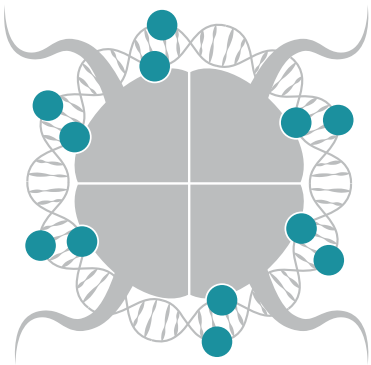
Analyzing DNA Methylation Signatures of Cell Identity

Dissertation

zur Erlangung des Grades
des Doktors der Naturwissenschaften
der Fakultät für Mathematik und Informatik
der Universität des Saarlandes

von
Fabian Müller



Saarbrücken, December 2016

Tag des Kolloquiums:	31. Mai 2017
Dekan:	Prof. Dr. Frank-Olaf Schreyer
Prüfungsausschuss:	
Vorsitzender:	Prof. Dr. Hans-Peter Lenhof
Berichterstatter:	Prof. Dr. Dr. Thomas Lengauer Dr. Christoph Bock Prof. Dr. Benedikt Brors
Akademischer Mitarbeiter:	Dr. Markus List

§1: *Et es wie et es.*
§2: *Et kütt wie et kütt.*
§3: *Et hät noch immer jot jejange.*
§4: *Wat fott es, es fott.*
§5: *Et bliev nix, wie et wor.*
§6: *Kenne mer nit, bruche mer nit, fott domet.*
§7: *Wat wellste maache?*
§8: *Maach et jot, ävver nit ze of.*
§9: *Wat sull dä Quatsch?*
§10: *Dringste eine met?*
§11: *Do laachs dich kapott.*

– Et kölsche Jrundjesetz,
which coincidentally spotlights characteristics of many regulatory processes in biology:

§1: observational evidence
§2: stochasticity
§3: evolution
§4: drift
§5: dynamics
§6: selection
§7: determinism
§8: efficiency
§9: complexity
§10: cooperativity
§11: ingenuity

Abstract

Although virtually all cells in an organism share the same genome, regulatory mechanisms give rise to hundreds of different, highly specialized cell types. Understanding these mechanisms has been in the limelight of epigenomic research. It is now evident that cellular identity is inscribed in the epigenome of each individual cell. Nonetheless, the precise mechanisms by which different epigenomic marks are involved in regulating gene expression are just beginning to be unraveled. Furthermore, epigenomic patterns are highly dynamic and subject to environmental influences. Any given cell type is defined by cell populations exhibiting epigenetic heterogeneity at different levels. Characterizing this heterogeneity is paramount in understanding the regulatory role of the epigenome.

Different epigenomic marks can be profiled using high-throughput sequencing, and global initiatives have started to provide a comprehensive picture of the human epigenome by assaying a multitude of marks across a broad panel of cell types and conditions. In particular, DNA methylation has been extensively studied for its gene-regulatory role in health and disease.

This thesis describes computational methods and pipelines for the analysis of DNA methylation data. It provides concepts for addressing bioinformatic challenges such as the processing of large, epigenome-wide datasets and integrating multiple levels of information in an interpretable manner. We developed RnBeads, an R package that facilitates comprehensive, interpretable analysis of large-scale DNA methylation datasets at the level of single CpGs or genomic regions of interest. With the epiRepeatR pipeline, we introduced additional tools for studying global patterns of epigenomic marks in transposons and other repetitive regions of the genome.

Blood-cell differentiation represents a useful model for studying trajectories of cellular differentiation. We developed and applied bioinformatic methods to dissect the DNA methylation landscape of the hematopoietic system. Here, we provide a broad outline of cell-type-specific DNA methylation signatures and phenotypic diversity reflected in the epigenomes of human mature blood cells. We also describe the DNA methylation dynamics in the process of immune memory formation in T helper cells. Moreover, we portrayed epigenetic fingerprints of defined progenitor cell types and derived computational models that were capable of accurately inferring cell identity. We used these models in order to characterize heterogeneity in progenitor cell populations, to identify DNA methylation signatures of hematopoietic differentiation and to infer the epigenomic similarities of blood cell types.

Finally, by interpreting DNA methylation patterns in leukemia and derived pluripotent cells, we started to discern how epigenomic patterns are altered in disease and explored how reprogramming of these patterns could potentially be used to restore a non-malignant state.

In summary, this work showcases novel methods and computational tools for the identification and interpretation of epigenetic signatures of cell identity. It provides a detailed view on the epigenomic landscape spanned by DNA methylation patterns in hematopoietic cells that enhances our understanding of epigenetic regulation in cell differentiation and disease.

Kurzfassung

Obwohl praktisch alle Zellen eines Organismus dieselbe Genomsequenz besitzen, führen diverse regulatorische Mechanismen dazu, dass sich hunderte verschiedene, hochspezialisierte Zelltypen entwickeln können. Diese Mechanismen zu verstehen ist Kernziel der Epigenomforschung. Das Epigenom einer Zelle spiegelt ihren Phänotyp und somit ihre Identität wider. Die genauen Mechanismen, die durch die verschiedenen epigenomischen Merkmale einer Zelle gesteuert werden, sind jedoch bisher weitestgehend unbekannt. Außerdem sind die zugrundeliegenden epigenomischen Muster dynamisch und können sich abhängig von ihrer Umgebung verändern. Verschiedene Zelltypen bestehen zudem aus heterogenen Populationen einzelner Zellen. Um die regulatorische Rolle des Epigenoms zu verstehen ist daher eine genaue Charakterisierung dieser Heterogenität unabdingbar.

Epigenomische Profile können mithilfe moderner Technologien, wie etwa der Hochdurchsatzsequenzierung der DNS, erzeugt werden. Globale Initiativen untersuchen inzwischen eine Vielzahl epigenomischer Modifikationen in einer breiten Spanne verschiedener Zelltypen und legen damit den Grundstein für ein umfassendes Bild des dynamischen humanen Epigenoms. Insbesondere stellt die Methylierung der DNS, welche mit der Genregulation in gesunden sowie erkrankten Zellen assoziiert ist, eines der am besten beschriebenen epigenomischen Merkmale dar.

Diese Arbeit beschreibt computergestützte Verfahren und Software-Pipelines für die Analyse von DNS-Methylierungsdaten. Herangehensweisen für bioinformatische Herausforderungen, wie etwa dem Umgang mit großen, heterogenen, epigenomweiten Datensätzen und der Datenintegration aus verschiedenen Informationsebenen, werden vorgestellt. Unsere RnBeads Software ermöglicht eine umfassende Analyse großer DNS-Methylierungsdatensätze auf Basis einzelner CpGs oder genomischer Regionen und stellt die Resultate in interpretierbarer Form dar. Des Weiteren stellt die epiRepeatR Pipeline Werkzeuge für die Untersuchung globaler epigenomischer Muster in Transposons und anderen repetitiven Abschnitten des Genoms bereit.

Das blutbildende System stellt ein nützliches Modell für die Beschreibung und Erforschung von Zelldifferenzierungsprozessen dar. Hier beschreiben wir die epigenomische Landschaft, die durch DNS-Methylierungsmuster in hämatopoetischen Zellen aufgespannt wird. Bioinformatische Methoden zur Analyse epigenomischer Muster in den Differenzierungsprozessen wurden erarbeitet und angewandt. Mithilfe dieser Methoden wurden zelltypspezifische Methylierungsprofile ausdifferenzierter Blutzellen identifiziert, welche die phänotypische Diversität der Zellen widerspiegeln. In einer vertiefenden Analyse wurde die Methylierungsdynamik während der Ausbildung des Immungedächtnisses in menschlichen T-Zellen offengelegt. Darüber hinaus konnten epigenetische Fingerabdrücke von Blutvorläuferzellen identifiziert und statistische Verfahren entwickelt werden, mit deren Hilfe Zellidentität abgeleitet werden kann. Diese Verfahren ermöglichen die Charakterisierung von Zellheterogenität in Populationen von Vorläuferzellen, die Herausstellung von Methylierungssignaturen der Zelldifferenzierung und die Quantifizierung der epigenomischen Ähnlichkeit zwischen Zelltypen.

Schließlich beschäftigt sich diese Arbeit mit der Beschreibung epigenomischer Muster, die in Krebszellen abnormal verändert sind und die sich durch Zellreprogrammierung einen pluripotenten, potenziell gutartigen Zustand zurückversetzen lassen.

Zu diesem Zweck wurden die Methylierungsprofile leukämischer Zellen und deren reprogrammierter Gegenstücken mit den entwickelten bioinformatischen Methoden ausgewertet.

Zusammenfassend beschreibt diese Dissertation neuartige Methoden und Softwarewerkzeuge zur Identifizierung und Interpretation epigenetischer Signaturen der Zellidentität. Sie zeichnet ein Bild der DNS-Methylierungslandschaft in menschlichen Blutzellen, welches zum Verständnis von epigenetischen Regulationsprozessen während Zelldifferenzierung und Krankheitsbildung beitragen kann.

Acknowledgments

There are many people from whose support this thesis has benefited greatly and who deserve special acknowledgment here.

First and foremost, I would like to express my utmost gratitude and admiration to Thomas Lengauer for his invaluable support and mentorship, and for providing me with all the liberty to pursue my projects. Thomas, you are a role model for scientific curiosity, enthusiasm, management, professionalism, fairness and humaneness. Furthermore, I am deeply indebted to Christoph Bock who kindled my interest in computational epigenetics. Thanks a million for your extraordinary guidance from near and far, your trust, impetus and motivating challenges. I would like to extend my special thanks to Benedikt Brors who agreed to review this thesis on short notice.

Jörn Walter has been an inexhaustible source of inspiration and advice for me. Thanks for the illuminating discussions, excursions into molecular biology and pushes into the right direction. I am also deeply grateful to Alex Meissner whose lab I had the privilege of joining for a research fellowship during the initial phase of my PhD. The time spent there has been highly instructive and formative. Thanks for providing such a rich and stimulating environment, fruitful discussions and advice.

I am immensely thankful for all collaboration partners and coauthors who have proven to me that extraordinary people can accomplish extraordinary research, when working together. I particularly thank Michael Ziller for smooth team work, council and friendship. Furthermore, it has been both a pleasure and an honor to work together with Pavlo Lutsik and my long-term office mate Yassen Assenov. It was never boring. Giovanni, thanks for the exciting project and putting up with the naive questions of a bioinformatician. Life is tough! Kudos to Matthias Farlik and Florian Halbritter for the immense productivity, their never-ending enthusiasm and a fantastic working chemistry. I also extend my gratitude to all members of DEEP and BLUEPRINT. Thanks for making my involvement in this amazing community such an enjoyable experience. Furthermore, I had the fortune of supervising exceptionally talented Bachelor and Master students. Michael, Nora and Alex, thanks for your great work and teaching me so much about teaching.

My colleagues and friends at MPII and Saarland University provided the most wonderful working environment. I am especially grateful to the past and present members of our computational epigenetics family: Michael, Markus, Marcel, Felipe, Peter, Karl, Pavlo, Yassen, Konstantin, Lars and Christoph. Countless pleasant and insightful conversations often made my days. In this context, I also would like to mention Daniel, Basti and Sven as well as the most noble order of the coffee table, Michael, Lisa, Markus, Peter, Matthias and Glenn. I would like to acknowledge all the members of the Meissner and Walter labs for teaching me so much about biology and enlightening me about the (harsh) reality of the bench world. Thanks also to Hans-Peter Lenhof for the constant guidance and encouragement, placing me on the path of bioinformatics.

Markus, Daniel, Alejandro and Peter proofread this thesis and I greatly appreciate all their helpful remarks and pointers. I would also like to thank Reviewer 3 for always challenging our research when providing feedback on our manuscripts. This work has benefited greatly from his/her insightful comments.

I highly value all the technical help from Achim Büch, Georg Friedrich and the MPII-IST as well as all the administrative support from Ruth Schneppen-Christmann, the true good soul of the group.

Finally, I owe deep gratitude to my friends and family, particularly my parents, my sister and grandparents. This work would not have been possible without their unconditional love and support.

Large parts of this work have been conducted within the context of the DEEP (BMBF Grant 01KU1216A) and BLUEPRINT (EU Grant HEALTH-F5-2011-282510) projects. Financial support has also been provided by basic funding of the Max Planck Society. During my research fellowship at Alexander Meissner's lab, I was supported by an Ambassadorial Scholarship of Rotary International and by Harvard University.

Contents

List of Figures

List of Tables

1 Introduction

A single fertilized oocyte can give rise to an entire human organism consisting of approximately 30 to 40 trillion cells which differ greatly in their morphology and behavior [Sender *et al.* 2016; Bianconi *et al.* 2013]. Embryogenesis constitutes a prime example for the differentiation processes that lead to this diversity. During this process individual tissues develop from three germ layers: neural tissues such as brain and spinal cord develop from the ectoderm, the endoderm gives rise to organs such as liver and kidney, and blood and muscle are derived from the mesoderm [Gilbert 2014]. It is therefore useful to categorize cells based on tissue localization, developmental state and function. The number of distinct cell types in healthy human tissues is projected to be around 200 [Alberts *et al.* 2008]. Additionally, other factors, such as disease and environmental influences, determine the state of a cell and cells of the same lineage can exhibit considerable heterogeneity. The term "cell type" is therefore typically defined in a context-specific manner rather than universally. Importantly, despite their differences in appearance, virtually all cells of an individual share the same genomic DNA sequence. How this genome is differentially packaged and interpreted is largely determined by the epigenome.

Goldberg *et al.* [2007] define epigenetics as "the study of any potentially stable and, ideally, heritable change in gene expression or cellular phenotype that occurs without changes in Watson-Crick base-pairing of DNA". Historically, the term has been coined by Conrad Waddington (1905-1975) [Waddington 1942]. He also proposed the metaphor of the "epigenetic landscape" [Waddington 1957]. In this concept, a differentiating cell traverses a landscape spanned by molecular factors (Figure 1.1). High grounds in this landscape represent pluripotent cell states. As cells differentiate, they follow downhill paths through several valleys in the landscape. These paths are interjected by numerous bifurcations which symbolize commitment to specific cell lineages. The topology of the epigenetic landscape is determined by molecular factors which regulate gene expression on many different levels. These factors are highly dynamic as cells change their state and position in the landscape. Significant progress has been made in unraveling their role of in gene regulation (cf. Section 2.1). Particularly, the packaging of the DNA and its modification play an essential role in providing or denying access for the cellular machinery reading the molecular blueprint inscribed in the DNA. This thesis focuses on analyzing and interpreting patterns of DNA methylation, which represents one of the most widely studied epigenomic features involved in gene regulation.

It has been shown that cells can be manipulated to change the position in the landscape corresponding to their current cell state: mammalian cells can be reprogrammed to a pluripotent cell state by means of nuclear transfer, cell fusion or the introduction of exogenous transcription factors (cf. Yamanaka and Blau [2010]). In Waddington's epigenetic landscape this corresponds to a change in position, moving from a valley to high ground. In particular, Induced Pluripotent Stem Cells (iPSCs), which can be

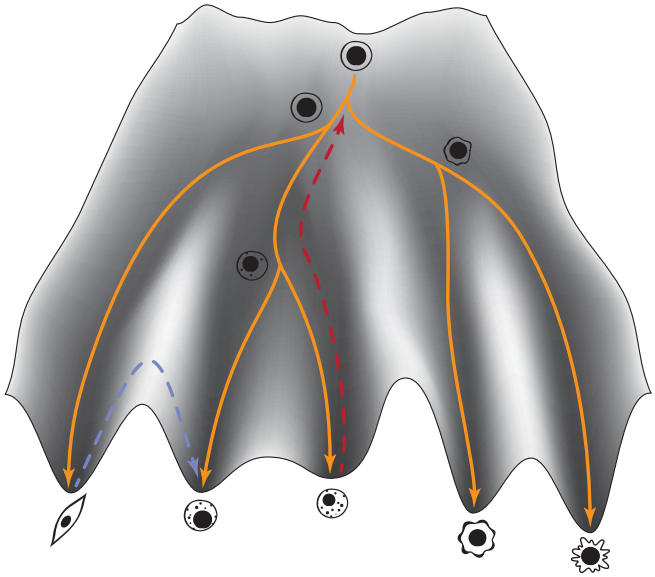


Figure 1.1: The epigenetic landscape of cell differentiation. Adapted from [Waddington 1957]. Differentiation paths are indicated by orange arrows. The dashed red arrow depicts induction of pluripotency via reprogramming, the dashed purple arrow transdifferentiation from one cell type to another.

obtained via overexpression of a defined set of transcription factors [Takahashi and Yamanaka 2006; Takahashi *et al.* 2007], have become a popular model for pluripotency and an important tool in molecular biology. Their expression and DNA methylation patterns have been extensively characterized and reprogrammed cells resemble the state of Embryonic Stem Cells (ESCs) [Mikkelsen *et al.* 2008; Guenther *et al.* 2010]. However, some traces of "epigenetic memory" persist, which manifest themselves in epigenetic and transcriptional patterns that resemble the cells of origin [Chin *et al.* 2009; Doi *et al.* 2009; Kim *et al.* 2010; Bock *et al.* 2011]. Furthermore, iPSCs bear great promise for regenerative medicine, provided that their carcinogenic potential can be overcome and optimized differentiation protocols become available. Cells of one type of tissue can also be stimulated to transdifferentiate to cells of another type in natural and artificial contexts. For instance, murine fibroblasts have successfully been transdifferentiated into neurons [Vierbuchen *et al.* 2010] and cardiomyocytes [Qian *et al.* 2012; Song *et al.* 2012].

