whether the statistics are primarily driven by overlaps with HOT regions or not. On top of that, more stringent filtering for ChIP-seq peaks on HOT regions, such as removing peaks without canonical motifs, might be necessary.

4 Discussion

This doctoral work has resulted in several articles, including a review on strategies for analyzing bisulfite sequencing data (Wreczycka, Gosdschan, et al. 2017), an article describing a computational pipeline for Bisulfite-seq from raw fastq files to extensive reports (Wurmus et al. 2018) which I used for the pre-processing of Bisulfite-seq reads of the neuroblastoma tumor samples analysed in Chapter 2, and Bisulfite-seq samples from NT2 cell line to investigate methylation dynamics on HOT regions and R-loops in Chapter 3, and finally, an article describing the properties of HOT regions (Wreczycka, Franke, et al. 2019). Findings on DNA methylation landscape and how it is associated with transcription factor network aberrancies in high-risk neuroblastoma, described in Chapter 2, are not published at the time of writing this thesis. The methods and findings presented in the above listed publications, and this dissertation contribute to the wide field of computational detection of biological functions of DNA methylation in the context of gene regulation, and integration of DNA methylation data with other omic data in cancer and healthy cells.

4.1 High-risk neuroblastoma methylation landscape

Epigenomics is a unique approach to cancer research since it can help in the identification of groups of patients with similar epigenetic changes and characteristics in their tumors. The best known epigenetic marker is DNA methylation, and many genes and other regulatory regions

affected by this epigenetic mark have been described in several different tumor types (Hovestadt, D. T. W. Jones, et al. 2014; Klughammer, Kiesel, Roetzer, Fortelny, Nemc, Nenning, Furtner, Nathan C Sheffield, et al. 2018a; Capper et al. 2018; M. A. Kerachian, Javadmanesh, et al. 2020).

In chapter 2, I investigated DNA methylation landscape of high-risk neuroblastomas. In the recent years, DNA methylation has been investigated as a biomarker in neuroblastoma using mainly methylation arrays technology and methyl-CpG-binding domain sequencing (Henrich et al. 2016; Olsson et al. 2016; Gómez et al. 2015; Mayol et al. 2012; Decock et al. 2012; Carén et al. 2011; Buckley et al. 2011; Charlet et al. 2017; Margetts et al. 2008). In our study, we applied Bisulfite sequencing on samples from neuroblastoma tumors to uncover in a single-nucleotide, genome-wide manner, the coding and non-coding regions affected by DNA methylation abnormalities linked specifically to high-risk subgroups. We confirmed previous findings by Henrich et al. 2016 of unsupervised discrimination between high-risk and low risk groups based on only DNA methylation, and MYCN-amplified (MNA) samples forming a distinct cluster. We detected differentially methylated regions (DMRs) between high-risk and lower-risk neuroblastoma subgroups. These regions are associated with genes exhibiting neuronal activity.

MYCN amplification is the best known and the most explored high-risk neuroblastoma biomarker. We showed that our MNA DMRs are associated with MYCN activity by enrichment of MYCN binding signal (Figure 2.6), and enrichment of MYCN, c-MYC, MAX, and AP-2 DNA motifs (Figure 2.5). However, in our analysis, we did not take into account genomic rearrangements. Recently, many genomic rearrangements (Richard P Koche et al. 2020), and specifically a region that includes amplified MYCN with a proximal enhancer driven by the adrenergic core regulatory circuit or distal chromosomal fragments harboring CRC-driven enhancers (Helmsauer et al. 2020) was linked to the formation of extrachromosomal circular DNA in neuroblastoma. These amplicons were shown to have a decreased level of DNA methylation using Nanopore sequencing (M. Jain et al. 2016). The oncogenic extrachromosomal DNA might function as mobile enhancers to globally amplify chromosomal transcription (Zhu et al. 2021). Such rearrangements have previously gone largely undetected or underestimated in whole genome sequencing analyses due to lack of integrative, sequencing-based methods identifying circular DNA in tumor samples. Perhaps, in the future, these amplicons might be associated with partially or lowly methylated domains, or even larger epigenetic alterations,

such as long range epigenetic activation domains. Additionally, methylation patterns of cell-free extrachromosomal circular DNA (Sin et al. 2021) and circulating cell-free DNA (Moss et al. 2018) in neuroblastoma (Andersson et al. 2020; Su et al. 2020), might give more insights into heterogeneity, prevention, early detection and management of this cancer.

Recent reports showed that neuroblastoma exhibits at least two subtypes of tumor cells corresponding to mesenchymal/neural crest-like cells and committed noradrenergic cells (Gartlgruber et al. 2020; Boeva et al. 2017; Groningen et al. 2017). Our integration analysis of genomic and DNA methylation data suggests that DNA methylation abnormalities in MYCN amplified neuroblastomas can be explained by TFs that are driven by MYCN and exhibit committed adrenergic signature, and bHLH binding TFs. Importantly, we were interested in whether high-risk non-MYCN-amplified samples share similar regulatory and methylation patterns to the MYCN-amplified samples (Figure 2.7). We found a few key TFs in neuroblastoma development (ARNTL, JUN, and TFAP2B) that are associated with DNA methylation in both MYCN and high-risk non-MYCN-amplified samples. Additionally, we found key TFs that exhibit mesenchymal signature in non-MYCN-amplified samples.

In cancer biology, it is important to take into account and quantify the extent to which microenvironment influences tumor signatures (Blavier, R.-M. Yang, and DeClerck 2020), and decompose signals originating from tumor cells from normal infiltrating cells (e.g. mature Schwann cells in neuroblastoma (Gartlgruber et al. 2020)). Neuroblastoma trascriptome, exome, and other "-ome" studies have been performed extensively from bulk tissues, revealing a wealth of potential targets that have been evaluated over past decades (Peifer et al. 2015), but even though they reveal increasing number of findings on neuroblastoma developmental programs and the trajectory of its differentiation (Banerjee et al. 2020), they are still limited in comparison to single-cell technologies (Kashima et al. 2020; W. Chen et al. 2021). Thus, in recent years, by using the resolution of single cells to perform precise comparison of neuroblastoma to normal cells, a few single-cell transcriptomics studies have provided more insights into the origins of neuroblastoma (Kameneva et al. 2021; Kildisiute et al. 2021; Dong et al. 2020). In the future, larger-scale analysis, and more integrative approaches combining transcriptomic with epigenomic studies are likely to define more nuances in high-risk neuroblastomas.

4.2 Modeling genomic and epigenomic signals through regulatory DNA motifs

Integration of any type of sequencing data, bulk or single-cell, comes with it's own challenges. In terms of reconstruction of gene regulatory networks, the biggest issue consist of reducing the dimensionality, so that models can be meaningfully fitted to the data. On other hand, it is important to incorporate all available data and relevant prior biological information, and formulate the models in terms of concrete biological mechanisms.

Cell type-specific gene expression programs are maintained in large parts through the distinct binding of TFs. There are an estimated 1,639 TFs in the human genome (Lambert et al. 2018), and they bind in sequence-specific manner to DNA motifs, typically in a 1000-fold or greater preference as compared to other sequences (Geertz, Shore, and Maerkl 2012). Mis-regulation of TF expression or binding is associated with a variety of diseases, such as developmental disorders and cancer (T. I. Lee and Young 2013). Our knowledge on how TF binding is affected by chromatin state, and the combinatorial interactions between TFs and their cofactors is still limited. Therefore, quantitative models of genome-wide regulatory dynamics are needed, and computational approaches can provide models that help guide time consuming and costly experimental efforts.

In chapter 2, we applied a type of multivariate linear regression method to model how DNA motif occurrences within given regulatory genomic regions (explanatory variable) explain epigenomic signals in these regions (response variable) to obtain concrete predictions on the key TF regulators acting in their system, their activities, their genome-wide targets (Figure 1.9). Previously, similar methods were applied, using either gene expression from RNA-seq or chromatin marks from ChIP-seq (T. F. Consortium and Center 2009; Piotr J Balwierz et al. 2014; Osmanbeyoglu et al. 2014; Madsen et al. 2018; Lederer et al. 2020). The novelty of our approach was to use differentially methylated regions as input sequences instead of list of genes, and DNA methylation as an input signal variable instead of gene expression or ChIP-seq signal. Since our input sequences were rich in CG dinucleotides (e.g. CpGs islands), we applied higher-order background models (order-3 Markov model) implemented in motifcounter (Kopp and Vingron 2017) to count DNA motif occurrences in differentially methylated regions. Our approach led us to predict TF binding sites not only in proximal promoters (T. F. Consortium

and Center 2009; Piotr J Balwierz et al. 2014; Osmanbeyoglu et al. 2014; Lederer et al. 2020) or proximal and distal elements in two step manner (Madsen et al. 2018), but focus on the effects of proximal and distal regulatory regions at the same time. The dynamics of input signals and input regulatory regions are themselves also controlled by complex dependencies of regulatory sites on the genome. Therefore, we assigned each differentially methylated region to a nearby gene, but in the future, with more data available from techniques analysing spatial organization of chromatin in a cell (such as chromosome conformation capture techniques (Kempfer and Pombo 2019), more accurate reconstruction of interactions between regulatory regions and genes will be possible. Additionally, these linear regression methods assume that TFs act either as an activators or repressors, whereas it is clear that some TFs can act as an activator on some targets and as a repressor on others. Considering higher degree of transcription factor dependencies could be an extension to be considered in the future, such as by using deep leraning approaches (Kopp, Monti, et al. 2020; Ronen, Hayat, and Akalin 2019)

4.3 Immunoprecipitation blues

To find where TFs bind to genomic loci of interest, ChIP-seq is the most popular choice, which involves cross-linking, then shearing chromatin, immunoprecipitation with an antibody that recognizes a TF of interest, followed by sequencing. ChIP-seq has been called a "cornerstone of epigenetics research" since the field moved from gene-specific epigenetics to the genome-wide approaches of epigenomics.

However, as I described in Chapter 3, ChIP-seq experiments don't always work as planned. L. Teytelman et al. 2013 and D. Park et al. 2013 were first to discover that in ChIP-seq, TFs often appear to bind genomic regions that are unreasonable to their function and by now, the existence of HOT regions is fairly well established (Partridge et al. 2020; Xu et al. 2020; Gheorghe et al. 2018; Tosti et al. 2018). This issue can be extended to more techniques that use immunoprecipitation (IP), such as DNA immunoprecipitation followed by sequencing (DIP-seq) (L. Shen et al. 2013) or methylated DNA immunoprecipitation sequencing (MeDIP-Seq) (Weber et al. 2005), in which pitfalls also lie with the question of sequencing depth, library preparation (cross-linking and chromatin fragmentation), but mostly with antibody specificity (Marx 2019).

In each of these techniques spurious sites are mostly at promoters, where many proteins bind with disordered domains, and non-specific antibody-binding could be a likely a culprit. To role out these sites, usually knockout or knockdown experiments are recommended, but they might not be feasible if the factor under investigation is essential.

Even though labs try to validate antibodies using western blots and immunofluorescence, the problem of non-specificity remains. Additionally, often times, the antibody that works in one validation assay does not work in the other. Other methods such as techniques that are antibody-free can be considered, or check for independent replication in other labs. However, the latter might not remove HOT regions, because enrichment of a ChIP-seq signal with completely unrelated antibodies can be found as well, such as anti-GFP, or as we showed a ChIP-seq signal on HOT regions from TFs which genes were knocked-out. This error might be pervasive and common enough to hide in plain sight in most of the IP experiments, because of the consistency of the error across studies.

4.4 HOT R-loops

The past decade has revolutionized our thinking about regulatory RNAs. Although the biology of some non-coding RNAs is well established (Statello et al. 2021), an understanding of the regulatory functions of other non-coding RNAs remains elusive. One way by which RNA may reveal regulatory functions in the genome, in a sequence-specific manner, is through the formation of R-loops. R-loops accumulate at repetitive sequences, such as transposable elements, ribosomal DNA, centromeres and telomeres, and they are also prominent at promoter regions that harbor CpG islands. Furthermore, there is evidence that R-loop formation may play a functional role in protecting against DNA methylation (Nadel et al. 2015; Paul A Ginno et al. 2012; Ross et al. 2010). Dysregulation of R-loop metabolism undermine genome maintenance, replicative senescence, and epigenetic stability (Niehrs and Luke 2020). Recent studies indicate involvement of R-loops in cancer cells, such as in inducing replication stress (Kotsantis et al. 2016), and an interaction of R-loops with oncogenes or tumor suppressors, such as breast cancer susceptibility proteins type 2 (BRCA2) and type 1 (BRCA1). BRCA2 interacts with RNase H2, and binds directly to RNA–DNA hybrids in vitro at DNA double-strand breaks (D'Alessandro et al. 2018). BRCA1 suppresses R-loop formation upon MYCN activation

in high-risk neuroblastoma samples and is associated with significant hypomethylation of BRCA1, and its neighbourhood (Herold et al. 2019). In this case, if the release of RNAPII from transcriptional pause sites (pause release) fails, MYCN recruits BRCA1 to promoter-proximal regions.

In Chapter 3, I described our novel observation that HOT regions can be associated with R-loops due to sequence-specific features, DRIP-seq and RDIP-seq enrichment on HOT regions, and hypo-methylation. However, current methods to detect R-loops are also not noise-free. As mentioned in Chapter 3, the dominant strategies to detect R-loops rely on the immunoprecipitation of chromatin containing R-loops using a monoclonal antibody, S9.6, specific for DNA:RNA hybrids (Boguslawski et al. 1986), however recently it has been shown that they suffer from off-target binding (Smolka et al. 2020). To overcome low-resolution, and no clear boundaries of the R-loop peaks from these techniques, Dumelie and Jaffrey 2017 created bisulfite DNA-RNA immunoprecipitation (bis-DRIP) approach that involves treating genomic DNA with bisulfite under non-denaturing conditions and combines it later with immunoprecipitation using the S9.6 antibody. While the DNA template strand is hybridised with the transcript and therefore protected against bisulfite treatment, in the non-template DNA single strand bisulfite causes cytosine-to-uracil conversions (Yu et al. 2003). Then, sequencing of both strands reveals the strand-specific location of R-loops at near-nucleotide resolution. Other strategies take advantage of the natural affinity of RNase H enzyme for DNA-RNA hybrids, and DNA-RNA hybrids bound by RNase H are enriched by affinity purification (Paul A Ginno et al. 2012; L. Chen et al. 2017), or to utilize RNase H to guide micrococcal nuclease to R-loops (Yan et al. 2019).

More practically, properties of R-loops used in these experiments might be useful to take into account when designing future ChIP-seq experiments in order to remove HOT regions created due to R-loops formation. Pre-treatment with RNase H, or RNase III as suggested by Smolka et al. 2020 might correct HOT regions issue. Interestingly, R-loops have a specific nucleotide composition (GC skew), and a DNA strand that is a part of DNA-RNA hybrid might be more methylated than non-template DNA single strand, which could be further examined to predict R-loops, and consequently at least a subset of HOT regions.

4.5 Final remarks

Combining experimental and computational approaches allows modelling of the principles of gene regulation and DNA methylation dynamics and help us to understand where and why imbalances may occur.

In terms of disease-specific aberrations of DNA methylation, the integration of genomic and epigenetic sequencing data provides an evidence that DNA methylation is dysregulated in high-risk neuroblastomas, and it is tightly connected with crucial transcription factor machinery specific to neuroblastoma. Characterization of the progression of cellular states as early as human embryogenesis has provided insights into the origin of this pediatric disease (Kameneva et al. 2021). Epigenetic studies will complement it in the future in order to uncover its heterogeneity, and serve to suggest novel avenues for prevention and treatment of this disease.

My second conclusion is that genomic regions that have been reported to be the targets of an extreme abundance of transcription factors, and treated mainly as technical artifacts of ChIP-seq, provide more biological insights into gene regulation processes than we initially have thought, namely the association of antibodies used in ChIP-seq experiments with DNA secondary structures in lowly methylated regions. Firstly, it is crucial for better understanding of antibodies properties, their specificity, and necessary improvement of ChIP-seq protocols, and also other protocols that employ an immunoprecipitation step. Secondly, R-loops that were long considered unfavorable by-products of transcription that interfered with the transcription process itself, are now known to be involved in gene regulation mechanisms, and might interfere with antibodies that bind on HOT regions. Perhaps soon, more will be uncovered about specific properties of R-loops in normal cells and in disease states, as in an example of neuroblastoma (Herold et al. 2019).

A

Supplementary Material for Chapter 2

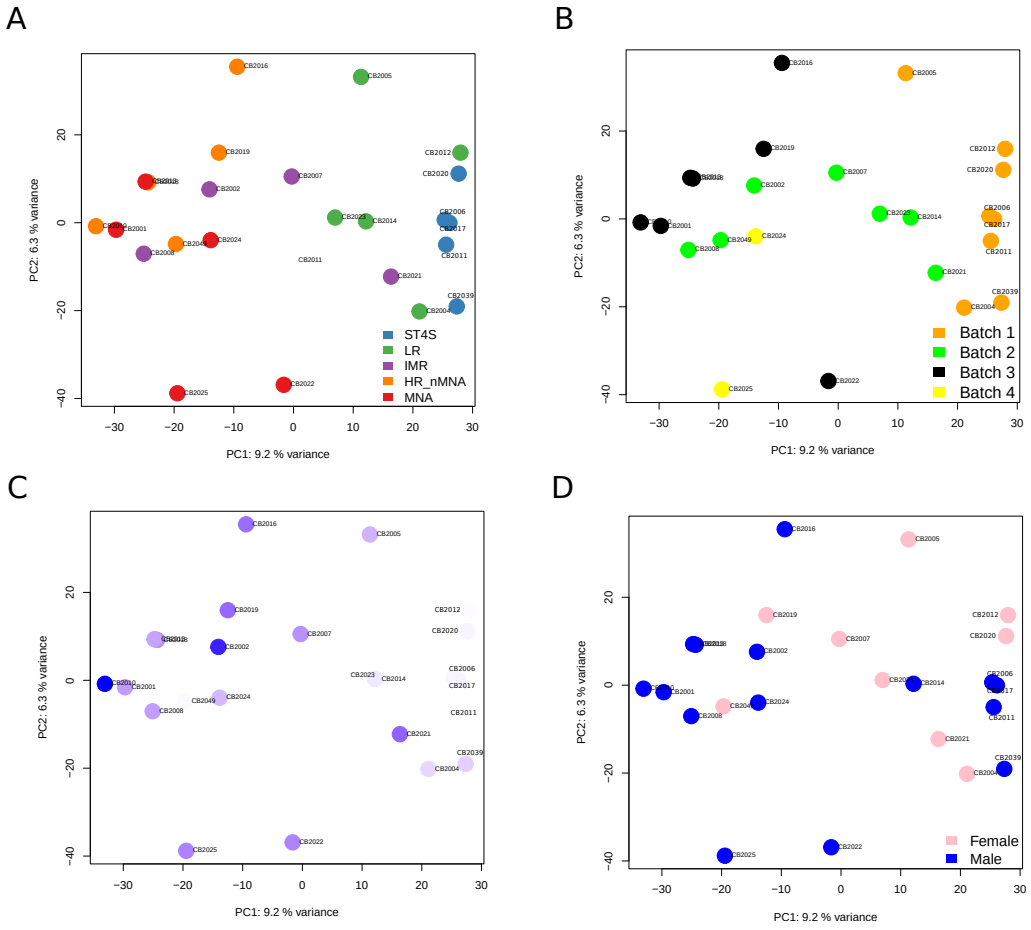


Figure A.1: PCA analysis on 24 Neuroblastoma WGBS samples. A PCA plot color-coded by neuroblastoma risk groups (A), by sequencing batches (B), age [days] (C), gender (D).

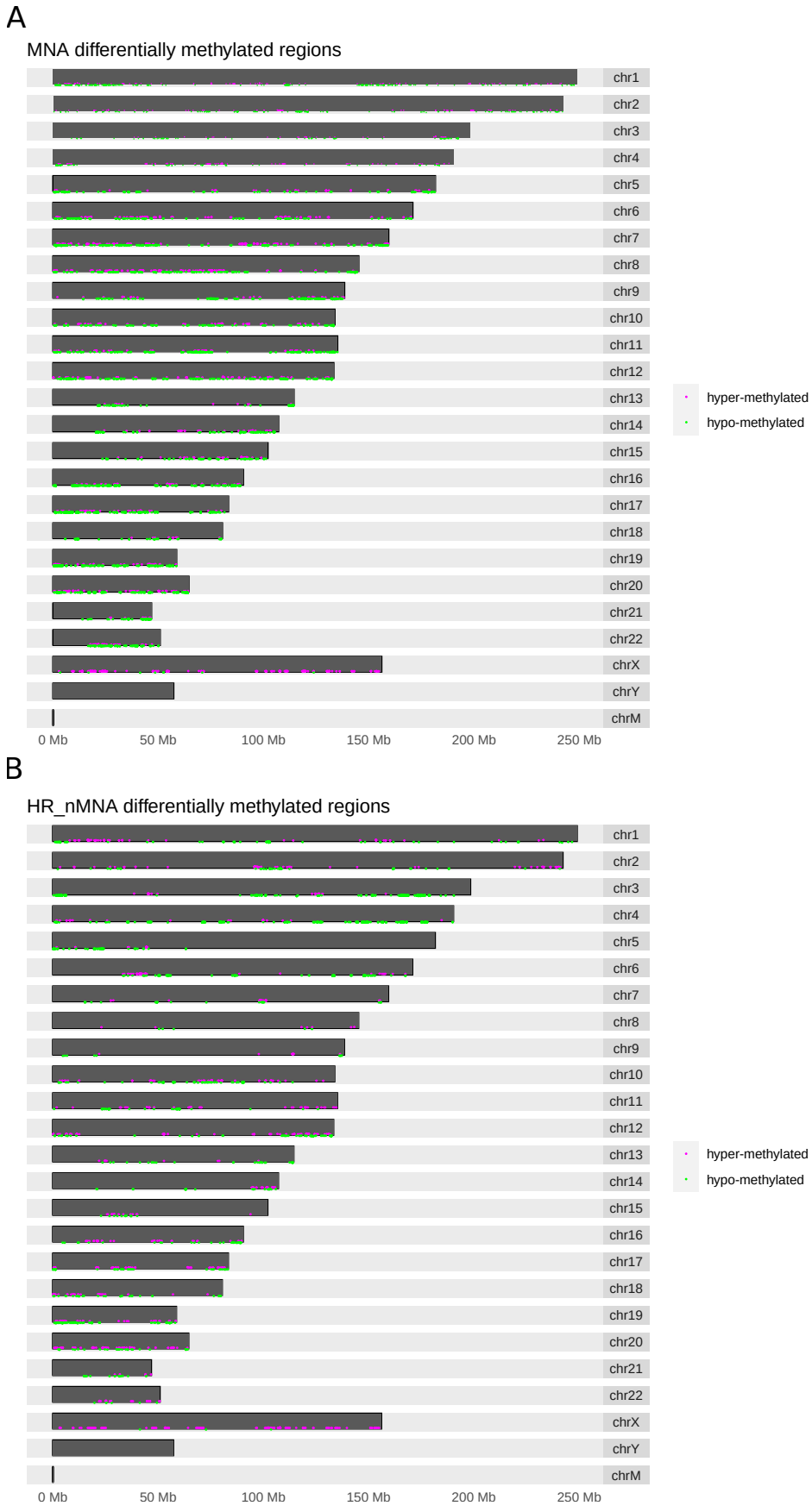


Figure A.2: Ideograms depicting hyper- and hypo-methylated MNA (A), and HR_nMNA (B) DMRs correlated with expression of nearby genes.

A

	term.name	term.id	domain	p.value
1	circulatory system process	GO:0003013	BP	1.98e−02
2	positive regulation of nitrogen compound metabolic process	GO:0051173	BP	5.12e−08
3	actin cytoskeleton organization	GO:0030036	BP	9.24e−05
4	positive regulation of RNA metabolic process	GO:0051254	BP	8.14e−07
5	response to ammonium ion	GO:0060359	BP	2.65e−03
6	cell differentiation	GO:0030154	BP	1.33e−18
7	positive regulation of macromolecule biosynthetic process	GO:0010557	BP	1.17e−06
8	response to endogenous stimulus	GO:0009719	BP	2.59e−05
9	response to organic substance	GO:0010033	BP	7.29e−06
10	negative regulation of transcription by RNA polymerase II	GO:0000122	BP	1.61e−04
11	intracellular	GO:0005622	CC	8.06e−10
12	cell junction	GO:0030054	CC	3.01e−03
13	membrane−bounded organelle	GO:0043227	CC	2.37e−06
14	bounding membrane of organelle	GO:0098588	CC	9.85e−03
15	synapse part	GO:0044456	CC	1.56e−04
16	whole membrane	GO:0098805	CC	3.28e−02
17	sequence−specific DNA binding	GO:0043565	MF	1.78e−05
18	protein binding	GO:0005515	MF	4.98e−14
19	transcription regulatory region DNA binding	GO:0044212	MF	4.24e−06
20	transcription regulatory region sequence−specific DNA binding	GO:0000976	MF	2.39e−04
21	small GTPase binding	GO:0031267	MF	3.51e−02
22	Ras guanyl−nucleotide exchange factor activity	GO:0005088	MF	1.96e−02
23	Cocaine addiction	KEGG:05030	keg	4.36e−02
24	Estrogen signaling pathway	KEGG:04915	keg	1.10e−02
25	Rap1 signaling pathway	KEGG:04015	keg	3.45e−02
26	Cushing syndrome	KEGG:04934	keg	2.73e−02
27	NCAM signaling for neurite out−growth	REAC:R−HSA−375165	rea	3.82e−03
28	Neuronal System	REAC:R−HSA−112316	rea	1.40e−02
29	Activated NTRK2 signals through FYN	REAC:R−HSA−9032500	rea	2.65e−02

B

	term.name	term.id	domain	p.value
1	generation of neurons	GO:0048699	BP	2.52e−02
2	Calcium signaling pathway	KEGG:04020	keg	2.75e−02
3	Neuroactive ligand−receptor interaction	KEGG:04080	keg	2.88e−02

C

	term.name	term.id	domain	p.value
1	neural precursor cell proliferation	GO:0061351	BP	1.35e−03
2	nervous system development	GO:0007399	BP	1.22e−07
3	neuron projection morphogenesis	GO:0048812	BP	1.75e−03
4	stem cell proliferation	GO:0072089	BP	6.76e−03
5	X−linked inheritance	HP:0001417	hp	8.99e−06
6	Intellectual disability	HP:0001249	hp	4.18e−03
7	Rap1 signaling pathway	KEGG:04015	keg	6.94e−03
8	5−Phosphoribose 1−diphosphate biosynthesis	REAC:R−HSA−73843	rea	8.90e−03

Figure A.3: List of enriched Gene ontology terms of genes correlated with only MNA DMRs (A), only HR_nMNA (B) and both types of DMRs (C) by gProfiler2 (Raudvere et al. 2019).

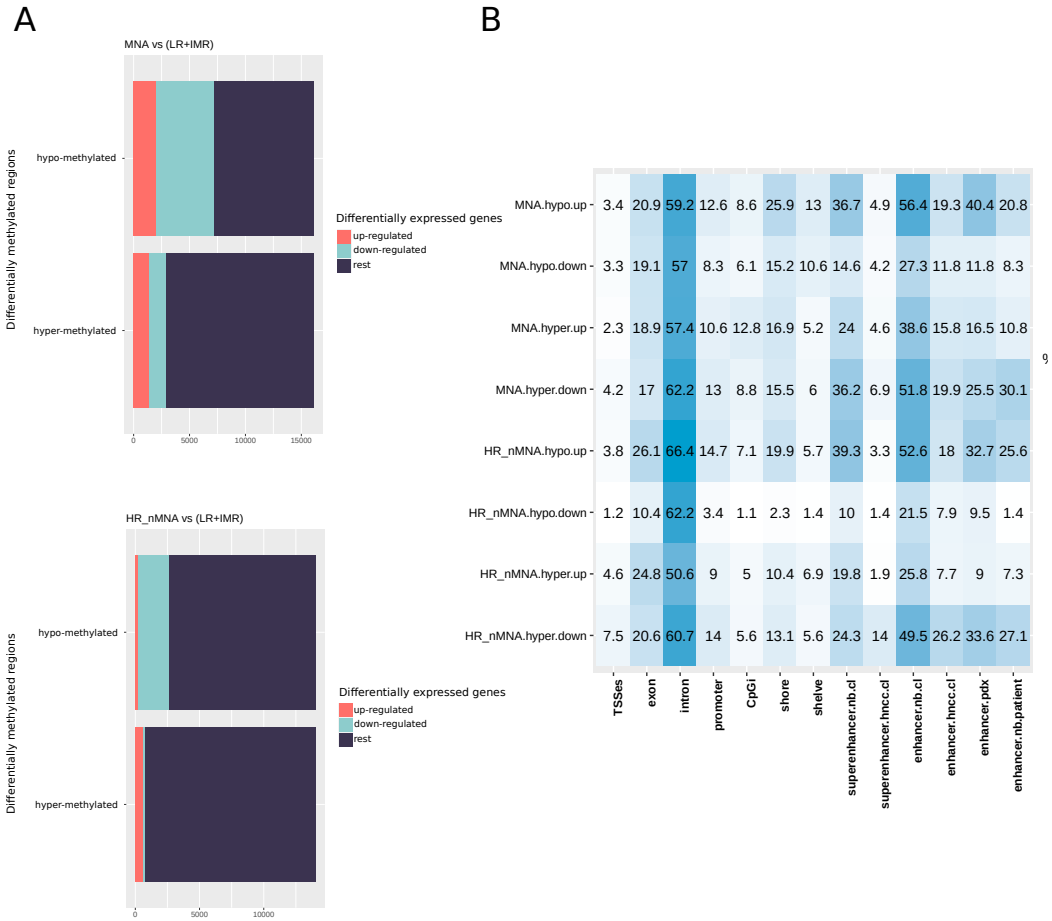


Figure A.4: (A) Bar plots show number of down- and up-regulated genes correlated to hypo- and hyper-methylated MNA or HR_nMNA DMRs. (B) Percent of overlapping hypo- and hyper-methylated MNA or HR_nMNA DMRs with down- and up-regulated correlated genes.

	gene.hgnc	gene.ensg	gene.family	DMR.type	diff.expression	diff.methylation
1	ALX1	ENSG00000180318	homeodomain_proteins; transcription_factors	MNA	up−regulated	hypo−methylated
2	ARC	ENSG00000198576	transcription_factors	MNA	up−regulated	hypo−methylated
3	ASCL1	ENSG00000139352	transcription_factors	MNA	up−regulated	hypo−methylated
4	BMP7	ENSG00000101144	cytokines_and_growth_factors	MNA	up−regulated	hypo−methylated
5	CD101	ENSG00000134256	cell_differentiation_markers	MNA	up−regulated	hypo−methylated
6	ELK1	ENSG00000126767	transcription_factors	MNA	up−regulated	hypo−methylated
7	FGF5	ENSG00000138675	cytokines_and_growth_factors	MNA	up−regulated	hypo−methylated
8	GSC	ENSG00000133937	homeodomain_proteins; transcription_factors	MNA	up−regulated	hypo−methylated
9	IRAK1	ENSG00000184216	protein_kinases	MNA	up−regulated	hypo−methylated
10	IRX3	ENSG00000177508	homeodomain_proteins; transcription_factors	MNA	up−regulated	hypo−methylated
11	MAP3K15	ENSG00000180815	protein_kinases	MNA	up−regulated	hypo−methylated
12	MDK	ENSG00000110492	cytokines_and_growth_factors	MNA	up−regulated	hypo−methylated
13	MELK	ENSG00000165304	protein_kinases	MNA	up−regulated	hypo−methylated
14	MYRF	ENSG00000124920	transcription_factors	MNA	up−regulated	hypo−methylated
15	NFIB	ENSG00000147862	oncogenes; transcription_factors; translocated_cancer_genes	MNA	up−regulated	hypo−methylated
16	PAX5	ENSG00000196092	homeodomain_proteins; oncogenes; transcription_factors; translocated_cancer_genes; tumour_supressors	MNA	up−regulated	hypo−methylated
17	PDGFRA	ENSG00000134853	cell_differentiation_markers; cytokines_and_growth_factors; oncogenes; protein_kinases; translocated_cancer_genes	MNA	up−regulated	hypo−methylated
18	PRDM14	ENSG00000147596	transcription_factors	MNA	up−regulated	hypo−methylated
19	SCML2	ENSG00000102098	transcription_factors	MNA	up−regulated	hypo−methylated
20	SCRT2	ENSG00000215397	transcription_factors	MNA	up−regulated	hypo−methylated
21	SIX2	ENSG00000170577	homeodomain_proteins; transcription_factors	MNA	up−regulated	hypo−methylated
22	SIX3	ENSG00000138083	homeodomain_proteins; transcription_factors	MNA	up−regulated	hypo−methylated
23	SMAD3	ENSG00000166949	transcription_factors; tumour_supressors	MNA	up−regulated	hypo−methylated
24	TBX1	ENSG00000184058	transcription_factors	MNA	up−regulated	hypo−methylated
25	TEAD4	ENSG00000197905	transcription_factors	MNA	up−regulated	hypo−methylated
26	TERT	ENSG00000164362	oncogenes	MNA	up−regulated	hypo−methylated
27	TFAP2D	ENSG00000008197	transcription_factors	MNA	up−regulated	hypo−methylated
28	TRIP13	ENSG00000071539	oncogenes; transcription_factors; tumour_supressors	MNA	up−regulated	hypo−methylated
29	TSSK6	ENSG00000178093	protein_kinases	MNA	up−regulated	hypo−methylated
30	TWIST1	ENSG00000122691	transcription_factors	MNA	up−regulated	hypo−methylated
31	ZNF217	ENSG00000171940	transcription_factors	MNA	up−regulated	hypo−methylated