Inheritance of epigenetic regulation has been observed mitotically as well as meiotically. While it is evident that epigenetic patterns can be stably maintained across cell divisions, transgenerational epigenetic inheritance is less well characterized in mammals. Most patterns are reset during germ cell development as well as at the stage of embryonic pre-implantation and they are only reestablished later during embryonic development [Heard and Martienssen 2014]. How the epigenetic signal can be transmitted across generations is therefore currently an open question in the scientific community. Nonetheless, studies have described the effect of environmental influences on descendant generations in mammals in the contexts of metabolism, longevity, response to stress as well as predisposition for diseases like diabetes and imprinting disorders [Wei

et al. 2015; Heard and Martienssen 2014; Grossniklaus *et al.* 2013]. For instance, individuals prenatally exposed to malnutrition during the Dutch famine of 1944–1945 exhibited lower birth weights and an increased risk of diabetes and cardiovascular disease than unexposed individuals and these effects could potentially be propagated to their offspring [Painter *et al.* 2008; Veenendaal *et al.* 2013]. Studies investigating the association of food abundance and mortality due to diabetes and cardiovascular disease in the Swedish population of Överkalix provide further evidence for sex-specific, epigenetic inheritance that is transmitted through the male germline [Kaati *et al.* 2002; Pembrey *et al.* 2006]. However, it is important to note that evidence for effects that span more than two generations and thus would be indicative of germ-cell-based inheritance has been scarce so far [Heard and Martienssen 2014; Grossniklaus *et al.* 2013]. The molecular background of transgenerational epigenetic inheritance is just beginning to be investigated. For instance, DNA methylation patterns associated with prenatal malnutrition have been identified in the Dutch famine cohort [Heijmans *et al.* 2008; Tobi *et al.* 2014]. Molecular mechanisms likely to play a role include genomic imprinting and regulation of repetitive elements in the genome [Grossniklaus *et al.* 2013].

Thesis Scope and Outline

This thesis spotlights epigenomic[1] patterns that are characteristic of cell identity. The hematopoietic system has proven particularly suitable and relevant for studying these patterns, since it comprises a large variety of cell types and corresponding epigenomes. It thus provides a widely studied framework for cell differentiation, that is also employed by many research initiatives. This work focuses on the analysis of DNA methylation profiles (methylomes) in various blood-related cells. We explore how DNA methylation signatures can be used to infer relations between the epigenomes of different cell populations. To this end, we developed, implemented and applied computational methods for the detailed analysis of DNA methylation data. Software tools for the comprehensive characterization of methylomes in relation to one another are an integral part of this work. Our models shed light on the DNA methylation dynamics in differentiating cells and in disease and contribute to our understanding of the factors responsible for the topography of the epigenetic landscape. The remainder of this thesis is structured as follows:

Chapter 2 establishes a background of regulatory concepts and epigenetic mechanisms involved in health and disease. It provides a general overview of the different levels of epigenetic gene regulation and outlines technology used to generate epigenome maps. It also introduces large, global efforts in mapping more than a thousand epigenomes. Because the algorithms and statistical methods described in this thesis are highly diverse, this work does not contain a dedicated background chapter on computational methods. Instead, these methods are introduced in the respective sections in later chapters.

Chapter 3 presents computational tools and pipelines that we developed for data preprocessing and high-level analysis of genome-wide DNA methylation data: a pipeline for quantifying methylation levels from reads obtained from bisulfite sequencing experiments is described in Section 3.1. Section 3.2 introduces RnBeads, a software package

[1] In this thesis, the term "epigenetic" describes (sets of) individual or specific characteristics, changes, mechanisms or processes while "epigenomic" has a more global notion and refers to the collective of all epigenetic events in a given entity (such as a single cell, cell type or organism).

for the comprehensive, start-to-finish analysis of genome-wide DNA methylation data at the resolution of single CpGs and genomic regions. Furthermore, we developed the epiRepeatR pipeline for quantifying DNA methylation and other epigenomic marks in subfamilies of repetitive DNA and applied it to characterize epigenetic regulation of repetitive elements in human blood cells (Section 3.3).

Chapter 4 dissects the *in vivo* DNA methylation dynamics during human hematopoietic differentiation, employing the methods described in Chapter 3. Section 4.1 provides biological background on the human blood system and on its epigenetic regulation. In order to obtain a global characterization of the epigenomic landscape involved in hematopoiesis, we analyzed the methylomes of a large panel of differentiated blood cell types and delineated lineage specific patterns (Section 4.2). Section 4.3 focuses on the part of the hematopoietic hierarchy that pertains to T helper cells and describes DNA methylation changes indicative of a linear progression during T cell memory formation. Section 4.4 constitutes the main part of Chapter 4 and is devoted to a detailed account of lineage-specific methylation signatures in hematopoietic stem cells and early progenitors of blood. Using genome-wide data from low-input DNA methylation profiling, we derive statistical models for inferring lineage propensities. Application of these models facilitate insights into within-cell-type heterogeneity and data-driven lineage reconstruction.

In order to provide a perspective on hematopoietic diseases, Chapter 5 focuses on *in vitro* changes in DNA methylation associated with inducing a pluripotent cell state in human leukemia cells by means of cellular reprogramming. We show that reprogrammed leukemia cells exhibit reduced oncogenic signatures and resemble a epigenomic cell state reminiscent of ESCs.

Finally, Chapter 6 summarizes this work and puts it into the perspective of a model-driven interpretation of Waddington's epigenetic landscape.

2

Biological Background

This chapter introduces biological aspects of epigenetic regulation in health and disease. Section 2.1 provides a general introduction to epigenetic regulatory mechanisms and Section 2.2 outlines how epigenetic patterns are perturbed in human disease. Methods for epigenome profiling are presented in Section 2.3 and Section 2.4. Section 2.5 concludes this chapter with an introduction of national and international epigenome mapping consortia that employ these methods in order to chart the epigenomes of hundreds of cell types, tissues and individuals.

2.1 The Regulatory Role of the Epigenome

In the metaphase of the cell cycle, the $2 \times 3.2 \times 10^9$ basepairs (bp) of human DNA[1], which, if expanded, would correspond to a string of 2 meters in length, are compacted by a factor of approximately 10,000 to fit into the nucleus of a cell [Allis *et al.* 2007]. At any given point in time, only parts of the genome are accessible to the transcription machinery. The dynamic and specific selection of these parts requires an efficient indexing strategy in which the epigenome plays a central role. Epigenetic cues are involved in signaling the opening and compaction of the DNA's scaffold as well as in the recruitment of the cellular machinery responsible for reading the genome. Protein complexes which sometimes consist of more than 100 units are required to act in in a highly regulated fashion in order to facilitate the transcription of DNA to Messenger RNA (mRNA). It is therefore not surprising that a large fraction of the genome is comprised of regulatory elements whose purpose is to orchestrate the processes of transcriptional activation and repression: the ENCODE consortium assigned functional and regulatory roles to 80.4 % of the human genome [ENCODE Project Consortium 2012]. In contrast, it is estimated that less than 2 % of the mammalian genome are protein-coding.

2.1.1 Chromatin Organization

The entirety of DNA and associated proteins is called chromatin. Chromatin is organized in hierarchical structures (Figure 2.1). The nucleosome constitutes the lowest level in this hierarchy. It comprises 147 bp of DNA wrapped around a core particle in 1.7 turns [Luger *et al.* 1997], fixated by hydrogen bonds, hydrophobic interactions and salt bridges (Figure 2.2). The core particle is an octamer consisting of two units of each of the histone proteins H3, H4, H2A and H2B. Nucleosomes interlocked by the linker histone H1 arrange in a "beads-on-a-string" structure which is further compacted to fibers at a rate of 50 fold or more [Bell *et al.* 2011] (Figure 2.1). The fibers are organized into chromosomal domains involving loop structures. These Topological Associated Domains

[1] The factor of 2 corresponds to the diploid nature of the mammalian genome

(TADs) have recently been characterized by charting genome-wide chromatin interactions (reviewed in [Sexton and Cavalli 2015]) and have been shown to occupy distinct regions in the nucleosome. On an even coarser level, the chromosome is composed of tissue-specific compartments of two types as determined by Eigenvalue analysis of chromatin interaction maps [Lieberman-Aiden *et al.* 2009; Dekker *et al.* 2013]. Historically, the term "chromatin" refers to the staining properties of DNA. Highly compacted DNA appears darker in staining experiments and is referred to as heterochromatin. In contrast, euchromatin designates open, accessible chromatin and is associated with lighter staining. Telomeric and centromeric chromosomal regions are generally highly heterochromatic across cell types. Gene-poor regions in the periphery of the nucleus, so-called Lamina-Associated Domains (LADs), are also associated with a closed chromatin signature. In other genomic regions, DNA accessibility is highly dynamic and varies in time and between cell types. Different epigenomic marks, such as DNA methylation and histone modifications, are associated with accessible and compacted chromatin (Figure 2.3). The following sections outline these associations. Chromatin remodeling complexes belonging to the ISWI and SWI/SNF families can rearrange nucleosomes along the DNA in an ATP dependent fashion and thus facilitate the transition from euchromatin to heterochromatin and vice versa [Allis *et al.* 2007]. Additionally, histone octamers can be incorporated or removed entirely or the core histone proteins may be exchanged by histone variants which affect nucleosome remodeling and chromatin accessibility.

2.1.2 Histone Modifications

The N-terminal domains of the histone proteins are structurally disordered and are therefore called "tails" (Figure 2.2). They are subject to a plethora of post-translational modifications (reviewed in [Kouzarides 2007]). Different histone modifications are associated with accessible and compacted chromatin (see Figure 2.3). Various enzymes specifically catalyze the dynamic deposition and removal of these marks. Modifications to entire chromatin regions can be placed and removed within minutes in response to certain cell stimuli.

Acetyl groups can be attached and detached to lysine residues by Histone Acetyltransferases (HATs) and Histone Deacetylases (HDACs) respectively. Acetylation leads to a reduction in the positive charge of the histone and thus to a weakened contact with the negatively charged DNA backbone. Therefore, it is generally associated with open chromatin. Furthermore, arginines and lysines in the histone tails are subject to methylation, which is introduced by Histone Methyltransferases (HMTs) and is removed by Histone Demethylases (HDMs). Here, one, two or three methyl-groups can be covalently bound to the same residue and the corresponding states are referred to as mono-, di- and trimethylation respectively. Lysine methylation is one of the most widely-studied epigenomic marks and is associated with different regulatory roles depending on the residue's position and degree of methylation. For instance, trimethylation of the lysine (K) residues located at the fourth position from the N-terminal tail of histone H3 (termed H3K4me3 according to nomenclature) is generally associated with initiation of gene transcription, and occurs in the vicinity of Transcription Start Sites (TSSs) [Mikkelsen *et al.* 2007]. On the other hand, trimethylation of H3 histone proteins at the lysine 27 residue (H3K27me3; a modification that is catalyzed by the Polycomb repressive complex 2 (PRC2) involving Polycomb Group (PcG) protein units) across a gene region is generally associated with repressed transcription [Mikkelsen *et al.* 2007]. Interestingly,

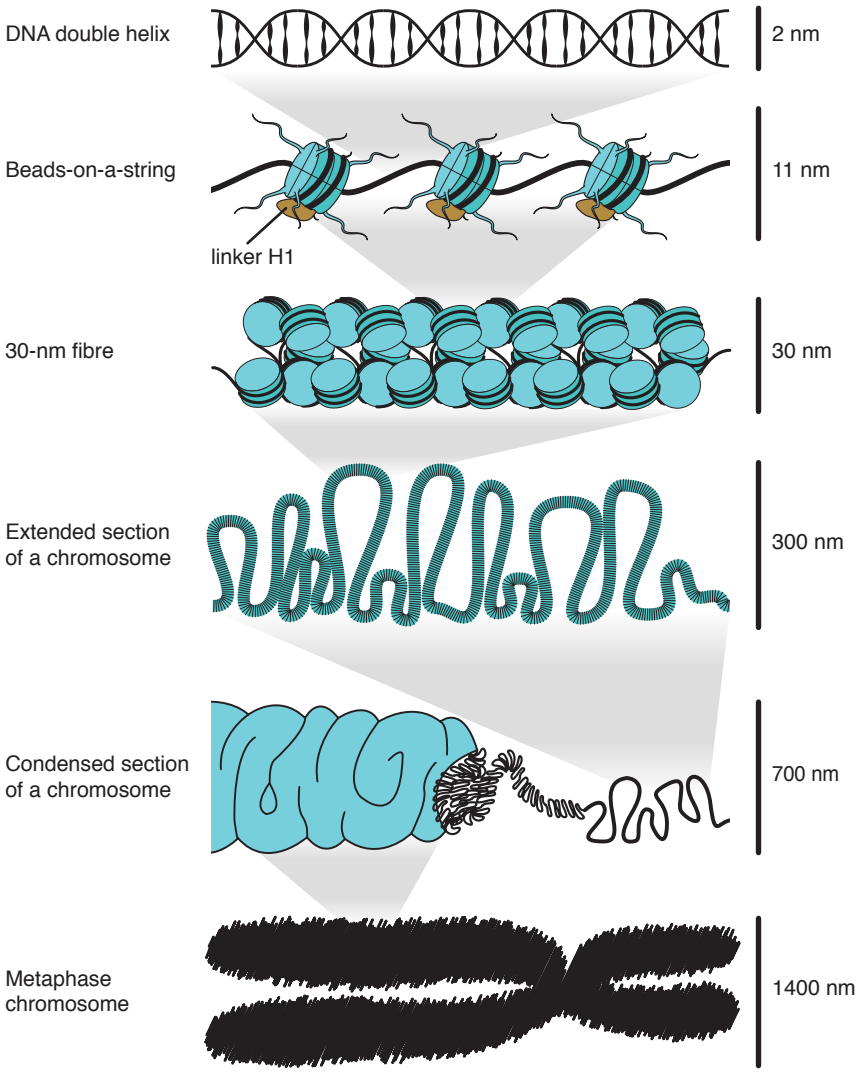


Figure 2.1: Levels of chromatin organization. Adapted from [Alberts *et al.* 2008].

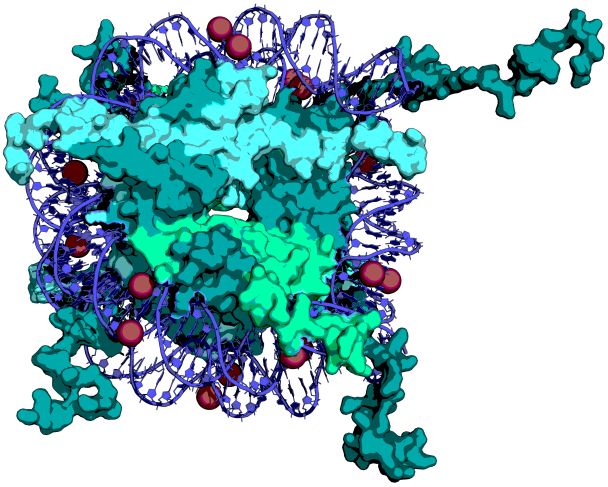


Figure 2.2: Crystal structure of the nucleosome. 147 basepairs of DNA (blue) wrap around a core octamer. Different histone proteins are colored in different shades of green and turquoise. DNA methylation is schematically indicated by magenta spheres. The image was rendered from Protein Data Bank structure `1kx5`.

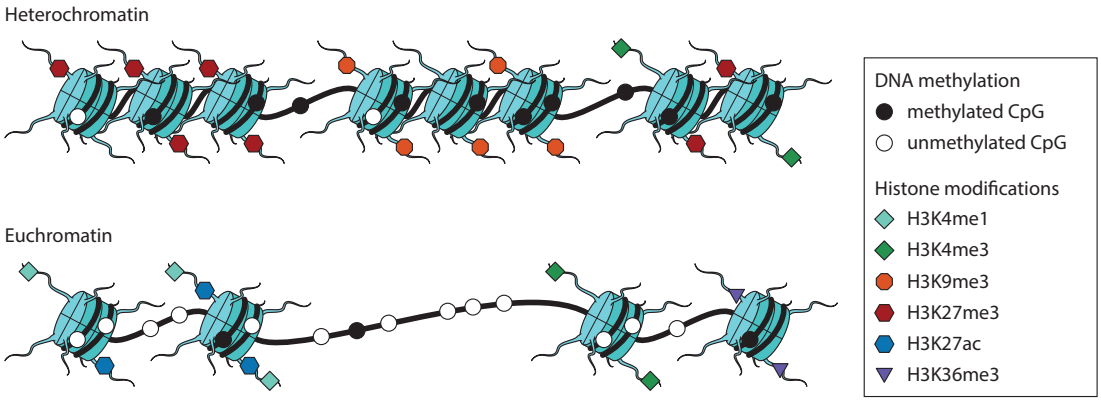


Figure 2.3: Epigenomic marks associated with euchromatin and heterochromatin.

in Embryonic Stem Cells (ESCs) both marks can co-occur within the same genomic region [Bernstein *et al.* 2006]. These bivalent domains frequently associate with with the promoters of transcription factors which are involved in embryonic development and which are typically expressed at low levels in ESCs. They represent a poised chromatin state subject to rapid activation or inactivation upon differentiation. Further examples of informative histone modifications include H3K36me3, which is attributed with a role in transcriptional elongation, and H3K9me3, which is thought to be involved in stable gene silencing and the formation of pericentric heterochromatin [Kouzarides 2007]. Modifications such as H3K9me3 and H3K27me3 occur across large domains in the genome and are thus referred to as "broad" histone marks whereas other modifications, including H3K4me3 exhibit a more localized or "narrow" distribution.

With the exception of acetylation, histone modifications generally do not influence chromatin accessibility directly through physiochemical properties, but various regulatory roles have been ascribed to certain histone modifications that facilitate the recruitment of non-histone protein complexes to chromatin. These complexes incorporate domains responsible for reading designated chromatin patterns. For instance, protein domains termed chromo, tudor and PHD are capable of recognizing different degrees of lysine methylation with varying specificity. Lysine acetylation is typically read by proteins carrying a bromodomain. The plethora of different marks in different contexts suggests that the simple black-and-white picture of histone marks with activating or repressing effects requires further refinement. The dynamic interplay between modifications and their readers and writers leads to the concept of a "histone code" [Jenuwein and Allis 2001] involved in transcriptional regulation. Signatures of co-occurring histone modifications have been summarized into "chromatin states" and have been associated with regulatory elements (cf. Section 2.1.5). They can be identified on a whole-genome scale using computational approaches implementing Hidden Markov Models (HMMs) [Ernst and Kellis 2010; Ernst *et al.* 2011; Hoffman *et al.* 2012; Mammana and H.-R. Chung 2015]. Recently, the application of statistical learning techniques to predict the presence of epigenomic marks from DNA sequence motifs indicates a strong interrelation of histone modifications and DNA methylation with the underlying sequence [Whitaker *et al.* 2015].

2.1.3 DNA Methylation

In many eukaryotes, cytosine bases in the DNA are frequently modified by the addition of a methyl groups to the carbon 5 atoms (5mC). In 1975, a regulatory role of this modification as a heritable, epigenetic mark has been proposed [Holliday and Pugh 1975; Riggs 1975]. In mammals, methylation of cytosines occurs predominantly in the context of CpG dinucleotides [Bird 2002; Jones 2012]. CpG methylation is typically symmetric, i.e. both cytosines on the two complementary strands are methylated. While non-CpG methylation is common in plants, it only occurs at low levels in certain mammalian cell types such as ESCs and neuronal cells [Lister *et al.* 2009; Lister *et al.* 2013; Ziller *et al.* 2011]. Methylated cytosines exhibit increased deamination rates compared to unmethylated bases and are thus frequently mutated. Therefore, CpG dinucleotides have been progressively lost during the course of evolution and are depleted relative to other sequence contexts. For instance, the human genome contains only approximately 28 million CpGs compared to roughly 400 million CpAs [Lander *et al.* 2001; Venter *et al.* 2001]. Mammalian DNA also contains stretches of DNA with a particularly high

frequency of CpG dinucleotides and with a low abundance of DNA methylation. Approximately 29,000 of these CpG Islands (CGIs) are present in the human genome [Bird 1986; Gardiner-Garden and Frommer 1987]. They span regions ranging in size from 200 bp up to a few kilobases and contain highly conserved sequences. They also coincide with the promoters of about half of all human genes. In contrast to these canonically unmethylated regions, the vast majority of CpGs in the mammalian genome (70 % to 80 %) are methylated [Bird 1986; Schübeler 2015]. DNA sequence poses an important determinant of DNA methylation of surrounding genomic regions, as has been shown by statistical inference [Whitaker *et al.* 2015; Bock *et al.* 2006] as well as by experimental assays [Lienert *et al.* 2011].

Employing a gene-centric view, the level of CGI methylation in the vicinity of TSSs is generally anti-correlated with the expression level of the corresponding gene while methylation in the gene body s positively associated with transcriptional elongation [Jones 2012; Schübeler 2015]. Furthermore, DNA methylation is implicated with roles in stabilizing the genome by means of silencing transposable elements [Jones 2012] (cf. Section 3.3).

DNA methylation patterns are established by DNA Methyltransferases (DNMTs) (reviewed in [Bestor 2000]) and can be stably maintained throughout the cell cycle. The maintenance methyltransferase DNMT1 preferentially methylates hemimethylated DNA and is therefore responsible for propagating DNA methylation across cell divisions. *De novo* DNA methylation is catalyzed by the DNMT3A and DNMT3B methyltransferases. The related DNA methyltransferase 3-Like (DNMT3L) does not contain a catalytic domains, forms complexes with DNMT3A and is associated with roles in embryonic development. Both methylated and unmethylated CpGs can be read by respective DNA binding proteins. Proteins containing a Methyl-CpG Binding Domain (MBD) specifically bind methylated DNA and are believed to be involved in repressing transcription [Schübeler 2015]. In contrast, stretches of DNA containing unmethylated CpGs can be recognized by proteins such as those containing CXXC domains and various histone demethylases.

DNA methylation patterns are highly dynamic in differentiating cells. During early embryonic and germ cell development, the mammalian genome becomes globally demethylated — with the exception of a few selected imprinted genes and transposons [Z. D. Smith and Meissner 2013]. The high rate at which this loss of methylation occurs suggests a mechanism for active DNA demethylation that complements passive depletion that is due to incomplete maintenance of methylation during replication [Kohli and Zhang 2013; Schübeler 2015]. Active demethylation can be catalyzed by enzymes of the ten-eleven translocation (TET) family which convert 5-Methylcytosine (5mC) to 5-Hydroxymethylcytosine (5hmC) and subsequently to 5-Formylcytosine (5fC) and 5-Carboxylcytosine (5caC). The resulting bases can be excised via thymine-DNA glycosylase (TDG) enzymes and are re-substituted by unmethylated cytosines. As tissues develop in the early embryo, DNA methylation patterns are reestablished subsequent to demethylation in a highly coordinated fashion involving the *de novo* methyltransferases. It has been shown that DNMT1, DNMT3A and DNMT3B are indispensable for normal development [E. Li *et al.* 1992; Okano *et al.* 1999], which underlines the essential role of DNA methylation in the process.

2.1.4 Non-Coding RNA

Eukaryotic cells harbor an extraordinary diversity of Non-Coding RNAs (ncRNAs) of which new classes are constantly discovered. It is difficult to assign distinct roles and definitions to these RNA classes, since they are involved in similar regulatory functions, participate in overlapping regulatory pathways and, in part, employ the same molecular machinery. Therefore, ncRNAs are often classified according to their length.

Small RNAs are involved in RNA Interference (RNAi) pathways that can downregulate gene expression on the transcriptional or post-transcriptional level [Morris and Mattick 2014]. Small Interfering RNAs (siRNAs), which have primarily been studied in plants and fungi, are around 21-22 nucleotides (nt) in length in their processed form. Their maturation involves processing of double-stranded RNA via Dicer enzymes, loading into Argonaute complexes and transport to the cytoplasm. They can act in a post-translational manner by binding to target mRNA by perfect base complementarity which triggers its degradation. Similar pathways of maturation apply to Micro RNAs (miRNAs), which, in contrast to siRNAs, recognize mRNA by imperfect base-pairing and therefore can target multiple mRNA sequences. Rather than degrading mRNA directly, miRNAs are involved in inhibiting translation or destabilizing their target through shortening of its Poly(A) tail. They also play a role in transcriptional gene silencing involving other epigenetic mechanisms [Morris and Mattick 2014]. A third class of small RNA involved in RNAi are piRNAs (26-31 nt in length), which are named according to their interaction with PIWI protein domains. They have been associated with silencing transposable elements in the germ cell lineage.

Our genome also encodes more than 9,200 different ncRNAs longer than 200 nt whose genes are located in intronic, intergenic regions or on the antisense strand of protein-coding genes [Derrien *et al.* 2012; Morris and Mattick 2014]. The genes for these Long Non-Coding RNAs (lncRNAs) often span several kilobases in the genome and contain introns themselves. lncRNAs have been implicated in an extraordinary number of regulatory processes. They have been shown to interact with epigenetic modifiers such as PRC2, trithorax complexes (which catalyze methylation of H3K4) as well as DNMTs and can also act as scaffolds for the assembly of molecular components. Examples hinting at their crucial role in development include the XIST lncRNA which coats the inactive X chromosome in females and recruits epigenetic silencing via PRC2 [Allis *et al.* 2007]. Furthermore, various lncRNAs involved in genomic imprinting have been described.

In addition, transcribed RNA itself has been attributed with a regulatory role, even when it is not translated: RNA elements could potentially compete for miRNA binding and thus regulate each other if they carry specific patterns of similar miRNA response elements [Salmena *et al.* 2011]. It has been hypothesized that these Competing Endogenous RNAs (ceRNAs) could act as decoys or "sponges" in order to regulate miRNA targeting of other transcripts. The recently discovered Circular RNAs (circRNAs) constitute one class of these sponges [Memczak *et al.* 2013; Hansen *et al.* 2013].

2.1.5 Transcription Factors and Gene Regulatory Elements

In order to assemble and initiate the transcriptional machinery at the promoter region of a gene, hundreds of individual proteins and subunits must act in concert. It is estimated that approximately eight percent of mammalian proteins are involved in transcriptional regulation [Alberts *et al.* 2008]. DNA binding proteins termed Transcription Factors (TFs) are key players in the recruitment and stabilization of the RNA polymerase

(RNA Polymerase II (RNAPII)). They establish accessibility of DNA by recruiting proteins responsible for the decompaction of chromatin [Spitz and Furlong 2012]. However, TFs can also negatively influence the rate of transcription. Their importance in cellular development is elucidated by the fact that the expression of only four TFs is necessary to induce a state of pluripotency from differentiated cells [Takahashi and Yamanaka 2006; Takahashi *et al.* 2007].

TFs contain different DNA-interacting domains and typically bind motifs comprising 6 to 12 nt of DNA with varying degrees of specificity [Spitz and Furlong 2012]. Only few TFs can bind to a closed chromatin structure. In addition to nucleotide sequence and chromatin state, the affinity of TFs to their respective Transcription Factor Binding Sites (TFBSs) is influenced by DNA methylation. While it is generally assumed that DNA methylation inhibits transcription factor binding, several factors capable of binding to methylated DNA in CpG-poor regions have been identified [Schübeler 2015]. Events responsible for the initiation of transcription are not restricted to promoter regions but also occur at distal regulatory elements tens of kilobases upstream or downstream of genes [Ong and V. G. Corces 2011]. Distal regulatory elements containing clusters of TFBSs, responsible for increased transcription are termed enhancers. Stabilized by the mediator complex and cohesin, they physically contact the promoter regions via loop structures [Gilbert 2014] (Figure 2.4). Active enhancers are characterized by the presence of the histone modifications H3K4me1, H3K4me2, H3K27ac as well as DNaseI hypersensitivity and occupancy of p300 proteins [Ong and V. G. Corces 2011]. They associate with nucleosome-poor regions and histone variants H2A.Z and H3.3 [Ong and V. G. Corces 2011]. Distal regulatory elements frequently coincide with Low-Methylated Regions (LMRs) in mouse ESCs and are closely linked to TF occupancy [Stadler *et al.* 2011]. Furthermore, enhancers themselves can also be transcribed into Enhancer RNA (eRNA) whose expression correlates with mRNA expression of nearby genes [Ong and V. G. Corces 2011]. In general, there is no clear one-to-one relationship between promoters and enhancers: multiple enhancers can physically contact the same promoter in a context-specific manner and an enhancer can be associated with multiple promoters. Multiple promoters and enhancers can co-localize in nuclear subcompartments associated with open chromatin and high transcriptional activity, so-called transcription factories [Ong and V. G. Corces 2011].

Other distal gene regulatory elements termed silencers are responsible reduced transcriptional activity. Insulators are generally marked by CTCF occupancy and represent barriers for chromatin interactions, preventing the spreading of epigenomic marks from one genomic region to another.

The interplay of these diverse regulatory elements is tightly regulated in time and space. It is therefore not surprising that our understanding of the precise regulatory mechanisms of these complex interactions is still limited.

2.2 The Epigenome of Human Disease

Epigenome dysregulation has been linked to imprinting disorders, neurodegenerative conditions such as Alzheimer's disease, inflammatory and autoimmune disorders (e.g. rheumatoid arthritis and systemic lupus erythematosus) and metabolic diseases like diabetes [Heyn and Esteller 2012]. Notably, epigenomic aberrations associated with cancer have been intensively studied in recent years and have the potential of contributing to the understanding of carcinogenesis as well as clinical detection, diagnosis and prognosis.

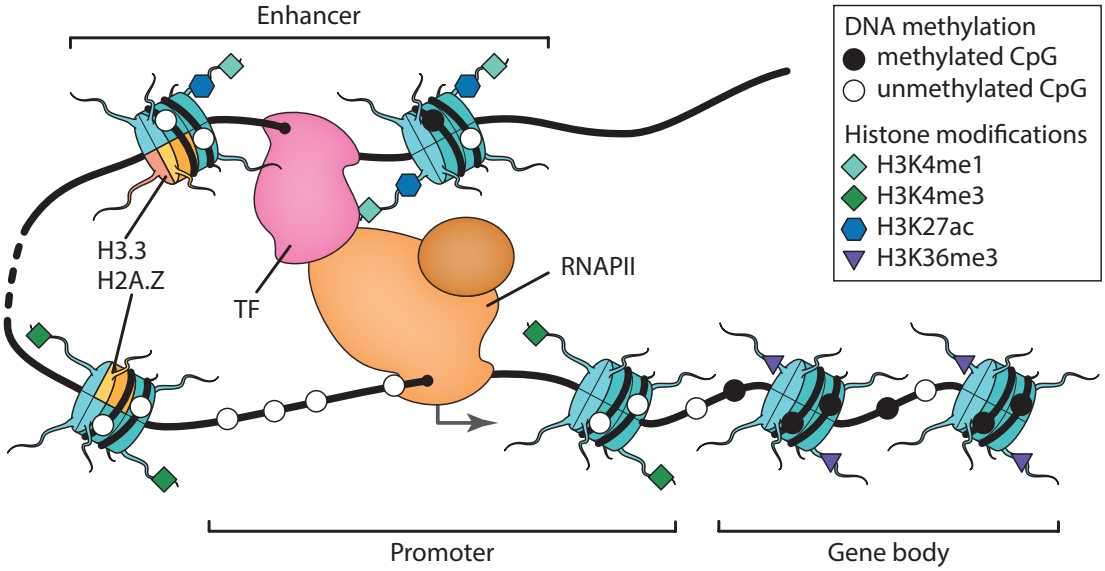


Figure 2.4: Epigenetic regulation of active gene expression. RNAPII is recruited to the promoter region of a gene. Active promoters frequently carry the H3K4me3 histone modification as well as low levels of DNA methylation and are associated with Nucleosome-Depleted Regions (NDRs). Transcriptional elongation is associated with H3K36me3 and DNA methylation in gene bodies. Active enhancers (marked by histone acetylation and H3K4me1) can be located dozens of kilobases upstream or downstream of a gene and contact promoter regions via loop structures. TFs can localize in promoter regions or distal elements in order to regulate the assembly of the transcriptional machinery.

Many cancer-associated mutations occur in the context of enzymes involved in modifying epigenome patterns. For instance, DNMT3A is frequently mutated in Acute Myeloid Leukemia (AML) [Baylin and Jones 2011; Schübeler 2015]. TET2, IDH1, IDH2 are often mutated in leukemia and glioma. These mutation events correlate with increased DNA methylation at specific loci. Changes in DNA methylation patterns represent well-characterized epigenomic aberrations in cancer. During transformation towards a malignant cell state, specific, CpG-rich regions become hypermethylated, i.e. they exhibit higher methylation levels in cancer cells compared to normal cells. These regions often coincide with promoters of genes associated with tumor suppressor activity and cell signaling [Bergman and Cedar 2013; Baylin and Jones 2011; Jones 2012]. CpG island hypermethylation events in glioma, colorectal cancer and other tumor types have also been described as CpG Island Methylator Phenotype (CIMP) and have been linked to DNA mutations and clinical outcome [Baylin and Jones 2011]. Many cancer-related DNA methylation changes occur in putative regulatory regions, such as enhancers and silencers. This is in accordance with recent studies focusing on chromatin accessibility that described the heterogeneity between cells in acute myeloid leukemia [M. R. Corces *et al.* 2016] and that characterized disease subtypes in chronic lymphocytic leukemia [Rendeiro *et al.* 2016]. Furthermore, aberrant DNA methylation in cancer is also associated with the silencing of miRNAs and other ncRNAs [Baylin and Jones 2011]. Importantly, it has been shown that DNA domains hypermethylated in the context of colorectal cancer and other tumor types are frequently bound by PcG protein complexes and their associated marks (H3K27me3 and bivalent domains) in ESCs [Ohm *et al.* 2007;

Schlesinger *et al.* 2007; Widschwendter *et al.* 2007]. It has therefore been postulated that DNA methylation in these regions is responsible for permanently establishing a stem-cell-like state of self-renewal in tumor cells [Baylin and Jones 2011]. Globally, large genomic domains exhibit lower methylation in cancer than in normal cells. The role of these hypomethylation events is currently not well understood. However, they frequently occur in lamin-bound and late-replicating regions [Berman *et al.* 2012]. Furthermore, the activation of retrotransposons by global demethylation could contribute to genome destabilization by increasing the rate at which mutations and structural variations occur [Sharma *et al.* 2010]. Hypomethylation can also lead to loss of imprinting and thus to the aberrant expression of genes associated with cell expansion and growth [Sharma *et al.* 2010].

Interestingly, Polycomb-related hypermethylation events have also been linked to aging [Teschendorff *et al.* 2010]. Recent studies have also identified variability in DNA methylation when comparing newborns and centenarians [Heyn *et al.* 2012]. Furthermore, "epigenetic clocks" that can accurately predict human age were derived from DNA methylation signatures of a few hundred CpGs [Horvath 2013]. Notably, when applied to tumor cells these clocks indicated the presence of DNA methylation signatures associated with accelerated aging.