Figure A.5: A list of MNA upregulated expressed genes correlated with nearby hypo-methylated MNA DMRs (proximal extension to upstream 5.0kb, and downstream 1.0kb), and that are part of gene families from the MSigDB gene sets (*MSigDB gene sets* n.d.; Liberzon et al. 2011; Subramanian et al. 2005), and oncogene and tumor suppressor list from the OncoKB Cancer Gene List (Chakravarty et al. 2017).

	gene.hgnc	gene.ensg	gene.family	DMR.type	diff.expression	diff.methylation
1	ANKK1	ENSG00000170209	protein_kinases	MNA	down−regulated	hyper−methylated
2	ARNTL	ENSG00000133794	transcription_factors	MNA	down−regulated	hyper−methylated
3	BCL6	ENSG00000113916	oncogenes; transcription_factors; translocated_cancer_genes	MNA	down−regulated	hyper−methylated
4	BHLHE41	ENSG00000123095	transcription_factors	MNA	down−regulated	hyper−methylated
5	BMP8B	ENSG00000116985	cytokines_and_growth_factors	MNA	down−regulated	hyper−methylated
6	BTBD3	ENSG00000132640	transcription_factors	MNA	down−regulated	hyper−methylated
7	CASZ1	ENSG00000130940	transcription_factors	MNA	down−regulated	hyper−methylated
8	CBX7	ENSG00000100307	transcription_factors	MNA	down−regulated	hyper−methylated
9	CCL2	ENSG00000108691	cytokines_and_growth_factors	MNA	down−regulated	hyper−methylated
10	CD53	ENSG00000143119	cell_differentiation_markers	MNA	down−regulated	hyper−methylated
11	CD6	ENSG00000013725	cell_differentiation_markers	MNA	down−regulated	hyper−methylated
12	CD69	ENSG00000110848	cell_differentiation_markers	MNA	down−regulated	hyper−methylated
13	CD9	ENSG00000010278	cell_differentiation_markers	MNA	down−regulated	hyper−methylated
14	CHD5	ENSG00000116254	transcription_factors	MNA	down−regulated	hyper−methylated
15	CHGB	ENSG00000089199	cytokines_and_growth_factors	MNA	down−regulated	hyper−methylated
16	CLEC10A	ENSG00000132514	cell_differentiation_markers	MNA	down−regulated	hyper−methylated
17	CRH	ENSG00000147571	cytokines_and_growth_factors	MNA	down−regulated	hyper−methylated
18	EDN1	ENSG00000078401	cytokines_and_growth_factors	MNA	down−regulated	hyper−methylated
19	EPHB3	ENSG00000182580	protein_kinases	MNA	down−regulated	hyper−methylated
20	ERRFI1	ENSG00000116285	tumour_supressors	MNA	down−regulated	hyper−methylated
21	FOXF2	ENSG00000137273	transcription_factors	MNA	down−regulated	hyper−methylated
22	GCG	ENSG00000115263	cytokines_and_growth_factors	MNA	down−regulated	hyper−methylated
23	IKBKE	ENSG00000263528	oncogenes; protein_kinases	MNA	down−regulated	hyper−methylated
24	IKZF2	ENSG00000030419	transcription_factors	MNA	down−regulated	hyper−methylated
25	IL6R	ENSG00000160712	cell_differentiation_markers	MNA	down−regulated	hyper−methylated
26	IL7	ENSG00000104432	cytokines_and_growth_factors	MNA	down−regulated	hyper−methylated
27	IL7R	ENSG00000168685	cell_differentiation_markers; oncogenes	MNA	down−regulated	hyper−methylated
28	IRF8	ENSG00000140968	transcription_factors; tumour_supressors	MNA	down−regulated	hyper−methylated
29	ITGB4	ENSG00000132470	cell_differentiation_markers	MNA	down−regulated	hyper−methylated
30	JAZF1	ENSG00000153814	oncogenes; translocated_cancer_genes	MNA	down−regulated	hyper−methylated
31	KL	ENSG00000133116	cytokines_and_growth_factors	MNA	down−regulated	hyper−methylated
32	MAFB	ENSG00000204103	oncogenes; transcription_factors; translocated_cancer_genes	MNA	down−regulated	hyper−methylated
33	MEIS1	ENSG00000143995	homeodomain_proteins; transcription_factors	MNA	down−regulated	hyper−methylated
34	MEOX1	ENSG00000005102	homeodomain_proteins; transcription_factors	MNA	down−regulated	hyper−methylated
35	MRC1	ENSG00000260314	cell_differentiation_markers	MNA	down−regulated	hyper−methylated

Figure A.6: A list (first 35 entries) of MNA downregulated expressed genes correlated with nearby hyper-methylated MNA DMRs (proximal extension to upstream 5.0kb, and downstream 1.0kb), and that are part of gene families from the MSigDB gene sets (*MSigDB gene sets* n.d.; Liberzon et al. 2011; Subramanian et al. 2005), and oncogene and tumor suppressor list from the OncoKB Cancer Gene List (Chakravarty et al. 2017).

	gene.hgnc	gene.ensg	gene.family	DMR.type	diff.expression	diff.methylation
36	NEUROD6	ENSG00000164600	transcription_factors	MNA	down−regulated	hyper−methylated
37	NGFR	ENSG00000064300	cell_differentiation_markers	MNA	down−regulated	hyper−methylated
38	NPY	ENSG00000122585	cytokines_and_growth_factors	MNA	down−regulated	hyper−methylated
39	NTRK3	ENSG00000140538	oncogenes; protein_kinases; translocated_cancer_genes	MNA	down−regulated	hyper−methylated
40	PDE4DIP	ENSG00000178104	oncogenes; translocated_cancer_genes	MNA	down−regulated	hyper−methylated
41	PDIK1L	ENSG00000175087	protein_kinases	MNA	down−regulated	hyper−methylated
42	PDLIM5	ENSG00000163110	transcription_factors	MNA	down−regulated	hyper−methylated
43	PRDM16	ENSG00000142611	oncogenes; transcription_factors; translocated_cancer_genes	MNA	down−regulated	hyper−methylated
44	PRKCB	ENSG00000166501	protein_kinases	MNA	down−regulated	hyper−methylated
45	PRKCZ	ENSG00000067606	protein_kinases	MNA	down−regulated	hyper−methylated
46	PTPRT	ENSG00000196090	tumour_supressors	MNA	down−regulated	hyper−methylated
47	RORA	ENSG00000069667	transcription_factors	MNA	down−regulated	hyper−methylated
48	RORB	ENSG00000198963	transcription_factors	MNA	down−regulated	hyper−methylated
49	RUNX3	ENSG00000020633	transcription_factors	MNA	down−regulated	hyper−methylated
50	SEMA3E	ENSG00000170381	cytokines_and_growth_factors	MNA	down−regulated	hyper−methylated
51	SEMA3G	ENSG00000010319	cytokines_and_growth_factors	MNA	down−regulated	hyper−methylated
52	SFRP1	ENSG00000104332	tumour_supressors	MNA	down−regulated	hyper−methylated
53	SIK1	ENSG00000142178	protein_kinases	MNA	down−regulated	hyper−methylated
54	SIRPA	ENSG00000198053	cell_differentiation_markers	MNA	down−regulated	hyper−methylated
55	SOX1	ENSG00000182968	transcription_factors	MNA	down−regulated	hyper−methylated
56	SOX8	ENSG00000005513	transcription_factors	MNA	down−regulated	hyper−methylated
57	STYK1	ENSG00000060140	protein_kinases	MNA	down−regulated	hyper−methylated
58	TBX20	ENSG00000164532	transcription_factors	MNA	down−regulated	hyper−methylated
59	TCF7L2	ENSG00000148737	transcription_factors; tumour_supressors	MNA	down−regulated	hyper−methylated
60	TFAP2A	ENSG00000137203	transcription_factors	MNA	down−regulated	hyper−methylated
61	TFAP2C	ENSG00000087510	transcription_factors	MNA	down−regulated	hyper−methylated
62	TNFRSF1B	ENSG00000028137	cell_differentiation_markers	MNA	down−regulated	hyper−methylated
63	TNFRSF9	ENSG00000049249	cell_differentiation_markers	MNA	down−regulated	hyper−methylated
64	TP63	ENSG00000073282	oncogenes; transcription_factors; tumour_supressors	MNA	down−regulated	hyper−methylated
65	TRPM6	ENSG00000119121	protein_kinases	MNA	down−regulated	hyper−methylated
66	UCN3	ENSG00000178473	cytokines_and_growth_factors	MNA	down−regulated	hyper−methylated
67	VENTX	ENSG00000151650	homeodomain_proteins; transcription_factors	MNA	down−regulated	hyper−methylated
68	ZBTB16	ENSG00000109906	oncogenes; transcription_factors; translocated_cancer_genes	MNA	down−regulated	hyper−methylated
69	ZFAND3	ENSG00000156639	transcription_factors	MNA	down−regulated	hyper−methylated
70	ZNF208	ENSG00000160321	transcription_factors	MNA	down−regulated	hyper−methylated

Figure A.7: A list (36-70 entries) of MNA downregulated expressed genes correlated with nearby hyper-methylated MNA DMRs (proximal extension to upstream 5.0kb, and downstream 1.0kb), and that are part of gene families from the MSigDB gene sets (*MSigDB gene sets* n.d.; Liberzon et al. 2011; Subramanian et al. 2005), and oncogene and tumor suppressor list from the OncoKB Cancer Gene List (Chakravarty et al. 2017).

	gene.hgnc	gene.ensg	gene.family	DMR.type	diff.expression	diff.methylation
1	BARX2	ENSG00000043039	homeodomain_proteins; transcription_factors	HR_nMNA	up−regulated	hypo−methylated
2	CLEC11A	ENSG00000105472	cytokines_and_growth_factors	HR_nMNA	up−regulated	hypo−methylated
3	DKK1	ENSG00000107984	cytokines_and_growth_factors	HR_nMNA	down−regulated	hyper−methylated
4	EPHA4	ENSG00000116106	protein_kinases	HR_nMNA	up−regulated	hypo−methylated
5	FGF10	ENSG00000070193	cytokines_and_growth_factors	HR_nMNA	up−regulated	hypo−methylated
6	FOXJ1	ENSG00000129654	transcription_factors	HR_nMNA	up−regulated	hypo−methylated
7	KLF6	ENSG00000067082	transcription_factors; tumour_supressors	HR_nMNA	down−regulated	hyper−methylated
8	NRIP1	ENSG00000180530	transcription_factors	HR_nMNA	down−regulated	hyper−methylated
9	RPS6KA4	ENSG00000162302	oncogenes; protein_kinases	HR_nMNA	up−regulated	hypo−methylated
10	VIP	ENSG00000146469	cytokines_and_growth_factors	HR_nMNA	down−regulated	hyper−methylated

Figure A.8: A list of HR_MNA downregulated expressed genes correlated with nearby hypermethylated HR_MNA DMRs (proximal extension to upstream 5.0kb, and downstream 1.0kb), and that are part of gene families from the MSigDB gene sets (*MSigDB gene sets* n.d.; Liberzon et al. 2011; Subramanian et al. 2005), and oncogene and tumor suppressor list from the OncoKB Cancer Gene List (Chakravarty et al. 2017)

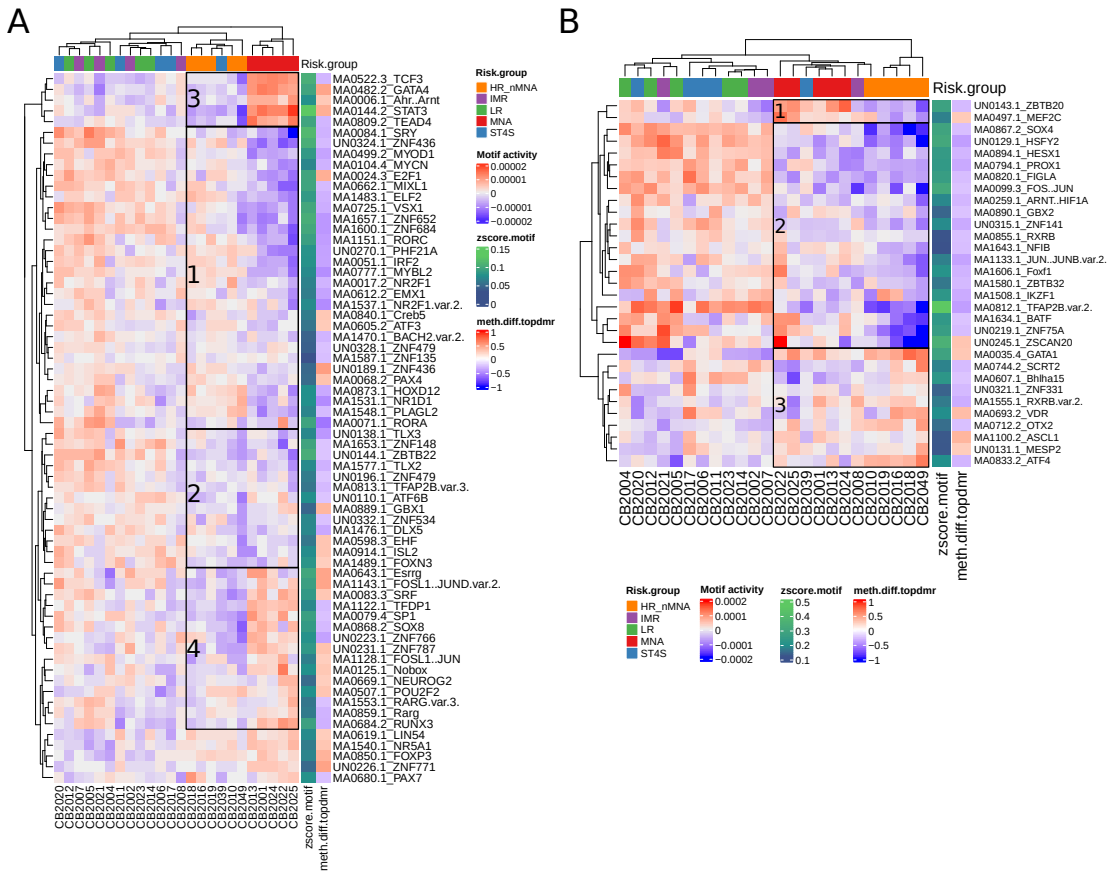


Figure A.9: Heatmaps represents motif activity values of DNA motifs from the MNA (A) and HR_nMNA networks (B). Top annotation indicate sample risk group (Risk.group). The right annotation indicate z-score of a DNA motif (zscore.motif), and average methylation difference of a top target MNA DMR of a corresponding motif (meth.diff.topdmr).

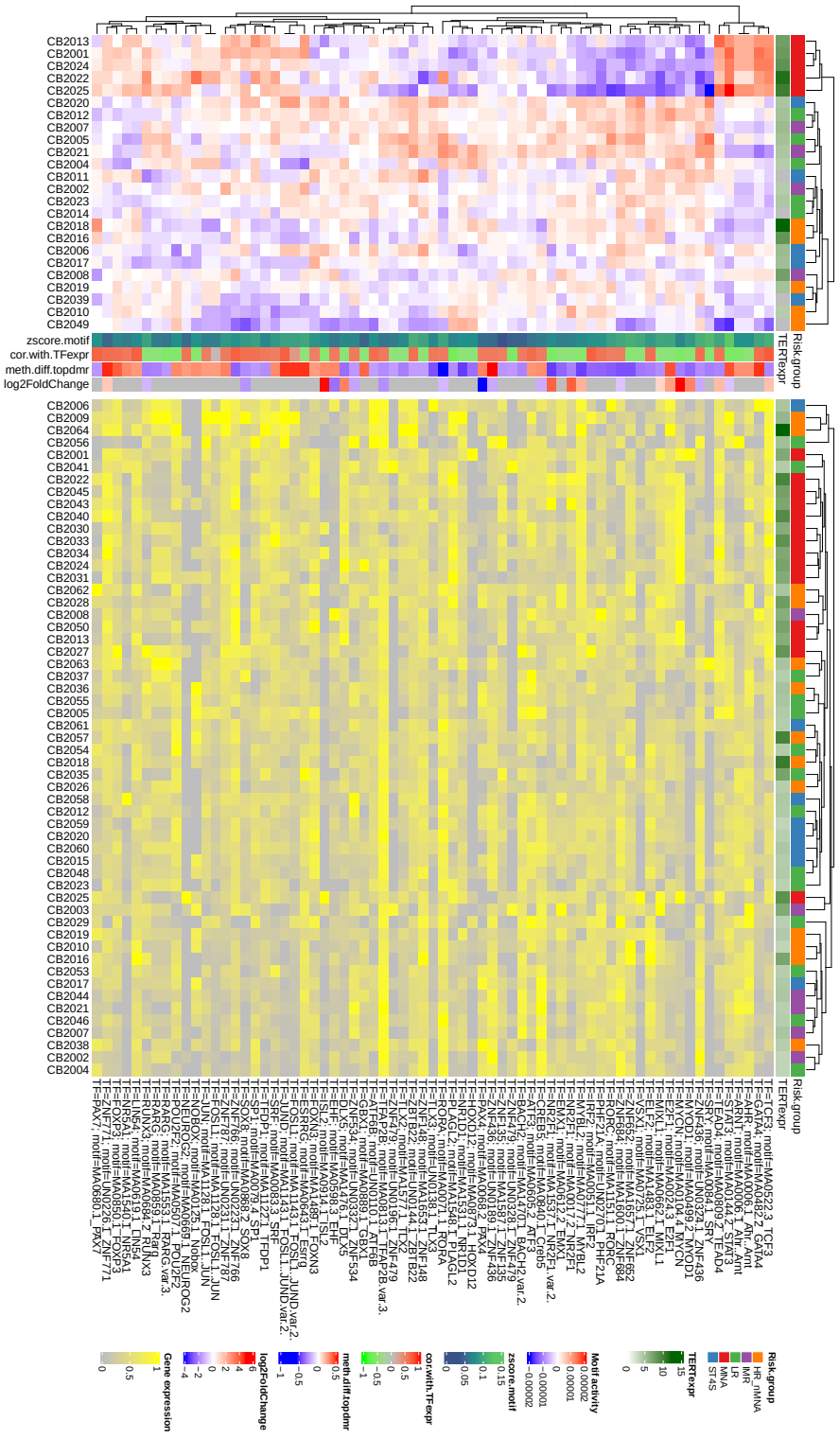


Figure A.10: The upper heatmap represents motif activity values, and the lower heatmap represents gene expression (normalized to 0-1 range) of TFs that bind to DNA motifs in MNA network. Rows of both heatmaps are clustered according to motif activity. Top annotation indicate sample risk groups (Risk.group), expression of TERT gene (TERT.expr), and the right annotation indicate z-score of a DNA motif (zscore.motif), a Pearson correlation between motif activities and TFs expression across all samples (cor.with.TFexpr), average methylation difference of a top target MNA DMR of a corresponding motif (meth.diff.topdmr), log2 fold change of significantly differentially expressed genes between MNA vs low+intermediate risk groups (log2FoldChange).

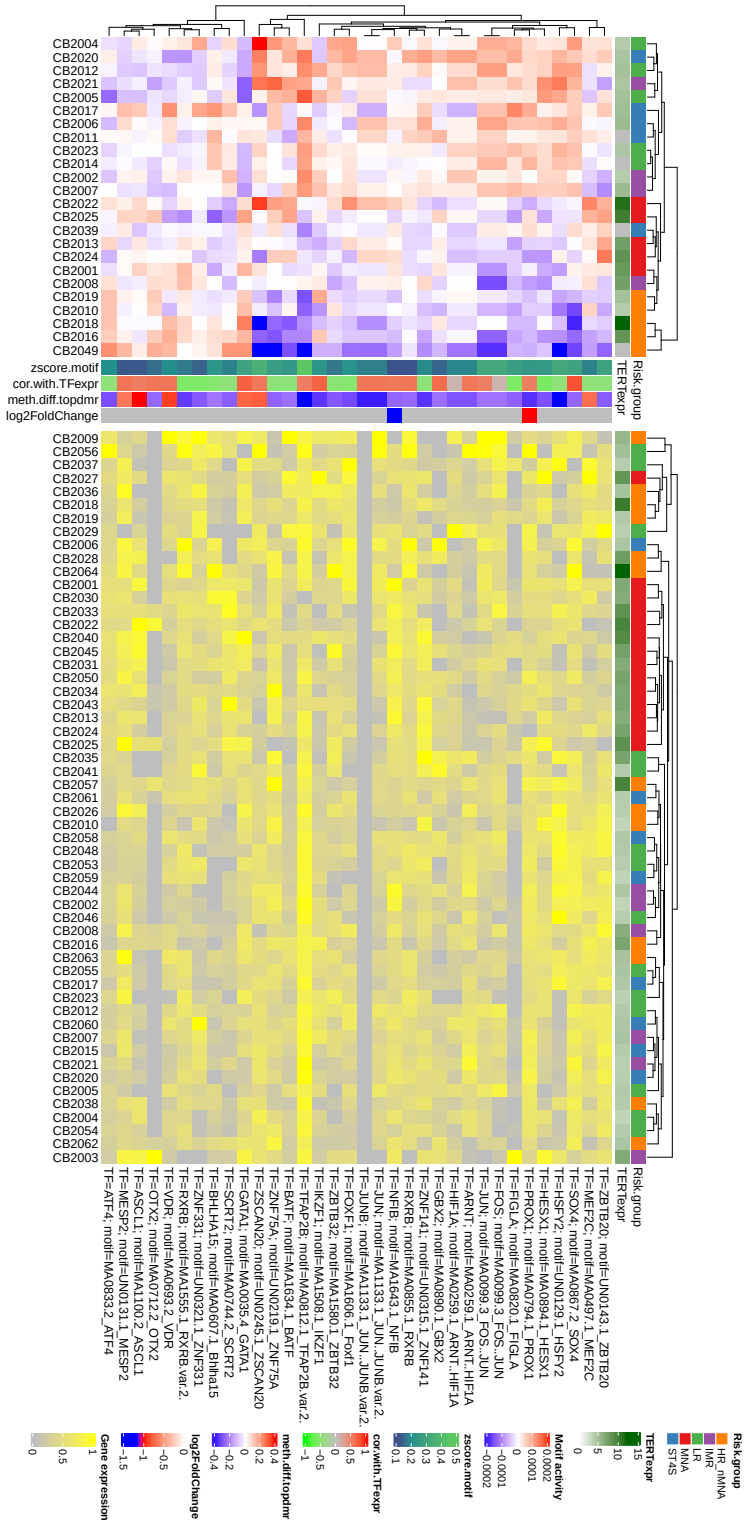


Figure A.11: HR_nMNA network represented as motif activity and gene expression heatmaps as described in Figure A.10.

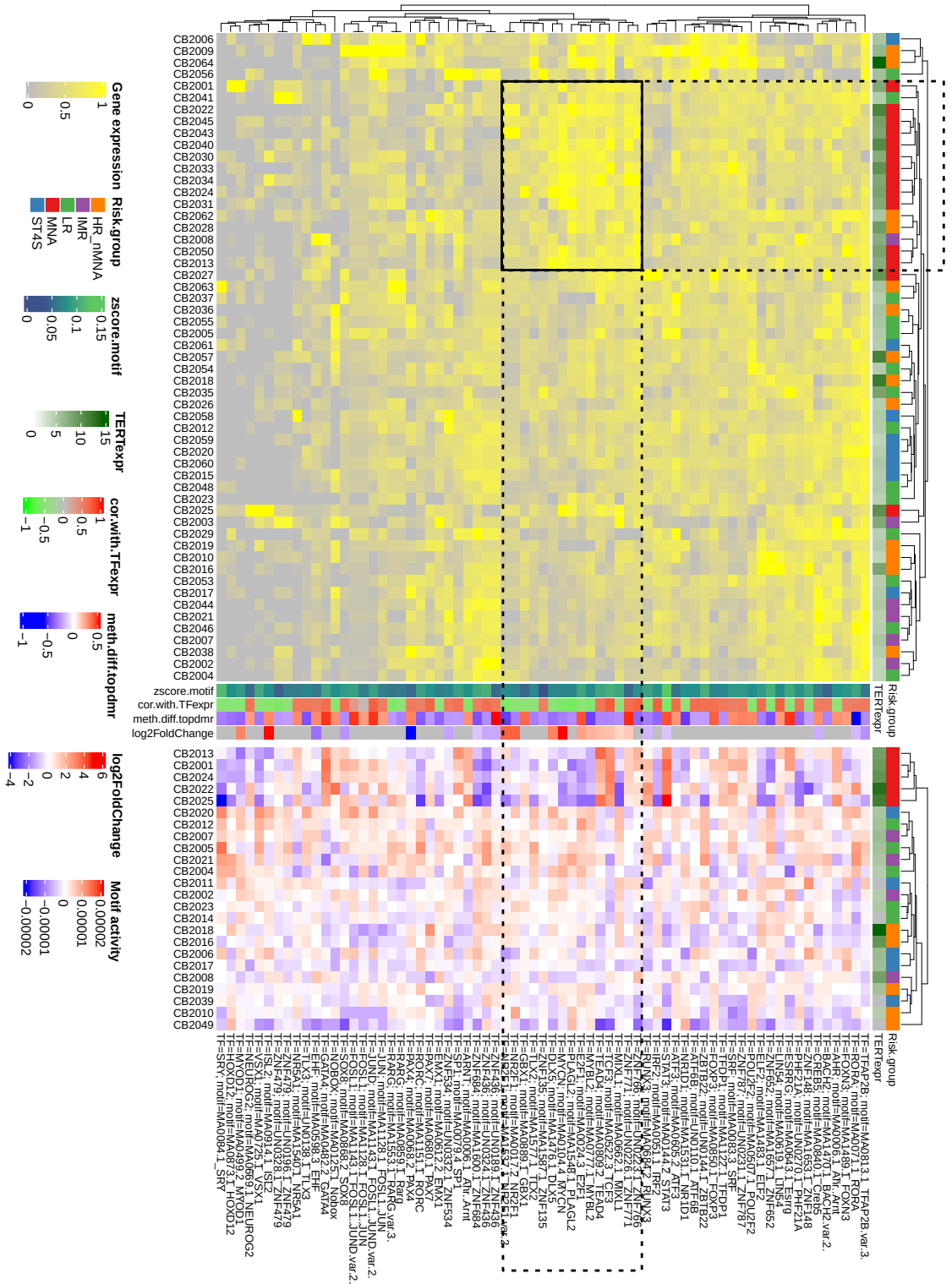


Figure A.12: The upper heatmap represents gene expression (normalized to 0-1 range) of TFs that bind to DNA motifs in MNA network, and the lower heatmap represents motif activity values. Rows of both heatmaps are clustered according to gene expression. Top annotation indicate sample risk groups (Risk.group), expression of TERT gene (TERT.expr), and the right annotation indicate z-score of a DNA motif (zscore.motif), a Pearson correlation between motif activities and TFs expression across all samples (cor.with.TFexpr), average methylation difference of a top target MNA DMR of a corresponding motif (meth.diff.topdmr), log2 fold change of significantly differentially expressed genes between MNA vs low+intermediate risk groups (log2FoldChange).

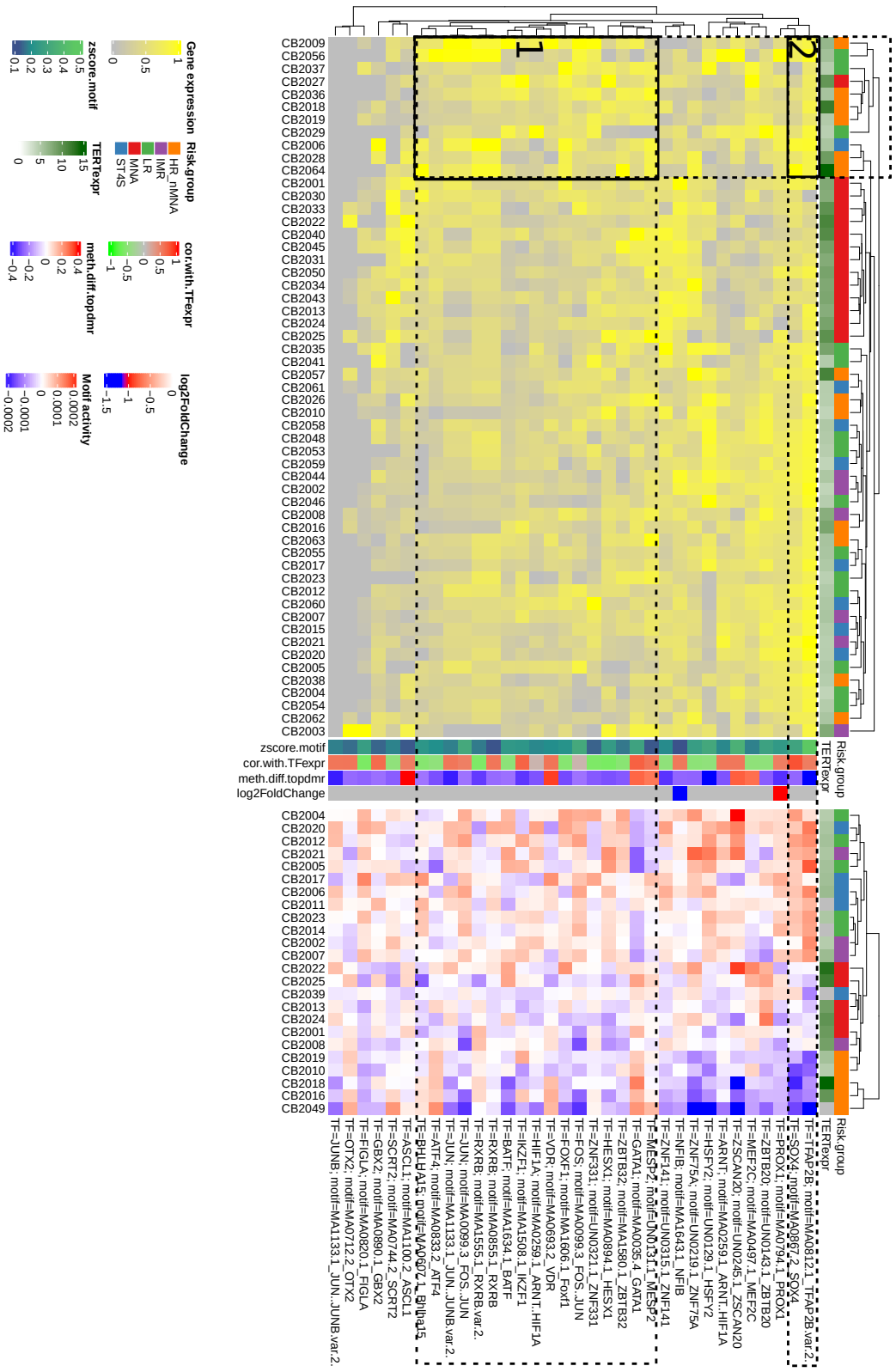
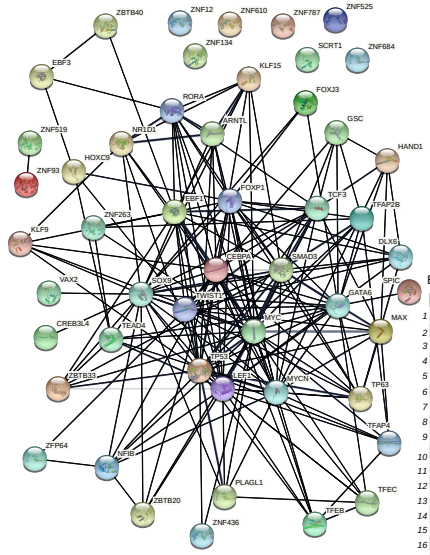


Figure A.13: HR_nMNA network represented as gene expression and motif activity heatmaps as described in Supplementary Figure A.12. Cluster 1 corresponds to DNA motifs that activity is negative in HR_nMNA samples and expression of genes coding TF that bind to them is the highest in HR_nMNA samples. In cluster 2 DNA motifs have negative activity and the gene expression is high in HR_nMNA samples and low in MNA samples.