In the context of clinical detection, diagnosis and prognosis of disease, epigenetic biomarkers are becoming increasingly important [Baylin and Jones 2011; Bock and Lengauer 2012; Bock, Halbritter, *et al.* 2016]. The cell material used for screening for these biomarkers can be obtained from biopsies from affected tissue, or alternatively from accessible body fluids in a non-invasive fashion. For instance, DNA methylation levels of tumor suppressor genes can be detected from blood serum and saliva [Heyn and Esteller 2012]. However, due to the low specificity of the applied assays and confounding factors, only few reproducible markers have been discovered so far. One example is the hypermethylation of the *MGMT* gene promoter which can be used for detecting gliomas as well as head and neck cancers [Heyn and Esteller 2012]. *MGMT* promoter methylation measured in glioblastoma biopsies is also capable of accurately predicting the tumor's response to therapies, which are based on alkylating agents [Hegi *et al.* 2005; Mikeska *et al.* 2007; Heyn and Esteller 2012]. Recently, the identification of epigenetic biomarkers has also been facilitated by Epigenome-Wide Association Studies (EWAS) [Rakyan *et al.* 2011] which pinpoint the genomic location of disease-related, epigenetic variability using large patient cohorts. High-throughput assays such as the Illumina Infinium HumanMethylation450 BeadChip (450K) (cf. Section 2.4) and methods based on bisulfite sequencing (cf. Section 2.3) are particularly suited for these types of studies. Assays for biomarker validation are becoming increasingly standardized and are approaching clinical application [Bock, Halbritter, *et al.* 2016].

In principle, epigenetic modifications are reversible. Therefore, therapies based on reverting to a "normal" epigenetic state bear great clinical promise. For instance, chromatin remodeling factors that are frequently mutated in cancer pose potential drug targets [Heyn and Esteller 2012]. More specifically, drugs currently approved by the US Food and Drug Administration (FDA) for cancer therapy include DNMT and HDAC inhibitors. As cancer therapies increasingly take pharmacogenomic properties into account, selecting an appropriate combination therapy for treatment based on the patients and the tumor's characteristic will be increasingly important. Computational methods are already applied in clinical treatment of Human Immunodeficiency Virus (HIV) and

similar methods could be applicable to cancer [Bock and Lengauer 2012]. In these approaches, epigenetic signatures of the tumor and healthy patient tissue clearly represent important markers to gauge tumor resistance and treatment selection.

2.3 Sequencing Technology and Computational Methods for Epigenome Profiling

With the advent of novel sequencing technologies, coined Next Generation Sequencing (NGS) [Metzker 2010], in the middle of the last decade, sequencing costs have decreased dramatically and sequencing entire mammalian genomes is now affordable. For instance, sequencing an entire human genome in a matter of only a few days only costs slightly more than 1,000 US dollars [National Human Genome Research Institute 2016] (although clinical genome sequencing is still substantially more expensive due to high quality requirements and the need for data interpretation). In the last years, technical innovations and developments in experimental protocols have led to the establishment of sequencing as a versatile tool with applications far beyond the sequencing of genomes. In particular, protocols for epigenome mapping have undergone rapid development (Figure 2.5). All of these protocols implement the following basic steps. Initially, an epigenetic signal is translated into DNA sequence. This step can involve chemically altering the DNA sequence itself or enriching for sequence fragments with certain properties. Next, sequencing libraries are constructed from the resulting fragments which are then subjected to NGS. Illumina's sequencing platforms ("HiSeq" and previously "GenomeAnalyzer") are most widely applied and millions to billions of short sequencing reads are typically produced in a single run. Finally, the sequencing readout is translated into profiles of epigenomic marks. In species whose genome has already been sequenced and assembled, this typically involves mapping the sequencing reads to the reference genome and deriving a measure of epigenetic signal strength at each genomic locus.

Recent advantages in technology and protocols allow for the profiling of single cells rather than bulk cell populations. Deep characterization of the heterogeneity of individual cells as well as the profiling of small and rare cell populations become feasible using these assays [Shapiro *et al.* 2013].

2.3.1 Quantifying the Expression of Large and Small RNAs

In order to quantify the expression of large and small RNAs, different RNA-Sequencing (RNA-seq) protocols are employed. In these protocols, the reverse transcriptase enzyme is used to convert RNA into cDNA from which the sequencing libraries are constructed. Different protocols have been established in order to extract RNA from specific cell compartments. Typically, RNA is extracted from the nucleus, the cytosol or from the entire cell (total RNA). The different species of RNA can be separated using size selection. A typical threshold for segregating large from small RNAs is 200 nt. Protocols involving polyA selection specifically enrich for mRNA. Processing the sequencing data involves mapping the reads to the reference genome, identifying expressed genes and isoforms and the quantification of their expression levels based on the number of reads assigned to a given entity [Garber *et al.* 2011]. In downstream analyses, differential expression between groups of samples can be quantified. A plethora of tools and pipelines have been developed for these steps. A selection of these tools is highlighted in [Garber *et al.* 2011].

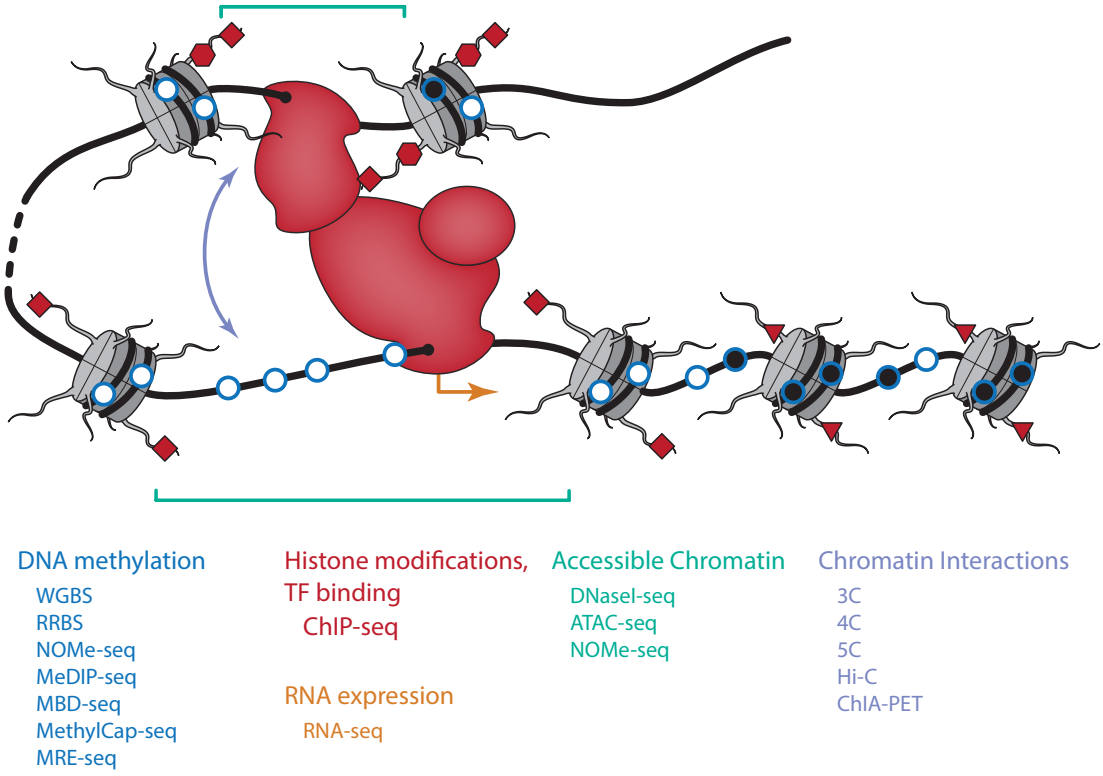


Figure 2.5: Sequencing technology for profiling epigenomic features. The same schematic as in Figure 2.4 is shown. Colors indicate different epigenomic features and a selection of corresponding sequencing technologies that can be used to quantify them.

Recently, it has become possible to profile single cells using RNA-seq, allowing for the characterization of cell heterogeneity and the roles of small cell populations in development [Shapiro *et al.* 2013], but also introducing new challenges in data analysis [Stegle *et al.* 2015].

2.3.2 Determining the Localization of Transcription Factors and Histone Modifications

In order to determine the genomic positions of Transcription Factors or histones carrying specific modifications, Chromatin Immunoprecipitation-Sequencing (ChIP-seq) can be used [P. J. Park 2009]. Here, the presence of DNA binding factors is estimated from the abundance of DNA fragments to which they are bound. To this end, using formaldehyde, all binding factors are cross-linked to the DNA at the position where they associate. Subsequently, the DNA is fragmented and DNA-protein complexes are selectively enriched by immunoprecipitation using factor-specific antibodies. Finally, the cross-linking is undone and the resulting purified DNA is subjected to standardized sequencing library construction. In practice, antibodies can vary in terms of binding strength and cross-reactivity to other factors. However, due to efforts in validating antibody sensitivity and specificity, protocols have become reasonably well standardized and adequate quality control measures have been established [Landt *et al.* 2012] — particularly for the immunoprecipitation of histone modifications. Recent adaptations

of the protocol require only very low amounts of input material and thus enable the profiling of rare cell populations [Lara-Astiaso *et al.* 2014; Schmidl *et al.* 2015].

The resulting sequencing reads can be aligned to the reference genome assembly using standard genome alignment methods. Counting the number of reads mapping to a given genomic position results in an enrichment signal for the respective factor. This signal is typically normalized to the so-called input signal[2], which is derived from an analogous sequencing experiment using the same cell material, but omitting the immunoprecipitation step. This results in a quantitative measure for the relative abundance of a specific factor at each genomic position. Genomic regions which are particularly enriched can be identified using peak-calling algorithms. These algorithms vary in their distributional assumptions and employed statistical models [Koohy *et al.* 2014]. Guidelines and standards for data quality control and processing pipelines are now emerging [Landt *et al.* 2012].

2.3.3 Charting DNA Methylation

Sequencing-based methods for profiling DNA methylation can be broadly categorized into protocols based on selective enrichment, specific cleavage by restriction enzymes or treatment with sodium bisulfite. Furthermore, array-based methods are widely applied (cf. Section 2.4). Benchmarking studies show that there is generally high agreement between the various protocols [Bock *et al.* 2010; R. A. Harris *et al.* 2010]. However, they differ greatly in the number of assayed cytosines, the amount of sequencing required and therefore also in their experimental costs.

In addition to profiling DNA binding factors, methods based on immunoprecipitation have also been successfully applied for genome-wide profiling of DNA methylation. Here, methylated DNA is selectively enriched by an antibody specific to methylated DNA, as applied in Methylated DNA Immunoprecipitation Sequencing (MeDIP-seq). Alternatively, protocols involving methyl-binding proteins (MBD-seq, MethylCap-seq, etc.) can be employed [Plongthongkum *et al.* 2014; Laird 2010]. Enrichment of methylated DNA is then quantified in a similar fashion as in ChIP-seq experiments. Other methods such as Methylation-Sensitive Restriction Enzyme Sequencing (MRE-seq) are based on the methylation-sensitive or insensitive digestion of DNA by endonucleases [Laird 2010]. A drawback of protocols based on selective enrichment or endonuclease cleavage is that the resulting data is of low genomic resolution and provides a notion of relative enrichment rather than an absolute quantification CpG methylation. However, bioinformatic algorithms have been devised that can infer single-CpG resolution methylation levels from these assays [Laird 2010; Stevens *et al.* 2013].

Methods based on the treatment of DNA with sodium bisulfite represent the current gold standard for quantifying DNA methylation. This chemical leads to the conversion of unmethylated cytosines to uracil while methylated cytosines remain intact. It can therefore be used to translate the methylation status into base sequence: after bisulfite-treatment, DNA fragmentation and subsequent Polymerase Chain Reaction (PCR) amplification, methylated cytosines are read as `C` (cytosine) while unmethylated cytosines are read as `T` (thymine) when they are sequenced. Given a mapping to a corresponding reference, the methylation information for each cytosine in each sequencing read can be inferred from the sequence context. For previously unsequenced genomes, bioinformatic methods such as ReFreeDMA can be used, which assemble relevant parts of the reference genome directly from sequencing reads [Klughammer *et al.* 2015].

[2] sometimes also referred to as Whole-Cell Extract (WCE)

A variety of protocols employing this principle have emerged. Of these, Whole Genome Bisulfite Sequencing (WGBS) is the most comprehensive method and assays the methylation status of nearly all 28 million CpGs in the human genome [Lister *et al.* 2009; Lister *et al.* 2011; Ziller *et al.* 2013]. Ideally, several hundred million short reads are sequenced in order to obtain an average depth of 30 reads per CpG. Reduced Representation Bisulfite Sequencing (RRBS) offers a more cost-efficient alternative. Here, restriction enzymes are used for DNA fragmentation. Typically, enzymes are selected in such a way that they enrich genomic regions with high CpG content [Meissner *et al.* 2008; Z. D. Smith *et al.* 2009; Gu *et al.* 2010]. For instance, MspI with its `CCGG` restriction site is most commonly used. Thereby, 2-3 million CpGs can be profiled by sequencing only approximately 1 % of the human genome.

Due to the damaging effects of treating DNA with sodium bisulfite, large quantities of input material are lost during the corresponding step of library preparation. Nonetheless, contemporary protocols have significantly reduced the amount of input material required for sequencing and therefore allow for epigenome-wide profiling at single-cytosine resolution from only a few or even single cells. Certain low-input protocols apply a strategy in which bisulfite treatment is applied before the adapter ligation step [Miura *et al.* 2012] and they have recently been employed for the generation of genome-wide DNA methylation data [Farlik *et al.* 2015; Smallwood *et al.* 2014]. In the case of RRBS, the amount of required input material has also been greatly reduced [Guo *et al.* 2013; Z. D. Smith *et al.* 2014] and high-throughput protocols allow for the simultaneous profiling of large sample numbers. It is important to note that 5hmC is indistinguishable from 5mC when using bisulfite-based protocols. However, derived protocols employing different chemical treatments or enrichment methods can be used to quantify oxidized forms of methylcytosine (5hmC, 5fC and 5caC) [Booth *et al.* 2012; Plongthongkum *et al.* 2014].

Bisulfite-incurred base conversions lead to imperfect matching to the reference. Furthermore, because the bulk of cytosines in the genome are converted and thymines are thus overrepresented, resulting reads exhibit reduced sequencing complexity. Therefore, aligning sequencing reads from bisulfite experiments is challenging (cf. Section 3.1 for details).

In the methylation calling step, methylation levels for each cytosine in the genome are extracted (cf. Section 3.1). Typically, the fraction of methylated read-cytosines among all reads covering a particular cytosine is quantified. Current methods aim at improving accuracy by employing additional steps such as local realignment and estimating allelic distributions. They can also take into account genetic variation and infer genotype calls for non-cytosine nucleotides [Liu *et al.* 2012].

Furthermore, contiguous regions with different methylation levels can be identified using segmentation algorithms and thresholding [Burger *et al.* 2013]. The resulting Unmethylated Regions (UMRs) and Low-Methylated Regions (LMRs) are characterized by near absence of methylation and low methylation levels, respectively. Partially Methylated Domains (PMDs) display disordered methylation patterns resulting in slightly lower average methylation levels than the typically highly methylated genome-wide background.

A plethora of algorithms and implementations for the identification of differentially methylated CpGs and regions between groups of samples exists. A table of selected methods and software tools for this purpose can be found in Appendix B. They vary greatly in their applied (statistical) methodology and in their output. Some methods quantify differential methylation only on the level of individual cytosines while other

methods identify Differentially Methylated Regions (DMRs). Finding DMRs can be based either on predefined genomic regions of interest or on heuristic definitions that rely on locally consistent methylation differences. Segmentation approaches can also be used for their *de novo* detection. Statistical methods for determining differential methylation frequently employ *t*-tests, Wilcoxon rank-sum tests, Fisher's Exact Tests, hierarchical linear modeling, mixture modeling or beta-binomial models.

2.3.4 Assessing Accessible Chromatin

Gene regulation requires access to sequence elements such as promoters, enhancers and insulators. Identifying regions in the DNA which are located in accessible chromatin is therefore crucial for the characterization of transcription factor binding events and other regulatory processes on the genome-scale. Preferential cleavage by nuclease enzymes in accessible DNA compared to condensed chromatin has inspired sequencing-based protocols for profiling accessible chromatin: library preparation for DNaseI-seq and MNase-seq involves the digestion of DNA with the respective nuclease (DNaseI or MNase) [Zentner and Henikoff 2014]. Quantifying the number of aligned sequencing reads allows for the accurate, high-resolution profiling of nucleosomes and non-histone, DNA-bound proteins by MNase-seq and the aligned read count provides a direct estimate of DNA accessibility by DNaseI-seq. From the latter signal, DNaseI hypersensitive sites are typically identified by similar peak-calling algorithms as employed in ChIP-seq. Recently, it has been shown that regions of open chromatin can be pinpointed using a transposase that preferentially cuts in accessible chromatin (Assay for Transposase-Accessible Chromatin using Sequencing (ATAC-seq)) [Buenrostro *et al.* 2013]. In these protocols, transposase is used for DNA fragmentation and adapter ligation. No additional digestion step is required which results in a simplified procedure for sequencing library preparation compared to DNaseI-seq. Furthermore, the amount of input material required by ATAC-seq is also reduced and current implementations of the protocol allow for the profiling of single cells [Buenrostro *et al.* 2015; Cusanovich *et al.* 2015]. In a different method, termed Nucleosome Occupancy and Methylome-Sequencing (NOMe-seq), input DNA is treated with the M.CviPI methylase that specifically methylates cytosines in GpC context in regions of accessible chromatin [Kelly *et al.* 2012]. Using bisulfite-sequencing and steps analogous to those of WGBS, open chromatin can be identified at single GpC resolution in a quantitative manner. In this way, NOMe-seq enables simultaneous assessment of chromatin accessibility and DNA methylation levels for cytosines out of GpC context at the experimental cost of WGBS.