A



B

Molecular Function (Gene Ontology)

	#term ID	term description	count in network	false discovery rate	strength
1	GO:0043425	bHLH transcription factor binding	6 of 29	2.39e−09	1.91
2	GO:0001222	transcription corepressor binding	2 of 11	2.20e−03	1.85
3	GO:0070888	E−box binding	7 of 43	3.38e−10	1.80
4	GO:0016251	RNA polymerase II general transcription initiation factor activity	2 of 14	3.10e−03	1.75
5	GO:0070742	C2H2 zinc finger domain binding	2 of 16	3.90e−03	1.69
6	GO:0070491	repressing transcription factor binding	5 of 60	3.89e−06	1.51
7	GO:0043027	cysteine−type endopeptidase inhibitor activity involved in apoptotic process	2 of 26	8.40e−03	1.48
8	GO:0001103	RNA polymerase II repressing transcription factor binding	2 of 29	1.00e−02	1.43
9	GO:0003707	steroid hormone receptor activity	3 of 59	2.30e−03	1.30
10	GO:0001228	DNA−binding transcription activator activity, RNA polymerase II−specific	20 of 408	1.01e−19	1.28
11	GO:0008013	beta−catenin binding	4 of 82	3.40e−04	1.28
12	GO:0000978	RNA polymerase II proximal promoter sequence−specific DNA binding	23 of 492	1.76e−22	1.26
13	GO:0001085	RNA polymerase II transcription factor binding	5 of 121	9.54e−05	1.21
14	GO:0000977	RNA polymerase II transcription regulatory region sequence−specific DNA binding	26 of 647	3.95e−24	1.20
15	GO:0001227	DNA−binding transcription repressor activity, RNA polymerase II−specific	9 of 229	5.64e−08	1.19
16	GO:0004879	nuclear receptor activity	2 of 50	2.39e−02	1.19
17	GO:0000976	transcription regulatory region sequence−specific DNA binding	31 of 829	1.50e−28	1.17
18	GO:1990837	sequence−specific double−stranded DNA binding	27 of 747	4.59e−24	1.15
19	GO:0043621	protein self−association	2 of 55	2.75e−02	1.15
20	GO:0043565	sequence−specific DNA binding	35 of 1047	1.47e−31	1.12

Biological Process (Gene Ontology)

	#term ID	term description	count in network	false discovery rate	strength
1	GO:0014036	neural crest cell fate specification	2 of 2	5.40e−04	2.59
2	GO:0060486	Clara cell differentiation	2 of 4	1.10e−03	2.29
3	GO:0090403	oxidative stress−induced premature senescence	2 of 4	1.10e−03	2.29
4	GO:0060510	type II pneumocyte differentiation	2 of 5	1.40e−03	2.19
5	GO:1901203	positive regulation of extracellular matrix assembly	2 of 5	1.40e−03	2.19
6	GO:2000794	regulation of epithelial cell proliferation involved in lung morphogenesis	2 of 7	2.20e−03	2.05
7	GO:0032909	regulation of transforming growth factor beta2 production	2 of 8	2.50e−03	1.99
8	GO:0048340	paraxial mesoderm morphogenesis	2 of 10	3.40e−03	1.89
9	GO:2001269	positive regulation of cysteine−type endopeptidase activity involved in apoptotic signaling pathway	2 of 10	3.40e−03	1.89
10	GO:0006978	DNA damage response, signal transduction by p53 class mediator resulting in transcription of p21 class mediator	3 of 17	2.70e−04	1.84
11	GO:1902894	negative regulation of pri−miRNA transcription by RNA polymerase II	2 of 13	5.00e−03	1.78
12	GO:0060487	lung epithelial cell differentiation	3 of 20	3.90e−04	1.77
13	GO:0002053	positive regulation of mesenchymal cell proliferation	4 of 28	2.38e−05	1.75
14	GO:0014898	cardiac muscle hypertrophy in response to stress	2 of 14	5.40e−03	1.75
15	GO:0033151	V(D)J recombination	2 of 14	5.40e−03	1.75
16	GO:0010464	regulation of mesenchymal cell proliferation	5 of 36	1.36e−06	1.74
17	GO:0043923	positive regulation by host of viral transcription	2 of 17	7.10e−03	1.66
18	GO:2000773	negative regulation of cellular senescence	2 of 17	7.10e−03	1.66
19	GO:1902893	regulation of pri−miRNA transcription by RNA polymerase II	4 of 35	5.24e−05	1.65
20	GO:0042093	T−helper cell differentiation	3 of 26	7.20e−04	1.65

C

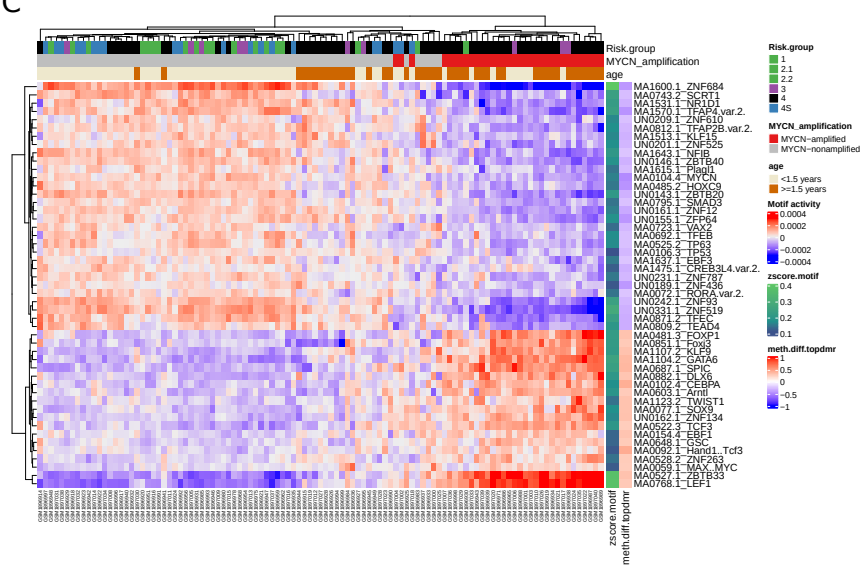
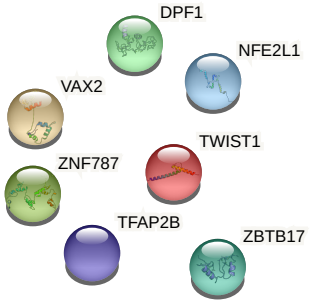


Figure A.14: (A) Regulatory PPI network based on motif activity results using MNA DMRs based on methylation microarray data from Henrich et al. 2016. (B) Biological process and molecular function Gene Ontology terms of the result TFs. (C) A heatmap of motif activity values for each DNA motif (rows) and a sample (columns). The heatmap annotations indicate stage according to INSS classification (Risk.group), presence of MYCN amplification, age, z-score of a DNA motif (zscore.motif) and average methylation difference of a top target MNA DMR of a corresponding motif (meth.diff.topdmr).

A



B

Molecular Function (Gene Ontology)

	#term ID	term description	count in network	false discovery rate	strength
1	GO:0001227	DNA−binding transcription repressor activity, RNA polymerase II−specific	2 of 229	9.30e−03	1.39
2	GO:0000978	RNA polymerase II proximal promoter sequence−specific DNA binding	4 of 492	8.46e−05	1.36
3	GO:0000977	RNA polymerase II transcription regulatory region sequence−specific DNA binding	5 of 647	5.06e−05	1.33
4	GO:0001228	DNA−binding transcription activator activity, RNA polymerase II−specific	3 of 408	1.40e−03	1.31
5	GO:0046982	protein heterodimerization activity	3 of 519	2.40e−03	1.21
6	GO:0003682	chromatin binding	2 of 501	3.81e−02	1.05
7	GO:0003712	transcription coregulator activity	2 of 534	4.11e−02	1.02
8	GO:0000981	DNA−binding transcription factor activity, RNA polymerase II−specific	6 of 1633	5.13e−05	1.01
9	GO:0003677	DNA binding	6 of 2457	1.30e−04	0.83

Biological Process (Gene Ontology)

	#term ID	term description	count in network	false discovery rate	strength
1	GO:0007398	ectoderm development	2 of 20	7.00e−04	2.45
2	GO:0048048	embryonic eye morphogenesis	2 of 33	1.30e−03	2.23
3	GO:0035137	hindlimb morphogenesis	2 of 36	1.40e−03	2.19
4	GO:0035136	forelimb morphogenesis	2 of 39	1.50e−03	2.16
5	GO:0048592	eye morphogenesis	3 of 145	5.10e−04	1.76
6	GO:0048593	camera−type eye morphogenesis	2 of 109	7.40e−03	1.71
7	GO:0060041	retina development in camera−type eye	2 of 136	1.01e−02	1.61
8	GO:0043010	camera−type eye development	3 of 292	1.80e−03	1.46
9	GO:0048598	embryonic morphogenesis	3 of 545	7.60e−03	1.19
10	GO:0000122	negative regulation of transcription by RNA polymerase II	4 of 809	1.50e−03	1.14
11	GO:0006366	transcription by RNA polymerase II	3 of 784	1.60e−02	1.03
12	GO:0045944	positive regulation of transcription by RNA polymerase II	4 of 1104	4.10e−03	1.01
13	GO:0006351	transcription, DNA−templated	7 of 2569	4.30e−04	0.88
14	GO:0009653	anatomical structure morphogenesis	5 of 1992	2.80e−03	0.85
15	GO:0009888	tissue development	4 of 1626	1.19e−02	0.84
16	GO:0006357	regulation of transcription by RNA polymerase II	6 of 2633	9.20e−04	0.80
17	GO:0006355	regulation of transcription, DNA−templated	7 of 3661	4.30e−04	0.73
18	GO:0007399	nervous system development	4 of 2206	3.00e−02	0.70
19	GO:0007275	multicellular organism development	6 of 4726	1.04e−02	0.55
20	GO:0048731	system development	5 of 4144	4.13e−02	0.53

C

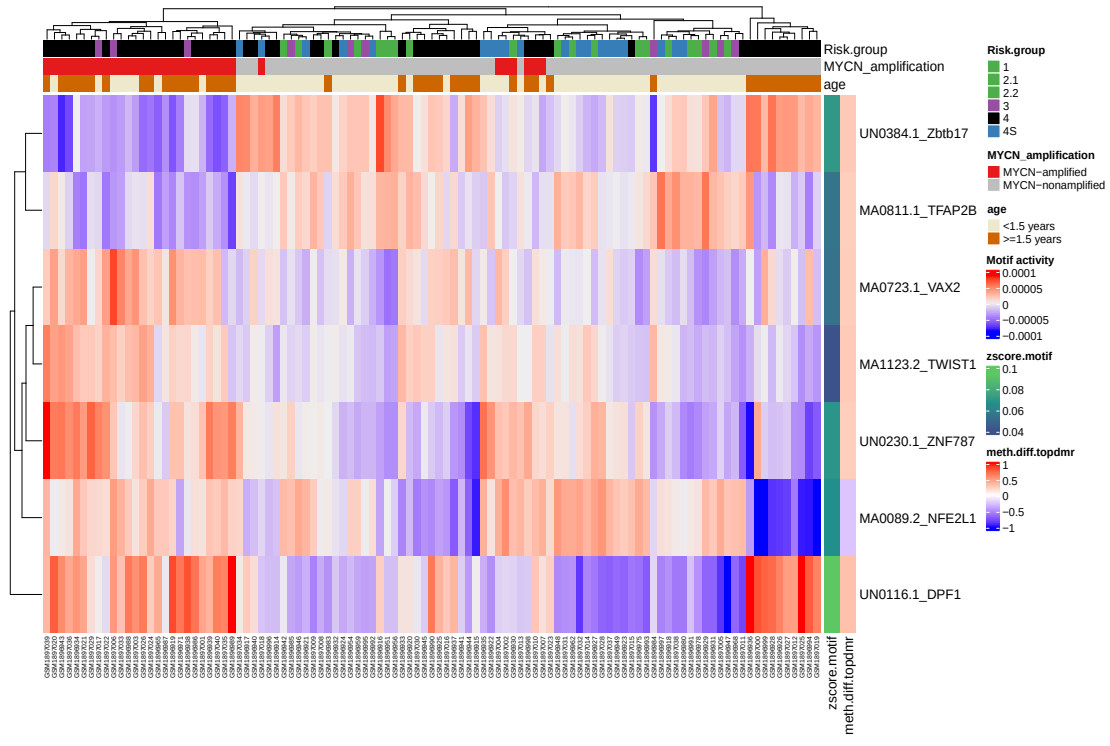


Figure A.15: (A) Regulatory PPI network based on motif activity results using HR_nMNA DMRs based on methylation microarray data from Henrich et al. 2016. (B) Biological process and molecular function Gene Ontology terms of the result TFs. (C) A heatmap of motif activity values for each DNA motif (rows) and a sample (columns). The heatmap annotations indicate stage according to INSS classification (Risk.group), presence of MYCN amplification, age, z-score of a DNA motif (zscore.motif) and average methylation difference of a top target HR_nMNA DMR of a corresponding motif (meth.diff.topdmr).

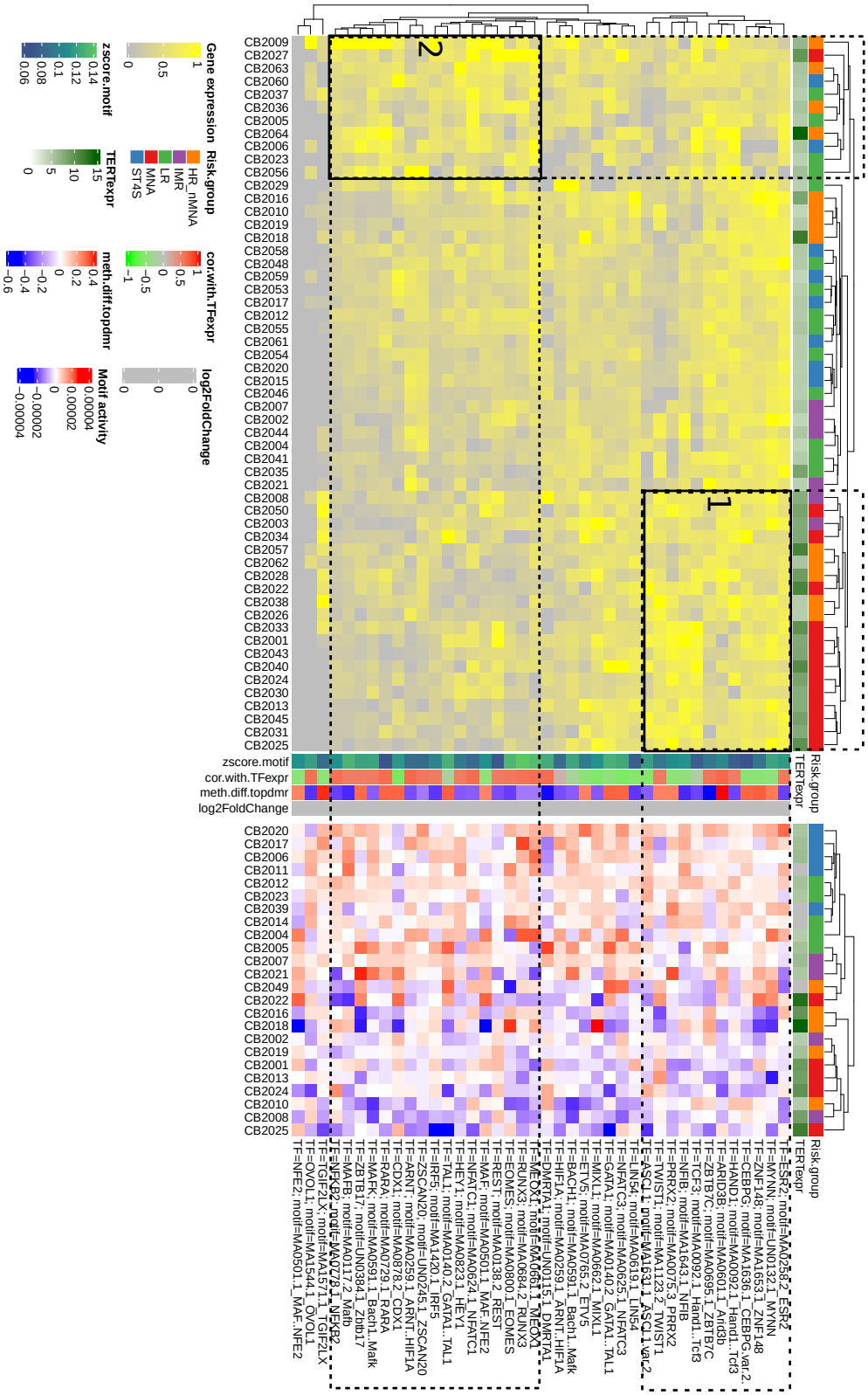
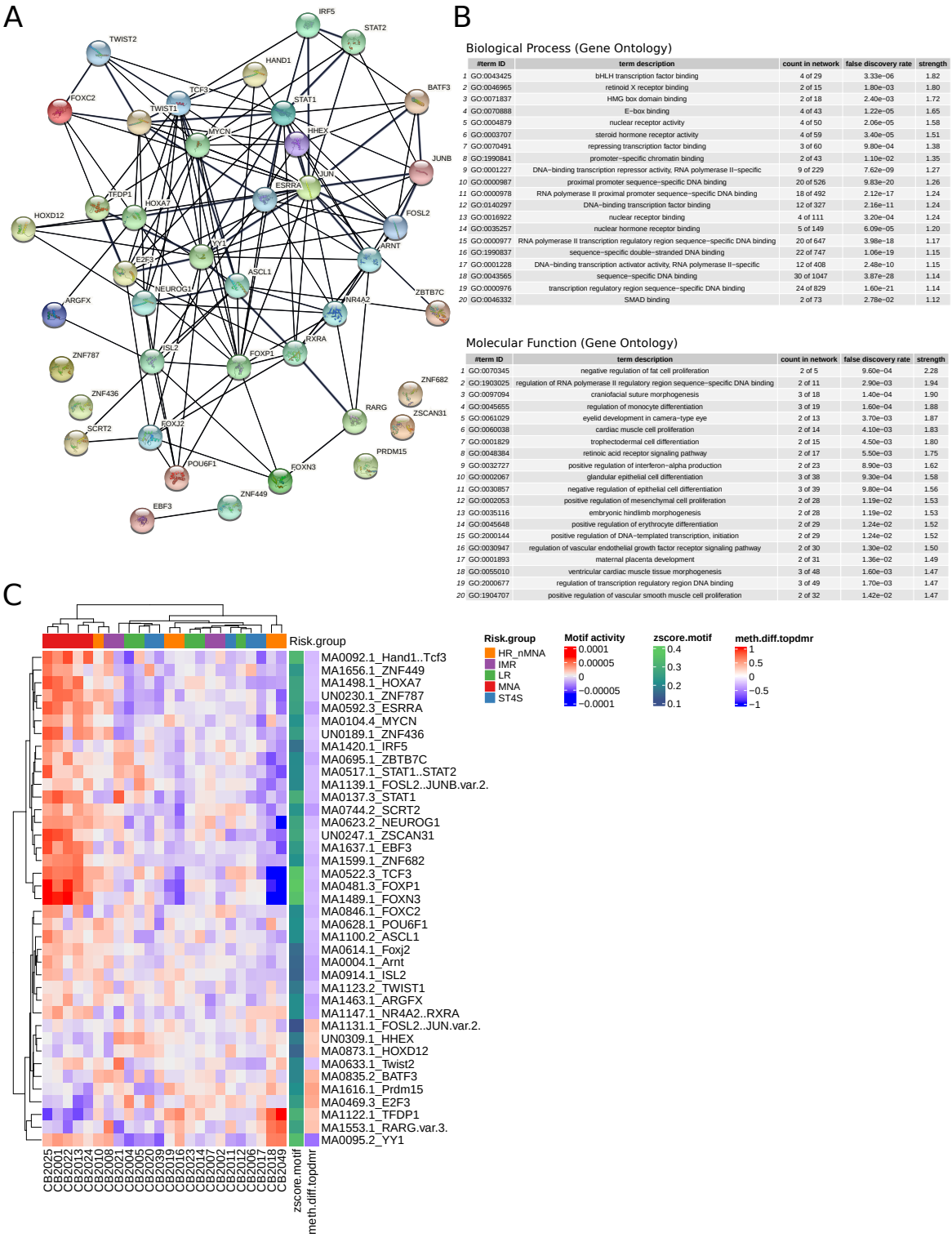


Figure A.16: HR_nMNA with high versus low TERT expression network represented as gene expression and motif activity heatmaps as described in Supplementary Figure A.12. Clusters 1 and 2 depict DNA motifs that activity is negative in MNA and HR_nMNA samples. Cluster 1 contains DNA motifs that TFs have relatively high expression in MNA and HR_nMNA samples. TFs in cluster 2 show high expression in subset of HR_nMNA samples with low TERT expression.



	#term ID	term description	count in network	false discovery rate	strength
1	GO:0043425	bHLH transcription factor binding	4 of 29	3.33e−06	1.82
2	GO:0046965	retinoid X receptor binding	2 of 15	1.80e−03	1.80
3	GO:0071837	HMG box domain binding	2 of 18	2.40e−03	1.72
4	GO:0070888	E−box binding	4 of 43	1.22e−05	1.65
5	GO:0004879	nuclear receptor activity	4 of 50	2.06e−05	1.58
6	GO:0003707	steroid hormone receptor activity	4 of 59	3.40e−05	1.51
7	GO:0070491	repressing transcription factor binding	3 of 60	9.80e−04	1.38
8	GO:1990841	promoter−specific chromatin binding	2 of 43	1.10e−02	1.35
9	GO:0001227	DNA−binding transcription repressor activity, RNA polymerase II−specific	9 of 229	7.62e−09	1.27
10	GO:0000987	proximal promoter sequence−specific DNA binding	20 of 526	9.83e−20	1.26
11	GO:0000978	RNA polymerase II proximal promoter sequence−specific DNA binding	18 of 492	2.12e−17	1.24
12	GO:0140297	DNA−binding transcription factor binding	12 of 327	2.16e−11	1.24
13	GO:0016922	nuclear receptor binding	4 of 111	3.20e−04	1.24
14	GO:0035257	nuclear hormone receptor binding	5 of 149	6.09e−05	1.20
15	GO:0000977	RNA polymerase II transcription regulatory region sequence−specific DNA binding	20 of 647	3.98e−18	1.17
16	GO:1990837	sequence−specific double−stranded DNA binding	22 of 747	1.06e−19	1.15
17	GO:0001228	DNA−binding transcription activator activity, RNA polymerase II−specific	12 of 408	2.48e−10	1.15
18	GO:0043565	sequence−specific DNA binding	30 of 1047	3.87e−28	1.14
19	GO:0000976	transcription regulatory region sequence−specific DNA binding	24 of 829	1.60e−21	1.14
20	GO:0046332	SMAD binding	2 of 73	2.78e−02	1.12

	#term ID	term description	count in network	false discovery rate	strength
1	GO:0070345	negative regulation of fat cell proliferation	2 of 5	9.60e−04	2.28
2	GO:1903025	regulation of RNA polymerase II regulatory region sequence−specific DNA binding	2 of 11	2.90e−03	1.94
3	GO:0097094	craniofacial suture morphogenesis	3 of 18	1.40e−04	1.90
4	GO:0045655	regulation of monocyte differentiation	3 of 19	1.60e−04	1.88
5	GO:0061029	eyelid development in camera−type eye	2 of 13	3.70e−03	1.87
6	GO:0060038	cardiac muscle cell proliferation	2 of 14	4.10e−03	1.83
7	GO:0001829	trophectodermal cell differentiation	2 of 15	4.50e−03	1.80
8	GO:0048384	retinoic acid receptor signaling pathway	2 of 17	5.50e−03	1.75
9	GO:0032727	positive regulation of interferon−alpha production	2 of 23	8.90e−03	1.62
10	GO:0002067	glandular epithelial cell differentiation	3 of 38	9.30e−04	1.58
11	GO:0030857	negative regulation of epithelial cell differentiation	3 of 39	9.80e−04	1.56
12	GO:0002053	positive regulation of mesenchymal cell proliferation	2 of 28	1.19e−02	1.53
13	GO:0035116	embryonic hindlimb morphogenesis	2 of 28	1.19e−02	1.53
14	GO:0045648	positive regulation of erythrocyte differentiation	2 of 29	1.24e−02	1.52
15	GO:2000144	positive regulation of DNA−templated transcription, initiation	2 of 29	1.24e−02	1.52
16	GO:0030947	regulation of vascular endothelial growth factor receptor signaling pathway	2 of 30	1.30e−02	1.50
17	GO:0001893	maternal placenta development	2 of 31	1.36e−02	1.49
18	GO:0055010	ventricular cardiac muscle tissue morphogenesis	3 of 48	1.60e−03	1.47
19	GO:2000677	regulation of transcription regulatory region DNA binding	3 of 49	1.70e−03	1.47
20	GO:1904707	positive regulation of vascular smooth muscle cell proliferation	2 of 32	1.42e−02	1.47

Figure A.17: (A) Regulatory PPI network based on motif activity results using MNA versus HR_nMNA DMRs. (B) Biological process and molecular function Gene Ontology terms of the result TFs. (C) A heatmap of motif activity values for each DNA motif (rows) and a sample (columns). The heatmap annotations indicate risk groups (Risk.group), z-score of a DNA motif (zscore.motif) and average methylation difference of a top target HR_nMNA DMR of a corresponding motif (meth.diff.topdmr).

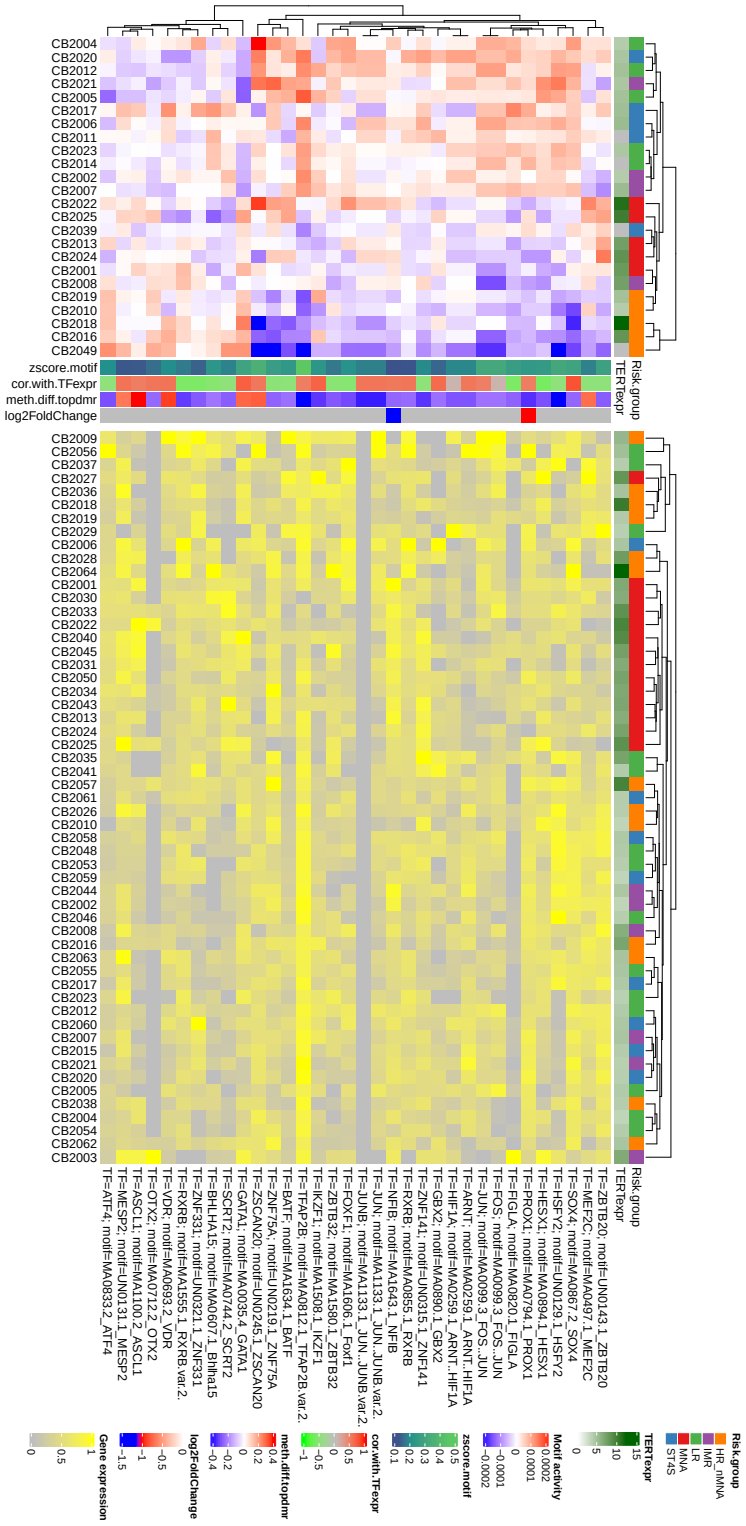
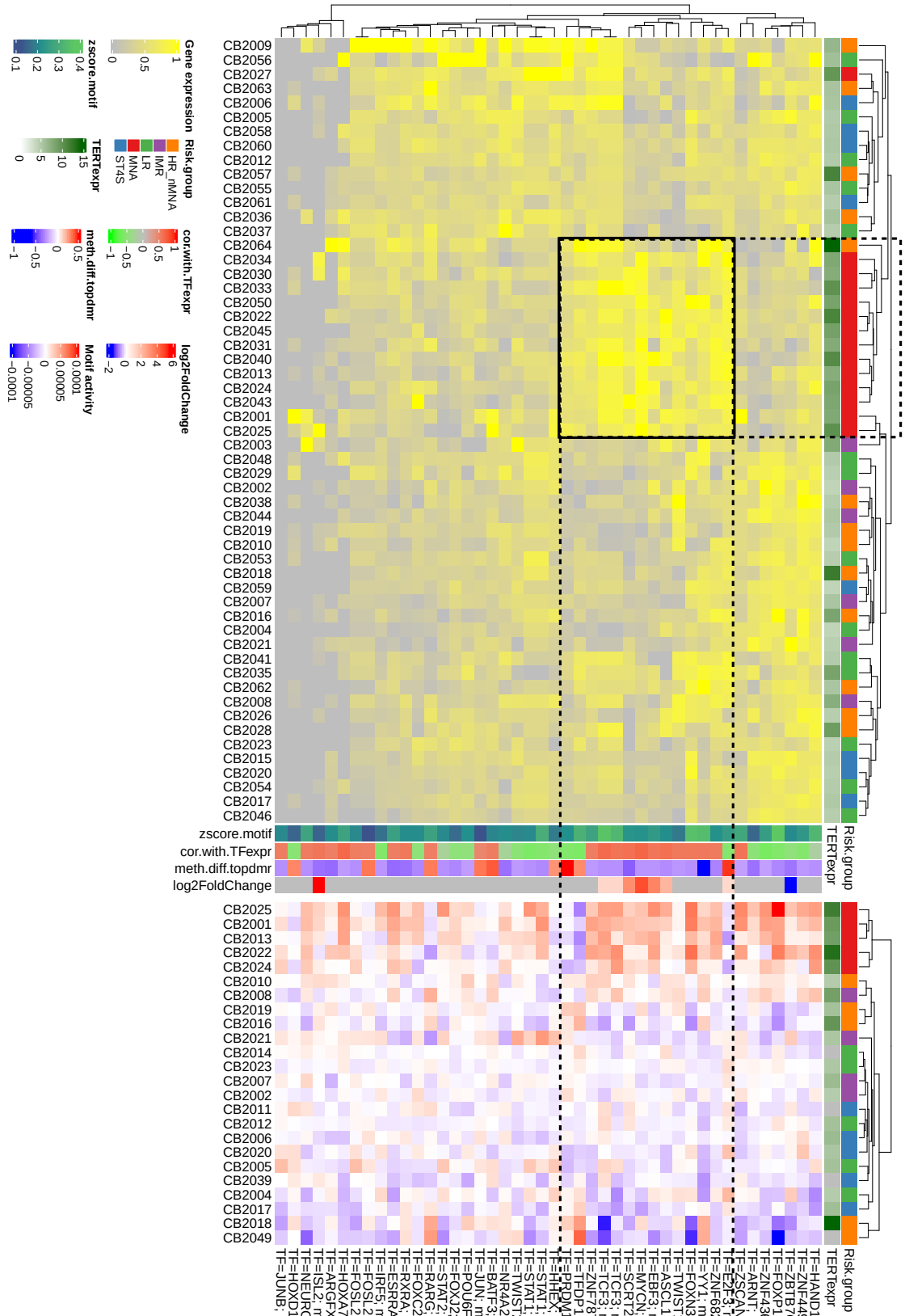
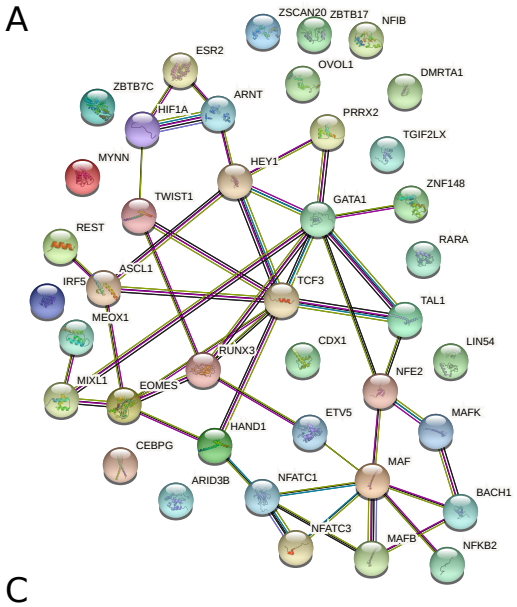


Figure A.18: HR_nMNA network represented as gene expression and motif activity heatmaps as described in Figure A.10.


Gene expression
Risk.group
HR_nMNA
IMR
LR
MNA
ST4S
zscore.motif
TERTexpr
cor.with.TFexpr
meth.diff.topdmr
log2FoldChange
Motif activity
CB2009
CB2056
CB2027
CB2063
CB2006
CB2005
CB2058
CB2060
CB2012
CB2057
CB2055
CB2061
CB2036
CB2037
CB2064
CB2034
CB2030
CB2033
CB2050
CB2022
CB2045
CB2031
CB2040
CB2013
CB2024
CB2043
CB2001
CB2025
CB2003
CB2048
CB2029
CB2002
CB2038
CB2044
CB2019
CB2010
CB2053
CB2018
CB2059
CB2007
CB2016
CB2004
CB2021
CB2041
CB2035
CB2062
CB2008
CB2026
CB2028
CB2023
CB2015
CB2020
CB2054
CB2017
CB2046
CB2025
CB2001
CB2013
CB2022
CB2024
CB2010
CB2008
CB2019
CB2016
CB2021
CB2014
CB2023
CB2007
CB2002
CB2011
CB2012
CB2006
CB2020
CB2005
CB2039
CB2004
CB2017
CB2018
CB2049
TF=HAND1; motif=MA0092.1_Hand1..Tcf3
TF=ZNF449; motif=MA1656.1_ZNF449
TF=ZBTB7C; motif=MA0695.1_ZBTB7C
TF=FOXP1; motif=MA0481.3_FOXP1
TF=ZNF436; motif=UN0189.1_ZNF436
TF=ARNT; motif=MA0004.1_Arnt
TF=ZSCAN31; motif=UN0247.1_ZSCAN31
TF=E2F3; motif=MA0469.1_E2F3
TF=ZNF682; motif=MA1595.1_ZNF682
TF=YY1; motif=MA0095.2_YY1
TF=FOXN3; motif=MA1489.1_FOXN3
TF=TWIST1; motif=MA1123.2_TWIST1
TF=ASCL1; motif=MA1100.2_ASCL1
TF=EBF3; motif=MA1637.1_EBF3
TF=MYCN; motif=MA0104.4_MYCN
TF=SCRT2; motif=MA0744.2_SCRT2
TF=TCF3; motif=MA0092.1_Hand1..Tcf3
TF=TCF3; motif=MA0522.3_TCF3
TF=ZNF787; motif=UN0230.1_ZNF787
TF=TFDP1; motif=MA1122.1_TFDP1
TF=HHEX; motif=UN0309.1_HHEX
TF=PRDM15; motif=MA1616.1_Prdm15
TF=STAT1; motif=MA0137.3_STAT1
TF=STAT1; motif=MA0517.1_STAT1..STAT2
TF=TWIST2; motif=MA0633.1_Twist2
TF=NR4A2; motif=MA1147.1_NR4A2..RXRA
TF=BATF3; motif=MA0835.2_BATF3
TF=JUN; motif=MA1131.1_FOSL2..JUN.var.2.
TF=POU6F1; motif=MA0628.1_POU6F1
TF=FOXI2; motif=MA0614.1_Foxi2
TF=STAT2; motif=MA0517.1_STAT1..STAT2
TF=RARG; motif=MA1553.1_RARG.var.3.
TF=FOXC2; motif=MA0846.1_FOXC2
TF=ESRRA; motif=MA0592.3_ESRRA
TF=RXRA; motif=MA1147.1_NR4A2..RXRA
TF=IRF5; motif=MA1420.1_IRF5
TF=FOSL2; motif=MA1131.1_FOSL2..JUN.var.2.
TF=FOSL2; motif=MA1139.1_FOSL2..JUNB.var.2.
TF=HOXA7; motif=MA1498.1_HOXA7
TF=ARGFX; motif=MA1463.1_ARGFX
TF=ISL2; motif=MA0914.1_ISL2
TF=NEUROG1; motif=MA0623.2_NEUROG1
TF=HOXD12; motif=MA0873.1_HOXD12
TF=JUNB; motif=MA1139.1_FOSL2..JUNB.var.2.
Risk.group
TERTexpr



B

Biological Process (Gene Ontology)

	#term ID	term description	count in network	false discovery rate	strength
1	GO:0043425	bHLH transcription factor binding	4 of 29	3.97e−06	1.83
2	GO:0070888	E−box binding	5 of 43	3.27e−07	1.75
3	GO:0001223	transcription coactivator binding	2 of 20	4.20e−03	1.69
4	GO:0001221	transcription cofactor binding	3 of 36	3.50e−04	1.61
5	GO:0000987	proximal promoter sequence−specific DNA binding	26 of 526	6.26e−30	1.38
6	GO:0000978	RNA polymerase II proximal promoter sequence−specific DNA binding	24 of 492	3.69e−27	1.38
7	GO:0001228	DNA−binding transcription activator activity, RNA polymerase II−specific	18 of 408	5.85e−19	1.33
8	GO:0001227	DNA−binding transcription repressor activity, RNA polymerase II−specific	10 of 229	3.05e−10	1.33
9	GO:1990837	sequence−specific double−stranded DNA binding	31 of 747	1.23e−34	1.31
10	GO:0000977	RNA polymerase II transcription regulatory region sequence−specific DNA binding	27 of 647	1.87e−29	1.31
11	GO:0001085	RNA polymerase II transcription factor binding	5 of 121	3.23e−05	1.31
12	GO:0004879	nuclear receptor activity	2 of 50	2.08e−02	1.29
13	GO:0000976	transcription regulatory region sequence−specific DNA binding	31 of 829	2.26e−33	1.26
14	GO:0003707	steroid hormone receptor activity	2 of 59	2.70e−02	1.22
15	GO:0043565	sequence−specific DNA binding	35 of 1047	1.44e−37	1.21
16	GO:0070491	repressing transcription factor binding	2 of 60	2.72e−02	1.21
17	GO:0008134	transcription factor binding	18 of 610	5.08e−16	1.16
18	GO:0042826	histone deacetylase binding	3 of 110	7.10e−03	1.13
19	GO:0002039	p53 binding	2 of 73	3.83e−02	1.13
20	GO:0140297	DNA−binding transcription factor binding	8 of 327	2.28e−06	1.08

Molecular Function (Gene Ontology)

	#term ID	term description	count in network	false discovery rate	strength
1	GO:0030221	basophil differentiation	2 of 2	2.80e−04	2.69
2	GO:1905064	negative regulation of vascular smooth muscle cell differentiation	2 of 8	1.70e−03	2.09
3	GO:0061140	lung secretory cell differentiation	2 of 9	2.10e−03	2.04
4	GO:0046886	positive regulation of hormone biosynthetic process	2 of 10	2.40e−03	1.99
5	GO:0043619	regulation of transcription from RNA polymerase II promoter in response to oxidative stress	2 of 11	2.80e−03	1.95
6	GO:0045647	negative regulation of erythrocyte differentiation	2 of 11	2.80e−03	1.95
7	GO:1903025	regulation of RNA polymerase II regulatory region sequence−specific DNA binding	2 of 11	2.80e−03	1.95
8	GO:0010002	cardioblast differentiation	2 of 12	3.20e−03	1.91
9	GO:0060218	hematopoietic stem cell differentiation	2 of 12	3.20e−03	1.91
10	GO:0021781	glial cell fate commitment	2 of 13	3.40e−03	1.88
11	GO:1902894	negative regulation of pri−miRNA transcription by RNA polymerase II	2 of 13	3.40e−03	1.88
12	GO:0035162	embryonic hemopoiesis	3 of 20	1.70e−04	1.87
13	GO:2000678	negative regulation of transcription regulatory region DNA binding	3 of 21	1.80e−04	1.84
14	GO:0021527	spinal cord association neuron differentiation	2 of 14	3.80e−03	1.84
15	GO:0045648	positive regulation of erythrocyte differentiation	4 of 29	8.73e−06	1.83
16	GO:0001829	trophectodermal cell differentiation	2 of 15	4.20e−03	1.81
17	GO:0030502	negative regulation of bone mineralization	2 of 16	4.70e−03	1.79
18	GO:0030949	positive regulation of vascular endothelial growth factor receptor signaling pathway	2 of 16	4.70e−03	1.79
19	GO:0046885	regulation of hormone biosynthetic process	3 of 25	2.90e−04	1.77
20	GO:1902036	regulation of hematopoietic stem cell differentiation	3 of 26	3.10e−04	1.75

C

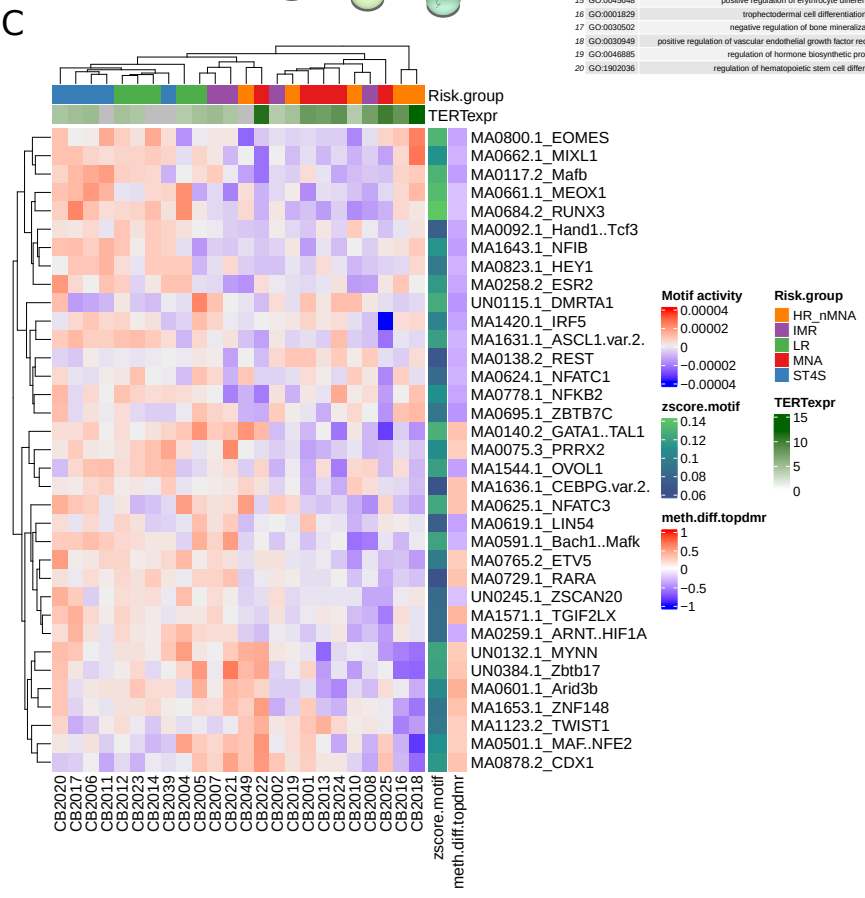


Figure A.20: (A) Regulatory PPI network based on motif activity results using HR_nMNA high versus low TERT DMRs. (B) Biological process and molecular function Gene Ontology terms of the result TFs. (C) A heatmap of motif activity values for each DNA motif (rows) and a sample (columns). The heatmap annotations indicate risk groups (Risk.group), z-score of a DNA motif (zscore.motif) and average methylation difference of a top target HR_nMNA high versus low TERT DMR of a corresponding motif (meth.diff.topdmr).

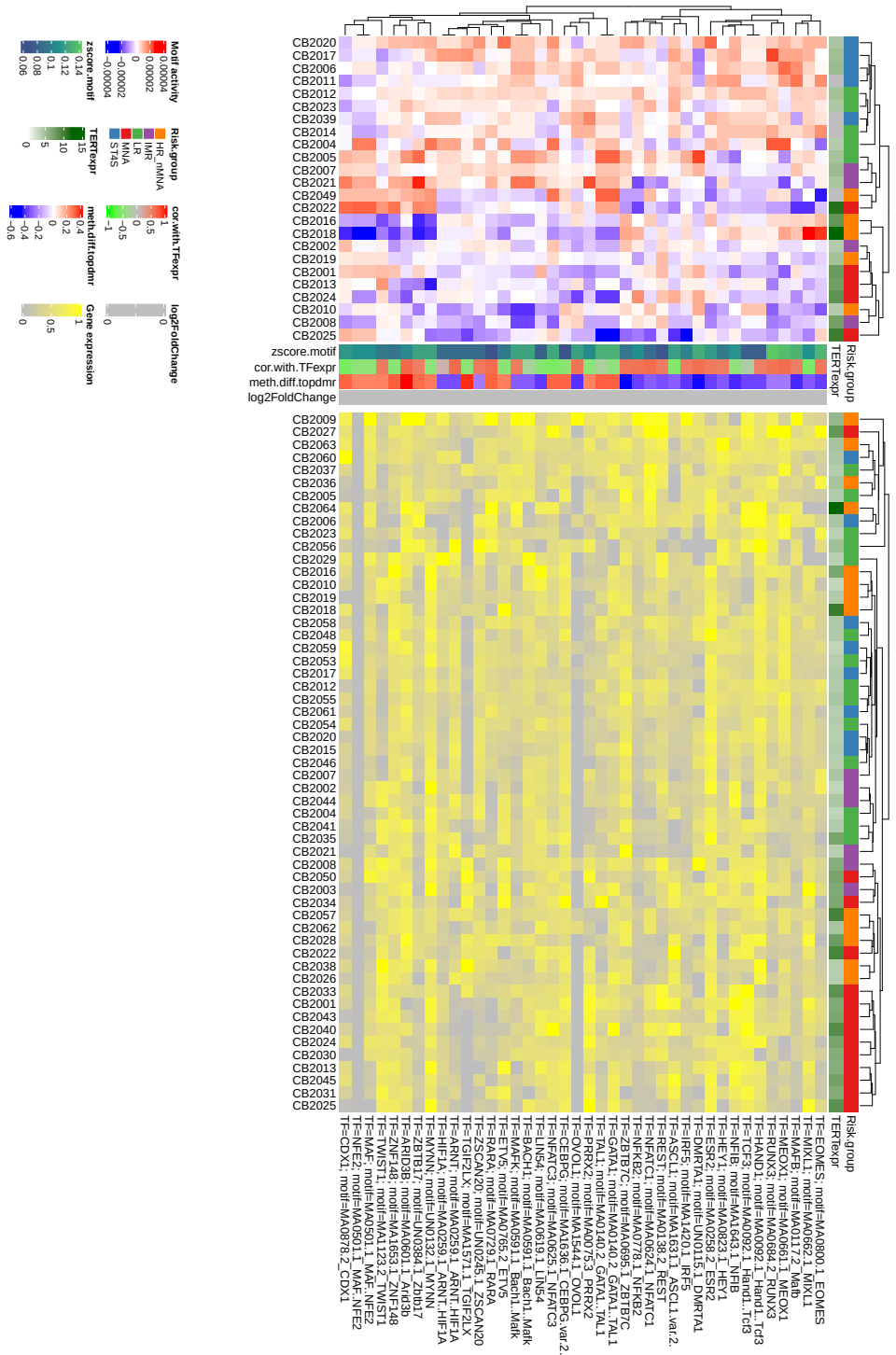


Figure A.21: The upper heatmap represents motif activity values, and the lower heatmap represents gene expression (normalized to 0-1 range) of TFs that bind to DNA motifs in HR_nMNA high versus low TERT network. Rows of both heatmaps are clustered according to motif activity. Top annotation indicate sample risk groups (Risk.group), expression of TERT gene (TERT.expr), and the right annotation indicate z-score of a DNA motif (zscore.motif), a Pearson correlation between motif activities and TFs expression across all samples (cor.with.TFexpr), average methylation difference of a top target MNA DMR of a corresponding motif (meth.diff.topdmr), log2 fold change of significantly differentially expressed genes between HR_nMNA high versus low TERT risk groups (log2FoldChange).

	chr	start	end	meth.diff	motif.name	correlation.with.gene	correlated.gene	genes.ovrlp.DMR
1	chr1	1601367	1602289	0.2661	UN0332.1_ZNF534	0.5097744	FNDC10	
2	chr1	2822872	2823882	0.2988	MA0868.2_SOX8	0.5090367	MMEL1	
3	chr1	2822872	2823882	0.2988	MA0813.1_TFAP2B.var.3.	0.5090367	MMEL1	
4	chr1	2825808	2826008	0.4462	MA1657.1_ZNF652	0.5225910	MMEL1	
5	chr1	2825808	2826008	0.4462	MA1600.1_ZNF684	0.5225910	MMEL1	
6	chr1	2825808	2826008	0.4462	UN0138.1_TLX3	0.5225910	MMEL1	
7	chr1	2827143	2828806	0.3327	MA1483.1_ELF2	0.5075307	MMEL1	
8	chr1	2827143	2828806	0.3327	MA1577.1_TLX2	0.5075307	MMEL1	
9	chr1	2828817	2829128	0.4210	MA1151.1_RORC	0.5481934	MMEL1	
10	chr1	2861505	2861830	0.3990	UN0223.1_ZNF766	0.5090367	MMEL1	
11	chr1	20309196	20314413	0.3975	MA0612.2_EMX1	0.7424707	VWA5B1	VWA5B1;
12	chr1	20315783	20320618	0.3967	MA0051.1_IRF2	0.8072298	VWA5B1	VWA5B1;
13	chr1	24360440	24369389	0.2776	MA1587.1_ZNF135	0.5517020	GRHL3	GRHL3;STPG1
14	chr1	75125449	75126837	−0.4170	MA0144.2_STAT3	0.5187558	LHX8	
15	chr1	88990902	88991447	0.3740	MA1476.1_DLX5	0.5624060	GTF2B	KYAT3;RBMXL1
16	chr1	88991521	88991882	0.3941	MA0499.2_MYOD1	0.6631579	GTF2B	KYAT3;RBMXL1
17	chr1	111885892	111885983	0.4988	MA1531.1_NR1D1	0.5699248	KCND3	KCND3
18	chr1	111885892	111885983	0.4988	MA1537.1_NR2F1.var.2.	0.5699248	KCND3	KCND3
19	chr1	116941610	116941797	−0.2944	MA0083.3_SRF	0.5368421	CD101	PTGFRN
20	chr2	23524327	23524584	0.3200	MA1470.1_BACH2.var.2.	0.5278195	KLHL29	KLHL29;
21	chr2	99970012	99970260	−0.4626	MA0643.1_Esrrg	0.5849624	REV1	AFF3
22	chr2	99970012	99970260	−0.4626	MA1143.1_FOSL1..JUND.var.2.	0.5849624	REV1	AFF3
23	chr2	99970012	99970260	−0.4626	MA1143.1_FOSL1..JUND.var.2.	0.5849624	REV1	AFF3
24	chr2	99970012	99970260	−0.4626	MA0482.2_GATA4	0.5849624	REV1	AFF3
25	chr2	100046885	100047222	0.3498	MA0725.1_VSX1	0.5368421	REV1	AFF3
26	chr2	109026663	109031199	−0.3011	MA0598.3_EHF	0.5067669	SH3RF3	
27	chr2	109026663	109031199	−0.3011	MA0914.1_ISL2	0.5067669	SH3RF3	
28	chr2	116815269	116816817	0.3316	MA0104.4_MYCN	0.5578947	DDX18	
29	chr2	116873094	116874414	0.2927	MA1122.1_TFDP1	0.5714286	DDX18	
30	chr2	229560315	229561079	0.3142	MA0809.2_TEAD4	0.7293233	PID1	DNER
31	chr2	236159561	236159631	−0.4747	UN0226.1_ZNF771	0.5067669	GBX2	
32	chr3	184247827	184248075	0.3120	MA0522.3_TCF3	0.7233083	ECE2	ALG3
33	chr4	1623207	1623432	0.3266	MA0017.2_NR2F1	0.5097744	NKX1−1	FAM53A
34	chr5	24561028	24561902	0.2956	MA1548.1_PLAGL2	0.5864662	CDH10	CDH10;
35	chr5	24561028	24561902	0.2956	UN0324.1_ZNF436	0.5864662	CDH10	CDH10;
36	chr6	37820623	37822126	−0.2716	MA0605.2_ATF3	0.6556391	BTBD9	ZFAND3
37	chr6	57960241	57960297	0.4014	UN0270.1_PHF21A	0.5157895	PRIM2	LINC00680;
38	chr6	112254648	112255191	−0.3930	MA0850.1_FOXP3	0.5458647	LAMA4	LAMA4;LAMA4−AS1;
39	chr8	55388639	55391423	0.3710	MA0873.1_HOXD12	0.7007519	TGS1	XKR4
40	chr8	66166517	66173631	0.3248	MA0859.1_Rarg	0.5172932	CRH	TRIM55
41	chr8	78801515	78804122	−0.4173	MA0024.3_E2F1	0.5699248	IL7	IL7
42	chr8	98972418	98972620	−0.4985	UN0189.1_ZNF436	0.5578947	VPS13B	VPS13B−DT
43	chr9	132814681	132815052	0.3311	MA0084.1_SRY	0.7428571	DDX31	AK8
44	chr9	132823423	132823875	0.2925	MA0662.1_MIXL1	0.7293233	DDX31	AK8
45	chr10	99834796	99835462	−0.2544	MA0669.1_NEUROG2	0.6225564	DNMBP	ABCC2
46	chr11	71064735	71065211	0.2513	UN0196.1_ZNF479	0.5067669	SHANK2	SHANK2
47	chr11	72590070	72590779	−0.3144	MA0068.2_PAX4	0.6195489	CLPB	PDE2A
48	chr12	3004784	3005224	−0.2543	MA1128.1_FOSL1..JUN	0.6060150	TSPAN9	TEAD4
49	chr12	3004784	3005224	−0.2543	MA1128.1_FOSL1..JUN	0.6060150	TSPAN9	TEAD4
50	chr12	3004784	3005224	−0.2543	MA1489.1_FOXN3	0.6060150	TSPAN9	TEAD4
51	chr12	3260056	3260201	−0.3126	UN0231.1_ZNF787	0.5278195	TSPAN9	TSPAN9
52	chr12	3378615	3383732	0.2506	MA1653.1_ZNF148	0.5714286	PRMT8	PRMT8
53	chr12	3378615	3383732	0.2506	UN0110.1_ATF6B	0.5714286	PRMT8	PRMT8
54	chr12	3378615	3383732	0.2506	UN0144.1_ZBTB22	0.5714286	PRMT8	PRMT8
55	chr12	102401108	102403767	−0.2823	MA0006.1_Ahr..Arnt	0.6270677	PMCH	HELLPAR;LINC02456;IGF1
56	chr12	102401108	102403767	−0.2823	MA0006.1_Ahr..Arnt	0.6270677	PMCH	HELLPAR;LINC02456;IGF1
57	chr13	28510532	28518665	0.2642	MA0079.4_SP1	0.6360902	POMP	
58	chr14	23353566	23354836	0.2697	MA0680.1_PAX7	0.6090226	SLC22A17	
59	chr14	45713161	45716040	−0.2869	MA0840.1_Creb5	0.5233083	MIS18BP1	LINC02303
60	chr15	80584638	80584901	−0.2873	MA0125.1_Nobox	0.5233083	ABHD17C	ARNT2
61	chr16	9948348	9949314	0.4877	MA0777.1_MYBL2	0.5082707	GRIN2A	GRIN2A
62	chr16	88245253	88245597	−0.3745	MA0889.1_GBX1	0.5248120	ZNF469	
63	chr17	6209019	6209663	0.2644	UN0328.1_ZNF479	0.6812030	WSCD1	
64	chr19	30335317	30335682	0.3353	MA0684.2_RUNX3	0.6932331	ZNF536	ZNF536
65	chr19	35131370	35131547	0.5946	MA0071.1_RORA	0.5789474	LGI4	LGI4
66	chr22	34529014	34531011	0.2654	MA1553.1_RARG.var.3.	0.7654135	LARGE1	
67	chrX	71322535	71323050	−0.2610	MA0619.1_LIN54	0.6075188	TAF1	
68	chrX	96429596	96429849	−0.3017	MA0507.1_POU2F2	0.5578947	DIAPH2	
69	chrX	96843236	96843815	−0.2978	MA1540.1_NR5A1	0.5473684	DIAPH2	DIAPH2

Figure A.22: Top target MNA DMRs from the MNA regulatory network. Each row corresponds to a predicted target DMR that was the most significant from the difference of the log-likelihood of the model in which only the occurrences for the motif (motif.name) in the DMRs were removed, and the full model. Columns correspond to genomic coordinates of target DMRs, average methylation difference between MNA and low + intermediate risk samples within DMRs (meth.diff), name of the motif that was removed in the model to detect target DMRs (motif.name), correlation between average % DNA methylation in a DMR and gene expression of nearby gene to the DMR (correlation.with.gene), name of the correlated gene (correlated.gene), and names of genes that overlap with the DMR (genes.ovrlp.DMR).

	chr	start	end	meth.diff	motif.name	correlation.with.gene	correlated.gene	genes.ovrlp.DMR
1	chr1	2800053	2800542	0.2504	MA0607.1_Bhlha15	0.5496994	MMEL1	TTC34
2	chr1	2855229	2855353	0.2843	MA0820.1_FIGLA	0.5527115	MMEL1	
3	chr1	2855402	2855456	0.4175	MA0812.1_TFAP2B.var.2.	0.5240970	MMEL1	
4	chr1	2855402	2855456	0.4175	UN0129.1_HSFY2	0.5240970	MMEL1	
5	chr1	32925676	32926348	−0.3819	MA1100.2_ASCL1	0.6573632	TMEM54	
6	chr1	101803719	101806864	0.2780	MA1643.1_NFIB	0.6526316	OLFM3	OLFM3
7	chr1	101983903	101985996	0.3955	MA1133.1_JUN..JUNB.var.2.	0.6466165	OLFM3	OLFM3
8	chr1	101983903	101985996	0.3955	MA1133.1_JUN..JUNB.var.2.	0.6466165	OLFM3	OLFM3
9	chr1	102098623	102100148	0.3645	UN0315.1_ZNF141	0.7654135	OLFM3	
10	chr1	102761036	102761696	0.2819	UN0219.1_ZNF75A	0.5744361	OLFM3	
11	chr2	10041998	10042771	−0.3001	UN0245.1_ZSCAN20	0.5187970	KLF11	
12	chr3	2970811	2971089	0.3692	MA0794.1_PROX1	0.5789474	IL5RA	CNTN4
13	chr3	96612749	96613550	0.2692	MA0855.1_RXRB	0.5112782	MTRNR2L12	
14	chr3	96629420	96631576	0.2715	MA0712.2_OTX2	0.6195489	EPHA6	
15	chr3	178845029	178849768	0.2920	MA0744.2_SCRT2	0.5518797	KCNMB2	KCNMB2−AS1
16	chr4	1472719	1475433	0.2576	MA1634.1_BATF	0.5278195	NKX1−1	
17	chr4	144592447	144597290	0.2594	MA0890.1_GBX2	0.5218045	GYPA	
18	chr4	156365441	156367449	0.2757	MA0867.2_SOX4	0.5669173	CTSO	
19	chr4	175226250	175226667	0.3527	MA1555.1_RXRB.var.2.	0.5563910	ADAM29	
20	chr4	175237163	175239384	0.2945	UN0143.1_ZBTB20	0.6466165	ADAM29	
21	chr5	19774367	19778360	0.3181	MA0833.2_ATF4	0.5924812	CDH18	CDH18
22	chr5	20317785	20320109	0.3609	MA1508.1_IKZF1	0.5654135	CDH18	CDH18;CDH18−AS1
23	chr6	50442243	50444028	0.3162	MA0894.1_HESX1	0.7263158	DEFB112	
24	chr6	50573121	50574190	0.3278	MA1606.1_Foxf1	0.5849624	DEFB112	
25	chr6	54697628	54703766	0.3112	MA0259.1_ARNT..HIF1A	0.5142857	FAM83B	
26	chr6	54697628	54703766	0.3112	MA0259.1_ARNT..HIF1A	0.5142857	FAM83B	
27	chr14	105900655	105900690	0.3008	MA1580.1_ZBTB32	0.5127820	TMEM121	
28	chr15	63087526	63087590	0.2547	MA0099.3_FOS..JUN	0.5007519	LACTB	
29	chr15	63087526	63087590	0.2547	MA0099.3_FOS..JUN	0.5007519	LACTB	
30	chr16	55657684	55659709	0.2993	UN0321.1_ZNF331	0.5142857	SLC6A2	SLC6A2
31	chr16	66426354	66426732	−0.2736	MA0497.1_MEF2C	0.5082707	BEAN1	
32	chrX	54248069	54248174	−0.2886	MA0035.4_GATA1	0.5924812	FAM120C	WNK3
33	chrX	111829085	111830156	−0.3428	MA0693.2_VDR	0.5639098	ALG13	TRPC5
34	chrX	129929191	129929570	−0.2595	UN0131.1_MESP2	0.5353383	BCORL1	;UTP14A

Figure A.23: Top target HR_nMNA DMRs from the HR_nMNA regulatory network. Columns and rows are explained in Figure A.22.

	chr	start	end	meth.diff	motif.name	correlation.with.gene	correlated.gene	genes.ovrlp.DMR
1	chr1	102863310	102863935	0.3319	MA1599.1_ZNF682	0.5473684	COL11A1	
2	chr2	178865778	178865843	−0.4742	MA1616.1_Prdm15	0.5699248	TTN	;CCDC141
3	chr2	199360096	199364864	0.3812	MA1147.1_NR4A2..RXRA	0.7398496	SATB2	SATB2
4	chr2	199360096	199364864	0.3812	MA1147.1_NR4A2..RXRA	0.7398496	SATB2	SATB2
5	chr3	2025611	2026148	0.3709	MA0695.1_ZBTB7C	0.5473684	CNTN4	
6	chr3	124115411	124115575	0.2849	MA1637.1_EBF3	0.6120301	UMPS	KALRN
7	chr3	124115411	124115575	0.2849	MA0104.4_MYCN	0.6120301	UMPS	KALRN
8	chr3	131361925	131362384	0.6147	MA0095.2_YY1	0.5624060	NUDT16	;NUDT16L2P
9	chr4	44406560	44408014	0.4309	MA0744.2_SCRT2	0.7142857	KCTD8	KCTD8
10	chr4	46512753	46515466	0.2956	MA0517.1_STAT1..STAT2	0.5984962	GABRA2	
11	chr4	46512753	46515466	0.2956	MA0517.1_STAT1..STAT2	0.5984962	GABRA2	
12	chr4	76628871	76628952	0.2938	MA0623.2_NEUROG1	0.5338346	SHROOM3	SHROOM3
13	chr4	186728518	186730701	0.2942	MA0092.1_Hand1..Tcf3	0.5308271	FAT1	
14	chr4	186728518	186730701	0.2942	MA0092.1_Hand1..Tcf3	0.5308271	FAT1	
15	chr5	23246634	23247550	0.3821	MA0914.1_ISL2	0.5398496	CDH12	
16	chr5	24919697	24921559	0.2934	MA1489.1_FOXN3	0.5082707	CDH10	
17	chr5	25622573	25623767	0.3126	MA0137.3_STAT1	0.5007519	CDH10	
18	chr5	129503576	129503992	0.3979	MA1463.1_ARGFX	0.6054224	ADAMTS19	ADAMTS19;
19	chr5	129503576	129503992	0.3979	MA0592.3_ESRRA	0.6054224	ADAMTS19	ADAMTS19;
20	chr5	129650542	129652267	0.3226	MA0846.1_FOXC2	0.6641574	ADAMTS19	ADAMTS19;
21	chr5	129650542	129652267	0.3226	MA0614.1_Foxj2	0.6641574	ADAMTS19	ADAMTS19;
22	chr5	129650542	129652267	0.3226	MA0481.3_FOXP1	0.6641574	ADAMTS19	ADAMTS19;
23	chr8	49897575	49898491	0.2932	MA0522.3_TCF3	0.5203008	SNTG1	
24	chr8	49983113	49985424	0.2974	MA1139.1_FOSL2..JUNB.var.2.	0.6135338	SNTG1	SNTG1
25	chr8	49983113	49985424	0.2974	MA1139.1_FOSL2..JUNB.var.2.	0.6135338	SNTG1	SNTG1
26	chr8	108693395	108695308	0.2790	MA1123.2_TWIST1	0.5774436	TMEM74	TMEM74
27	chr8	108700084	108701139	0.2768	MA1656.1_ZNF449	0.6917293	TMEM74	TMEM74
28	chr8	112513699	112514761	0.2832	MA1420.1_IRF5	0.5142857	CSMD3	CSMD3
29	chr8	112657144	112657789	0.3065	MA1100.2_ASCL1	0.5969925	CSMD3	CSMD3
30	chr8	112660939	112661474	0.2612	MA0628.1_POU6F1	0.5834586	CSMD3	CSMD3
31	chr8	112861722	112863156	0.2529	UN0247.1_ZSCAN31	0.6541353	CSMD3	CSMD3
32	chr8	112863766	112865362	0.2671	MA1498.1_HOXA7	0.6781955	CSMD3	CSMD3
33	chr8	112901526	112904220	0.2739	UN0230.1_ZNF787	0.5729323	CSMD3	CSMD3
34	chr8	113897020	113898003	0.2711	MA0633.1_Twist2	0.5187970	CSMD3	
35	chr11	127175524	127176035	−0.2922	MA0873.1_HOXD12	0.6315789	KIRREL3	
36	chr12	132416530	132416607	−0.4376	MA0469.3_E2F3	0.5864662	GALNT9	
37	chr14	62686775	62688197	0.3776	MA0004.1_Arnt	0.6345865	KCNH5	
38	chr16	48107799	48109588	−0.3129	MA1553.1_RARG.var.3.	0.5591399	ABCC12	ABCC12
39	chr16	48125249	48146970	−0.2804	MA1122.1_TFDP1	0.5037079	ABCC12	ABCC12
40	chr17	40618262	40619724	−0.2625	UN0309.1_HHEX	0.5909774	CCR7	
41	chr17	42405335	42405603	−0.3251	MA1131.1_FOSL2..JUN.var.2.	0.5639098	STAT3	CAVIN1
42	chr17	42405335	42405603	−0.3251	MA1131.1_FOSL2..JUN.var.2.	0.5639098	STAT3	CAVIN1
43	chr19	13299090	13299167	0.2526	UN0189.1_ZNF436	0.5308271	IER2	CACNA1A
44	chr20	57800818	57801409	−0.3796	MA0835.2_BATF3	0.5669173	RAB22A	

Figure A.24: Top target MNA and HR_nMNA DMRs from the MNA vs HR_nMNA regulatory network. Columns and rows are explained in Figure A.22.

	chr	start	end	meth.diff	motif.name	correlation.with.gene	correlated.gene	genes.ovrlp.DMR
1	chr1	2815922	2822120	0.4002	MA1420.1_IRF5	0.5210849	MMEL1	
2	chr1	3557243	3557839	0.3872	MA0800.1_EOMES	0.5082707	MEGF6	MEGF6
3	chr1	69806126	69806457	−0.3082	MA0729.1_RARA	0.7458647	LRRC7	LRRC7
4	chr1	115561674	115562979	0.3980	MA0619.1_LIN54	0.5172932	VANGL1	
5	chr1	115640104	115640940	0.4624	MA0695.1_ZBTB7C	0.6511278	VANGL1	
6	chr2	115159224	115159330	−0.2579	MA0075.3_PRRX2	0.6330827	DPP10	DPP10;DPP10−AS1
7	chr3	7127822	7134678	−0.2597	MA0765.2_ETV5	0.6406015	GRM7	GRM7
8	chr3	7219103	7219679	−0.2992	UN0384.1_Zbtb17	0.6646617	GRM7	GRM7
9	chr4	120303437	120304675	−0.2544	MA1123.2_TWIST1	0.5308271	MAD2L1	
10	chr5	1446987	1454264	0.3336	MA0662.1_MIXL1	0.7962321	SLC6A3	
11	chr5	84396832	84398474	−0.3083	MA1636.1_CEBPG.var.2.	0.6406015	EDIL3	EDIL3−DT
12	chr5	84461958	84462667	−0.4041	MA0601.1_Arid3b	0.6375940	EDIL3	EDIL3−DT
13	chr5	84621191	84621991	−0.3193	MA1653.1_ZNF148	0.5804511	EDIL3	
14	chr5	85071077	85074930	−0.2527	MA0501.1_MAF..NFE2	0.5127820	EDIL3	
15	chr5	85071077	85074930	−0.2527	MA0501.1_MAF..NFE2	0.5127820	EDIL3	
16	chr7	31336327	31336481	−0.3092	MA0878.2_CDX1	0.8004923	NEUROD6	
17	chr7	31336327	31336481	−0.3092	MA0625.1_NFATC3	0.8004923	NEUROD6	
18	chr9	135780889	135782350	0.3546	MA0138.2_REST	0.8511278	KCNT1	KCNT1
19	chr10	127069073	127069182	0.3840	MA0258.2_ESR2	0.5654135	DOCK1	DOCK1
20	chr10	127069073	127069182	0.3840	MA0778.1_NFKB2	0.5654135	DOCK1	DOCK1
21	chr12	100945730	100950777	0.2688	UN0245.1_ZSCAN20	0.6406015	ANO4	ANO4
22	chr12	132322023	132322174	0.3266	MA1631.1_ASCL1.var.2.	0.7263158	GALNT9	GALNT9
23	chr12	132322023	132322174	0.3266	MA0591.1_Bach1..Mafk	0.7263158	GALNT9	GALNT9
24	chr12	132322023	132322174	0.3266	MA0591.1_Bach1..Mafk	0.7263158	GALNT9	GALNT9
25	chr18	3898015	3898093	−0.3822	MA1571.1_TGIF2LX	0.7127820	TGIF1	DLGAP1
26	chr18	4364640	4365941	−0.2685	UN0132.1_MYNN	0.7609023	DLGAP1	DLGAP1
27	chr18	4606002	4607211	−0.3140	MA0140.2_GATA1..TAL1	0.6360902	DLGAP1	
28	chr18	4606002	4607211	−0.3140	MA0140.2_GATA1..TAL1	0.6360902	DLGAP1	
29	chr18	11637255	11637373	0.3751	MA0092.1_Hand1..Tcf3	0.7368421	GNAL	NPIPB1P
30	chr18	11637255	11637373	0.3751	MA0092.1_Hand1..Tcf3	0.7368421	GNAL	NPIPB1P
31	chr18	11691896	11708965	0.3265	MA0624.1_NFATC1	0.8631579	GNAL	GNAL
32	chr18	55300167	55303982	0.2691	MA0684.2_RUNX3	0.7293233	CCDC68	TCF4
33	chr19	1801308	1801579	0.3574	MA0259.1_ARNT..HIF1A	0.6120301	ATP8B3	ATP8B3
34	chr19	1801308	1801579	0.3574	MA0259.1_ARNT..HIF1A	0.6120301	ATP8B3	ATP8B3
35	chr19	43436051	43437507	0.2693	MA0661.1_MEOX1	0.5647597	LYPD3	
36	chr20	62663175	62663402	0.4583	UN0115.1_DMRTA1	0.7323308	NTSR1	SLCO4A1;SLCO4A1−AS1
37	chrX	65699250	65699735	0.3357	MA0823.1_HEY1	0.5248120	MSN	MSN
38	chrX	65709646	65710325	0.4085	MA1544.1_OVOL1	0.6090226	MSN	MSN
39	chrX	65729515	65730261	0.4219	MA0117.2_Mafb	0.6360902	MSN	MSN
40	chrX	65729515	65730261	0.4219	MA1643.1_NFIB	0.6360902	MSN	MSN

Figure A.25: Top target HR_nMNA DMRs from the HR_nMNA TERT high vs low expression regulatory network. Columns and rows are explained in Figure A.22.