As regions of open chromatin frequently coincide with transcription factor occupancy, one of the goals for downstream analysis is the identification of TF footprints. They are characterized by small "dips" in regions with high DNaseI signal or local drops in GpC methylation in the case of NOMe-seq and enable the identification transcription factors likely to bind these regions via their sequence motifs.

2.3.5 Mapping Chromatin Interactions and Higher Order Architecture

Physical contacts between genomic regions can be profiled using Chromosome Conformation Capture (3C) and its derived protocols (reviewed in [Dekker *et al.* 2013]). In these assays, nuclear DNA is first cross-linked to its scaffold of structural molecules using formaldehyde and the accessible chromatin is subsequently digested. DNA fragments

which are in close proximity are cross-linked to the same molecules and therefore associate preferentially compared to fragments located at further distances. In a ligation step, the fragments are joined together to form hybrid molecules which represent sequences that are not necessarily adjacent in the one-dimensional genome sequence. By identifying the genomic positions of the fragments that form such hybrids and quantifying the relative frequency of ligation events from resulting sequencing data, spatial proximity can be deduced. There are different strategies that can be employed for this task: in classical 3C, interactions of individual loci are assessed using PCR. In Circularized Chromosome Conformation Capture (4C), inverse PCR is applied to amplify fragments ligated to an individual anchor region, thereby generating genome-wide interaction profiles for that anchor. Carbon-Copy Chromosome Conformation Capture (5C) employs a many-by-many strategy in order to determine the interactions between two potentially large sets of loci using a multiplexed variation of PCR termed "multiplex ligation-mediated amplification" and sequencing. Hi-C [Lieberman-Aiden *et al.* 2009] offers an unbiased approach generating truly genome-wide interaction maps. Here, sequencing libraries are prepared from the ligated and biotin-labeled fragments and interactions are quantified via the bioinformatic identification of sequence junctions. However, as Hi-C is dependent on sequencing depth in order to capture all possible interactions (i.e. the number of reads required is quadratic in the size of the genome), high-resolution maps can be quite costly to obtain. Therefore, protocols based on sequence capturing have been developed that increase resolution by focusing on select interactions [Mifsud *et al.* 2015]. Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) combines fragment ligation with an immunoprecipitation step for a protein of interest (e.g. a transcription factor), followed by sequencing and thus identifies genome-wide, protein-associated interactions [Fullwood and Ruan 2009].

2.4 Array-Based Methods for Quantifying DNA Methylation

Microarrays facilitate standardized, high-throughput DNA methylation profiling of large sample cohorts at relatively low experimental costs and therefore represent important tools in EWAS. This section describes the Illumina Infinium assay series, which constitutes the most widely used microarray-based assay for DNA methylation analysis. These arrays are spotted with stretches of DNA (probes) which represent defined genomic regions. Fragmented and bisulfite-converted genomic DNA is hybridized to these probes. Signal intensities representative of the binding frequencies can be measured for each probe from the fluorescence emitted during the extension of the probe sequences using labeled nucleotides. These intensities, which are indicative of either methylated or unmethylated cytosines, are used to quantify DNA methylation levels as β-values (or M-values).

Currently, most available datasets employ the Illumina Infinium HumanMethylation450 BeadChip (450K), which covers 482,421 CpGs in the human genome [Bibikova *et al.* 2011]. In comparison, its predecessor, the Illumina Infinium HumanMethylation27 BeadChip (27K), assayed slightly more than 27,000 CpGs [Bibikova *et al.* 2009]. The 450K also contains 3091 non-CpG, 65 Single-Nucleotide Variation (SNV) probes as well as sets of probes designed for quality control purposes. The selection of the CpG probes is biased towards genomic regions associated with high CpG content and annotated genes [Bibikova *et al.* 2011; Sandoval *et al.* 2011]. The Illumina Infinium MethylationEPIC BeadChip (EPIC) [Moran *et al.* 2015], the most recent release of the platform,

covers 853,307 CpGs and particularly increases the coverage in regulatory regions of the genome. On the 450K and the EPIC arrays, two probe-designs exist: type I probes employ two different sequences for each cytosine, that exhibit basepair complementarity to either the bisulfite-converted or the unconverted case in order to measure the unmethylated and methylated signal, respectively. Type II probes harness a single-probe design which incorporates degenerate bases for cytosines and different fluorescence dyes for nucleotides complementary to the converted and unconverted cytosines.

Data processing involves normalization of methylation levels, which corrects for different probe types and locations within a single array as well as for between-array variation [Bock 2012]. Further experimental biases and batch effects are identified and corrected for in computational pipelines [Bock 2012] (cf. Section 3.2). As in the case of sequencing-based methods, pinpointing differentially methylated sites and regions constitutes an important type of analysis (cf. Appendix B for applicable methods).

2.5 Global Efforts for Epigenome Mapping

Since the publication of the first draft sequence of the human genome in 2001 [Lander *et al.* 2001; Venter *et al.* 2001], the scientific community's interest in the genome-wide mapping of gene regulatory elements and epigenomic marks has grown steadily. Several large-scale endeavors have been launched with this task in mind.

The first of these projects, termed **Encyclopedia of DNA Elements (ENCODE)** [ENCODE Project Consortium 2012], was initiated by the National Human Genome Research Institute (NHGRI). It launched its pilot phase in 2003, initially mapping functional elements in approximately 1 % of the human genome [ENCODE Project Consortium 2004]. The ensuing scale-up phase (2007 to 2012) encompassed the genome-wide identification of regulatory and functional elements in 147 cell types, mainly by means of charting epigenomic marks and transcription factors using NGS technology [ENCODE Project Consortium 2012]. Assayed cell types mostly comprised *in vitro* cultured pluripotent, fetal and adult cell lines, but also a few samples derived from primary tissues. Expression profiles were generated using RNA-seq and resulted in a catalog of transcribed and protein-coding regions. ChIP-seq was employed to map the distribution of transcription factors and histone modifications. Genome-scale DNA methylation was measured using mainly the RRBS protocol. Finally, regions of open chromatin, TF footprints and chromatin interaction maps were generated harnessing methods such as DNaseI-seq, ChIA-PET, 3C and 5C. In an effort for data integration, functional roles have been attributed to 80.4 % of the human genome in one cell line or another [ENCODE Project Consortium 2012]. Due to the large number of cell types and assays, it is not surprising that the resulting "data matrix" still contains gaps to be filled. The project currently focuses on developing and applying methods for integrative data analysis for annotating functional elements in the genome from the available data. In 2007, the project has been expanded to other model organisms such as the worm *Caenorhabditis elegans* and the fruit fly (*Drosophila melanogaster*) in the **modENCODE** project [Celniker *et al.* 2009].

The **NIH Roadmap Epigenomics Mapping Consortium (REMC)** [Roadmap Epigenomics Consortium *et al.* 2015] started in 2008. Features charted in this project include DNA methylation (assayed by WGBS, RRBS, MeDIP-seq and MRE-seq), histone modifications (ChIP-seq), open chromatin (DNaseI-seq) and RNA expression (RNA-seq). As the project reached completion, 111 reference maps derived from cell types of various

tissues and organs have been characterized and published [Roadmap Epigenomics Consortium *et al.* 2015]. Epigenomic, cell-type-dependent variability was described, specific regulatory elements were identified and links to genetic variation were established. Using statistical learning methods, entire datasets were reliably imputed at high resolution [Ernst and Kellis 2015] thereby filling the gaps in the project's data matrix.

The EU-funded **BLUEPRINT** project [Adams *et al.* 2012] aimed at charting the epigenomes of more than 100 different blood-related samples in health and disease. Launched in 2011 and terminating in 2016, a large number of epigenomic reference maps have been generated using RNA-seq, histone modification ChIP-seq, DNaseI-seq and WGBS. Project goals included the identification of epigenetic mechanisms involved in hematopoietic differentiation as well as diseases like leukemia and diabetes. Furthermore, relations of the epigenome and the genome were established in large patient cohorts and novel experimental technologies were developed in collaboration with industry partners.

The German contribution to generating epigenomic reference maps is called **DEEP ("Deutsches Epigenom Programm")**[3] and has a strong focus on human metabolic, immune and inflammatory diseases. During the course of this 5-year project which was started in 2012, approximately 90 epigenomes comprising DNA methylation, histone modification, open chromatin and expression of long and short RNAs will be generated. DEEP's objective is to characterize cell types such as adipocytes, hepatocytes, fibroblasts, epithelial cells, macrophages, monocytes and T cells in healthy individuals as well as in the context of inflammatory bowel diseases, rheumatoid arthritis, liver steatosis and adipogenesis.

The ENCODE, REMC, BLUEPRINT and DEEP projects as well as consortia from Canada (Canadian Epigenetics, Environment and Health Research Consortium), South Korea (Korea National Institute of Health), Japan ("Core Research for Evolutional Science and Technology" program), Singapore (The Genome Institute of Singapore) and Hong Kong (Hong Kong University of Science and Technology) comprise the **International Human Epigenome Consortium (IHEC)** [Stunnenberg *et al.* 2016]. Launched in 2010, this global umbrella project aims at generating more than 1,000 reference epigenomes, each comprising expression quantification by RNA-seq, ChIP-seq profiles for at least six well-studied histone modifications (H3K4me1, H3K4me3, H3K9me3, H3K27me3, H3K27ac and H3K36me3), chromatin accessible regions and DNA methylation patterns profiled by WGBS.

Ultimately, these projects represent key players in contributing to the standardization of assays and methods and have greatly helped in paving the way in establishing common practices in epigenome analysis and scientific communication in general.

[3] http://www.deutsches-epigenom-programm.de

3

Pipelines for Comprehensive DNA Methylome Analysis

This chapter introduces software tools for DNA methylation analysis. I developed the BISEQMETHCALLING *tool for the quantification of methylation levels from aligned bisulfite sequencing reads. It was incorporated into a pipeline which I co-developed with Natalie Jäger and Christoph Bock for the initial processing of bisulfite sequencing data. This epigenome processing pipeline was used for validating the RRBSMAP alignment software [Xi et al. 2012] and employed in various data analysis settings (e.g. [Ziller et al. 2011; Ziller et al. 2013]).*

Together with Yassen Assenov and Pavlo Lutsik, I developed the RNBEADS *software package [Assenov et al. 2014] (equal contributions). The project was supervised by Jörn Walter, Thomas Lengauer and Christoph Bock. Corresponding text and figures in this chapter have been adapted from the publication. I had a leading role in the overall software design of the package as well as in writing the manuscript, the package vignette and in designing corresponding figures. During package development, I was mainly responsible for implementing the analysis of bisulfite sequencing data, methods for exploratory data analysis, data export, methods for covariate inference, adjusting for confounding factors and detecting differential methylation. Furthermore, I conducted a performance benchmarking study and applied the package to large-scale, publicly available datasets, generating a methylome resource. Using* RNBEADS *we contributed DNA methylation analyses to the studies of Sandoval et al. [2013], Planello et al. [2014], Wallner et al. [2016] and Amabile et al. [2015].*

A software suite I developed for characterizing DNA methylation in consensus sequences of repetitive elements was applied and contributed analyses to Bock et al. [2010], Tobi et al. [2014] and Deplus et al. [2014]. A redesigned production version of the software is currently under active development.

As outlined in the previous chapter, DNA methylation constitutes an important epigenomic mark involved in regulating chromatin structure and RNA expression. Computational pipelines for the processing of bisulfite sequencing data typically involve aligning the sequencing reads to a reference genome and identifying methylation levels at individual cytosines. Downstream analysis such as identifying domains with consistent methylation patterns, characterizing within-sample and between-sample variability and determining differential methylation between groups of samples represent typical tasks in many studies focusing on DNA methylation. Various software tools for these individual tasks exist (reviewed in [Krueger *et al.* 2012] and [Bock 2012]).

This chapter focuses on software development for key steps in the analysis of genome-wide DNA methylation data: Section 3.1 discusses analysis steps involved in processing reads originating from bisulfite sequencing experiments and introduces the BISEQMETHCALLING tool for quantifying methylation levels at individual cytosines. Section 3.2 describes the RNBEADS software package that we developed for the comprehensive analysis of DNA methylation data from various assays. Finally, Section 3.3 highlights the importance of epigenetic regulation of repetitive genomic elements and outlines a pipeline implementing methods that exploit sequencing data to estimate global profiles of epigenomic marks in subfamilies of these elements.

3.1 Quantifying DNA Methylation Using Bisulfite Sequencing

As described in Chapter 2.3, bisulfite sequencing provides quantitative readouts of DNA methylation levels for each assayed CpG. Here, we outline the processing steps involved in analyzing this type of data and present a tool that we developed for quantifying DNA methylation levels from aligned sequencing reads.

3.1.1 Processing of Bisulfite Sequencing Data

Before DNA methylation signatures can be interpreted, a series of bioinformatic preprocessing steps is required in order to derive methylation levels from the raw sequencing reads. These steps involve:

1. Read preprocessing
2. Alignment
3. Methylation calling
4. Quality Control (QC)

Segments of reads that exhibit low per-base quality scores or low sequence complexity as well as sequences pertaining to adapters used for library preparation could introduce false measurements. Therefore, they are removed during read preprocessing. Several tools that can identify and trim such segments from sequencing reads are available [Krueger *et al.* 2012]. Moreover, tools exist that provide a general overview on sequencing data quality. They can help with the identification of biases and artifacts in the sequence and quality composition of the sequencing readout. For instance, the FASTQC toolkit[1] enables data analysts to inspect sequencing reads for enriched sequence motifs as well as overall sequence complexity and generates reports summarizing per-base quality scores.

Correct alignment of bisulfite-converted reads to the reference genome is not straightforward: first, due to the induced nucleotide substitution, reads do not map to the reference by perfect base complementarity. Second, bisulfite conversion reduces the sequence complexity of resulting reads (due to the relative depletion of cytosines). This can lead to their incorrect placement in the genome. Third, cytosine methylation results in an asymmetry between the two reference strands of the genome and therefore increase the complexity of the alignment task. Specifically, four different orientations are possible in which sequencing reads can align to the reference sequence (Figure 3.1). Additionally, bisulfite conversion is not always 100 percent efficient: typically, less than two percent of all unmethylated cytosines remain cytosines and unspecific effects might

[1] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

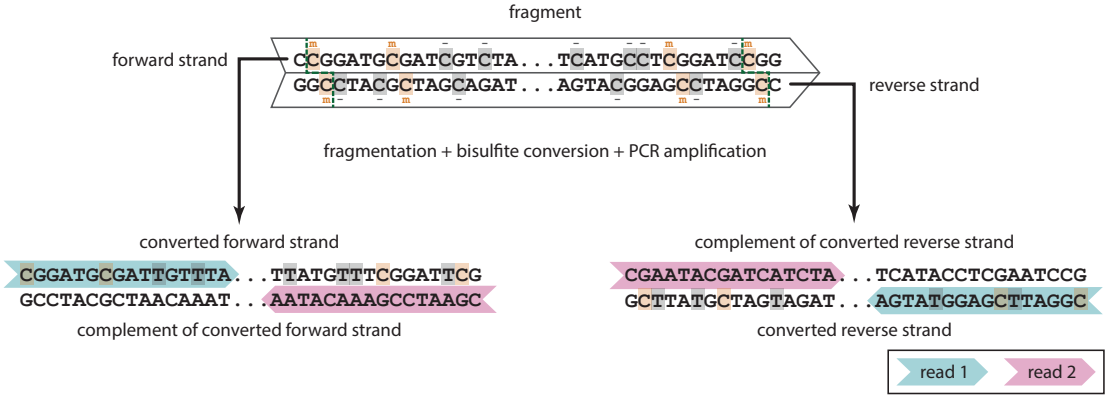


Figure 3.1: Orientations of reads in bisulfite sequencing experiments. Methylated cytosines (orange background) are unaffected by bisulfite conversion while unmethylated cytosines (grey background) are converted to uracils, which results in a thymine readout after subsequent PCR amplification. Colors denote read mates in paired-end sequencing. For single-end sequencing only reads labeled "read 1" are applicable. In paired-end sequencing mate reads are sequenced from the start and end positions of the same template strand. In directional bisulfite libraries, only the four depicted read orientations are valid for the individual mate reads ("read 1" and "read 2"). Green, dashed zig-zag lines indicate the digestion sites of the MspI endonuclease frequently applied in RRBS protocols.

lead to the conversion of methylated cytosines after bisulfite treatment. However, these under- and over-conversion events are relatively rare and experimental quality can be assessed by measuring methylation levels in spike-in control sequences, which originate from phage DNA and are assayed along with the target DNA. Other challenges which are inherent to the general problem of short-sequence alignment also need to be addressed. These include repetitive sequences in the genome, sequencing errors as well as genotypes that deviate from the reference. Despite or because of the complexity of bisulfite alignment, an abundance of mapping programs is available [Krueger *et al.* 2012; Bock 2012; Kunde-Ramamoorthy *et al.* 2014]. They implement two alternative paradigms for dealing with the above challenges: Wild-card aligners do not penalize C to T mismatches during read mapping. Available implementations include BSMAP [Xi and W. Li 2009], its derivative RRBSMAP[2] [Xi *et al.* 2012], GSNAP [Wu and Nacu 2010], Last [Frith *et al.* 2012], Pash [Coarfa *et al.* 2010], RMAP [A. D. Smith *et al.* 2009] and segemehl [Otto *et al.* 2012]. In contrast, three-letter aligners, such as Bismark [Krueger and Andrews 2011], methylCtools [Hovestadt *et al.* 2014], BRAT [E. Y. Harris *et al.* 2010], BS-Seeker [P.-Y. Chen *et al.* 2010], BatMeth [Lim *et al.* 2012] and MethylCoder [Pedersen *et al.* 2011], convert all cytosines in the sequencing reads as well as the forward and reverse strand of the reference genome to thymines and employ standard mapping algorithms on this reduced base-space. In a post-alignment step, duplicate reads, i.e. reads with the same sequence, mapping to identical genomic positions are usually marked in the resulting data files. Such reads are likely to represent PCR amplification artifacts. They can be identified using tools such as picard[3].