B

Supplementary Material for Chapter 3

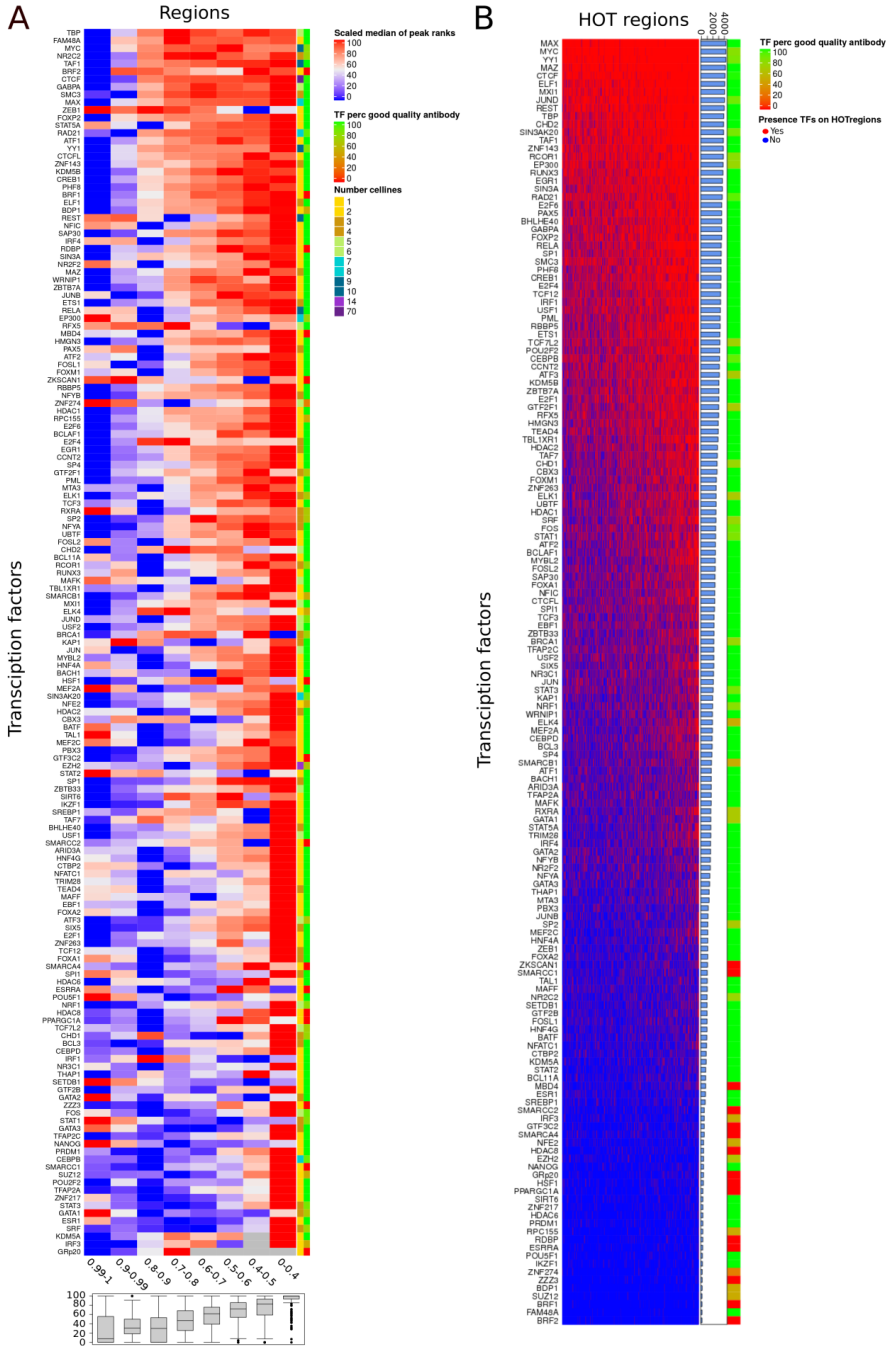


Figure B.1: A) HOT peaks are top rank TF peaks for majority of TFs. A heatmap shows median of ranks of TFs per HOT region (the highest rank is 1, the lowest 100). In comparison to less HOT regions, HOT region have the lowest median of ranks (see boxplots to the right of the heatmap). Annotation bars above the heatmap show percentage of good quality of antibodies (quality is per experiment is either good or caution) and number of cell lines per TF. B) Distribution of different TFs on HOT regions. Heatmap shows presence/absence of TFs on HOT regions. Hot regions are sorted from the most HOT regions on top to the less HOT regions in bottom. Annotation bar of quality of antibodies is as in S1A and the barplot indicates number of HOT regions present in each TF. The scatterplot to the right from the heatmap shows number of TFs per HOT region.

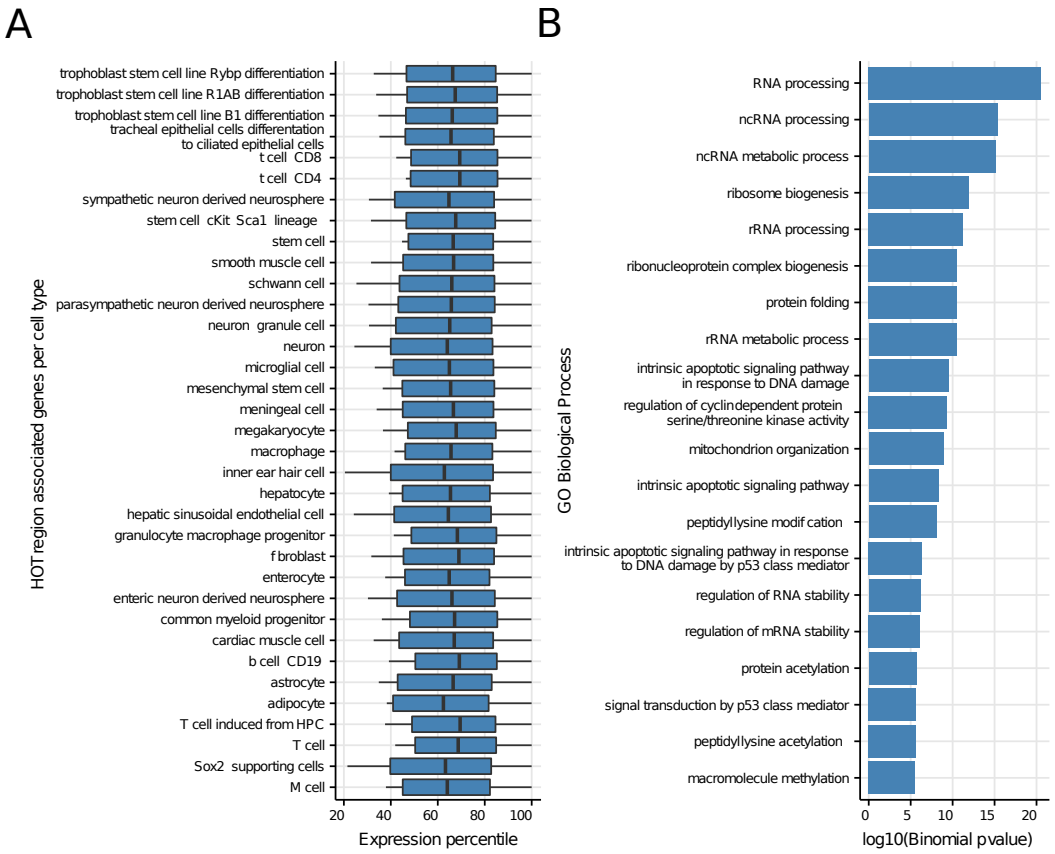
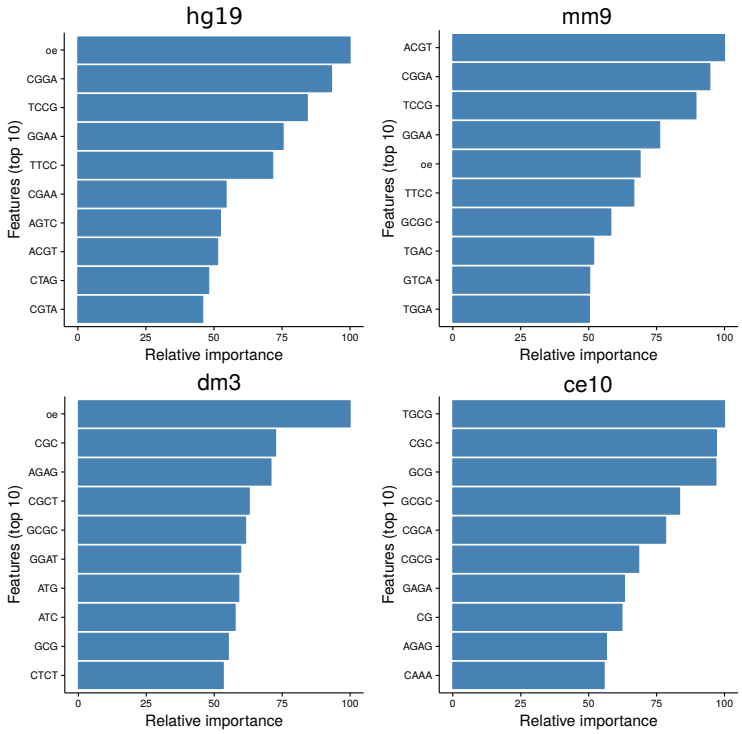


Figure B.2: A) Expression profiles of genes associated with HOT regions across cell types in mouse (Expression Atlas EBI databases using fantom5 CAGE expression). The genes are stably expressed in all 35 cell types between 40th and 80th percentile. B) Functional enrichment analysis with Gene Ontology and and KEGG pathway on genes associated with murine HOT regions. HOT regions are significantly enriched for terms that relate to housekeeping functions.

A



B

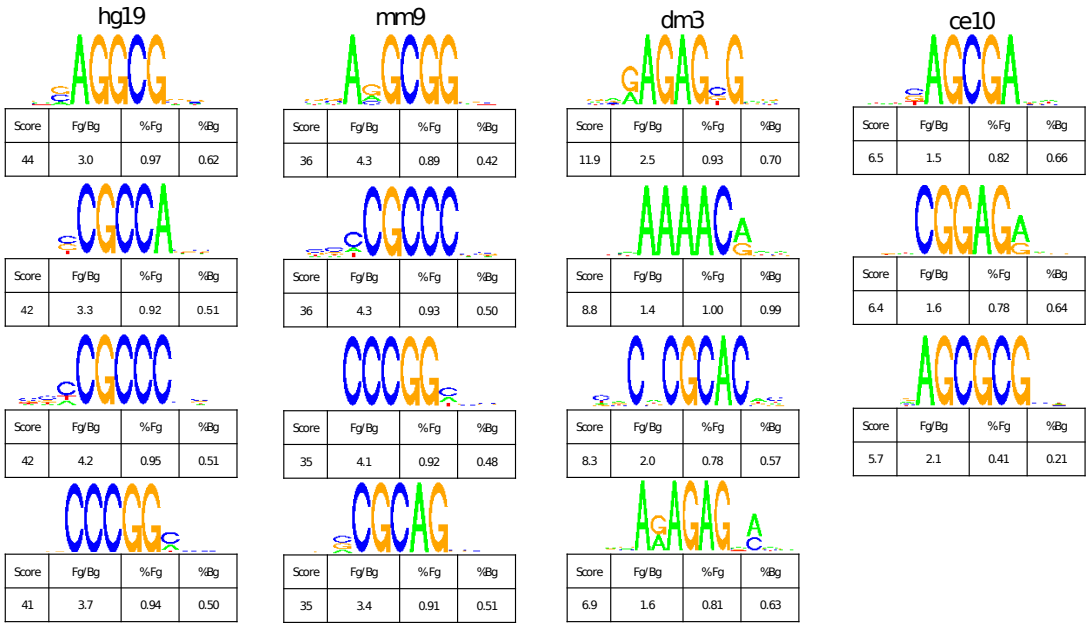


Figure B.3: A) Top 10 features selected from the predictive model of "hotness" of genomic regions using a penalized multivariate regression method for *H.sapiens*, *M.musculus*, *D.melanogaster*, and *C.elegans*. B) Discriminative motif discovery results between HOT and non-HOT regions in *H.sapiens*, *M.musculus*, *D.melanogaster*, and *C.elegans*. MotifRG was used to find longer sequence patterns that could discriminate between HOT and COLD regions. Motif discovery resulted in short, mostly G(CG) rich motifs in all four organisms. Top four most enriched motifs are shown for each organism, except *C.elegans*, for which there were three statistically significant motifs. Table under each motif contains the enrichment statistics for the corresponding motif: Score - motifRG calculated score, Fg/Bg, log2 ratio of percentage of foreground sequences (HOT regions) that contained the motif Vs the percentage of background sequences(non-HOT regions) containing the motif; %Fg - percentage of foreground sequences containing the motif; %Bg - percentage of background sequences containing the motif.

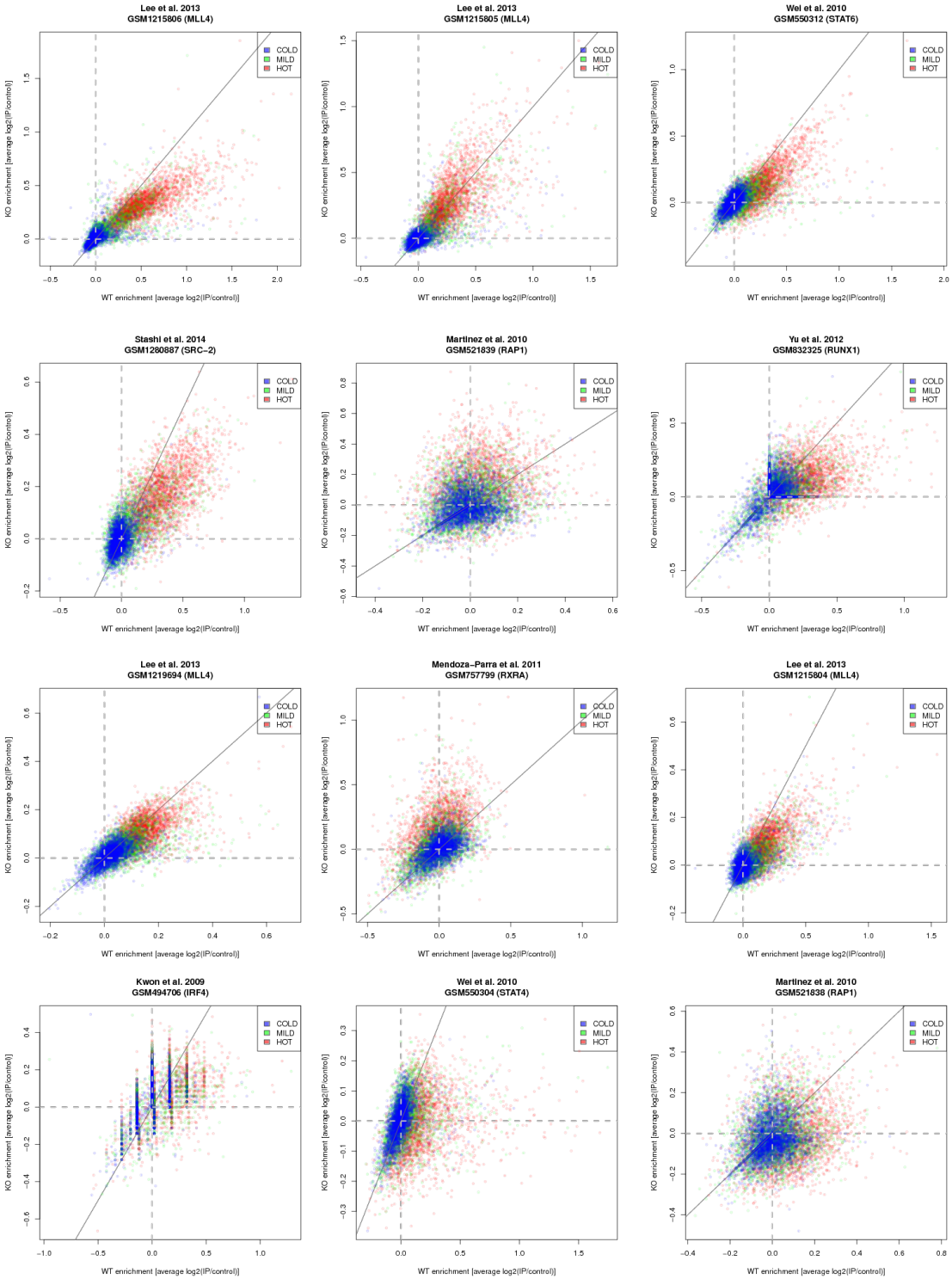


Figure B.4: Most of wild-type (WT) and knock-out (KO) ChIP-seq samples have scores that show strong correlation on HOT regions. Scatterplots show the relationship between the ChIP enrichment in WT and KO experiments for KO samples shown in Figure 3A (besides TF E2A for which we haven't found a WT experiment). Signal strength is measured as log2 RPKM ChIP / Input. The color on figures designates HOT (red), MILD (green), and COLD (blue) regions. Each dot represents one transcription factor WT peak.

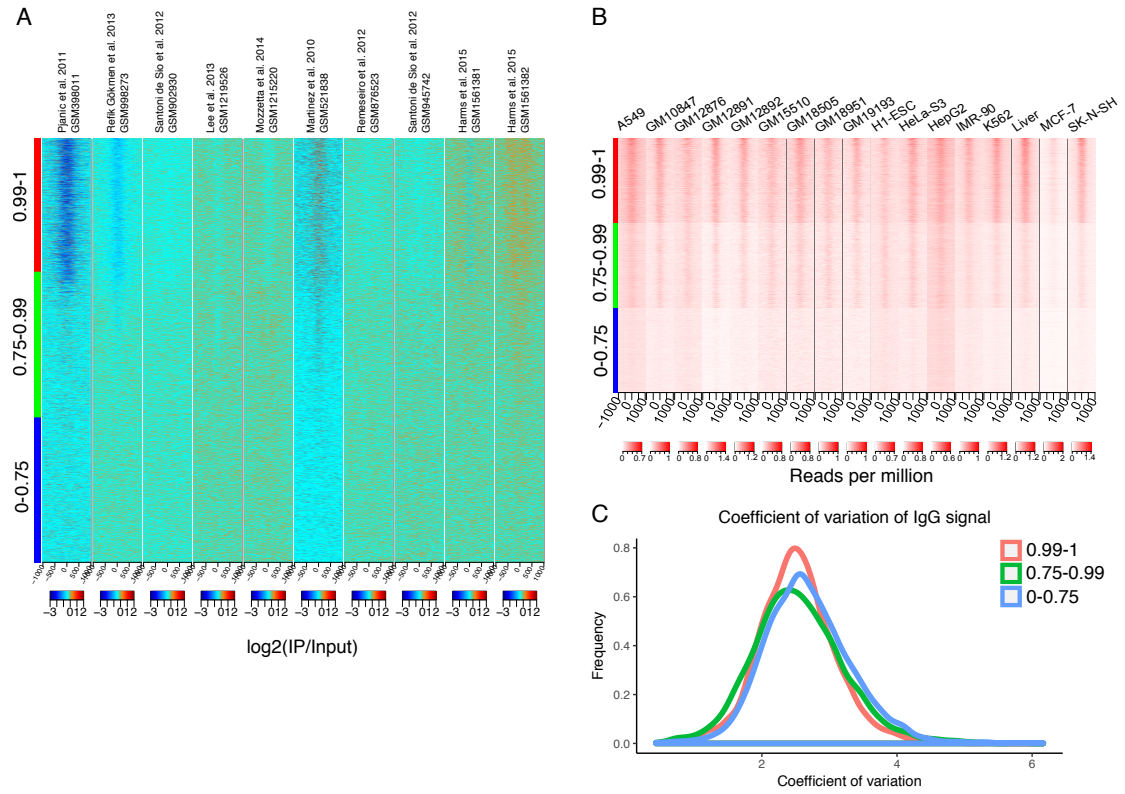


Figure B.5: A) Signal profile of KO experiments where the log2 IP/Input was less than 0, around regions with decreasing transcription factor density. Specific antibodies have low signal over hot regions. B) Signal profile (measured as reads per million) of the IgG control antibody (ENCAB000AOJ) in multiple cell lines around regions with decreasing transcription factor density. IgG controls show an increased signal on HOT regions (0.99-1), when compared to MILD (0.75-0.99) or COLD (0-0.75), however, the signal intensity is weak and cell type dependent. C) Distribution of the coefficient of variation of the IgG signal for each region in B). Each color represents regions of decreasing transcription factor occupancy density (red -HOT, green - MILD, blue - COLD). If IgG showed a consistently high signal in over HOT regions in multiple cell types, the distribution of the coefficient of variation should have a right skew, when compared to the distributions over MILD and COLD regions. This is however not the case - HOT regions show the same amount of variation as do MILD and COLD regions.

Figure B.6: Average methylation level on genomic regions (HOT regions, CpG islands that are not associated with HOT regions and other genomic regions with lower TF occupancy) for 37 human cell lines derived from the Roadmap Epigenomics database.

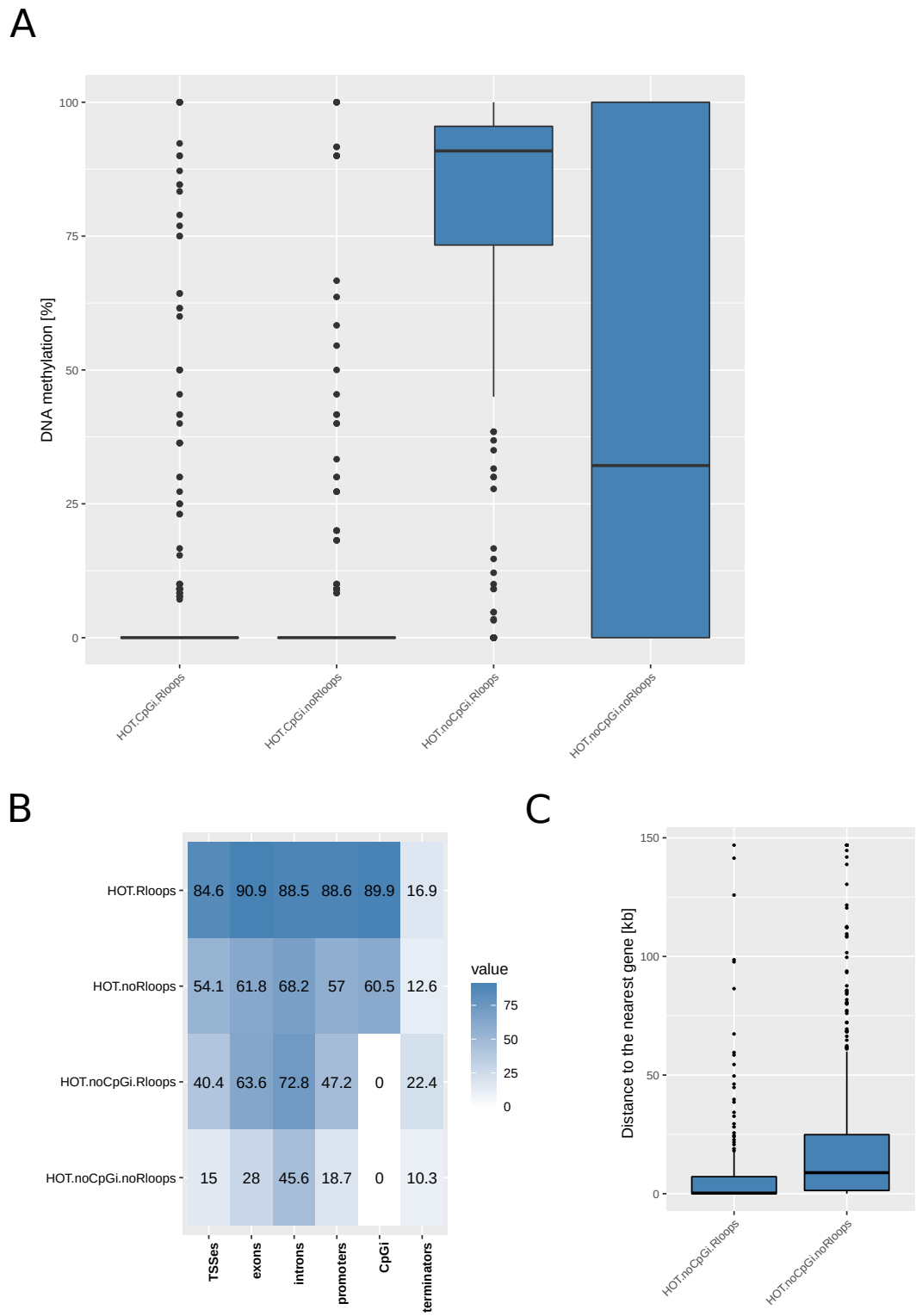


Figure B.7: DNA methylation patters on HOT regions and R-loops. (A) Percent DNA methylation on HOT regions that overlap CpG islands and R-loops (HOT.CpGi.Rloops), and don't overlap R-loops (HOT.CpGi.noRloops), and respectively HOT regions that don't overlap CpG islands (HOT.noCpGi.noRloops and HOT.noCpGi.noRloops). (B) Annotation of HOT regions and R-loops with genomic regions - TSSes, exons, introns, promoters (2.5kb around TSSes), CpG islands and terminators (regions within 1kb of the end of an annotated transcript) (C) Distance (absolute value) from HOT regions to the nearest gene that don't overlap with CpG-islands, and overlap (suffix .Rloops), or don't overlap with R-loops (suffix .noRloops).

GEO id (signal)	name of KO gene	number of uniquely mapped reads (signal)	GEO id (control)	number of uniquely mapped reads (control)	conditional KO	Cell line/tissue	Type of control	References
GSM494705	IRF4pre	2343111	NA	NA	no	CD4+ T cells	NA	(Kwon et al. 2009)
GSM494706	IRF4post	3222235	GSM494699	373256	no	CD4+ T cells	IgG	(Kwon et al. 2009)
GSM550312	Stat6	8639338	GSM550319	6621836	no	primary CD4+ T	IgG	(Wei et al. 2010)
GSM550304	Stat4	8040660	GSM550319	6621836	no	primary CD4+ T	IgG	(Wei et al. 2010)
GSM546535	E2A1h	5211214	GSM546540	9229303	yes	pre-pro-B cell	IgG	(Lin et al. 2010)
GSM546536	E2A6h	1952009	GSM546540	9229303	yes	pre-pro-B cell	IgG	(Lin et al. 2010)
GSM398011	NFI	2810251	GSM398012	4132910	no	MEF (mouse embryonic fibroblasts)	Input	(Pjanic et al. 2011)
GSM521838	RAPrep1	9927806	GSM521840	12388176	yes	MEF (mouse embryonic fibroblasts)	Input	(Martinez et al. 2010)
GSM521839	RAPrep2	6773623	GSM521840	12388176	yes	MEF (mouse embryonic fibroblasts)	Input	(Martinez et al. 2010)
GSM742022	Gata3	1381117	NA	NA	yes	CD8 cells	NA	(Wei et al. 2011)
GSM757799	rxra	3872108	GSM757813	46598249	no	F9 embryonal carcinoma cells	Input	(Mendoza-Parra et al. 2011)
GSM832325	runx1	8285703	GSM832328	10455598	yes	primary thymocyte	IgG	(Yu et al. 2012)
GSM876523	SA1	18115696	GSM800623	20871798	no	MEF (mouse embryonic fibroblasts)	Input	(Remesiro et al. 2012)
GSM1561381	Prdm16rep1	40334445	GSM1561389	26865948	yes	BAT (brown adipose tissue)	Input	(Harms et al. 2015)
GSM1561382	Prdm16rep2	43379852	GSM1561389	26865948	yes	BAT (brown adipose tissue)	Input	(Harms et al. 2015)
GSM1280887	SRC-2	43206659	GSM1280888	43445145	yes	liver	Input	(Stashi et al. 2014)
GSM1399547	Suz12	35957212	NA	NA	no	mES (embryonic stem cells)	NA	(Riising et al. 2014)
GSM998273	Tbet	17194840	GSM998271	21752412	yes	Th1	Input	(Refik Gökmen et al. 2013)
GSM1215806	MLL4_d7	16803820	GSM1215808	20589343	yes	immortalized brown preadipocytes	Input	(Lee et al. 2013)
GSM1215805	MLL4_d2	19794786	GSM1215808	20589343	yes	immortalized brown preadipocytes	Input	(Lee et al. 2013)
GSM1215804	MLL4_d0	13948901	GSM1215808	20589343	yes	immortalized brown preadipocytes	Input	(Lee et al. 2013)
GSM878199	Nkx3	14609846	NA	NA	-	prostate	NA	(Anderson et al. 2012)
GSM1514932	PPARa_GW7647	39870165	NA	NA	-	liver	NA	(Lee et al. 2014)
GSM1514931	PPARa_Vehicle	37900437	NA	NA	-	liver	NA	(Lee et al. 2014)
GSM1334041	Atf3_HDL_CpG	5653371	NA	NA	-	BMDM (bone marrow derived macrophage)	NA	(Krebs et al. 2014)
GSM1334041	Atf3_HDL_CpG	18170793	NA	NA	-	BMDM (bone marrow derived macrophage)	NA	(Krebs et al. 2014)
GSM1334041	Atf3_HDL_CpG	10275645	NA	NA	-	BMDM (bone marrow derived macrophage)	NA	(Krebs et al. 2014)
GSM1334039	Atf3_HDL	5622991	NA	NA	-	BMDM (bone marrow derived macrophage)	NA	(Krebs et al. 2014)
GSM1334039	Atf3_HDL	23565419	NA	NA	-	BMDM (bone marrow derived macrophage)	NA	(Krebs et al. 2014)
GSM1334039	Atf3_HDL	18268214	NA	NA	-	BMDM (bone marrow derived macrophage)	NA	(Krebs et al. 2014)
GSM1334037	Atf3_Unstim	1555793	NA	NA	-	BMDM (bone marrow derived macrophage)	NA	(Krebs et al. 2014)
GSM1334037	Atf3_Unstim	8150131	NA	NA	-	BMDM (bone marrow derived macrophage)	NA	(Krebs et al. 2014)
GSM1334037	Atf3_Unstim	21570766	NA	NA	-	BMDM (bone marrow derived macrophage)	NA	(Krebs et al. 2014)
GSM1334037	Atf3_Unstim	25562425	NA	NA	-	BMDM (bone marrow derived macrophage)	NA	(Krebs et al. 2014)
GSM1215220	G9aGLP	45873732	GSM1215221	17016836	no	mES (embryonic stem cells)	Input	(Mozzetta et al. 2014)
GSM1219326	MLL4_EBFbeta_Cre	16389950	GSM1219532	19967393	yes	immortalized brown preadipocytes	Input	(Lee et al. 2013)
GSM1219694	MLL4_myocytes	13142751	GSM1219710	13882475	yes	MyoD-induced immortalized brown preadipocytes	Input	(Lee et al. 2013)
GSM1219693	MLL4_preadipocytes	11087308	GSM1219709	16924206	yes	immortalized brown preadipocytes	Input	(Lee et al. 2013)
GSM873425	Chop	2131887	NA	NA	no	mouse embryonic fibroblast	NA	(Han et al. 2013)
GSM873427	Atf4	11470110	NA	NA	no	mouse embryonic fibroblast	NA	(Han et al. 2013)
GSM945742	KAP1_KAP1	14892018	GSM945743	18925856	yes	T cell progenitors	Input	(Santoni de Sio et al. 2012)
GSM902930	KAP1	15278519	GSM902931	17023238	yes	B cell enriched splenocytes	Input	(Santoni de Sio et al. 2012)
GSM890127	Rev-erb	76026482	NA	NA	no	liver	NA	(Bugge et al. 2012)

Table B.1: Sequencing quality table of KO ChIP-seq samples.

RunID	Library layout	Reference	Technique	Cell line	Treatment	Type	Organism
SRR2028291	PAIRED	Julie Nadel et al. Epigenetics and Chromatin 2015	RDIPseq	IMR90		signal	human
SRR2028292	PAIRED	Julie Nadel et al. Epigenetics and Chromatin 2015	RDIPseq	IMR90		input	human
SRR2028293	PAIRED	Julie Nadel et al. Epigenetics and Chromatin 2015	RDIPseq	HEK293T		signal	human
SRR2028294	PAIRED	Julie Nadel et al. Epigenetics and Chromatin 2015	DRIPseq	HEK293T		input	human
SRR1272444	SINGLE	Lim YW et al. Elife 2015	DRIPseq	Primary fibroblast		signal	human
SRR1295464	SINGLE	Lim YW et al. Elife 2015	DRIPseq	Primary fibroblast		signal	human
SRR1272443	SINGLE	Lim YW et al. Elife 2015	DRIPseq	Primary fibroblast		input	human
SRR2075681	SINGLE	Sanz LA et al. Mol Cell 2016	DRIPseq	NT2		signal	human
SRR2075682	SINGLE	Sanz LA et al. Mol Cell 2016	DRIPseq	NT2		signal	human
SRR2075684	SINGLE	Sanz LA et al. Mol Cell 2016	DRIPseq	NT2	RNaseH	signal	human
SRR2075685	SINGLE	Sanz LA et al. Mol Cell 2016	DRIPseq	K562		signal	human
SRR3993997	PAIRED	Zeller et al. Nature Genetics 2016	DRIPseq		RNaseH	signal	worm
SRR3993997	PAIRED	Zeller et al. Nature Genetics 2016	DRIPseq		RNaseH	signal	worm
SRR3993995	PAIRED	Zeller et al. Nature Genetics 2016	DRIPseq			signal	worm
SRR3993995	PAIRED	Zeller et al. Nature Genetics 2016	DRIPseq			signal	worm
SRR3993993	PAIRED	Zeller et al. Nature Genetics 2016	DRIPseq			input	worm
SRR3993993	PAIRED	Zeller et al. Nature Genetics 2016	DRIPseq			input	worm
SRR3504389	SINGLE	Wahba et al. Genes and Dev. 2016	DRIP-seq			signal	yeast
SRR3095355	SINGLE	Hänsel-Hertsch et al. Nature Genetics 2016	BG4seq	HaCaT		signal	human
SRR3095357	SINGLE	Hänsel-Hertsch et al. Nature Genetics 2016	BG4seq	HaCaT		input	human
SRR3095318	SINGLE	Hänsel-Hertsch et al. Nature Genetics 2016	BG4seq	NHEK		signal	human
SRR3095305	SINGLE	Hänsel-Hertsch et al. Nature Genetics 2016	BG4seq	NHEK		input	human
SRR3095306	SINGLE	Hänsel-Hertsch et al. Nature Genetics 2016	BG4seq	NHEK		signal	human
SRR3095307	SINGLE	Hänsel-Hertsch et al. Nature Genetics 2016	BG4seq	NHEK		signal	human
SRR3095308	SINGLE	Hänsel-Hertsch et al. Nature Genetics 2016	BG4seq	NHEK		signal	human
SRR3095309	SINGLE	Hänsel-Hertsch et al. Nature Genetics 2016	BG4seq	NHEK		signal	human

Table B.2: Description of DRIP/RDIP-seq and G4 ChIP–seq (BG4seq) samples.

gene_name	paralogue	percent_identity	ensembl_link
Stat6	Stat1	24.5	https://www.ensembl.org/Homo_sapiens/Gene/Compara_Paralog?db=core;g=ENSG00000166888;r=12:57095408-57132139
Stat6	Stat2	18.33	
Stat6	Stat3	23.51	
Stat6	Stat4	24.6	
Stat6	Stat5A	34.63	
Stat6	Stat5b	36.64	
Stat4	Stat1	52.19	
Stat4	Stat2	30.26	
Stat4	Stat3	46.49	
Stat4	Stat5A	21.98	
Stat4	Stat5b	26.85	
Stat4	Stat6	27.23	
E2a	No	NA	https://www.ensembl.org/Mus_musculus/Gene/Summary?db=core;g=ENSMUSG00000020167;r=10:80409514-80433647
Nfic	Nfix	52.43	http://www.ensembl.org/Mus_musculus/Gene/Compara_Paralog?g=ENSMUSG00000055053;r=10:81396186-81455635
Nfic	Nfia	47.37	
Nfic	Nfib	52.65	
Rap1	No	NA	http://www.ensembl.org/Mus_musculus/Gene/Summary?g=ENSMUSG00000068798;r=3:105727267-105801336
Gata3	No	NA	http://www.ensembl.org/Mus_musculus/Gene/Summary?db=core;g=ENSMUSG00000015619;r=2:9857078-9890034
Rxra	No	NA	http://www.ensembl.org/Mus_musculus/Gene/Summary?db=core;g=ENSMUSG00000015846;r=2:27676440-27762957
Runx1	No	NA	http://www.ensembl.org/Mus_musculus/Gene/Summary?db=core;g=ENSMUSG00000022952;r=16:92601466-92826149
Sa1	Stag3	46.77	http://www.ensembl.org/Mus_musculus/Gene/Compara_Paralog?db=core;g=ENSMUSG00000037286;r=9:100597798-100959375
Sa1	Stag2	70.35	http://www.ensembl.org/Mus_musculus/Gene/Compara_Paralog?db=core;g=ENSMUSG00000037286;r=9:100597798-100959375
Trib1	Stk40	20.71	http://www.ensembl.org/Mus_musculus/Gene/Compara_Paralog?db=core;g=ENSMUSG00000032501;r=15:59648350-59657099
Trib1	Trib2	47.18	
Trib1	Trib3	57.73	
Prdm16	No	NA	http://www.ensembl.org/Mus_musculus/Gene/Summary?db=core;g=ENSMUSG00000039410;r=4:154316125-154636873
Src-2	Ncoa1	41.34	http://www.ensembl.org/Mus_musculus/Gene/Summary?g=ENSMUSG00000005886;r=1:13139105-13374083
Src-2	Ncoa3	37.22	
Suz12	No	NA	http://www.ensembl.org/Mus_musculus/Gene/Summary?db=core;g=ENSMUSG00000017548;r=11:79993106-80034123
Tbet	No	NA	http://www.ensembl.org/Mus_musculus/Gene/Summary?db=core;g=ENSMUSG00000001444;r=11:97098071-97115331;t=ENSMUST00000001484
Mll4	No	NA	http://www.ensembl.org/Mus_musculus/Gene/Summary?db=core;g=ENSMUSG00000048154;r=15:98831669-98871204
Nfatc1	No	NA	http://www.ensembl.org/Mus_musculus/Gene/Summary?db=core;g=ENSMUSG00000027544;r=2:168476410-168601657
Nkx3	No	NA	http://www.ensembl.org/Mus_musculus/Gene/Summary?db=core;g=ENSMUSG00000022061;r=14:69190638-69194662;t=ENSMUST00000022646
Ppara	No	NA	http://www.ensembl.org/Mus_musculus/Gene/Summary?db=core;g=ENSMUSG00000022383;r=15:85734983-85802819
Atf3	No	NA	http://www.ensembl.org/Mus_musculus/Gene/Summary?db=core;g=ENSMUSG00000026628;r=1:191170296-191218039
G9a	No	NA	http://www.ensembl.org/Mus_musculus/Gene/Summary?db=core;g=ENSMUSG00000013787;r=17:34898469-34914052
Chop	No	NA	http://www.ensembl.org/Mus_musculus/Gene/Summary?db=core;g=ENSMUSG00000116429;r=10:127290829-127293753;t=ENSMUST00000230446
Atf4	Atf5	28.98	http://www.ensembl.org/Mus_musculus/Gene/Summary?db=core;g=ENSMUSG00000042406;r=15:80255184-80257541
Kap1	No	NA	http://www.ensembl.org/Mus_musculus/Gene/Summary?db=core;g=ENSMUSG00000005566;r=7:12999114-13031035
Rev-erb	No	NA	http://www.ensembl.org/Mus_musculus/Gene/Summary?db=core;g=ENSMUSG00000021775;r=14:18204054-18239127

Table B.3: Description of knock-out TFs and their paralogs.

References

"A promoter-level mammalian expression atlas" (Mar. 2014). In: *Nature* 507.7493, pp. 462–470. DOI: 10.1038/nature13182. URL: https://doi.org/10.1038/nature13182.

Abe, Masanobu et al. (Feb. 2005). "CpG island methylator phenotype is a strong determinant of poor prognosis in neuroblastomas". en. In: *Cancer Res.* 65.3, pp. 828–834.

Ackermann, Sandra et al. (Nov. 2014). "FOXP1inhibits cell growth and attenuates tumorigenicity of neuroblastoma". In: *BMC Cancer* 14.1. DOI: 10.1186/1471-2407-14-840. URL: https://doi.org/10.1186/1471-2407-14-840.

Ahn, M. et al. (Oct. 2007). "Regulation of NaV1.2 Channels by Brain-Derived Neurotrophic Factor, TrkB, and Associated Fyn Kinase". In: *Journal of Neuroscience* 27.43, pp. 11533–11542. DOI: 10.1523/jneurosci.5005-06.2007. URL: https://doi.org/10.1523/jneurosci.5005-06.2007.

Akalin, Altuna (Dec. 2020). *Computational Genomics With R*. en. CRC Press.

Akalin, Altuna, Vedran Franke, et al. (Apr. 2015). "Genomation: a toolkit to summarize, annotate and visualize genomic intervals". en. In: *Bioinformatics* 31.7, pp. 1127–1129.

Akalin, Altuna, Matthias Kormaksson, et al. (Oct. 2012). "methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles". en. In: *Genome Biol.* 13.10, R87.

Albertson, Donna G (Aug. 2006). "Gene amplification in cancer". en. In: *Trends Genet.* 22.8, pp. 447–455.

Allen, Benjamin L and Dylan J Taatjes (Mar. 2015). "The Mediator complex: a central integrator of transcription". en. In: *Nat. Rev. Mol. Cell Biol.* 16.3, pp. 155–166.

Allis, C David, C David Allis, and Thomas Jenuwein (2016). *The molecular hallmarks of epigenetic control*.

Andersson, Daniel et al. (Apr. 2020). "Circulating cell-free tumor DNA analysis in pediatric cancers". In: *Molecular Aspects of Medicine* 72, p. 100819. DOI: 10.1016/j.mam.2019.09.003. URL: https://doi.org/10.1016/j.mam.2019.09.003.

Ang, Cheen Euong et al. (Jan. 2019). "The novel lncRNA lnc-NR2F1 is pro-neurogenic and mutated in human neurodevelopmental disorders". In: *eLife* 8. DOI: 10.7554/elife.41770. URL: https://doi.org/10.7554/elife.41770.

Anzalone, Andrew V., Luke W. Koblan, and David R. Liu (June 2020). "Genome editing with CRISPR–Cas nucleases, base editors, transposases and prime editors". In: *Nature Biotechnology* 38.7, pp. 824–844. DOI: 10.1038/s41587-020-0561-9. URL: https://doi.org/10.1038/s41587-020-0561-9.

Arab, Khelifa et al. (Jan. 2019). "GADD45A binds R-loops and recruits TET1 to CpG island promoters". In: *Nature Genetics* 51.2, pp. 217–223. DOI: 10.1038/s41588-018-0306-6. URL: https://doi.org/10.1038/s41588-018-0306-6.

Aran, Dvir, Sivan Sabato, and Asaf Hellman (2013). "DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes". In: *Genome Biology* 14.3, R21. DOI: 10.1186/gb-2013-14-3-r21. URL: https://doi.org/10.1186/gb-2013-14-3-r21.

Argelaguet, Ricard, Damien Arnol, et al. (May 2020). "MOFA: a statistical framework for comprehensive integration of multi-modal single-cell data". In: *Genome Biology* 21.1. DOI:

10.1186/s13059-020-02015-1. URL: https://doi.org/10.1186/s13059-020-02015-1.

Argelaguet, Ricard, Britta Velten, et al. (June 2018). "Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets". In: *Molecular Systems Biology* 14.6. DOI: 10.15252/msb.20178124. URL: https://doi.org/10.15252/msb.20178124.

Assenov, Yassen et al. (2014). *Comprehensive analysis of DNA methylation data with RnBeads*.

Aydin, Begüm et al. (May 2019). "Proneural factors Ascl1 and Neurog2 contribute to neuronal subtype identities by establishing distinct chromatin landscapes". In: *Nature Neuroscience* 22.6, pp. 897–908. DOI: 10.1038/s41593-019-0399-y. URL: https://doi.org/10.1038/s41593-019-0399-y.

B., Bushnell (n.d.). *BBMap*. URL: http://sourceforge.net/projects/bbmap/.

Babraham, Bioinformatics (2018a). URL: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

— (2018b). *fastqc*. URL: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Balwierz, P. J. et al. (Feb. 2014). "ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs". In: *Genome Research* 24.5, pp. 869–884. DOI: 10.1101/gr.169508.113. URL: https://doi.org/10.1101/gr.169508.113.

Balwierz, Piotr J et al. (May 2014). "ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs". en. In: *Genome Res.* 24.5, pp. 869–884.

Banerjee, Deblina et al. (July 2020). "Lineage specific transcription factor waves reprogram neuroblastoma from self-renewal to differentiation". In: DOI: 10.1101/2020.07.23.218503. URL: https://doi.org/10.1101/2020.07.23.218503.

Bannister, Andrew J and Tony Kouzarides (Mar. 2011). "Regulation of chromatin by histone modifications". en. In: *Cell Res.* 21.3, pp. 381–395.

Barbas, S. M. et al. (Mar. 1995). "Human autoantibody recognition of DNA." In: *Proceedings of the National Academy of Sciences* 92.7, pp. 2529–2533. DOI: 10.1073/pnas.92.7.2529. URL: https://doi.org/10.1073/pnas.92.7.2529.

Barillot, Emmanuel et al. (Aug. 2012). *Computational Systems Biology of Cancer*. CRC Press. DOI: 10.1201/b12677. URL: https://doi.org/10.1201/b12677.

Barski, Artem et al. (2007). *High-Resolution Profiling of Histone Methylations in the Human Genome*.

Bednar, J et al. (1998). *Nucleosomes, linker DNA, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin*.

Berdasco, Maríéa and Manel Esteller (Feb. 2019). "Clinical epigenetics: seizing opportunities for translation". en. In: *Nat. Rev. Genet.* 20.2, pp. 109–127.

Berman, Benjamin P et al. (Nov. 2011). "Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains". en. In: *Nat. Genet.* 44.1, pp. 40–46.

Bert, Saul A et al. (Jan. 2013). "Regional activation of the cancer genome by long-range epigenetic remodeling". en. In: *Cancer Cell* 23.1, pp. 9–22.

Bhattacharyya, Anamitra, Alastair I. H. Murchie, and David M. J. Lilley (Feb. 1990). "RNA bulges and the helical periodicity of double-stranded RNA". In: *Nature* 343.6257, pp. 484–487. DOI: 10.1038/343484a0. URL: https://doi.org/10.1038/343484a0.

Bibikova, Marina (2016). *DNA Methylation Microarrays*.

Bird, Adrian P (1986). *CpG-rich islands and the function of DNA methylation*.

Blackwood, E. and R. Eisenman (Mar. 1991). "Max: a helix-loop-helix zipper protein that forms a sequence-specific DNA-binding complex with Myc". In: *Science* 251.4998, pp. 1211–1217. DOI: 10.1126/science.2006410. URL: https://doi.org/10.1126/science.2006410.

Blavier, Laurence, Ren-Ming Yang, and Yves A. DeClerck (Oct. 2020). "The Tumor Microenvironment in Neuroblastoma: New Players, New Mechanisms of Interaction and New Perspectives". In: *Cancers* 12.10, p. 2912. DOI: 10.3390/cancers12102912. URL: https://doi.org/10.3390/cancers12102912.

Bock, Christoph (2012). *Analysing and interpreting DNA methylation data*.

Boeva, Valentina et al. (2017). *Heterogeneity of neuroblastoma cell identity defined by transcriptional circuitries*.

Boguslawski, Sophie J. et al. (May 1986). "Characterization of monoclonal antibody to DNA · RNA and its application to immunodetection of hybrids". In: *Journal of Immunological Methods* 89.1, pp. 123–130. DOI: 10.1016/0022-1759(86)90040-2. URL: https://doi.org/10.1016/0022-1759(86)90040-2.

Bonvouloir, Nadia et al. (Aug. 2001). "Molecular Cloning, Tissue Distribution, and Chromosomal Localization of MMEL2, a Gene Coding for a Novel Human Member of the Neutral Endopeptidase-24.11 Family". In: *DNA and Cell Biology* 20.8, pp. 493–498. DOI: 10.1089/104454901316976127. URL: https://doi.org/10.1089/104454901316976127.

Borghini, Silvia et al. (Mar. 2006). "The TLX2 homeobox gene is a transcriptional target of PHOX2B in neural-crest-derived cells". In: *Biochemical Journal* 395.2, pp. 355–361. DOI: 10.1042/bj20051386. URL: https://doi.org/10.1042/bj20051386.

Boying Gong (2019). *MethCP*. DOI: 10.18129/B9.BIOC.METHCP. URL: https://bioconductor.org/packages/MethCP.

Boyle, Alan P et al. (2014). *Comparative analysis of regulatory information and circuits across distant species*.

Brady, Samuel W. et al. (Oct. 2020). "Pan-neuroblastoma analysis reveals age- and signature-associated driver alterations". In: *Nature Communications* 11.1. DOI: 10.1038/s41467-020-18987-4. URL: https://doi.org/10.1038/s41467-020-18987-4.

Braun, Ralph P. and Jeremy S. Lee (1986). "Variations in duplex DNA conformation detected by the binding of monoclonal autoimmune antibodies". In: *Nucleic Acids Research* 14.12, pp. 5049–5065. DOI: 10.1093/nar/14.12.5049. URL: https://doi.org/10.1093/nar/14.12.5049.

Brenet, Fabienne et al. (Jan. 2011). "DNA Methylation of the First Exon Is Tightly Linked to Transcriptional Silencing". In: *PLoS ONE* 6.1. Ed. by Nina Papavasiliou, e14524. DOI: 10.1371/journal.pone.0014524. URL: https://doi.org/10.1371/journal.pone.0014524.

Brinkman, Arie B et al. (Apr. 2019). "Partially methylated domains are hypervariable in breast cancer and fuel widespread CpG island hypermethylation". en. In: *Nat. Commun.* 10.1, p. 1749.

Brodeur, G M et al. (June 1984). "Amplification of N-myc in untreated human neuroblastomas correlates with advanced disease stage". en. In: *Science* 224.4653, pp. 1121–1124.

Brodeur, G. et al. (June 1984). "Amplification of N-myc in untreated human neuroblastomas correlates with advanced disease stage". In: *Science* 224.4653, pp. 1121–1124. DOI: 10.1126/science.6719137. URL: https://doi.org/10.1126/science.6719137.

Buckley, Patrick G et al. (2011). *Genome-wide DNA methylation analysis of neuroblastic tumors reveals clinically relevant epigenetic events and large-scale epigenomic alterations localized to telomeric regions*.

Bulger, Michael and Mark Groudine (Feb. 2011). "Functional and mechanistic diversity of distal transcription enhancers". en. In: *Cell* 144.3, pp. 327–339.

Bulyk, Martha L. (n.d.). "Protein Binding Microarrays for the Characterization of DNA–Protein Interactions". In: *Analytics of Protein–DNA Interactions*. Springer Berlin Heidelberg, pp. 65–85. DOI: 10.1007/10_025. URL: https://doi.org/10.1007/10_025.

Burger, Lukas et al. (July 2013). "Identification of active regulatory regions from DNA methylation data". In: *Nucleic Acids Research* 41.16, e155–e155. DOI: 10.1093/nar/gkt599. URL: https://doi.org/10.1093/nar/gkt599.

CAPONI, L. et al. (Nov. 2002). "Anti-ribosomal antibodies from lupus patients bind DNA". In: *Clinical & Experimental Immunology* 130.3, pp. 541–547. DOI: 10.1046/j.1365-2249.2002.02014.x. URL: https://doi.org/10.1046/j.1365-2249.2002.02014.x.

Capper, David et al. (Mar. 2018). "DNA methylation-based classification of central nervous system tumours". In: *Nature* 555.7697, pp. 469–474. DOI: 10.1038/nature26000. URL: https://doi.org/10.1038/nature26000.

Carén, Helena et al. (Feb. 2011). "Identification of epigenetically regulated genes that predict patient outcome in neuroblastoma". en. In: *BMC Cancer* 11, p. 66.

Celniker, Susan E et al. (2009). *Unlocking the secrets of the genome*.

CERUTTI, M et al. (Feb. 2005). "A viral DNA-binding domain elicits anti-DNA antibodies of different specificities". In: *Molecular Immunology* 42.3, pp. 327–333. DOI: 10.1016/j.molimm.2004.09.003. URL: https://doi.org/10.1016/j.molimm.2004.09.003.

Chakravarty, Debyani et al. (Nov. 2017). "OncoKB: A Precision Oncology Knowledge Base". In: *JCO Precision Oncology* 1, pp. 1–16. DOI: 10.1200/po.17.00011. URL: https://doi.org/10.1200/po.17.00011.

Charlet, Jessica et al. (2017). *Genome-wide DNA methylation analysis identifiesMEGF10as a novel epigenetically repressed candidate tumor suppressor gene in neuroblastoma*.

Chen, Liang et al. (Nov. 2017). "R-ChIP Using Inactive RNase H Reveals Dynamic Coupling of R-loops with Transcriptional Pausing at Gene Promoters". In: *Molecular Cell* 68.4, 745–757.e5. DOI: 10.1016/j.molcel.2017.10.008. URL: https://doi.org/10.1016/j.molcel.2017.10.008.

Chen, R. A.- J. et al. (Mar. 2014). "Extreme HOT regions are CpG-dense promoters in C. elegans and humans". In: *Genome Research* 24.7, pp. 1138–1146. DOI: 10.1101/gr.161992.113. URL: https://doi.org/10.1101/gr.161992.113.

Chen, Wanze et al. (Mar. 2021). "Genome-wide molecular recording using Live-seq". In: DOI: 10.1101/2021.03.24.436752. URL: https://doi.org/10.1101/2021.03.24.436752.

Chen, Xi et al. (2017). *Mocap: large-scale inference of transcription factor binding sites from chromatin accessibility*.

Cheng, J M et al. (June 1993). "Preferential amplification of the paternal allele of the N-myc gene in human neuroblastomas". en. In: *Nat. Genet.* 4.2, pp. 191–194.

Ciabrelli, Filippo and Giacomo Cavalli (2015). *Chromatin-Driven Behavior of Topologically Associating Domains*.

Clapier, Cedric R et al. (July 2017). "Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes". en. In: *Nat. Rev. Mol. Cell Biol.* 18.7, pp. 407–422.

Cohen, Netta Mendelson, Ephraim Kenigsberg, and Amos Tanay (May 2011). "Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection". en. In: *Cell* 145.5, pp. 773–786.

Consortium, The ENCODE Project (Sept. 2012). "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414, pp. 57–74. DOI: 10.1038/nature11247. URL: https://doi.org/10.1038/nature11247.

Consortium, The FANTOM and Riken Omics Science Center (Apr. 2009). "The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line".

In: *Nature Genetics* 41.5, pp. 553–562. DOI: 10.1038/ng.375. URL: https://doi.org/10.1038/ng.375.

Coolen, Marcel W et al. (Mar. 2010). "Consolidation of the cancer genome into domains of repressive chromatin by long-range epigenetic silencing (LRES) reduces transcriptional plasticity". en. In: *Nat. Cell Biol.* 12.3, pp. 235–246.

Cristini, Agnese et al. (May 2018). "RNA/DNA Hybrid Interactome Identifies DXH9 as a Molecular Player in Transcriptional Termination and R-Loop-Associated DNA Damage". In: *Cell Reports* 23.6, pp. 1891–1905. DOI: 10.1016/j.celrep.2018.04.025. URL: https://doi.org/10.1016/j.celrep.2018.04.025.

D'Alessandro, Giuseppina et al. (Dec. 2018). "BRCA2 controls DNA:RNA hybrid level at DSBs by mediating RNase H2 recruitment". In: *Nature Communications* 9.1. DOI: 10.1038/s41467-018-07799-2. URL: https://doi.org/10.1038/s41467-018-07799-2.

Daca-Roszak, Patrycja et al. (2015). *Impact of SNPs on methylation readouts by Illumina Infinium HumanMethylation450 BeadChip Array: implications for comparative population studies*.

Danilova, Nadia (2020). *p53, Guardian of the Genome*.

Day, Jeremy J and J David Sweatt (2010). *DNA methylation and memory formation*.

Deaton, Aimée M and Adrian Bird (2011). "CpG islands and the regulation of transcription". In: *Genes & development* 25.10, pp. 1010–1022.

Decaesteker, Bieke et al. (Nov. 2018). "TBX2 is a neuroblastoma core regulatory circuitry component enhancing MYCN/FOXM1 reactivation of DREAM targets". In: *Nature Communications* 9.1. DOI: 10.1038/s41467-018-06699-9. URL: https://doi.org/10.1038/s41467-018-06699-9.

Decock, Anneleen et al. (Oct. 2012). "Genome-wide promoter methylation analysis in neuroblastoma identifies prognostic methylation biomarkers". en. In: *Genome Biol.* 13.10, R95.

Deocharan, B et al. (Dec. 2002). "Antigenic triggers and molecular targets for anti-double-stranded DNA antibodies". In: *Lupus* 11.12, pp. 865–871. DOI: 10.1191/0961203302lu308rr. URL: https://doi.org/10.1191/0961203302lu308rr.

Depuydt, Pauline et al. (Oct. 2018). "Meta-mining of copy number profiles of high-risk neuroblastoma tumors". en. In: *Sci Data* 5, p. 180240.

Desai, D. D. et al. (Aug. 1993). "Antigen-specific induction of antibodies against native mammalian DNA in nonautoimmune mice". In: *J Immunol* 151.3, pp. 1614–1626.

Detchokul, S et al. (Dec. 2014). "Tetraspanins as regulators of the tumour microenvironment: implications for metastasis and therapeutic strategies". In: *British Journal of Pharmacology* 171.24, pp. 5462–5490. DOI: 10.1111/bph.12260. URL: https://doi.org/10.1111/bph.12260.

Dhaeseleer, Patrik (Apr. 2006). "What are DNA sequence motifs?" In: *Nature Biotechnology* 24.4, pp. 423–425. DOI: 10.1038/nbt0406-423. URL: https://doi.org/10.1038/nbt0406-423.

Dhingra Priyanka et al. (2017). *Integration Of Genetic And Epigenetic Alterations With Tissue-Specific Network Reveals Regulatory Drivers Of Prostate Cancer*. DOI: 10.5281/ZENODO.800729. URL: https://zenodo.org/record/800729.

Ditlevsen, Dorte Kornerup et al. (2008). "NCAM-induced intracellular signaling revisited". In: *Journal of Neuroscience Research* 86.4, pp. 727–743. DOI: 10.1002/jnr.21551. URL: https://doi.org/10.1002/jnr.21551.

Dixon, Jesse R et al. (Apr. 2012). "Topological domains in mammalian genomes identified by analysis of chromatin interactions". en. In: *Nature* 485.7398, pp. 376–380.

Doi, Akiko et al. (Dec. 2009). "Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts". en. In: *Nat. Genet.* 41.12, pp. 1350–1353.

Dolzhenko, Egor and Andrew D Smith (June 2014). "Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments". en. In: *BMC Bioinformatics* 15, p. 215.

Dong, Rui et al. (Nov. 2020). "Single-Cell Characterization of Malignant Phenotypes and Developmental Trajectories of Adrenal Neuroblastoma". In: *Cancer Cell* 38.5, 716–733.e6. DOI: 10.1016/j.ccell.2020.08.014. URL: https://doi.org/10.1016/j.ccell.2020.08.014.

Dreidax, Daniel et al. (2013). *Low p14ARF expression in neuroblastoma cells is associated with repressed histone mark status, and enforced expression induces growth arrest and apoptosis*.

Dumelie, Jason G and Samie R Jaffrey (Oct. 2017). "Defining the location of promoter-associated R-loops at near-nucleotide resolution using bisDRIP-seq". In: *eLife* 6. DOI: 10.7554/elife.28306. URL: https://doi.org/10.7554/elife.28306.

Duquette, M. L. (July 2004). "Intracellular transcription of G-rich DNAs induces formation of G-loops, novel structures containing G4 DNA". In: *Genes & Development* 18.13, pp. 1618–1629. DOI: 10.1101/gad.1200804. URL: https://doi.org/10.1101/gad.1200804.

Durbin, Adam D et al. (Sept. 2018). "Selective gene dependencies in MYCN-amplified neuroblastoma include the core transcriptional regulatory circuitry". en. In: *Nat. Genet.* 50.9, pp. 1240–1246.

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

Entz-Werlé, Natacha et al. (Nov. 2005). "Frequent genomic abnormalities at TWIST in human pediatric osteosarcomas". en. In: *Int. J. Cancer* 117.3, pp. 349–355.

Ernst, Jason et al. (Mar. 2011). "Mapping and analysis of chromatin state dynamics in nine human cell types". In: *Nature* 473.7345, pp. 43–49. DOI: 10.1038/nature09906. URL: https://doi.org/10.1038/nature09906.

Esteller, Manel (2002). *CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future*.

Evgeniou, Theodoros and Massimiliano Pontil (Aug. 2004). "Regularized Multi–Task Learning". In: *Conference Proceedings. In Proceedings of the Tenth Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*. URL: http://www0.cs.ucl.ac.uk/staff/M.Pontil/reading/mt-kdd.pdf.

Feinberg, A P and B Vogelstein (Jan. 1983). "Hypomethylation distinguishes genes of some human cancers from their normal counterparts". en. In: *Nature* 301.5895, pp. 89–92.

Feng, Hao, Karen N Conneely, and Hao Wu (2014). *A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data*.

Fernandez, Agustin F et al. (Feb. 2012). "A DNA methylation fingerprint of 1628 human samples". en. In: *Genome Res.* 22.2, pp. 407–419.

Fischle, Wolfgang, Yanming Wang, and C David Allis (2003). *Histone and chromatin cross-talk*.

Fleischer, Thomas et al. (Aug. 2014). "Genome-wide DNA methylation profiles in progression to in situ and invasive carcinoma of the breast with impact on gene transcription and prognosis". en. In: *Genome Biol.* 15.8, p. 435.

Fornes, Oriol et al. (Nov. 2019). "JASPAR 2020: update of the open-access database of transcription factor binding profiles". In: *Nucleic Acids Research*. DOI: 10.1093/nar/gkz1001. URL: https://doi.org/10.1093/nar/gkz1001.

Fraga, Mario F et al. (2005). *Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer*.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1, pp. 1–22. URL: https://www.jstatsoft.org/v33/i01/.

Frigola, Jordi et al. (May 2006). "Epigenetic remodeling in colorectal cancer results in coordinate gene suppression across an entire chromosome band". en. In: *Nat. Genet.* 38.5, pp. 540–549.

Frith, Martin C, Ryota Mori, and Kiyoshi Asai (2012). *A mostly traditional approach improves alignment of bisulfite-converted DNA*.

Frommer, M et al. (Mar. 1992). "A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 89.5, pp. 1827–1831.

Füllgrabe, J, E Kavanagh, and B Joseph (Apr. 2011). "Histone onco-modifications". In: *Oncogene* 30.31, pp. 3391–3403. DOI: 10.1038/onc.2011.121. URL: https://doi.org/10.1038/onc.2011.121.

Furey, Terrence S (2012). *ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions*.

Furlong, Eileen E. M. and Michael Levine (Sept. 2018). "Developmental enhancers and chromosome topology". In: *Science* 361.6409, pp. 1341–1345. DOI: 10.1126/science.aau0320. URL: https://doi.org/10.1126/science.aau0320.

Gama-Castro, Socorro et al. (Jan. 2016). "RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond". en. In: *Nucleic Acids Res.* 44.D1, pp. D133–43.

Gartlgruber, Moritz et al. (Dec. 2020). "Super enhancers define regulatory subtypes and cell identity in neuroblastoma". In: *Nature Cancer* 2.1, pp. 114–128. DOI: 10.1038/s43018-020-00145-w. URL: https://doi.org/10.1038/s43018-020-00145-w.

Gaynor, B. et al. (Mar. 1997). "Peptide inhibition of glomerular deposition of an anti-DNA antibody". In: *Proceedings of the National Academy of Sciences* 94.5, pp. 1955–1960. DOI: 10.1073/pnas.94.5.1955. URL: https://doi.org/10.1073/pnas.94.5.1955.