[2] RRBSMAP has now been integrated into the BSMAP software
[3] http://broadinstitute.github.io/picard/

From the aligned reads, methylation levels, i.e. the fraction of methylated cytosines among all reads covering a particular reference cytosine can be extracted. In essence, this is a counting exercise. Nonetheless, additional read and sequence information such as duplicate reads, read orientation, sequence complexity, mapping and base qualities need to be taken into account. In this work, BISEQMETHCALLING, a software tool that addresses these challenges, is presented. Other available software for methylation calling includes BISMARK [Krueger and Andrews 2011] and BIS-SNP [Liu *et al.* 2012]. BISMARK implements an integrated framework for bisulfite alignment and methylation calling. BIS-SNP is a tool for the quantification of methylation levels from aligned reads using Bayesian inference. It incorporates base quality recalibration steps and infers genomic variation from the reads (excluding C to T mutations), reporting genotype information additionally to methylation levels. In addition to these programs for the processing of genome-wide bisulfite data, software tools for the start-to-finish analysis of locus-specific bisulfite data exist. For instance, the BIQ ANALYZER tool provides an all-in-one software solution for read preprocessing, alignment, methylation calling and data visualization [Bock *et al.* 2005; Lutsik *et al.* 2011].

Finally, during all steps of the outlined process, quality control is a critical component. With a few exceptions (such as FASTQC) standardized tools for this task are not readily available. Therefore, best practices include the manual inspection of quality statistics, custom plots and data exploration in tools like genome browsers.

3.1.2 A Pipeline for Quantifying DNA Methylation from Bisulfite Sequencing Reads

We developed the BISEQMETHCALLING tool for the quantification of DNA methylation from aligned bisulfite sequencing reads. The tool offers extensive summary statistics and other outputs that facilitate quality control and genome-browser-based visualization of the results. Through various processing options it is possible to flexibly account for a variety of special cases and intricacies of methylation calling. For instance, several read filtering criteria can be applied and various features specific to the RRBS protocol are implemented.

Implementation Details

BISEQMETHCALLING was implemented using the PYTHON programming language and comprises approximately 4,000 lines of code. Aligned reads in the SAM/BAM format [H. Li *et al.* 2009] constitute the input to BISEQMETHCALLING. The software is typically applied to output data from the MAQ [H. Li *et al.* 2008] or (RR)BSMAP [Xi and W. Li 2009; Xi *et al.* 2012] tools, but can be adapted to the output of any bisulfite aligner. The program iterates over all input reads and identifies valid base calls for each cytosine contained in the reference sequence. Here, a position is considered valid if the base-quality of the cytosine and other bases in the respective motif exceed a parameter-defined threshold. Reads are discarded if they do not pass certain user-defined criteria: reads can be filtered based on the number of alignment mismatches, whether a read is marked as PCR duplicate, overall base-quality and sequence complexity (i.e. reads with unreasonably long stretches of the same nucleotide in their sequence can be excluded). If properly annotated, domains of reads pertaining to sequencing adapters can be excluded from the analysis. When considering paired-end data from directional sequencing protocols, only two read orientations are considered valid for each read mate (see Figure 3.1). BISEQMETHCALLING can identify and discard reads that do not fit one of these configurations. Furthermore, improperly paired reads, i.e. mate reads mapping to different chromosomes or mapping further apart than the expected fragment size can

be excluded. Optionally, only the mate in a pair with better alignment quality can be retained for analysis. If mate reads overlap due to short fragment lengths, only information for the mate with higher alignment quality is retained for methylation calling in the overlap. Additional features pertain specifically to the RRBS protocol. In these protocols, sequencing adapters and the nucleotides used for filling in the restriction site typically employ methylated cytosines. This artificially introduced methylation is excluded by BISEQMETHCALLING when quantifying DNA methylation. For WGBS and RRBS data, methylation information for the genomic forward and reverse strand can be considered separately or can be combined. By default, processing is restricted to CpG dinucleotides, but reporting of methylation in non-CpG contexts can optionally be enabled.

The output of BISEQMETHCALLING consists of methylation information that is summarized across individual cytosines, sequencing reads and fragments (Figure 3.2). A custom BED format [Kent *et al.* 2010] is used to store genomic locations along with their annotation and facilitates data sharing and visualization in genome browsers. For each cytosine in the considered CpG or non-CpG context, the tool counts the numbers of conversion and non-conversion events. All sequencing reads that pass the filtering criteria are reported along with relevant cytosine positions and methylation levels. When reads originate from paired-end sequencing or are aligned to an *in silico* digested genome (in case of RRBS protocols), the genomic segments resulting from DNA fragmentation and size selection during sequencing library preparation can be identified. Mean methylation levels of individual reads and fragments are reported. Additionally, extensive statistics, which can be used for quality control, are provided for both genomic strands and cytosine contexts individually as well as in a combined manner. These statistics include counts of retained reads and reads discarded due to the filtering criteria. Additionally, methylation levels (mean and quantile values) inside and outside of considered sequence motif (CpG, CpA, etc.) as well as summaries of base and alignment quality are reported. The overall methylation levels outside of CpG context can be used as an estimate for the rate of overall bisulfite conversion. If provided with the sequence of spike-in controls, methylation levels in these sequences are quantified and serve as dedicated measures for the conversion rate.

The tool offers several convenience options. Multiple alignment input files can be specified which are then merged during runtime according to user-provided grouping criteria. For instance, alignment files originating from multiple sequencing lanes for the same sample can be merged. Furthermore, analyses can be restricted to predefined genomic regions of interest. This feature has proven particularly useful for more targeted sequencing libraries, computational parallelization or software testing. Processing can conveniently be restarted from certain savepoints in case of a failed analysis run. To improve the wallclock runtime, BISEQMETHCALLING can be configured to be executed distributed across nodes of a scientific compute cluster and thus allows for the high-throughput analysis of whole-genome bisulfite sequencing data.

Availability

Although there is currently no stand-alone version of the tool, BISEQMETHCALLING has been customized to be routinely employed by a number of laboratories. An adaptation of the source code is part of the supplementary data of [Ziller *et al.* 2013] and is available from GITHUB[4].

[4] https://github.com/epigen/biseqMethCalling

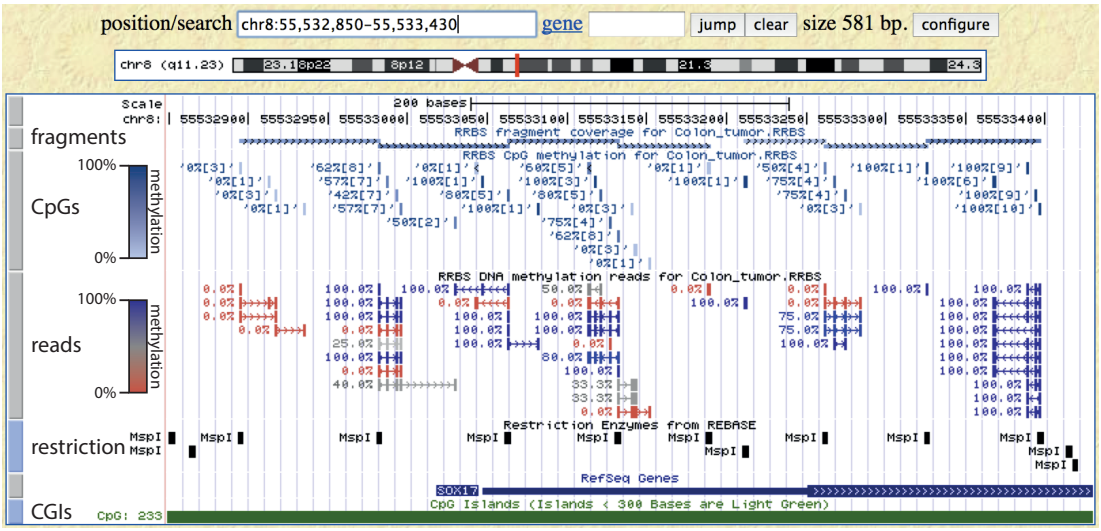


Figure 3.2: Genome browser view of fragments, reads and CpGs. The data was generated using RRBS in a human colon cancer sample. The browser view shows the *SOX17* promoter locus. RRBS restriction fragments, single CpGs and sequencing reads are shown as separate tracks. Mean methylation levels are labeled in percent values and are indicated by corresponding color scales. Read coverage for each CpG is displayed in square brackets. The data has been visualized using the UCSC Genome Browser [Kent *et al.* 2002].

3.1.3 Discussion

We have developed one of the first pipelines for quantifying DNA methylation levels from genome-wide bisulfite sequencing experiments. It has been established at the Meissner lab at the Department for Stem Cell and Regenerative Biology at Harvard University and the Broad Institute, but has also successfully been adapted for use in other laboratories. At the time of development, no software package for genome-wide methylation calling was publicly available and this pipeline represented one of the first tools for the processing of RRBS and WGBS data. Since the pipeline was first employed, other tools for methylation calling, such as Bismark [Krueger and Andrews 2011] and Bis-SNP [Liu *et al.* 2012], have emerged. Compared to these alternatives, biseqMethCalling provides extensive summary statistics and other outputs that facilitate quality control and genome-browser-based visualization of the results. The software was used to process human and murine methylation data for a number of studies (e.g. [Ziller *et al.* 2011; Ziller *et al.* 2013]), thus proving its utility for the standardized high-throughput processing of bisulfite sequencing data.

3.2 Comprehensive Analysis of DNA Methylation Data with RnBeads

High-throughput assays for DNA methylation profiling have enabled large-scale Epigenome-Wide Association Study (EWAS) and epigenome mapping projects. The number of assayed samples and methylation sites in these studies is steadily increasing, while access to bioinformatics support for many biomedical researchers in this area is still limited. Guidelines and good practices for DNA methylation analysis are now emerging [Bock 2012; Michels *et al.* 2013] and a plethora of bioinformatic tools, which

implement them, have been developed. However, the vast majority of these tools is dedicated to a particular task in methylation analysis or is only applicable to a specific experimental platform. Particularly the identification of differentially methylated regions has motivated various software developments (cf. Appendix B). We evaluated the features of 22 related software tools (supplementary information in [Assenov *et al.* 2014]) and concluded the demand for a pipeline that facilitates start-to-finish analysis of DNA methylation data and that can be employed by users with potentially little or no bioinformatics expertise.

With this goal in mind, we developed the RnBeads software package [Assenov *et al.* 2014] which establishes a user-friendly workflow for the analysis and interpretation of large-scale DNA methylation data that complies with emerging standards. RnBeads builds upon extensive prior research on bioinformatic and statistical methods for DNA methylation analysis. Based on our assessment of existing algorithms and software, we defined the following key features of RnBeads:

- support for all genome-scale and genome-wide DNA methylation assays that provide single-cytosine resolution
- extensive functionality for high-level DNA methylation analysis, including data visualization, quality control, exploratory analysis, handling of batch effects, correcting for tissue heterogeneity and differential methylation analysis
- analysis of DNA methylation at the level of individual CpGs as well as predefined genomic regions
- generation of interactive reports that allow users to select results and adjust parameters without having to rerun the analysis
- implementation of a standardized pipeline mode that is essentially self-configuring, with the option to adapt the workflow using custom parameter settings or custom scripts
- flexibility to run RnBeads on a personal computer, on a high-performance computing infrastructure, via a web-based service or in a cloud-computing environment, depending on the scale of the analysis
- usability without the need extensive programming knowledge
- sufficient performance to process the largest DNA methylation datasets that are currently available on a suitable scientific computing cluster (hundreds of RRBS or WGBS profiles or thousands of 450K profiles)
- reproducibility and sharing of results through automatic documentation of parameters and analysis methods in the RnBeads reports

The core workflow of RnBeads comprises data import, quality control, preprocessing and filtering, generation of genome browser tracks and data tables, optional inference of (confounding) covariates, exploratory analysis and differential DNA methylation analysis (Figure 3.3).

RnBeads supports any experimental protocol that provides single-basepair CpG methylation measurements. This includes array-based platforms such as Illumina's 450K, 27K and EPIC arrays as well as bisulfite-sequencing protocols (RRBS, WGBS, etc.). Notably, data obtained from enrichment-based methods can also be analyzed, provided that the data has been preprocessed with corresponding bioinformatic algorithms [Laird 2010; Stevens *et al.* 2013].

RnBeads is implemented using the R programming language and follows a modular design that supports automated pipeline workflows as well as flexible interactive analyses. The default RnBeads workflow is executed by invoking `rnb.run.analysis(...)`,

either in an interactive R session or via R's support for scripted analyses. Optionally, an XML configuration file can be provided in order to execute analyses with predefined parameter sets.

When used with default options on small- to medium-scale data sets, RnBeads is essentially self-configuring: it parses a user-provided sample annotation table and, using this annotation, executes the modules as shown in Figure 3.3. RnBeads workflows can also be fine-tuned using global configuration parameters, which are specified using the `rnb.options(...)` command. During the execution of an analysis, each step is tracked by extensive logging functionality. Upon successful module completion, an interactive analysis report is generated that comprises method descriptions, publication-quality diagrams and links to data tables. These reports use client-side scripting and the dynamic features of XHTML to enable interactive data exploration of precalculated results. RnBeads can also save analysis options and data objects, which makes it straightforward to rerun an analysis with the same parameters and to comply with the paradigm of reproducible research [R. C. Gentleman 2005]. Custom workflows can be designed by running analysis modules individually or by using R functions that operate directly on serialized `RnBSet` objects (these objects are instances of an R `S4` class and constitute the RnBeads representation of all DNA methylation and metadata within a given dataset).

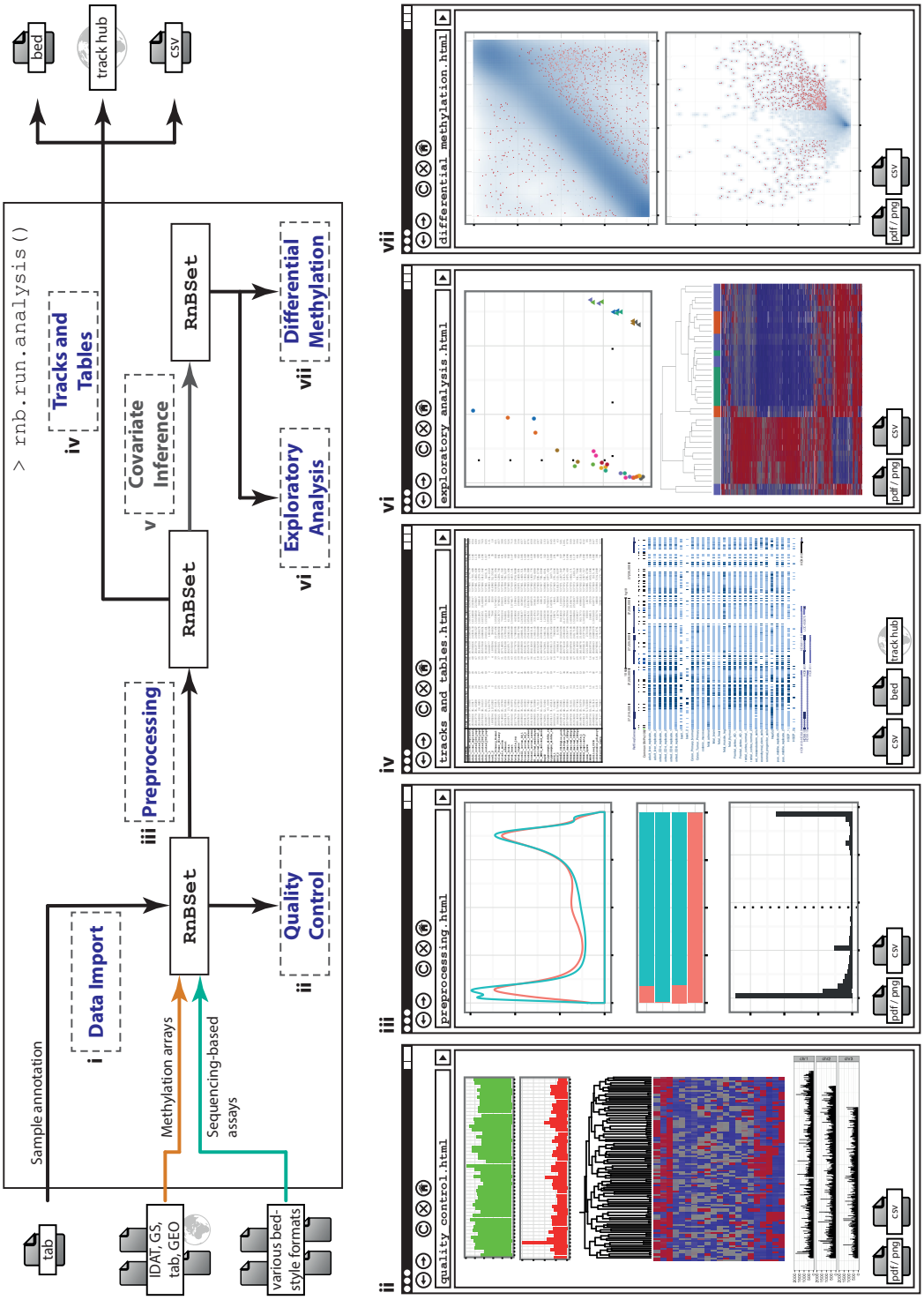
The following sections provide a detailed overview on RnBeads' analysis modules, implementation details and a use-case. The tool has also been extensively used for characterizing the DNA methylation datasets described in Chapters 4 and 5.