El-Gebali, Sara et al. (Oct. 2018). "The Pfam protein families database in 2019". In: *Nucleic Acids Research* 47.D1, pp. D427–D432. DOI: 10.1093/nar/gky995. URL: https://doi.org/10.1093/nar/gky995.

Geertz, M., D. Shore, and S. J. Maerkl (Sept. 2012). "Massively parallel measurements of molecular interaction kinetics on a microfluidic platform". In: *Proceedings of the National Academy of Sciences* 109.41, pp. 16540–16545. DOI: 10.1073/pnas.1206011109. URL: https://doi.org/10.1073/pnas.1206011109.

Gerstein, M. B. et al. (Dec. 2010). "Integrative Analysis of the Caenorhabditis elegans Genome by the modENCODE Project". In: *Science* 330.6012, pp. 1775–1787. DOI: 10.1126/science.1196914. URL: https://doi.org/10.1126/science.1196914.

Gerstein, Mark B et al. (2010). "Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project". In: *Science* 330.6012, pp. 1775–1787.

Gheorghe, Marius et al. (Dec. 2018). "A map of direct TF–DNA interactions in the human genome". In: *Nucleic Acids Research* 47.4, e21–e21. DOI: 10.1093/nar/gky1210. URL: https://doi.org/10.1093/nar/gky1210.

Gherardi, Samuele et al. (2013). "MYCN-mediated transcriptional repression in neuroblastoma: the other side of the coin". In: *Frontiers in Oncology* 3. DOI: 10.3389/fonc.2013.00042. URL: https://doi.org/10.3389/fonc.2013.00042.

Ginno, P. A. et al. (July 2013). "GC skew at the 5' and 3' ends of human genes links R-loop formation to epigenetic regulation and transcription termination". In: *Genome Research* 23.10,

pp. 1590–1600. DOI: 10.1101/gr.158436.113. URL: https://doi.org/10.1101/gr.158436.113.

Ginno, Paul A et al. (2012). "R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters". In: *Molecular cell* 45.6, pp. 814–825.

Gómez, Soledad et al. (2015). *DNA methylation fingerprint of neuroblastoma reveals new biological and clinical insights*.

gprofiler2 (n.d.). URL: https://cran.r-project.org/web/packages/gprofiler2.

Grandori, C et al. (1996). "Myc-Max heterodimers activate a DEAD box gene and interact with multiple E box-related sites in vivo." In: *EMBO J.* 15.16, pp. 4344–4357. DOI: 10.1016/j.celrep.2016.05.056. URL: https://doi.org/10.1016/j.celrep.2016.05.056.

Grant, Charles E, Timothy L Bailey, and William Stafford Noble (Apr. 2011). "FIMO: scanning for occurrences of a given motif". en. In: *Bioinformatics* 27.7, pp. 1017–1018.

Gröbner, Susanne N et al. (Mar. 2018). "The landscape of genomic alterations across childhood cancers". en. In: *Nature* 555.7696, pp. 321–327.

Groningen, Tim van et al. (Aug. 2017). "Neuroblastoma is composed of two super-enhancer-associated differentiation states". en. In: *Nat. Genet.* 49.8, pp. 1261–1266.

Grunseich, Christopher et al. (Feb. 2018). "Senataxin Mutation Reveals How R-Loops Promote Transcription by Blocking DNA Methylation at Gene Promoters". In: *Molecular Cell* 69.3, 426–437.e7. DOI: 10.1016/j.molcel.2017.12.030. URL: https://doi.org/10.1016/j.molcel.2017.12.030.

Gu, L et al. (Aug. 2011). "MDM2 regulates MYCN mRNA stabilization and translation in human neuroblastoma cells". In: *Oncogene* 31.11, pp. 1342–1353. DOI: 10.1038/onc.2011.343. URL: https://doi.org/10.1038/onc.2011.343.

Guccione, Ernesto et al. (July 2006). "Myc-binding-site recognition in the human genome is determined by chromatin context". en. In: *Nat. Cell Biol.* 8.7, pp. 764–770.

Guo, Shicheng et al. (Mar. 2017). "Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA". In: *Nature Genetics* 49.4, pp. 635–642. DOI: 10.1038/ng.3805. URL: https://doi.org/10.1038/ng.3805.

Hanahan, Douglas and Robert A Weinberg (2000). *The Hallmarks of Cancer*.

— (2011). *Hallmarks of Cancer: The Next Generation*.

Hänsel-Hertsch, Robert et al. (Sept. 2016). "G-quadruplex structures mark human regulatory chromatin". In: *Nature Genetics* 48.10, pp. 1267–1272. DOI: 10.1038/ng.3662. URL: https://doi.org/10.1038/ng.3662.

Hansen, Kasper D, Benjamin Langmead, and Rafael A Irizarry (2012). *BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions*.

Hansen, Kasper Daniel et al. (June 2011). "Increased methylation variation in epigenetic domains across cancer types". en. In: *Nat. Genet.* 43.8, pp. 768–775.

Harris, C C (1996). *Structure and Function of the p53 Tumor Suppressor Gene: Clues for Rational Cancer Therapeutic Strategies*.

Hasan, Md. Kamrul et al. (Dec. 2013). "ALK is a MYCN target gene and regulates cell migration and invasion in neuroblastoma". In: *Scientific Reports* 3.1. DOI: 10.1038/srep03450. URL: https://doi.org/10.1038/srep03450.

Hashimshony, Tamar et al. (June 2003). "The role of DNA methylation in setting up chromatin structure during development". en. In: *Nat. Genet.* 34.2, pp. 187–192.

Hatano, Masahiko et al. (Apr. 1997). "Ncx , a Hoxii related gene, is expressed in a variety of tissues derived from neural crest cells". In: *Anatomy and Embryology* 195.5, pp. 419–425. DOI: 10.1007/s004290050061. URL: https://doi.org/10.1007/s004290050061.

Hebestreit, Katja, Martin Dugas, and Hans-Ulrich Klein (2013). *Detection of significantly differentially methylated regions in targeted bisulfite sequencing data*.

Heinz, Sven et al. (2010). *Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities*.

Helmsauer, Konstantin et al. (Nov. 2020). "Enhancer hijacking determines extrachromosomal circular MYCN amplicon architecture in neuroblastoma". In: *Nature Communications* 11.1. DOI: 10.1038/s41467-020-19452-y. URL: https://doi.org/10.1038/s41467-020-19452-y.

Henikoff, Steven (Jan. 2008). "Nucleosome destabilization in the epigenetic regulation of gene expression". en. In: *Nat. Rev. Genet.* 9.1, pp. 15–26.

Henrich, Kai-Oliver et al. (Sept. 2016). "Integrative Genome-Scale Analysis Identifies Epigenetic Mechanisms of Transcriptional Deregulation in Unfavorable Neuroblastomas". en. In: *Cancer Res.* 76.18, pp. 5523–5537.

Herold, Steffi et al. (Mar. 2019). "Recruitment of BRCA1 limits MYCN-driven accumulation of stalled RNA polymerase". In: *Nature* 567.7749, pp. 545–549. DOI: 10.1038/s41586-019-1030-9. URL: https://doi.org/10.1038/s41586-019-1030-9.

Hoadley, Katherine A. et al. (Aug. 2014). "Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin". In: *Cell* 158.4, pp. 929–944. DOI: 10.1016/j.cell.2014.06.049. URL: https://doi.org/10.1016/j.cell.2014.06.049.

Hoebeeck, Jasmien et al. (2009). *Aberrant methylation of candidate tumor suppressor genes in neuroblastoma*.

Hoene, V et al. (Aug. 2009). "GATA factors in human neuroblastoma: distinctive expression patterns in clinical subtypes". In: *British Journal of Cancer* 101.8, pp. 1481–1489. DOI: 10.1038/sj.bjc.6605276. URL: https://doi.org/10.1038/sj.bjc.6605276.

Hon, Gary C et al. (Feb. 2012). "Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer". en. In: *Genome Res.* 22.2, pp. 246–258.

Hovestadt, Volker, David T W Jones, et al. (June 2014). "Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing". en. In: *Nature* 510.7506, pp. 537–541.

Hovestadt, Volker, Marc Remke, et al. (June 2013). "Robust molecular subgrouping and copy-number profiling of medulloblastoma from small amounts of archival tumour material using high-density DNA methylation arrays". en. In: *Acta Neuropathol.* 125.6, pp. 913–916.

Ikram, Fakhera et al. (Nov. 2015). "Transcription factor activating protein 2 beta (TFAP2B) mediates noradrenergic neuronal differentiation in neuroblastoma". In: *Molecular Oncology* 10.2, pp. 344–359. DOI: 10.1016/j.molonc.2015.10.020. URL: https://doi.org/10.1016/j.molonc.2015.10.020.

Irizarry, Rafael A et al. (2009). *The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores*.

Issa, Jean-Pierre (2004). *CpG island methylator phenotype in cancer*.

Jackson-Grusby, Laurie et al. (2001). *Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation*.

Jain, Dhawal et al. (June 2015). "Active promoters give rise to false positive 'Phantom Peaks' in ChIP-seq experiments". In: *Nucleic Acids Research* 43.14, pp. 6959–6968. DOI: 10.1093/nar/gkv637. URL: https://doi.org/10.1093/nar/gkv637.

Jain, Jugnu et al. (Apr. 1992). "Nuclear factor of activated T cells contains Fos and Jun". In: *Nature* 356.6372, pp. 801–804. DOI: 10.1038/356801a0. URL: https://doi.org/10.1038/356801a0.

Jain, Miten et al. (Dec. 2016). "Erratum to: The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community". In: *Genome Biology* 17.1. DOI: 10.1186/s13059-016-1122-x. URL: https://doi.org/10.1186/s13059-016-1122-x.

Jankowski, Aleksander, Jerzy Tiuryn, and Shyam Prabhakar (2016). *Romulus: robust multi-state identification of transcription factor binding sites from DNase-seq data*.

Jiao, Yinming, Martin Widschwendter, and Andrew E. Teschendorff (May 2014). "A systems-level integrative framework for genome-wide DNA methylation and gene expression data identifies differential gene expression modules under epigenetic control". In: *Bioinformatics* 30.16, pp. 2360–2366. DOI: 10.1093/bioinformatics/btu316. URL: https://doi.org/10.1093/bioinformatics/btu316.

Jin, Hulin, Jorge Sepúlveda, and Oscar R. Burrone (Jan. 2009). "Specific recognition of a dsDNA sequence motif by an immunoglobulin VH homodimer". In: *Protein Science* 13.12, pp. 3222–3229. DOI: 10.1110/ps.04921704. URL: https://doi.org/10.1110/ps.04921704.

Johnson, D S et al. (2007). *Genome-Wide Mapping of in Vivo Protein-DNA Interactions*.

Jones, P A (2001). *The Role of DNA Methylation in Mammalian Epigenetics*.

Jones, Peter A. (May 2012). "Functions of DNA methylation: islands, start sites, gene bodies and beyond". In: *Nature Reviews Genetics* 13.7, pp. 484–492. DOI: 10.1038/nrg3230. URL: https://doi.org/10.1038/nrg3230.

Jones, Peter A., Jean-Pierre J. Issa, and Stephen Baylin (Sept. 2016). "Targeting the cancer epigenome for therapy". In: *Nature Reviews Genetics* 17.10, pp. 630–641. DOI: 10.1038/nrg.2016.93. URL: https://doi.org/10.1038/nrg.2016.93.

Jubierre, Luz et al. (Apr. 2018). "Targeting of epigenetic regulators in neuroblastoma". en. In: *Exp. Mol. Med.* 50.4, p. 51.

Juergens, Rosalyn A. et al. (Nov. 2011). "Combination Epigenetic Therapy Has Efficacy in Patients with Refractory Advanced Non–Small Cell Lung Cancer". In: *Cancer Discovery* 1.7, pp. 598–607. DOI: 10.1158/2159-8290.cd-11-0214. URL: https://doi.org/10.1158/2159-8290.cd-11-0214.

Kalsi, Jatinderpal K. et al. (Mar. 1996). "Functional and modelling studies of the binding of human monoclonal anti-DNA antibodies to DNA". In: *Molecular Immunology* 33.4-5, pp. 471–483. DOI: 10.1016/0161-5890(95)00138-7. URL: https://doi.org/10.1016/0161-5890(95)00138-7.

Kamalakaran, Sitharthan et al. (2011). *DNA methylation patterns in luminal breast cancers differ from non-luminal subtypes and can identify relapse risk independent of other clinical variables*.

Kameneva, Polina et al. (Apr. 2021). "Single-cell transcriptomics of human embryos identifies multiple sympathoblast lineages with potential implications for neuroblastoma origin". In: *Nature Genetics*. DOI: 10.1038/s41588-021-00818-x. URL: https://doi.org/10.1038/s41588-021-00818-x.

Kandoth, Cyriac et al. (Oct. 2013). "Mutational landscape and significance across 12 major cancer types". en. In: *Nature* 502.7471, pp. 333–339.

Karanth, Santhosh et al. (June 2016). "FOXN3 Regulates Hepatic Glucose Utilization". In: *Cell Reports* 15.12, pp. 2745–2755. DOI: 10.1016/j.celrep.2016.05.056. URL: https://doi.org/10.1016/j.celrep.2016.05.056.

Karlic, R. et al. (Feb. 2010). "Histone modification levels are predictive for gene expression". In: *Proceedings of the National Academy of Sciences* 107.7, pp. 2926–2931. DOI: 10.1073/pnas.0909344107. URL: https://doi.org/10.1073/pnas.0909344107.

Kashima, Yukie et al. (Sept. 2020). "Single-cell sequencing techniques from individual to multiomics analyses". In: *Experimental & Molecular Medicine* 52.9, pp. 1419–1427. DOI: 10.1038/s12276-020-00499-2. URL: https://doi.org/10.1038/s12276-020-00499-2.

Kel, A E et al. (July 2003). "MATCH: A tool for searching transcription factor binding sites in DNA sequences". en. In: *Nucleic Acids Res.* 31.13, pp. 3576–3579.

Kellis, Manolis et al. (Apr. 2014). "Defining functional DNA elements in the human genome". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 111.17, pp. 6131–6138.

Kempfer, Rieke and Ana Pombo (Dec. 2019). "Methods for mapping 3D chromosome architecture". In: *Nature Reviews Genetics* 21.4, pp. 207–226. DOI: 10.1038/s41576-019-0195-2. URL: https://doi.org/10.1038/s41576-019-0195-2.

Kent, W. J. et al. (July 2010). "BigWig and BigBed: enabling browsing of large distributed datasets". In: *Bioinformatics* 26.17, pp. 2204–2207. DOI: 10.1093/bioinformatics/btq351. URL: https://doi.org/10.1093/bioinformatics/btq351.

Kerachian, Mohammad Amin, Ali Javadmanesh, et al. (Feb. 2020). "Crosstalk between DNA methylation and gene expression in colorectal cancer, a potential plasma biomarker for tracing this tumor". In: *Scientific Reports* 10.1. DOI: 10.1038/s41598-020-59690-0. URL: https://doi.org/10.1038/s41598-020-59690-0.

Kerachian, Mohammad Amin and Matin Kerachian (Jan. 2019). "Long interspersed nucleotide element-1 (LINE-1) methylation in colorectal cancer". en. In: *Clin. Chim. Acta* 488, pp. 209–214.

Kharchenko, Peter V, Michael Y Tolstorukov, and Peter J Park (2008). *Design and analysis of ChIP-seq experiments for DNA-binding proteins*.

Khoury, Amanda, Joanna Achinger-Kawecka, Saul A Bert, et al. (Jan. 2020a). "Constitutively bound CTCF sites maintain 3D chromatin architecture and long-range epigenetically regulated domains". en. In: *Nat. Commun.* 11.1, p. 54.

— (Jan. 2020b). "Constitutively bound CTCF sites maintain 3D chromatin architecture and long-range epigenetically regulated domains". In: *Nature Communications* 11.1. DOI: 10.1038/s41467-019-13753-7. URL: https://doi.org/10.1038/s41467-019-13753-7.

Kidder, Benjamin L, Gangqing Hu, and Keji Zhao (2011). *ChIP-Seq: technical considerations for obtaining high-quality data*.

Kildisiute, Gerda et al. (Feb. 2021). "Tumor to normal single-cell mRNA comparisons reveal a pan-neuroblastoma cancer cell". In: *Science Advances* 7.6, eabd3311. DOI: 10.1126/sciadv.abd3311. URL: https://doi.org/10.1126/sciadv.abd3311.

Kleinjan, D (1998). *Position effect in human genetic disease*.

Klughammer, Johanna, Barbara Kiesel, Thomas Roetzer, Nikolaus Fortelny, Amelie Nemc, Karl-Heinz Nenning, Julia Furtner, Nathan C Sheffield, et al. (Oct. 2018a). "The DNA methylation landscape of glioblastoma disease progression shows extensive heterogeneity in time and space". en. In: *Nat. Med.* 24.10, pp. 1611–1624.

Klughammer, Johanna, Barbara Kiesel, Thomas Roetzer, Nikolaus Fortelny, Amelie Nemc, Karl-Heinz Nenning, Julia Furtner, Nathan C. Sheffield, et al. (Aug. 2018b). "The DNA methylation landscape of glioblastoma disease progression shows extensive heterogeneity in time and space". In: *Nature Medicine* 24.10, pp. 1611–1624. DOI: 10.1038/s41591-018-0156-x. URL: https://doi.org/10.1038/s41591-018-0156-x.

Knudson Jr, A G (Apr. 1971). "Mutation and cancer: statistical study of retinoblastoma". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 68.4, pp. 820–823.

Koch, Alexander et al. (Apr. 2018). "Analysis of DNA methylation in cancer: location revisited". In: *Nature Reviews Clinical Oncology* 15.7, pp. 459–466. DOI: 10.1038/s41571-018-0004-4. URL: https://doi.org/10.1038/s41571-018-0004-4.

Koche, Richard P et al. (Jan. 2020). "Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma". en. In: *Nat. Genet.* 52.1, pp. 29–34.

Koche, Richard P. et al. (Dec. 2019). "Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma". In: *Nature Genetics* 52.1, pp. 29–34. DOI: 10.1038/s41588-019-0547-z. URL: https://doi.org/10.1038/s41588-019-0547-z.

Kolas, Nadine K. and Daniel Durocher (Apr. 2006). "DNA Repair: DNA Polymerase and Revi Break in". In: *Current Biology* 16.8, R296–R299. DOI: 10.1016/j.cub.2006.03.043. URL: https://doi.org/10.1016/j.cub.2006.03.043.

Kopp, Wolfgang, Remo Monti, et al. (July 2020). "Deep learning for genomics using Janggu". In: *Nature Communications* 11.1. DOI: 10.1038/s41467-020-17155-y. URL: https://doi.org/10.1038/s41467-020-17155-y.

Kopp, Wolfgang and Martin Vingron (Dec. 2017). "An improved compound Poisson model for the number of motif hits in DNA sequences". en. In: *Bioinformatics* 33.24, pp. 3929–3937.

Kotsantis, Panagiotis et al. (Oct. 2016). "Increased global transcription activity as a mechanism of replication stress in cancer". In: *Nature Communications* 7.1. DOI: 10.1038/ncomms13087. URL: https://doi.org/10.1038/ncomms13087.

Krebs, Wolfgang et al. (Nov. 2014). "Optimization of transcription factor binding map accuracy utilizing knockout-mouse models". In: *Nucleic Acids Research* 42.21, pp. 13051–13060. DOI: 10.1093/nar/gku1078. URL: https://doi.org/10.1093/nar/gku1078.

Krueger, Felix and Simon R Andrews (2011). *Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications*.

Kumar, Sanjeev et al. (May 2011). "p185, an Immunodominant Epitope, Is an Autoantigen Mimotope". In: *Journal of Biological Chemistry* 286.29, pp. 26220–26227. DOI: 10.1074/jbc.m111.224303. URL: https://doi.org/10.1074/jbc.m111.224303.

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

Kvon, Evgeny Z et al. (May 2012). "HOT regions function as patterned developmental enhancers and have a distinct cis-regulatory signature". en. In: *Genes Dev.* 26.9, pp. 908–913.

Kwok, Wai Kei et al. (2005). *Up-Regulation of TWIST in Prostate Cancer and Its Implication as a Therapeutic Target*.

Laird, Peter W (Apr. 2003). "The power and the promise of DNA methylation markers". en. In: *Nat. Rev. Cancer* 3.4, pp. 253–266.

Lakamp, Amanda S. and Michel M. Ouellette (2011). "A ssDNA Aptamer That Blocks the Function of the Anti-FLAG M2 Antibody". In: *Journal of Nucleic Acids* 2011, pp. 1–11. DOI: 10.4061/2011/720798. URL: https://doi.org/10.4061/2011/720798.

Lambert, Samuel A. et al. (Feb. 2018). "The Human Transcription Factors". In: *Cell* 172.4, pp. 650–665. DOI: 10.1016/j.cell.2018.01.029. URL: https://doi.org/10.1016/j.cell.2018.01.029.

Lander, Eric S (2011). *Initial impact of the sequencing of the human genome*.

Landt, Stephen G et al. (Sept. 2012). "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia". en. In: *Genome Res.* 22.9, pp. 1813–1831.

Lane, D P (1992). *p53, guardian of the genome*.

Langmead, Ben et al. (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome". In: *Genome Biology* 10.3, R25. DOI: 10.1186/gb-2009-10-3-r25. URL: https://doi.org/10.1186/gb-2009-10-3-r25.

Lawrence, Michael S et al. (July 2013). "Mutational heterogeneity in cancer and the search for new cancer-associated genes". en. In: *Nature* 499.7457, pp. 214–218.

Lázcoz, Paula et al. (Oct. 2006). "Frequent promoter hypermethylation of RASSF1A and CASP8 in neuroblastoma". en. In: *BMC Cancer* 6, p. 254.

Lederer, Simone et al. (May 2020). "Investigating the effect of dependence between conditions with Bayesian Linear Mixed Models for motif activity analysis". en. In: *PLoS One* 15.5, e0231824.

Lee, Qian Yi et al. (Mar. 2020). "Pro-neuronal activity of Myod1 due to promiscuous binding to neuronal genes". In: *Nature Cell Biology* 22.4, pp. 401–411. DOI: 10.1038/s41556-020-0490-3. URL: https://doi.org/10.1038/s41556-020-0490-3.

Lee, Shih-Han et al. (Aug. 2018). "Widespread intronic polyadenylation inactivates tumour suppressor genes in leukaemia". In: *Nature* 561.7721, pp. 127–131. DOI: 10.1038/s41586-018-0465-8. URL: https://doi.org/10.1038/s41586-018-0465-8.

Lee, Tong Ihn and Richard A. Young (Mar. 2013). "Transcriptional Regulation and Its Misregulation in Disease". In: *Cell* 152.6, pp. 1237–1251. DOI: 10.1016/j.cell.2013.02.014. URL: https://doi.org/10.1016/j.cell.2013.02.014.

Lentini, Antonio et al. (June 2018). "A reassessment of DNA-immunoprecipitation-based genomic profiling". In: *Nature Methods* 15.7, pp. 499–504. DOI: 10.1038/s41592-018-0038-7. URL: https://doi.org/10.1038/s41592-018-0038-7.

Li, H. (Sept. 2011). "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data". In: *Bioinformatics* 27.21, pp. 2987–2993. DOI: 10.1093/bioinformatics/btr509. URL: https://doi.org/10.1093/bioinformatics/btr509.

Li, H. et al. (June 2009). "The Sequence Alignment/Map format and SAMtools". In: *Bioinformatics* 25.16, pp. 2078–2079. DOI: 10.1093/bioinformatics/btp352. URL: https://doi.org/10.1093/bioinformatics/btp352.

Li, Sheng et al. (2013). *An optimized algorithm for detecting and annotating regional differential methylation*.

Li, X. (July 2006). "Cotranscriptional processes and their influence on genome stability". In: *Genes & Development* 20.14, pp. 1838–1847. DOI: 10.1101/gad.1438306. URL: https://doi.org/10.1101/gad.1438306.

Libertini, Emanuele et al. (June 2016). "Information recovery from low coverage whole-genome bisulfite sequencing". In: *Nature Communications* 7.1. DOI: 10.1038/ncomms11306. URL: https://doi.org/10.1038/ncomms11306.

Liberzon, A. et al. (May 2011). "Molecular signatures database (MSigDB) 3.0". In: *Bioinformatics* 27.12, pp. 1739–1740. DOI: 10.1093/bioinformatics/btr260. URL: https://doi.org/10.1093/bioinformatics/btr260.

Lim, Yoong Wearn et al. (July 2015). "Genome-wide DNA hypomethylation and RNA:DNA hybrid accumulation in Aicardi–Goutières syndrome". In: *eLife* 4. DOI: 10.7554/elife.08007. URL: https://doi.org/10.7554/elife.08007.

Lima, Walt F. et al. (Apr. 2016). "ViableRNaseH1knockout mice show RNaseH1 is essential for R loop processing, mitochondrial and liver function". In: *Nucleic Acids Research* 44.11, pp. 5299–5312. DOI: 10.1093/nar/gkw350. URL: https://doi.org/10.1093/nar/gkw350.

Linhares, Brian M, Jolanta Grembecka, and Tomasz Cierpicki (July 2020). "Targeting epigenetic protein–protein interactions with small-molecule inhibitors". In: *Future Medicinal Chemistry* 12.14, pp. 1305–1326. DOI: 10.4155/fmc-2020-0082. URL: https://doi.org/10.4155/fmc-2020-0082.

Lister, Ryan et al. (Nov. 2009). "Human DNA methylomes at base resolution show widespread epigenomic differences". en. In: *Nature* 462.7271, pp. 315–322.

Liu, Delong et al. (2007). *t(8;14;18): A 3-way chromosome translocation in two patients with Burkitt's lymphoma/leukemia*.

Locke, Warwick J et al. (Nov. 2019). "DNA Methylation Cancer Biomarkers: Translation to the Clinic". en. In: *Front. Genet.* 10, p. 1150.

Logan, Cairine et al. (July 1998). "Tlx-1andTlx-3Homeobox Gene Expression in Cranial Sensory Ganglia and Hindbrain of the Chick Embryo: Markers of Patterned Connectivity". In: *The Journal of Neuroscience* 18.14, pp. 5389–5402. DOI: 10.1523/jneurosci.18-14-05389.1998. URL: https://doi.org/10.1523/jneurosci.18-14-05389.1998.

Long, Hannah K., Sara L. Prescott, and Joanna Wysocka (Nov. 2016). "Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution". In: *Cell* 167.5, pp. 1170–1187. DOI: 10.1016/j.cell.2016.09.018. URL: https://doi.org/10.1016/j.cell.2016.09.018.

Love, Michael I, Wolfgang Huber, and Simon Anders (Dec. 2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15.12. DOI: 10.1186/s13059-014-0550-8. URL: https://doi.org/10.1186/s13059-014-0550-8.

Luger, Karolin et al. (Sept. 1997). "Crystal structure of the nucleosome core particle at 2.8AA resolution". In: *Nature* 389.6648, pp. 251–260. DOI: 10.1038/38444. URL: https://doi.org/10.1038/38444.

Madsen, Jesper Grud Skat et al. (Feb. 2018). "Integrated analysis of motif activity and gene expression changes of transcription factors". en. In: *Genome Res.* 28.2, pp. 243–255.

Mantovani, Fiamma, Licio Collavin, and Giannino Del Sal (Jan. 2019). "Mutant p53 as a guardian of the cancer cell". en. In: *Cell Death Differ.* 26.2, pp. 199–212.

Marchini, B et al. (1995). "Induction of anti-DNA antibodies in non autoimmune mice by immunization with a DNA-DNAase I complex". In: *Clinical and experimental rheumatology* 13.1, pp. 7–10. ISSN: 0392-856X. URL: http://europepmc.org/abstract/MED/7774106.

Margetts, C D E et al. (2008). *Evaluation of a functional epigenetic approach to identify promoter region methylation in phaeochromocytoma and neuroblastoma*.

Margueron, Raphaël and Danny Reinberg (2010). *Chromatin structure and the inheritance of epigenetic information*.

Marx, Vivien (Mar. 2019). "What to do about those immunoprecipitation blues". In: *Nature Methods* 16.4, pp. 289–292. DOI: 10.1038/s41592-019-0365-3. URL: https://doi.org/10.1038/s41592-019-0365-3.

Matthay, Katherine K et al. (Nov. 2016). "Neuroblastoma". en. In: *Nat Rev Dis Primers* 2, p. 16078.

Mayol, Gemma et al. (Nov. 2012). "DNA hypomethylation affects cancer-related biological functions and genes relevant in neuroblastoma pathogenesis". en. In: *PLoS One* 7.11, e48401.

McLean, Cory Y et al. (May 2010). "GREAT improves functional interpretation of cis-regulatory regions". In: *Nature Biotechnology* 28.5, pp. 495–501. DOI: 10.1038/nbt.1630. URL: https://doi.org/10.1038/nbt.1630.

Mészáros, Bálint et al. (Mar. 2017). "Degrons in cancer". In: *Science Signaling* 10.470, eaak9982. DOI: 10.1126/scisignal.aak9982. URL: https://doi.org/10.1126/scisignal.aak9982.

MethylDackel (n.d.). URL: https://github.com/dpryan79/MethylDackel.

Mikkelsen, Tarjei S et al. (2007). *Genome-wide maps of chromatin state in pluripotent and lineage-committed cells*.

Miranda, Tina Branscombe and Peter A. Jones (2007). "DNA methylation: The nuts and bolts of repression". In: *Journal of Cellular Physiology* 213.2, pp. 384–390. DOI: 10.1002/jcp.21224. URL: https://doi.org/10.1002/jcp.21224.

Moens, U. et al. (Dec. 1995). "In vivo expression of a single viral DNA-binding protein generates systemic lupus erythematosus-related autoimmunity to double-stranded DNA and histones." In: *Proceedings of the National Academy of Sciences* 92.26, pp. 12393–12397. DOI: 10.1073/pnas.92.26.12393. URL: https://doi.org/10.1073/pnas.92.26.12393.

Moens, Ugo et al. (Jan. 2002). "Green fluorescent protein modified to bind DNA initiates production of anti-DNA antibodies when expressed in vivo". In: *Molecular Immunology* 38.7, pp. 505–514. DOI: 10.1016/s0161-5890(01)00086-4. URL: https://doi.org/10.1016/s0161-5890(01)00086-4.

Mohammad, Helai P, Olena Barbash, and Caretha L Creasy (Mar. 2019). "Targeting epigenetic modifications in cancer therapy: erasing the roadmap to cancer". en. In: *Nat. Med.* 25.3, pp. 403–418.

Molenaar, Jan J et al. (Feb. 2012). "Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes". en. In: *Nature* 483.7391, pp. 589–593.

Moran, Sebastian et al. (Oct. 2016). "Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis". en. In: *Lancet Oncol.* 17.10, pp. 1386–1395.

Moss, Joshua et al. (Nov. 2018). "Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease". In: *Nature Communications* 9.1. DOI: 10.1038/s41467-018-07466-6. URL: https://doi.org/10.1038/s41467-018-07466-6.

Mossé, Yaël P et al. (2008). *Identification of ALK as a major familial neuroblastoma predisposition gene*.

MSigDB gene sets (n.d.). URL: https://www.gsea-msigdb.org/gsea/msigdb/gene_families.jsp.

Murphy, Derek M et al. (Dec. 2009). "Global MYCN transcription factor binding analysis in neuroblastoma reveals association with distinct E-box motifs and regions of DNA hypermethylation". en. In: *PLoS One* 4.12, e8154.

Nadel, Julie et al. (Nov. 2015). "RNA:DNA hybrids in the human genome have distinctive nucleotide characteristics, chromatin composition, and transcriptional relationships". In: *Epigenetics & Chromatin* 8.1. DOI: 10.1186/s13072-015-0040-6. URL: https://doi.org/10.1186/s13072-015-0040-6.

Nagai, Kozo (May 2001). "Molecular evolution of Sry and Sox gene". In: *Gene* 270.1-2, pp. 161–169. DOI: 10.1016/s0378-1119(01)00479-6. URL: https://doi.org/10.1016/s0378-1119(01)00479-6.

Neri, Francesco et al. (2017). *Intragenic DNA methylation prevents spurious transcription initiation*.

Niehrs, Christof and Brian Luke (Jan. 2020). "Regulatory R-loops as facilitators of gene expression and genome stability". In: *Nature Reviews Molecular Cell Biology* 21.3, pp. 167–178. DOI: 10.1038/s41580-019-0206-3. URL: https://doi.org/10.1038/s41580-019-0206-3.

Noushmehr, Houtan et al. (2010). *Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma*.

Nowell, P (1976). *The clonal evolution of tumor cell populations*.

Olsson, Maja et al. (2016). *Genome-wide methylation profiling identifies novel methylated genes in neuroblastoma tumors*.

Osmanbeyoglu, Hatice U. et al. (Sept. 2014). "Linking signaling pathways to transcriptional programs in breast cancer". In: *Genome Research* 24.11, pp. 1869–1880. DOI: 10.1101/gr.173039.114. URL: https://doi.org/10.1101/gr.173039.114.

Otto, Tobias (n.d.). "MYCN and Its Posttranslational Regulation in Neuroblastoma". In: *Pediatric and Adolescent Medicine*. S. Karger AG, pp. 47–58. DOI: 10.1159/000382085. URL: https://doi.org/10.1159/000382085.

Özdağ, Hilal et al. (Apr. 2006). "Differential expression of selected histone modifier genes in human solid cancers". In: *BMC Genomics* 7.1. DOI: 10.1186/1471-2164-7-90. URL: https://doi.org/10.1186/1471-2164-7-90.

Panning, B and R Jaenisch (1996). *DNA hypomethylation can activate Xist expression and silence X-linked genes*.

Park, Daechan et al. (Dec. 2013). "Widespread Misinterpretable ChIP-seq Bias in Yeast". In: *PLoS ONE* 8.12. Ed. by Ben Lehner, e83506. DOI: 10.1371/journal.pone.0083506. URL: https://doi.org/10.1371/journal.pone.0083506.

Park, Peter J (2009). *ChIP–seq: advantages and challenges of a maturing technology*.

Park, Yongseok et al. (2014). *MethylSig: a whole genome DNA methylation analysis pipeline*.

Parseghian, Missag Hagop (Dec. 2013). "Hitchhiker antigens: Inconsistent ChIP results, questionable immunohistology data, and poor antibody performance may have a common factor". In: *Biochemistry and Cell Biology* 91.6, pp. 378–394. DOI: 10.1139/bcb-2013-0059. URL: https://doi.org/10.1139/bcb-2013-0059.

Partridge, E. Christopher et al. (July 2020). "Occupancy maps of 208 chromatin-associated proteins in one human cell type". In: *Nature* 583.7818, pp. 720–728. DOI: 10.1038/s41586-020-2023-4. URL: https://doi.org/10.1038/s41586-020-2023-4.

Pedersen, Brent S. et al. (2014). *Fast and accurate alignment of long bisulfite-seq reads*. eprint: arXiv:1401.1129.

Peifer, Martin et al. (Oct. 2015). "Telomerase activation by genomic rearrangements in high-risk neuroblastoma". en. In: *Nature* 526.7575, pp. 700–704.

Petrakova, Natalia et al. (Apr. 2009). "Autoimmunogenicity of the helix-loop-helix DNA-binding domain". In: *Molecular Immunology* 46.7, pp. 1467–1480. DOI: 10.1016/j.molimm.2008.12.013. URL: https://doi.org/10.1016/j.molimm.2008.12.013.

Phillips, Jennifer E and Victor G Corces (June 2009). "CTCF: master weaver of the genome". en. In: *Cell* 137.7, pp. 1194–1211.

Picard (n.d.). URL: http://broadinstitute.github.io/picard.

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

Plank, Jennifer L. and Ann Dean (July 2014). "Enhancer Function: Mechanistic and Genome-Wide Insights Come Together". In: *Molecular Cell* 55.1, pp. 5–14. DOI: 10.1016/j.molcel.2014.06.015. URL: https://doi.org/10.1016/j.molcel.2014.06.015.

Pombo, Ana and Niall Dillon (Mar. 2015). "Three-dimensional genome architecture: players and mechanisms". In: *Nature Reviews Molecular Cell Biology* 16.4, pp. 245–257. DOI: 10.1038/nrm3965. URL: https://doi.org/10.1038/nrm3965.

Potzner, M. R. et al. (Feb. 2010). "Sequential requirement of Sox4 and Sox11 during development of the sympathetic nervous system". In: *Development* 137.5, pp. 775–784. DOI: 10.1242/dev.042101. URL: https://doi.org/10.1242/dev.042101.

Prebet, Thomas et al. (Apr. 2014). "Prolonged Administration of Azacitidine With or Without Entinostat for Myelodysplastic Syndrome and Acute Myeloid Leukemia With

Myelodysplasia-Related Changes: Results of the US Leukemia Intergroup Trial E1905". In: *Journal of Clinical Oncology* 32.12, pp. 1242–1248. DOI: 10.1200/jco.2013.50.3102. URL: https://doi.org/10.1200/jco.2013.50.3102.

Preter, Katleen De et al. (Jan. 2007). "Erratum to: Human fetal neuroblast and neuroblastoma transcriptome analysis confirms neuroblast origin and highlights neuroblastoma candidate genes". In: *Genome Biology* 8.1. DOI: 10.1186/gb-2007-8-1-401. URL: https://doi.org/10.1186/gb-2007-8-1-401.

Pugh, Trevor J et al. (Mar. 2013). "The genetic landscape of high-risk neuroblastoma". en. In: *Nat. Genet.* 45.3, pp. 279–284.

Raiber, Eun-Ang et al. (Oct. 2011). "A non-canonical DNA structure is a binding motif for the transcription factor SP1 in vitro". In: *Nucleic Acids Research* 40.4, pp. 1499–1508. DOI: 10.1093/nar/gkr882. URL: https://doi.org/10.1093/nar/gkr882.

Rao, X et al. (Sept. 2013). "CpG island shore methylation regulates caveolin-1 expression in breast cancer". en. In: *Oncogene* 32.38, pp. 4519–4528.

Raschella, Giuseppe et al. (July 1999). "Expression of Bmyb in Neuroblastoma Tumors Is a Poor Prognostic Factor Independent from MYCN Amplification". In: *Cancer research* 59 (14). DOI: 10.1093/hmg/ddv383. URL: https://cancerres.aacrjournals.org/content/59/14/3365.long.

Raudvere, Uku et al. (May 2019). "g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)". In: *Nucleic Acids Research* 47.W1, W191–W198. DOI: 10.1093/nar/gkz369. URL: https://doi.org/10.1093/nar/gkz369.

Reichlin, M et al. (Jan. 1994). "Lupus autoantibodies to native DNA cross-react with the A and D SnRNP polypeptides." In: *Journal of Clinical Investigation* 93.1, pp. 443–449. DOI: 10.1172/jci116980. URL: https://doi.org/10.1172/jci116980.

Reiff, T. et al. (Jan. 2010). "Neuroblastoma Phox2b Variants Stimulate Proliferation and Dedifferentiation of Immature Sympathetic Neurons". In: *Journal of Neuroscience* 30.3, pp. 905–915. DOI: 10.1523/jneurosci.5368-09.2010. URL: https://doi.org/10.1523/jneurosci.5368-09.2010.

Ren, Ping et al. (Oct. 2014). "ATF4 and N-Myc coordinate glutamine metabolism inMYCN-amplified neuroblastoma cells through ASCT2 activation". In: *The Journal of Pathology* 235.1, pp. 90–100. DOI: 10.1002/path.4429. URL: https://doi.org/10.1002/path.4429.

Rhie, Suhn Kyong et al. (Nov. 2016). "Identification of activated enhancers and linked transcription factors in breast, prostate, and kidney tumors by tracing enhancer networks using epigenetic traits". In: *Epigenetics & Chromatin* 9.1. DOI: 10.1186/s13072-016-0102-4. URL: https://doi.org/10.1186/s13072-016-0102-4.

Ribeiro, Diogo et al. (July 2016). "Regulation of Nuclear Hormone Receptors by MYCN-Driven miRNAs Impacts Neural Differentiation and Survival in Neuroblastoma Patients". In: *Cell Reports* 16.4, pp. 979–993. DOI: 10.1016/j.celrep.2016.06.052. URL: https://doi.org/10.1016/j.celrep.2016.06.052.

Ritchie, Matthew E et al. (2015). *limma powers differential expression analyses for RNA-sequencing and microarray studies*.

Robertson, Gordon et al. (2007). *Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing*.

Ronen, Jonathan, Sikander Hayat, and Altuna Akalin (Dec. 2019). "Evaluation of colorectal cancer subtypes and cell lines using deep learning". In: *Life Science Alliance* 2.6, e201900517. DOI: 10.26508/lsa.201900517. URL: https://doi.org/10.26508/lsa.201900517.

Ross, Jason P. et al. (Nov. 2010). "Recombinant mammalian DNA methyltransferase activity on model transcriptional gene silencing short RNA–DNA heteroduplex substrates". In: *Biochemical Journal* 432.2, pp. 323–332. DOI: 10.1042/bj20100579. URL: https://doi.org/10.1042/bj20100579.

Rubin, Jonathan D. et al. (Jan. 2020). "Transcription factor enrichment analysis (TFEA): Quantifying the activity of hundreds of transcription factors from a single experiment". In: DOI: 10.1101/2020.01.25.919738. URL: https://doi.org/10.1101/2020.01.25.919738.

Saint-André, Violaine et al. (Feb. 2016). "Models of human core transcriptional regulatory circuitries". In: *Genome Research* 26.3, pp. 385–396. DOI: 10.1101/gr.197590.115. URL: https://doi.org/10.1101/gr.197590.115.

Saito, Yutaka and Toutai Mituyama (2015). *Detection of differentially methylated regions from bisulfite-seq data by hidden Markov models incorporating genome-wide methylation level distributions*.

Salhab, Abdulrahman et al. (Sept. 2018). "A comprehensive analysis of 195 DNA methylomes reveals shared and cell-specific features of partially methylated domains". en. In: *Genome Biol.* 19.1, p. 150.

Sanger, F. and A.R. Coulson (May 1975). "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase". In: *Journal of Molecular Biology* 94.3, pp. 441–448. DOI: 10.1016/0022-2836(75)90213-2. URL: https://doi.org/10.1016/0022-2836(75)90213-2.

Santos-Pereira, José M. and Andrés Aguilera (Sept. 2015). "R loops: new modulators of genome dynamics and function". In: *Nature Reviews Genetics* 16.10, pp. 583–597. DOI: 10.1038/nrg3961. URL: https://doi.org/10.1038/nrg3961.

Sanz, Lionel A. et al. (July 2016). "Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals". In: *Molecular Cell* 63.1, pp. 167–178. DOI: 10.1016/j.molcel.2016.05.032. URL: https://doi.org/10.1016/j.molcel.2016.05.032.

Schlosberg, Christopher E., Nathan D. VanderKraats, and John R. Edwards (Feb. 2017). "Modeling complex patterns of differential DNA methylation that associate with gene expression changes". In: *Nucleic Acids Research* 45.9, pp. 5100–5111. DOI: 10.1093/nar/gkx078. URL: https://doi.org/10.1093/nar/gkx078.

Schönherr, C et al. (Mar. 2010). "Anaplastic lymphoma kinase activates the small GTPase Rap1 via the Rap1-specific GEF C3G in both neuroblastoma and PC12 cells". In: *Oncogene* 29.19, pp. 2817–2830. DOI: 10.1038/onc.2010.27. URL: https://doi.org/10.1038/onc.2010.27.

Schroeder, D I et al. (2011). *Large-scale methylation domains mark a functional subset of neuronally expressed genes*.

Schübeler, Dirk (2015). *Function and information content of DNA methylation*.

Schwab, Manfred et al. (Sept. 1983). "Amplified DNA with limited homology to myc cellular oncogene is shared by human neuroblastoma cell lines and a neuroblastoma tumour". In: *Nature* 305.5931, pp. 245–248. DOI: 10.1038/305245a0. URL: https://doi.org/10.1038/305245a0.

Sciascia, Sandra A. et al. (Jan. 2007). "Immunization of nonautoimmune mice with DNA binding domains of the largest subunit of RNA polymerase I results in production of anti-dsDNA and anti-Sm/RNP antibodies". In: *Autoimmunity* 40.1, pp. 38–47. DOI: 10.1080/08916930601185550. URL: https://doi.org/10.1080/08916930601185550.

Sean R. Eddy, Travis J. Wheeler (n.d.). *HHMER-Biological sequence analysis using profile hidden Markov models*. URL: http://hmmer.org/.

Sekiyama, Yoshiharu, Hitoshi Suzuki, and Toshifumi Tsukahara (Aug. 2011). "Functional Gene Expression Analysis of Tissue-Specific Isoforms of Mef2c". In: *Cellular and Molecular Neurobiology* 32.1, pp. 129–139. DOI: 10.1007/s10571-011-9743-9. URL: https://doi.org/10.1007/s10571-011-9743-9.

Shen, Li et al. (Apr. 2013). "Genome-wide Analysis Reveals TET- and TDG-Dependent 5-Methylcytosine Oxidation Dynamics". In: *Cell* 153.3, pp. 692–706. DOI: 10.1016/j.cell.2013.04.002. URL: https://doi.org/10.1016/j.cell.2013.04.002.

Shen, Ronglai, Adam B. Olshen, and Marc Ladanyi (Sept. 2009). "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis". In: *Bioinformatics* 25.22, pp. 2906–2912. DOI: 10.1093/bioinformatics/btp543. URL: https://doi.org/10.1093/bioinformatics/btp543.

Shim, Ki Shuk et al. (July 2006). "Bach2 is involved in neuronal differentiation of N1E-115 neuroblastoma cells". In: *Experimental Cell Research* 312.12, pp. 2264–2278. DOI: 10.1016/j.yexcr.2006.03.018. URL: https://doi.org/10.1016/j.yexcr.2006.03.018.

Shlyueva, Daria, Gerald Stampfel, and Alexander Stark (Apr. 2014). "Transcriptional enhancers: from properties to genome-wide predictions". en. In: *Nat. Rev. Genet.* 15.4, pp. 272–286.

Silva, Tiago C et al. (Oct. 2018). "ELMER v.2: an R/Bioconductor package to reconstruct gene regulatory networks from DNA methylation and transcriptome profiles". In: *Bioinformatics* 35.11. Ed. by Oliver Stegle, pp. 1974–1977. DOI: 10.1093/bioinformatics/bty902. URL: https://doi.org/10.1093/bioinformatics/bty902.

Simon, Noah et al. (2011). "Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent". In: *Journal of Statistical Software* 39.5, pp. 1–13. URL: https://www.jstatsoft.org/v39/i05/.

Sin, Sarah T K et al. (Feb. 2021). "Characteristics of Fetal Extrachromosomal Circular DNA in Maternal Plasma: Methylation Status and Clearance". In: *Clinical Chemistry*. DOI: 10.1093/clinchem/hvaa326. URL: https://doi.org/10.1093/clinchem/hvaa326.

Skourti-Stathaki, K. and N. J. Proudfoot (July 2014). "A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression". In: *Genes & Development* 28.13, pp. 1384–1396. DOI: 10.1101/gad.242990.114. URL: https://doi.org/10.1101/gad.242990.114.

Skourti-Stathaki, Konstantina, Nicholas J. Proudfoot, and Natalia Gromak (June 2011). "Human Senataxin Resolves RNA/DNA Hybrids Formed at Transcriptional Pause Sites to Promote Xrn2-Dependent Termination". In: *Molecular Cell* 42.6, pp. 794–805. DOI: 10.1016/j.molcel.2011.04.026. URL: https://doi.org/10.1016/j.molcel.2011.04.026.

Smith, Zachary D and Alexander Meissner (2013). *DNA methylation: roles in mammalian development*.

Smolka, John A. et al. (Jan. 2020). "Recognition of cellular RNAs by the S9.6 antibody creates pervasive artefacts when imaging RNA:DNA hybrids". In: DOI: 10.1101/2020.01.11.902981. URL: https://doi.org/10.1101/2020.01.11.902981.

Solomon, Mark J, Pamela L Larsen, and Alexander Varshavsky (1988). *Mapping proteinDNA interactions in vivo with formaldehyde: Evidence that histone H4 is retained on a highly transcribed gene*.

Song, Qiang et al. (2013). *A Reference Methylome Database and Analysis Pipeline to Facilitate Integrative and Comparative Epigenomics*.

Stadler, Michael B et al. (Dec. 2011). "DNA-binding factors shape the mouse methylome at distal regulatory regions". en. In: *Nature* 480.7378, pp. 490–495.

Statello, Luisa et al. (Jan. 2021). "Author Correction: Gene regulation by long non-coding RNAs and its biological functions". In: *Nature Reviews Molecular Cell Biology* 22.2, pp. 159–159. DOI: 10.1038/s41580-021-00330-4. URL: https://doi.org/10.1038/s41580-021-00330-4.

Steliarova-Foucher, Eva et al. (June 2017). "International incidence of childhood cancer, 2001-10: a population-based registry study". en. In: *Lancet Oncol.* 18.6, pp. 719–731.

Stephens, Philip J. et al. (Jan. 2011). "Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development". In: *Cell* 144.1, pp. 27–40. DOI: 10.1016/j.cell.2010.11.055. URL: https://doi.org/10.1016/j.cell.2010.11.055.

Stieglitz, Elliot et al. (2017). *Genome-wide DNA methylation is predictive of outcome in juvenile myelomonocytic leukemia*.

Stiller, C. A. and D. M. Parkin (Oct. 1992). "International variations in the incidence of neuroblastoma". In: *International Journal of Cancer* 52.4, pp. 538–543. DOI: 10.1002/ijc.2910520407. URL: https://doi.org/10.1002/ijc.2910520407.

Stirzaker, Clare et al. (2014). *Mining cancer methylomes: prospects and challenges*.

Stormo, Gary D. and Yue Zhao (Sept. 2010). "Determining the specificity of protein–DNA interactions". In: *Nature Reviews Genetics* 11.11, pp. 751–760. DOI: 10.1038/nrg2845. URL: https://doi.org/10.1038/nrg2845.

Strieder, Verena and Werner Lutz (Jan. 2003). "E2F Proteins Regulate MYCN Expression in Neuroblastomas". In: *Journal of Biological Chemistry* 278.5, pp. 2983–2989. DOI: 10.1074/jbc.m207596200. URL: https://doi.org/10.1074/jbc.m207596200.

Strobl-Mazzulla, Pablo H. and Marianne E. Bronner (Oct. 2012). "Epithelial to mesenchymal transition: New and old insights from the classical neural crest model". In: *Seminars in Cancer Biology* 22.5-6, pp. 411–416. DOI: 10.1016/j.semcancer.2012.04.008. URL: https://doi.org/10.1016/j.semcancer.2012.04.008.

Su, Yan et al. (Feb. 2020). "Increased plasma concentration of cell-free DNA precedes disease recurrence in children with high-risk neuroblastoma". In: *BMC Cancer* 20.1. DOI: 10.1186/s12885-020-6562-8. URL: https://doi.org/10.1186/s12885-020-6562-8.

Subramanian, A. et al. (Sept. 2005). "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles". In: *Proceedings of the National Academy of Sciences* 102.43, pp. 15545–15550. DOI: 10.1073/pnas.0506580102. URL: https://doi.org/10.1073/pnas.0506580102.

Sun, Deqiang et al. (2014). *MOABS: model based analysis of bisulfite sequencing data*.

Szklarczyk, Damian et al. (Nov. 2018). "STRING VII: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets". In: *Nucleic Acids Research* 47.D1, pp. D607–D613. DOI: 10.1093/nar/gky1131. URL: https://doi.org/10.1093/nar/gky1131.

Tan, Ge and Boris Lenhard (2016). *TFBSTools: an R/bioconductor package for transcription factor binding site analysis*.

Teschendorff, Andrew E. et al. (July 2015). "Correlation of Smoking-Associated DNA Methylation Changes in Buccal Cells With DNA Methylation Changes in Epithelial Cancer". In: *JAMA Oncology* 1.4, p. 476. DOI: 10.1001/jamaoncol.2015.1053. URL: https://doi.org/10.1001/jamaoncol.2015.1053.

Teytelman, L. et al. (Oct. 2013). "Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins". In: *Proceedings of the National Academy of Sciences* 110.46, pp. 18602–18607. DOI: 10.1073/pnas.1316064110. URL: https://doi.org/10.1073/pnas.1316064110.

Teytelman, Leonid et al. (Aug. 2009). "Impact of chromatin structures on DNA processing for genomic analyses". en. In: *PLoS One* 4.8, e6700.

Thangue, Nicholas B. La and David J. Kerr (Aug. 2011). "Predictive biomarkers: a paradigm shift towards personalized cancer medicine". In: *Nature Reviews Clinical Oncology* 8.10, pp. 587–596. DOI: 10.1038/nrclinonc.2011.121. URL: https://doi.org/10.1038/nrclinonc.2011.121.

Timp, Winston et al. (Aug. 2014). "Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors". en. In: *Genome Med.* 6.8, p. 61.

Torkov, Alica (2019). "Blinding the CYCLOPS - Neuroblastoma vulnerabilities unveiled by 1p loss". In: DOI: 10.11588/HEIDOK.00025783. URL: http://archiv.ub.uni-heidelberg.de/volltextserver/id/eprint/25783.

Tosti, Luca et al. (Mar. 2018). "Mapping transcription factor occupancy using minimal numbers of cells in vitro and in vivo". In: *Genome Research* 28.4, pp. 592–605. DOI: 10.1101/gr.227124.117. URL: https://doi.org/10.1101/gr.227124.117.

Toyota, M et al. (1999). *CpG island methylator phenotype in colorectal cancer*.

Tran, Trinh T et al. (Dec. 2003). "Specificity and immunochemical properties of anti-DNA antibodies induced in normal mice by immunization with mammalian DNA with a CpG oligonucleotide as adjuvant". In: *Clinical Immunology* 109.3, pp. 278–287. DOI: 10.1016/j.clim.2003.08.012. URL: https://doi.org/10.1016/j.clim.2003.08.012.

Turner, B. et al. (Oct. 2010). "iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence". In: *Database* 2010.0, baq023–baq023. DOI: 10.1093/database/baq023. URL: https://doi.org/10.1093/database/baq023.

Uhlen, Mathias et al. (Sept. 2016). "A proposal for validation of antibodies". In: *Nature Methods* 13.10, pp. 823–827. DOI: 10.1038/nmeth.3995. URL: https://doi.org/10.1038/nmeth.3995.

"UniProt: a worldwide hub of protein knowledge" (Nov. 2018). In: *Nucleic Acids Research* 47.D1, pp. D506–D515. DOI: 10.1093/nar/gky1049. URL: https://doi.org/10.1093/nar/gky1049.

Vaisvila, Romualdas et al. (Dec. 2019). "EM-seq: Detection of DNA Methylation at Single Base Resolution from Picograms of DNA". In: DOI: 10.1101/2019.12.20.884692. URL: https://doi.org/10.1101/2019.12.20.884692.

Valentijn, Linda J et al. (2015). *TERT rearrangements are frequent in neuroblastoma and identify aggressive tumors*.

VanderKraats, Nathan D. et al. (June 2013). "Discovering high-resolution patterns of differential DNA methylation that correlate with gene expression changes". In: *Nucleic Acids Research* 41.14, pp. 6816–6827. DOI: 10.1093/nar/gkt482. URL: https://doi.org/10.1093/nar/gkt482.

Venolia, L and S M Gartler (Mar. 1983). "Comparison of transformation efficiency of human active and inactive X-chromosomal DNA". en. In: *Nature* 302.5903, pp. 82–83.

Visel, Axel et al. (2009). *ChIP-seq accurately predicts tissue-specific activity of enhancers*.

Voynova, E N et al. (July 2005). "Breaking of tolerance to native DNA in nonautoimmune mice by immunization with natural protein/DNA complexes". In: *Lupus* 14.7, pp. 543–550. DOI: 10.1191/0961203305lu2165oa. URL: https://doi.org/10.1191/0961203305lu2165oa.

Waddington, C. H. (Mar. 1956). "Genetic Assimilation of the Bithorax Phenotype". In: *Evolution* 10.1, p. 1. DOI: 10.2307/2406091. URL: https://doi.org/10.2307/2406091.

Wahba, Lamia et al. (June 2016). "S1-DRIP-seq identifies high expression and polyA tracts as major contributors to R-loop formation". In: *Genes & Development* 30.11, pp. 1327–1338. DOI: 10.1101/gad.280834.116. URL: https://doi.org/10.1101/gad.280834.116.

Wahl, Simone et al. (Dec. 2016). "Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity". In: *Nature* 541.7635, pp. 81–86. DOI: 10.1038/nature20784. URL: https://doi.org/10.1038/nature20784.

Walsh, C. P. and T. H. Bestor (Jan. 1999). "Cytosine methylation and mammalian development". In: *Genes & Development* 13.1, pp. 26–34. DOI: 10.1101/gad.13.1.26. URL: https://doi.org/10.1101/gad.13.1.26.

Wang, C et al. (2012). *EZH2 Mediates Epigenetic Silencing of Neuroblastoma Suppressor Genes CASZ1, CLU, RUNX3, and NGFR*.

Wang, Lu et al. (Dec. 2019). "ASCL1 is a MYCN- and LMO1-dependent member of the adrenergic neuroblastoma core regulatory circuitry". en. In: *Nat. Commun.* 10.1, p. 5622.

Wang, Yiqiang, Jing Mi, and Xuetao Cao (July 2000). "Anti-DNA antibodies exhibit different binding motif preferences for single stranded or double stranded DNA". In: *Immunology Letters* 73.1, pp. 29–34. DOI: 10.1016/s0165-2478(00)00194-2. URL: https://doi.org/10.1016/s0165-2478(00)00194-2.

Wardle, Fiona C. and Haihan Tan (July 2015). "A ChIP on the shoulder? Chromatin immunoprecipitation and validation strategies for ChIP antibodies". In: *F1000Research* 4, p. 235. DOI: 10.12688/f1000research.6719.1. URL: https://doi.org/10.12688/f1000research.6719.1.

Waxman, Elisa A. (July 2019). "Bach2 is a potent repressor of Nrf2-mediated antioxidant enzyme expression in dopaminergic neurons". In: DOI: 10.1101/687590. URL: https://doi.org/10.1101/687590.

Weber, Michael et al. (July 2005). "Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells". In: *Nature Genetics* 37.8, pp. 853–862. DOI: 10.1038/ng1598. URL: https://doi.org/10.1038/ng1598.

Wei, Sung-Jen et al. (May 2020). "MYC transcription activation mediated by OCT4 as a mechanism of resistance to 13-cisRA-mediated differentiation in neuroblastoma". In: *Cell Death & Disease* 11.5. DOI: 10.1038/s41419-020-2563-4. URL: https://doi.org/10.1038/s41419-020-2563-4.

Weiss, W A (1997). *Targeted expression of MYCN causes neuroblastoma in transgenic mice*.

Weitzmann, M.N. and N. Savage (July 1994). "Cloning of an antibody binding DNA sequence: pitfalls of DNA/protein immunoprecipitation reactions". In: *Journal of Immunological Methods* 173.1, pp. 7–10. DOI: 10.1016/0022-1759(94)90276-3. URL: https://doi.org/10.1016/0022-1759(94)90276-3.

Wen, Bo et al. (Feb. 2009). "Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells". en. In: *Nat. Genet.* 41.2, pp. 246–250.

White, P. S. et al. (June 1995). "A region of consistent deletion in neuroblastoma maps within human chromosome 1p36.2-36.3." In: *Proceedings of the National Academy of Sciences* 92.12, pp. 5520–5524. DOI: 10.1073/pnas.92.12.5520. URL: https://doi.org/10.1073/pnas.92.12.5520.

Wijetunga, N. Ari et al. (Jan. 2017). "SMITE: an R/Bioconductor package that identifies network modules by integrating genomic and epigenomic information". In: *BMC Bioinformatics* 18.1. DOI: 10.1186/s12859-017-1477-3. URL: https://doi.org/10.1186/s12859-017-1477-3.

Wreczycka, Katarzyna, Vedran Franke, et al. (June 2019). "HOT or not: examining the basis of high-occupancy target regions". en. In: *Nucleic Acids Res.* 47.11, pp. 5735–5745.

Wreczycka, Katarzyna, Alexander Gosdschan, et al. (Nov. 2017). "Strategies for analyzing bisulfite sequencing data". en. In: *J. Biotechnol.* 261, pp. 105–115.

Wun, Hau-Ling et al. (Sept. 2001). "Molecular mimicry: anti-DNA antibodies may arise inadvertently as a response to antibodies generated to microorganisms". In: *International Immunology* 13.9, pp. 1099–1107. DOI: 10.1093/intimm/13.9.1099. URL: https://doi.org/10.1093/intimm/13.9.1099.

Wurmus, Ricardo et al. (Dec. 2018). "PiGx: reproducible genomics analysis pipelines with GNU Guix". en. In: *Gigascience* 7.12.

Xi, Yuanxin and Wei Li (2009). *BSMAP: whole genome bisulfite sequence MAPping program*.

Xie, Dan et al. (2013). *Dynamic trans-Acting Factor Colocalization in Human Cells*.

Xu, Jinrui et al. (Dec. 2020). "To mock or not: a comprehensive comparison of mock IP and DNA input for ChIP-seq". In: *Nucleic Acids Research* 49.3, e17–e17. DOI: 10.1093/nar/gkaa1155. URL: https://doi.org/10.1093/nar/gkaa1155.

Yan, Qingqing et al. (Oct. 2019). "Mapping Native R-Loops Genome-wide Using a Targeted Nuclease Approach". In: *Cell Reports* 29.5, 1369–1380.e5. DOI: 10.1016/j.celrep.2019.09.052. URL: https://doi.org/10.1016/j.celrep.2019.09.052.

Yang, Q (2004). *Association of Epigenetic Inactivation of RASSF1A with Poor Outcome in Human Neuroblastoma*.

Yang, Q et al. (2007). *Methylation of CASP8, DCR2, and HIN-1 in Neuroblastoma Is Associated with Poor Outcome*.

Yang, Xiaojing et al. (Oct. 2014). "Gene Body Methylation Can Alter Gene Expression and Is a Therapeutic Target in Cancer". In: *Cancer Cell* 26.4, pp. 577–590. DOI: 10.1016/j.ccr.2014.07.028. URL: https://doi.org/10.1016/j.ccr.2014.07.028.

Yao, Zizhen (2017). URL: https://www.bioconductor.org/packages//2.13/bioc/html/motifRG.html.

Yasuda, Kei et al. (July 2009). "Requirement for DNA CpG Content in TLR9-Dependent Dendritic Cell Activation Induced by DNA-Containing Immune Complexes". In: *The Journal of Immunology* 183.5, pp. 3109–3117. DOI: 10.4049/jimmunol.0900399. URL: https://doi.org/10.4049/jimmunol.0900399.

Yin, Yimeng et al. (May 2017). "Impact of cytosine methylation on DNA binding specificities of human transcription factors". In: *Science* 356.6337, eaaj2239. DOI: 10.1126/science.aaj2239. URL: https://doi.org/10.1126/science.aaj2239.

Yip, Kevin Y et al. (Sept. 2012). "Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors". en. In: *Genome Biol.* 13.9, R48.

Yu, Kefei et al. (Apr. 2003). "R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells". In: *Nature Immunology* 4.5, pp. 442–451. DOI: 10.1038/ni919. URL: https://doi.org/10.1038/ni919.

Yuan, Tian et al. (Feb. 2015). "An Integrative Multi-scale Analysis of the Dynamic DNA Methylation Landscape in Aging". In: *PLOS Genetics* 11.2. Ed. by John M. Greally, e1004996. DOI: 10.1371/journal.pgen.1004996. URL: https://doi.org/10.1371/journal.pgen.1004996.

Zeid, Rhamy et al. (Apr. 2018). "Enhancer invasion shapes MYCN-dependent transcriptional amplification in neuroblastoma". en. In: *Nat. Genet.* 50.4, pp. 515–523.

Zeineldin, Maged et al. (Feb. 2020). "MYCN amplification and ATRX mutations are incompatible in neuroblastoma". In: *Nature Communications* 11.1. DOI: 10.1038/s41467-020-14682-6. URL: https://doi.org/10.1038/s41467-020-14682-6.

Zeller, Peter et al. (Sept. 2016). "Histone H3K9 methylation is dispensable for Caenorhabditis elegans development but suppresses RNA:DNA hybrid-associated repeat instability". In: *Nature Genetics* 48.11, pp. 1385–1395. DOI: 10.1038/ng.3672. URL: https://doi.org/10.1038/ng.3672.

Zemach, A et al. (2010). *Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation*.

Zhang, Shihua et al. (Aug. 2012). "Discovery of multi-dimensional modules by integrative analysis of cancer genomic data". In: *Nucleic Acids Research* 40.19, pp. 9379–9391. DOI: 10.1093/nar/gks725. URL: https://doi.org/10.1093/nar/gks725.

Zhang, Wei et al. (Sept. 2010). "Specific cross-reaction of anti-dsDNA antibody with platelet integrin GPIIIa49-66". In: *Autoimmunity* 43.8, pp. 682–689. DOI: 10.3109/08916934.2010.506207. URL: https://doi.org/10.3109/08916934.2010.506207.

Zheng, Christina L et al. (Nov. 2014). "Transcription restores DNA repair to heterochromatin, determining regional mutation rates in cancer genomes". en. In: *Cell Rep.* 9.4, pp. 1228–1234.

Zhong, Shan, Xin He, and Ziv Bar-Joseph (Nov. 2013). "Predicting tissue specific transcription factor binding sites". en. In: *BMC Genomics* 14, p. 796.

Zhou, Vicky W, Alon Goren, and Bradley E Bernstein (Jan. 2011). "Charting histone modifications and the functional organization of mammalian genomes". en. In: *Nat. Rev. Genet.* 12.1, pp. 7–18.

Zhou, Wanding et al. (Apr. 2018). "DNA methylation loss in late-replicating domains is linked to mitotic cell division". en. In: *Nat. Genet.* 50.4, pp. 591–602.

Zhou, Wenlai et al. (Jan. 2008). "Histone H2A monoubiquitination represses transcription by inhibiting RNA polymerase II transcriptional elongation". en. In: *Mol. Cell* 29.1, pp. 69–80.

Zhou, Yingyao et al. (Apr. 2019). "Metascape provides a biologist-oriented resource for the analysis of systems-level datasets". In: *Nature Communications* 10.1. DOI: 10.1038/s41467-019-09234-6. URL: https://doi.org/10.1038/s41467-019-09234-6.

Zhu, Yanfen et al. (Apr. 2021). "Oncogenic extrachromosomal DNA functions as mobile enhancers to globally amplify chromosomal transcription". In: *Cancer Cell*. DOI: 10.1016/j.ccell.2021.03.006. URL: https://doi.org/10.1016/j.ccell.2021.03.006.

Ziller, Michael J. et al. (Aug. 2013). "Charting a dynamic DNA methylation landscape of the human genome". In: *Nature* 500.7463, pp. 477–481. DOI: 10.1038/nature12433. URL: https://doi.org/10.1038/nature12433.

Zink, Daniele, Andrew H. Fischer, and Jeffrey A. Nickerson (Sept. 2004). "Nuclear structure in cancer cells". In: *Nature Reviews Cancer* 4.9, pp. 677–687. DOI: 10.1038/nrc1430. URL: https://doi.org/10.1038/nrc1430.

Zou, Hui and Trevor Hastie (Apr. 2005). "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2, pp. 301–320. DOI: 10.1111/j.1467-9868.2005.00503.x. URL: https://doi.org/10.1111/j.1467-9868.2005.00503.x.