3.2.1 Analysis Modules in Detail

Data Import

First, the data import module of RnBeads parses a sample annotation table and uses the contained information to configure the analysis. This user-provided annotation should be in a tabular format that contains a row for each sample and columns describing sample characteristics. A column containing the file names for each sample's methylation data is required. Other columns may contain categorical information which are then parsed and used as sample annotation in later analysis steps. For instance, the color coding in the exploratory analysis and the grouping for differential methylation analysis is determined by these columns. Via RnBeads' analysis options the user can specify which of the columns are used in the analysis or RnBeads can automatically identify columns with categorical sample annotation. Second, the methylation information for

Figure 3.3 : *(On the next page)* RnBeads workflow for analyzing DNA methylation data. The workflow consists of seven modules (i-vii) and is essentially self-configuring on the basis of a sample annotation table provided by the user. Each module generates part of the RnBeads hypertext report, which includes method descriptions, diagrams and links to data tables. `RnBSet` objects store methylation and QC data and sample annotation in order to facilitate custom analysis workflows in R. Methylation data is exported for visualization using genome browsers and follow-up analyses using third-party software tools. Input data types include signal intensities (IDAT), Illumina GenomeStudio (GS), tab-delimited files (TAB), Gene Expression Omnibus (GEO) datasets and various BED-like formats.


> rnb.run.analysis()
Sample annotation
Methylation arrays
Sequencing-based assays
tab
IDAT, GS, tab, GEO
various bed-style formats
i Data Import
RnBSet
ii Quality Control
iii Preprocessing
RnBSet
v Covariate Inference
iv Tracks and Tables
RnBSet
vi Exploratory Analysis
vii Differential Methylation
bed
track hub
csv
ii
quality_control.html
iii
preprocessing.html
iv
tracks_and_tables.html
vi
exploratory_analysis.html
vii
differential_methylation.html
pdf / png
csv

each sample is parsed. A broad range of input formats is supported. For array-based analysis using the Infinium platforms, signal intensity data files (IDAT) can be read by RnBeads, which extracts methylation information from the signal intensities and performs data normalization. Alternatively, RnBeads can load precomputed data from Illumina GenomeStudio report files, directly from the Gene Expression Omnibus (GEO) database or import data from tabular formats. When IDAT files are loaded into RnBeads, the methylumi Bioconductor package is used for performing low-level data processing.

For sequencing-based methods, data preparation requires steps that are highly protocol-dependent, including sequence alignment and DNA methylation calling for single CpGs (cf. Section 3.1). These steps need to be executed using dedicated tools before loading the data into RnBeads. An RnBeads analysis starts with importing BED files or data tables that provide the number of methylated and unmethylated observations for each covered CpG. For example, the outputs of Bismark [Krueger and Andrews 2011] and Bis-SNP [Liu *et al.* 2012] as well as methylation calls obtained from the pipeline described in Section 3.1.2 can directly serve as inputs to RnBeads. Enrichment-based and restriction-enzyme-based assays require specialized algorithms for inferring DNA methylation levels at single-base-pair resolution. For these experiments, software tools such as MEDIPS [Chavez *et al.* 2010], MEDUSA [Wilson *et al.* 2012] and methylCRF [Stevens *et al.* 2013] generate DNA methylation tables that can be imported into RnBeads as BED files or in one of several other data file formats.

After the DNA methylation data has been loaded, RnBeads combines the data of all samples into a single `RnBSet` object that constitutes the basis for all further analysis steps. `RnBSet` objects store DNA methylation levels as β-values which are defined as

$$\beta = \frac{\max(M, 0)}{\max(M, 0) + \max(U, 0) + \epsilon}$$

for Illumina's array platforms. Here, M and U correspond to the measured, continuous-valued methylated and unmethylated intensity signal respectively and ϵ is a constant (typically set to 100). Bisulfite-sequencing-based methylation levels are simply described by the fraction of methylated cytosines compared to the number of reads covering a CpG. RnBeads can also derive M-values from β-values or methylation levels using the logit transformation:

$$M = \log_2 \frac{\beta}{1 - \beta}$$

In addition to storing methylation measurements for each CpG, these objects also contain aggregate methylation levels for predefined genomic regions which are computed by averaging over CpG methylation levels contained in a region. By default, these region definitions include promoters (defined as 1500 bp upstream and 500 bp downstream of the TSS), whole genes (transcription start to transcription end site), CpG islands and genomic tiling regions. In addition, custom region annotation can be supplied by the user (e.g. via BED files). Since data matrices become large when performing genome-wide analyses on large numbers of samples, RnBeads provides the option of maintaining these matrices on hard disk in order to reduce the demand on main memory. For this purpose, the efficient memory mapping procedures implemented in the ff package[5] are employed. RnBeads currently supports several human, mouse and rat genome assemblies via its auxiliary annotation packages RnBeads.hg19,

[5] http://cran.r-project.org/web/packages/ff/index.html

RnBeads.hg38, RnBeads.mm9, RnBeads.mm10 and RnBeads.rn5. A supplemental package, named RnBeadsAnnotationCreator can be used to generate RnBeads annotation packages for additional genomes.

Quality Control

RnBeads also can also assist the user in the identification of technical and biological biases that are common in large-scale DNA methylation datasets. These include technical assay failures, sample mix-ups and batch effects. Quality issues are highlighted in the resulting reports, but it is ultimately left to the user to handle them appropriately, for example by excluding samples with low quality, by resolving sample mix-ups using genotyping data or by correcting for batch effects using statistical methods. The detection of technical failures is assay-specific and differs between sequencing-based and microarray-based analyses. For Infinium array data, RnBeads generates summary plots for the microarray quality-control probes to monitor technical parameters such as bisulfite conversion efficiency and unspecific probe hybridization. For sequencing-based data sets, the quality assessment focuses on sequencing coverage, given that issues in bisulfite conversion and clonal read rates are typically already identified during the alignment and methylation calling steps. RnBeads also addresses the relatively common problem of sample mix-ups by graphical representation and clustering based on signal intensities of the genotyping probes that are present on the Infinium microarray. Samples with matching genotypes can thus be identified and checked for consistency with annotation records. In addition, RnBeads uses DNA methylation data to predict which samples were derived from male and female donors on the basis of their X-inactivation status and the presence or absence of measurements on the Y chromosome. This classifier highlights discrepancies between gender information from the sample annotation table and the biological sex of the analyzed samples, which are often indicative of sample mix-ups.

Preprocessing

To minimize the risk of measurement biases affecting the analysis, RnBeads implements a framework for rule-based filtering of samples, CpG sites and DNA methylation measurements. Filtering is performed in two steps to provide flexibility and to avoid biasing the normalization procedure of Infinium analyses with problematic samples. First, RnBeads removes low-quality data that could bias an analysis by discarding samples and CpGs that contain a substantial fraction of measurements with low technical quality (for example, a large detection p-value for Infinium data or low sequencing coverage in the case of bisulfite sequencing data) as well as CpGs and measurements that may be unreliable for other reasons. For example, RnBeads can remove Infinium probes overlapping Single-Nucleotide Polymorphisms (SNPs) that pose a high risk of influencing DNA methylation measurements. In a second filtering step that follows data normalization RnBeads discards those samples and CpGs that should be included in the normalization, since they contribute to the overall distributions of signal intensities and methylation levels, but that should not be included in subsequent analysis steps[6]. Examples include CpGs with too many missing values across samples or with zero variability in their methylation values. Furthermore, users can configure additional filtering rules

[6] Normalization is only applied to Infinium data. In case no normalization is performed, the order of filtering steps is irrelevant.

and define custom blacklists of CpGs that should always be excluded and/or whitelists of CpGs that should always be retained. The default filtering criteria were chosen relatively conservatively with the goal of reducing the risk of spurious or misleading results.

For Infinium data, RnBeads offers several alternative options for signal intensity-based normalization, which is an important step for reducing probe biases that could interfere with the analysis. RnBeads' default normalization method for Infinium data is SWAN [Maksimovic *et al.* 2012], which is implemented in the minfi package [Aryee *et al.* 2014] and provides a good balance of accuracy, robustness and run-time performance. Alternatively, RnBeads supports Illumina's standard normalization procedure as implemented in methylumi, the BMIQ normalization method [Teschendorff *et al.* 2013], and all modular normalization algorithms that are available in the wateRmelon package [Pidsley *et al.* 2013]. RnBeads also supports the background-correction techniques implemented in methylumi [Triche *et al.* 2013], which can be combined with normalization.

All filtering and normalization steps are tracked in the RnBeads report and plots visualize any changes in the global distribution of DNA methylation levels before and after preprocessing.

Tracks and Tables

RnBeads is able to export the preprocessed data in several formats that facilitate data visualization and ancillary analyses using third-party software. RnBeads can be configured to export methylation tracks in BIGBED and BIGWIG format and to summarize them into track hubs that can be loaded into various genome browsers, thus providing a common reference point for exploring the generated data tracks. Moreover, the software aggregates the preprocessed data in CSV and BED files that can be loaded and analyzed with custom scripts and web-based tools. In addition, sample-wise statistics, including the number of assayed CpGs and genomic regions, the number of assayed CpGs per region type, and the average read coverage (for sequencing data), are summarized in dedicated tables.

Exploratory Analysis

Global changes in DNA methylation can often be identified by visual inspection of normalized and quality-controlled DNA methylation data before in-depth analysis of differential DNA methylation. To facilitate this type of exploratory analysis, RnBeads visualizes sample-specific DNA methylation profiles at the single-CpG level and for genomic regions of interest. The global distribution of DNA methylation levels is summarized in density plots, which help identify samples and sample groups that deviate from the characteristic bimodal distribution of methylation levels with its clear-cut distinction between highly methylated loci and essentially unmethylated loci (for example, due to global gain or loss of DNA methylation). RnBeads also visualizes DNA methylation variation within and across sample groups, which facilitates the detection of hypervariable samples (for example, due to technical issues or biological effects such as high tissue heterogeneity). DNA methylation profiles are computed on the basis of single CpG measurements as well as using values aggregated in sets of predefined genomic regions, such as gene promoters or CpG islands. Furthermore, if the annotation includes information on biological or technical replicates, RnBeads calculates pairwise agreement between replicates and visualizes them as scatterplots, thereby providing a global assessment of reproducibility of experiments.

Hierarchically clustered heatmaps help to assess the presence of sample subgroups in the data set. This analysis is quantitatively supported by various distance metrics, by the calculation of silhouette statistics to identify the best fitting number of clusters as well as by systematic association testing between the obtained clusters and the user-provided sample annotation. Dimension reduction employing Principal Component Analysis (PCA) and Multidimensional Scaling (MDS) is also available. In combination with interactive sample coloring, this functionality provides a powerful way of visualizing associations between sample annotations and genome-wide DNA methylation profiles. The analysis of such patterns is also helpful for detecting batch effects, which can arise from technical confounders such as date and duration of sample processing, the person running the assay and sample origin. Batch effects are common in large-scale DNA methylation datasets, in particular among those generated with microarrays or with enrichment sequencing protocols such as MeDIP and MBD-seq. To systematically detect batch effects, RnBeads runs tests for significant association between user-provided sample annotation and the directions of largest variance identified in a PCA of the DNA methylation data. Statistical testing is also performed to identify significant associations among the sample annotations and with quality-control indicators such as bisulfite conversion rates and nonspecific binding (for Infinium data). In these comparisons, the appropriate statistical test is automatically selected based on the type of annotation data. Specifically, the selected tests are: Fisher's Exact Test for categorical data, Wilcoxon rank-sum test for continuous values in two sample groups, Kruskal-Wallis one-way analysis of variance for continuous values in multiple groups and Pearson correlation coupled with a permutation test for comparing continous values. All results are visualized in the RnBeads report, thus enabling a systematic assessment of associations between DNA methylation levels and sample annotations.

In addition, composite plots of DNA methylation levels around genes and other genomic regions are generated. These plots can help detect global changes in DNA methylation that affect gene promoters differently compared to intra- or intergenic regions. Finally, if provided with a list of custom genomic regions or genes of interest, genome browser views can be generated with RnBeads using functionality of the Gviz package [Hahne *et al.* 2012].

Differential DNA Methylation

After evaluating multiple approaches and their implementations for the identification of differential methylation (cf. Appendix B), we implemented the following strategy in RnBeads: different measures of differential methylation can be aggregated into combined ranks and differences can be analyzed at the level of individual CpGs (Algorithm 3.1) as well as by combining measurements across larger genomic regions (Algorithm 3.2). The latter approach increases statistical power and can result in interpretable sets of differentially methylated regions [Bock 2012]. Moreover, it reduces the susceptibility to differential coverage of individual CpGs between samples. The following paragraphs provide details on the computed differential methylation measures and their combination.

In each comparison defined by the sample annotation table, RnBeads computes p-values for all covered CpGs (π_is in Algorithm 3.1). By default, hierarchical linear models, implemented in the limma package [Smyth 2004] are employed for this task (limmaP(...) in Algorithm 3.1). limma was originally developed for identifying differential gene expression and employs the following linear relationship to model gene

expression levels based on sample features and corresponding coefficients:

$$\mathrm{E}(\boldsymbol{y}_g) = \boldsymbol{X}\boldsymbol{\beta}_g$$

Here, $\boldsymbol{y}_g$ denotes the vector of observed log-expression levels for gene g in n samples: $\boldsymbol{y}_g = (y_{g1}, \ldots, y_{gn})^T$. $\boldsymbol{X}$ denotes an $n \times p$ design matrix that encodes sample annotation for p covariates. This annotation includes the group membership of each sample for the target comparison (as binary variable) and can optionally contain other covariates to be included in the model. $\boldsymbol{\beta}_g$ denotes a vector of feature coefficients and is estimated by least squares:

$$\hat{\boldsymbol{\beta}}_g = (\boldsymbol{X}^T \boldsymbol{W}_g \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W}_g \boldsymbol{y}_g$$

where $\boldsymbol{W}_g$ is a diagonal matrix of known feature weights. In order to derive a p-value for differential expression, the null hypothesis that the coefficient for gene g and target comparison j does not deviate from zero is tested:

$$H_0 : \beta_{gj} = 0$$

In ordinary linear models this can be facilitated by the t-statistic:

$$t_{gj} = \frac{\hat{\beta}_{gj}}{s_g \sqrt{v_{gj}}}$$

where $s_g^2 = (\boldsymbol{y}_g - \boldsymbol{X}\hat{\boldsymbol{\beta}}_g)^T (\boldsymbol{y}_g - \boldsymbol{X}\hat{\boldsymbol{\beta}}_g)/d_g$ denotes the residual sample variance (with d_g residual degrees of freedom; usually $d_g = n - p$) and v_{gj} is the jth diagonal element of $(\boldsymbol{X}^T \boldsymbol{W}_g \boldsymbol{X})^{-1}$. Instead of the ordinary t-statistic, LIMMA defines a moderated t-statistic:

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{v_{gj}}}$$

where

$$\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}$$

Here, s_0^2 and d_0 denote prior estimates for variance and degrees of freedom. They represent hyperparameters shared by all genes and can be derived from the data using empirical Bayes methods (cf. [Smyth 2004] for details). The resulting hierarchical models[7] have been shown to lead to robust estimates of differential expression, even when the number of samples per group is small [Phipson *et al.* 2016]. Notably, the problem that the number of parameters to be estimated is much larger than the number of training samples is addressed by this mode of parameter sharing. In RnBeads, the LIMMA approach is directly applied to DNA methylation data by replacing log-expression ($\boldsymbol{y}_g$) with methylation M-values. They are derived from the fractional methylation levels y_{ci} (or β-values) of each CpG c and sample i by the logit function $M_{ci} = \log_2 \frac{y_{ci}}{1-y_{ci}}$ and reflect distributional properties of log-expression values considered in the model more closely than fractional methylation levels. We specifically selected the LIMMA method due to its sound statistical model, robustness to relatively small sample sizes, ease-of-use and computational performance. Alternatively, p-values can also be calculated using an ordinary two-sided t-test comparing the distributions of methylation levels between the

[7] Here, the term hierarchical refers to the fact that the models are composed of parameters shared between all genes and parameters pertaining to individual genes.

Algorithm 3.1: Site-level quantification of differential methylation between two groups of samples

Input:

Sample groups G_1, G_2 with $n_1 := |G_1|, n_2 := |G_2|, n := n_1 + n_2$
$p_S \in \mathbb{N}$ methylation sites
Methylation level matrix $\boldsymbol{Y} \in [0,1]^{p_S \times n}$

Output:

$\boldsymbol{\mu_1}, \boldsymbol{\mu_2} \in [0,1]^{p_S}$, vectors of group mean methylation levels for each site
$\boldsymbol{\delta} \in [-1,1]^{p_S}$, a vector of methylation level differences
$\boldsymbol{\theta} \in \mathbb{R}^{p_S}$, a vector of methylation ratio differences
$\boldsymbol{\pi} \in [0,1]^{p_S}$, a vector of differential methylation p-values
$\boldsymbol{\rho} \in [1, p_S]^{p_S}$, a vector of combined ranking scores for each site

for $i \in \{1, \ldots, p_S\}$ **do**
  $\mu_{1,i} \leftarrow \text{mean}\left(Y_{i,G_1}\right)$
  $\mu_{2,i} \leftarrow \text{mean}\left(Y_{i,G_2}\right)$
  $\delta_i \leftarrow \mu_{1,i} - \mu_{2,i}$
  $\theta_i \leftarrow \frac{\mu_{1,i}}{\mu_{2,i}}$
  $\pi_i \leftarrow \text{limmaP}\left(\tilde{\boldsymbol{Y}}, G_1, G_2, i\right)$ ▷ using the LIMMA package on the matrix of M-values
end for
$\boldsymbol{\rho}_\delta \leftarrow \text{rank}\left(|\boldsymbol{\delta}|\right)$
$\boldsymbol{\rho}_\theta \leftarrow \text{rank}\left(|\log\left(\boldsymbol{\theta}\right)|\right)$
$\boldsymbol{\rho}_\pi \leftarrow \text{rank}\left(-\boldsymbol{\pi}\right)$
for $i \in \{1, \ldots, p_S\}$ **do**
  $\rho_i \leftarrow \max\left(\{\rho_{\delta,i}, \rho_{\theta,i}, \rho_{\pi,i}\}\right)$
end for

two groups of samples or optionally the RefFreeEWAS method, which can be used to account for cell-type heterogeneity [Houseman *et al.* 2014] (see details below). In addition to the default, unpaired analysis, RnBeads also supports paired-sample analysis, which can substantially increase statistical power when analyzing matched pairs such as tumor versus normal tissue from the same individual or disease-discordant twins. The CpG-level p-values are adjusted for multiple testing using the False Discovery Rate (FDR) method [Benjamini and Hochberg 1995].

Furthermore, aggregate p-values are computed for predefined genomic regions. Here, the uncorrected, CpG-specific p-values within a given region are combined using an extension of Fisher's method that accounts for correlated p-values [Makambi 2003] (combineP $(\ldots)$ in Algorithm 3.2). Specifically, Fisher's method combines p-values $p_1, \ldots, p_m$ by computing the statistic

$$M_F = -2 \sum_{i=1}^{m} \ln(p_i)$$

When all the null hypothesis underlying the individual p_i are true, the resulting M_F follows a chi-square distribution with $2m$ degrees of freedom and the combined p-value can be obtained by testing its significance. However, Fisher's method does not account for dependencies in the p-values that are the result of correlated test statistics. The

Algorithm 3.2: Region-level quantification of differential methylation between two groups of samples

Input:

Sample groups G_1, G_2 with $n_1 := |G_1|, n_2 := |G_2|, n := n_1 + n_2$
$p_S, p_R \in \mathbb{N}$ the number of sites and regions respectively
$\phi : \{1, \ldots, p_R\} \to \mathbf{2}^{\{1,\ldots,p_S\}}$, a mapping of sites to each region
$\boldsymbol{\mu_1}, \boldsymbol{\mu_2} \in [0,1]^{p_S}$, group mean methylation levels for each site
$\boldsymbol{\pi} \in [-1,1]^{p_S}$, a vector of differential methylation p-values for each site

Output:

$\mathbf{M_1}, \mathbf{M_2}$, vectors of group mean methylation levels for each region
$\boldsymbol{\Delta} \in [-1,1]^{p_R}$, a vector of methylation differences for each region
$\boldsymbol{\Theta} \in \mathbb{R}^{p_R}$, a vector of methylation ratio differences for each region
$\boldsymbol{\Pi} \in [0,1]^{p_R}$, a vector of differential methylation p-values for each region
$\mathbf{P} \in [1, p_R]^{p_R}$, a vector of combined ranking scores for each region

---

for $j \in \{1, \ldots, p_R\}$ **do**
  $\mathrm{M}_{1,j} \leftarrow \mathrm{mean}\left(\mu_{1,\phi(j)}\right)$
  $\mathrm{M}_{2,j} \leftarrow \mathrm{mean}\left(\mu_{2,\phi(j)}\right)$
  $\Delta_j \leftarrow \mathrm{M}_{1,j} - \mathrm{M}_{2,j}$
  $\Theta_j \leftarrow \frac{\mathrm{M}_{1,j}}{\mathrm{M}_{2,j}}$
  $\Pi_j \leftarrow \mathrm{combineP}\left(\pi_{\phi(j)}\right)$ ▷ method for combining p-values
end for
$\mathbf{P}_{\boldsymbol{\Delta}} \leftarrow \mathrm{rank}\left(|\boldsymbol{\Delta}|\right)$
$\mathbf{P}_{\boldsymbol{\Theta}} \leftarrow \mathrm{rank}\left(|\log\left(\boldsymbol{\Theta}\right)|\right)$
$\mathbf{P}_{\boldsymbol{\Pi}} \leftarrow \mathrm{rank}\left(-\boldsymbol{\Pi}\right)$
for $j \in \{1, \ldots, p_R\}$ **do**
  $\mathrm{P}_j \leftarrow \max\left(\left\{\mathrm{P}_{\Delta,j}, \mathrm{P}_{\Theta,j}, \mathrm{P}_{\Pi,j}\right\}\right)$
end for

---

method introduced by Makambi [2003] therefore employs alternative distributional assumptions and use a weighted version of the test statistic:

$$M_{F,m} = -2 \sum_{i=1}^{m} \ln(p_i)\omega_i$$

where the ω_i are weights for individual p-values and sum to 1. Under the assumption of correlated p-values the statistic $\nu \frac{M_{F,m}}{\mathrm{E}(M_{F,m})}$ follows a chi-square distribution with ν degrees of freedom, which can be used to derive an aggregate p-value. Provided with an estimate for the pairwise correlations of p-values, ν can be computed from the given p-values and weights (cf. [Makambi 2003] for details). For combining p-values, RnBeads uses a estimated correlation coefficient of 0.8, which was empirically determined from methylation data, and uniform weights $\omega_i = \frac{1}{m}$. Subsequently, these aggregate p-values are subjected to multiple-testing correction using the FDR method.

In order to address the problem that minimal, but consistent differences tend to receive low p-values that reflect statistical but not biological significance, RnBeads ranks

the differentially methylated regions according to the combination of statistical significance and the degree of differential methylation. In Algorithms 3.1 and 3.2, the computed ranks are denoted by ρ_i and P_j for CpGs and predefined genomic regions, respectively. The degree of differential methylation is estimated as (i) the absolute difference in DNA methylation (δ_i and Δ_j) and (ii) the relative ratio of mean DNA methylation levels between sample groups (θ_i and Θ_j). These two measurements differ in their relevance for regions with low versus high DNA methylation levels and thus complement each other. In regions of the genome that exhibit DNA methylation values near 0 %, the DNA methylation ratio between sample groups tends to overestimate the effect size, and the absolute DNA methylation difference is a more appropriate measure. The opposite is true for high DNA methylation values near 100 %, where the relative ratio is the more stringent and appropriate measure of effect size.

In summary, RnBeads combines statistical testing with a priority ranking scheme that is based on the absolute and relative effect size of the differences between sample groups. It assigns a combined rank score for differential DNA methylation to each analyzed CpG site and genomic region (ρ_i and P_j in Algorithms 3.1 and 3.2). This combined rank is defined as the maximum (i.e. worst) of three individual rankings: (i) by absolute difference in mean DNA methylation levels ($\rho_{\delta,i}$ and $P_{\Delta,j}$), (ii) by the relative difference in mean DNA methylation levels, which is calculated as the absolute value of the logarithm of the quotient of mean DNA methylation levels ($\rho_{\theta,i}$ and $P_{\Theta,j}$) and (iii) by the CpG-based or region-based p-value ($\rho_{\pi,i}$ and $P_{\Pi,j}$; calculated as described above). The priority-ranked lists can be used directly for downstream analysis, such as manual inspection of the top-ranking regions in a genome browser or for web-based analysis using tools such as Galaxy [Giardine *et al.* 2005], GeneTrail2 [Stöckel *et al.* 2016] and EpiExplorer [Halachev *et al.* 2012]. In addition to the ranking of differential DNA methylation, RnBeads visualizes the observed differences using scatterplots and volcano plots, and it performs enrichment analysis for Gene Ontology (GO) terms associated with differentially methylated regions.

Covariate Inference

Even well-designed studies performed with accurate DNA methylation assays can include confounders and potential sources of batch effects. For example, the samples in an epigenome-wide association study may be collected in different countries, using different preprocessing steps or may stem from genetically distinct populations. Furthermore, many large cohort studies are currently conducted on whole blood, which is characterized by significant cellular heterogeneity. RnBeads implements a number of methods that can be used to mitigate such biases.

Batch effects arise from variation in the sample origin or sample handling [Leek *et al.* 2010] and their influence on the measurements can obscure biologically relevant differences. As long as the experimental design is chosen in such a way that the confounders exhibit an acceptable distribution across the phenotypes of interest in the dataset, RnBeads can correct for the resulting biases employing established statistical tools. To that end, known sources of potential batch effects should be documented in dedicated columns of the sample annotation table. These columns can then be specified as known confounders when performing the limma-based analysis of differential DNA methylation. RnBeads also integrates the Surrogate Variable Analysis (SVA) method implemented in the sva package [Leek *et al.* 2012] as an optional step of the workflow.

In SVA, so-called surrogate variables are identified from the singular value decomposition of the residual methylation matrix that results from regressing out known covariates [Leek and Storey 2007]. To be more precise, let $\boldsymbol{Y}$ an $m \times n$ matrix with methylation levels for m sites and in n samples. Then SVA models the methylation level for site $i = 1, \ldots, m$ in sample $j = 1, \ldots, n$ as

$$Y_{ij} = f_i(x_j) + \epsilon_{ij}$$

where $f_i(x_j)$ denotes a general function over the annotated sample information x_j. Parameters for $f_i(x_j)$ are typically estimated by linear regression methods. The term ϵ_{ij} denotes unmeasured sources of methylation variation. SVA performs singular value decomposition on the residual matrix $\boldsymbol{R}$, whose entries are defined by $R_{ij} = Y_{ij} - f_i(x_j)$: $\boldsymbol{R} = \boldsymbol{UDV}^T$. As a result, this decomposition captures the influence of unobserved factors on methylation variation in terms of an orthogonal basis of singular vectors. By means of statistical testing, singular vectors that are significantly associated with variation are identified and they are used for the construction of a set of significant surrogate variables (cf. [Leek and Storey 2007] for further details). The rationale behind this approach is that these surrogate variables estimate effects of unknown factors and they can be accounted for by controlling for them in differential methylation analysis. The latter can be accomplished by including the surrogate variables as additional covariates in the models used for the computation of p-values for differential methylation.

Other methods for batch-effect detection and correction have not yet been implemented in RnBeads, but can applied to `RnBSet` objects as part of custom workflows. For instance, similar to SVA, the ISVA method [Teschendorff *et al.* 2011] identifies surrogate variables, but employs independent component analysis for matrix decomposition and estimates the number of components by random matrix theory. Furthermore, ComBat [Johnson *et al.* 2007] provides an empirical Bayesian framework for correcting for known batch covariates.

DNA methylation differences between heterogeneous samples (such as blood, tumor tissue and most other types of tissue biopsies) can arise not only from cell-intrinsic differences in DNA methylation but also from differences in the cell-type composition between samples [Jaffe and Irizarry 2014; Teschendorff 2015]. RnBeads supports three alternative methods for handling cell-type heterogeneity in the context of analyzing differential DNA methylation.

First, for certain sample types such as whole blood, it is possible to purify reference populations of the most prevalent cell types contributing to sample heterogeneity. DNA methylation patterns from these references can then be used to quantify the cell composition of a heterogeneous sample [Houseman *et al.* 2012]. The resulting cell composition estimates can then be included as covariates in the limma-based analysis of differential DNA methylation. This method is most commonly used for EWAS performed on patient cohorts for which only whole-blood samples are available [Michels *et al.* 2013]. Suitable reference maps for sorted cells of the blood system have been generated for the 450K (e.g. [Reinius *et al.* 2012]). The model behind the method proposed in [Houseman *et al.* 2012] describes the following linear relationship between CpG methylation, sample phenotypes and cell type composition:

$$\boldsymbol{Y} = \boldsymbol{BX}^T + \boldsymbol{M\Omega}^T + \boldsymbol{E} \tag{3.1}$$

Here, $\boldsymbol{Y} \in [0,1]^{m \times n}$ is a matrix of methylation values for m CpGs and n samples. $\boldsymbol{X}$ is an $n \times d$ design matrix that contains sample annotation such as information on the

phenotype and known, potential confounders. $\boldsymbol{B}$ contains $m \times d$ regression coefficients representing direct effects of known covariates on methylation. The composition of k cell types for each sample is represented by the $n \times k$ matrix $\boldsymbol{\Omega}$ (k is fixed in advance). $\boldsymbol{M} \in [0,1]^{m \times k}$ contains methylation levels for the k cell types and is derived from the average methylation levels of purified samples for the given reference cell types. $\boldsymbol{E}$ is an $m \times n$ matrix of error terms. $\boldsymbol{\Omega}$ itself is assumed to be dependent on the sample covariates:

$$\boldsymbol{\Omega} = \boldsymbol{X}\boldsymbol{\Gamma} + \boldsymbol{\Xi} \tag{3.2}$$

where $\boldsymbol{\Gamma}$ contains coefficients describing the linear relationship between cell composition and sample covariates and $\boldsymbol{\Xi}$ is an error matrix. The goal of referenced-based estimation of cell-type heterogeneity is to estimate the unknown cell-type compositions $\boldsymbol{\Omega}$. In theory this problem can be solved using least squares estimation. However, Houseman *et al.* [2012] impose natural constraints on $\boldsymbol{\Omega}$, i.e. the composition values are in the interval $[0,1]$ and their sum is less than one for each sample, and solve the problem using a quadratic programming approach. In RnBeads the estimated contributions for the reference cell types are added to the sample annotation and they can be used as covariates in differential methylation analysis. Furthermore, they also provide useful annotation to be considered in exploratory analyses.

Second, a number of methods have been proposed for accounting for cell-type heterogeneity without the need for reference profiling [Houseman *et al.* 2014; Zou *et al.* 2014] and we have integrated them into the RnBeads pipeline. Extending their model for reference-based estimation, Houseman *et al.* [2014] propose the RefFreeEWAS method. In detail, substituting Equation 3.2 in Equation 3.1 yields

$$\boldsymbol{Y} = \underbrace{(\boldsymbol{B} + \boldsymbol{M}\boldsymbol{\Gamma}^T)}_{\boldsymbol{B}^*}\boldsymbol{X}^T + \underbrace{\boldsymbol{M}\boldsymbol{\Xi}^T + \boldsymbol{E}}_{\boldsymbol{E}^*}$$

RefFreeEWAS first estimates the coefficients $\boldsymbol{B}^*$ and residuals $\boldsymbol{E}^*$ of the unadjusted model and subsequently applies a singular value decomposition to the concatenation of the coefficient and residual matrices $\hat{\boldsymbol{R}} = [\hat{\boldsymbol{B}}^*, \hat{\boldsymbol{E}}^*]$ (which is in contrast to standard SVA that computes singular values for the residuals only). It associates the k largest singular values with cell-composition. They can be corrected for when computing association with a phenotype of interest. In RnBeads, the RefFreeEWAS method can be enabled as an alternative method for inferring p-values in the differential methylation analysis module.

Third, the FaST-LMM-EWASher software provides an alternative, reference-free approach for associating DNA methylation with a phenotype of interest and correcting methylation heterogeneity that is due to confounders [Zou *et al.* 2014]. It is based on linear mixed models which incorporate variation in the data explained by the first principal components:

$$\boldsymbol{x} = \beta_{Yj}\boldsymbol{y}_j + \boldsymbol{Z}^T\boldsymbol{\beta}_{Zj} + \sum_{l=1}^{L} \boldsymbol{A}_l\lambda_l v_l + \frac{1}{\sqrt{m}}\tilde{\boldsymbol{Y}}^T\boldsymbol{u} + \boldsymbol{\epsilon}_j \tag{3.3}$$

Here, $\boldsymbol{x}$ specifies the vector containing the phenotype of interest for each of n samples and $\boldsymbol{y}_j$ contains methylation levels at CpG j with β_{Yj} as the corresponding coefficient to be estimated. $\boldsymbol{Z}$ is a matrix with column vectors of d' known covariates with corresponding coefficients $\boldsymbol{\beta}_{Zj}$. $\tilde{\boldsymbol{Y}}$ is a matrix of methylation values for all m CpGs and n samples and has been standardized to have mean 0 and unit variance for each row (CpG). $\boldsymbol{u}$ is

a vector of random effects and accounts for confounding factors in the methylation matrix. These effects are assumed to be identically and independently distributed, sampled from a Gaussian with variance σ_u^2. $\boldsymbol{\epsilon}_j$ is a vector of Gaussian random noise with variance $\sigma_{\epsilon j}^2$. In [Zou *et al.* 2014], the authors show that standard linear mixed models only inadequately capture confounding when the sample size is large. Therefore, a key aspect of the model is to augment the standard linear mixed model with a term that explicitly accounts for confounding by modeling the variation along the first L principal components of the data (sum term in Equation 3.3). In this term, λ_l denotes the lth eigenvector of the data covariance matrix $\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^T$ with corresponding principal component $\boldsymbol{A}_l$. The effect corresponding to the lth principal component is captured by the coefficient v_l to be estimated. The sum term then corresponds to a low-rank approximation of confounding effects in the data: $\sum_{l=1}^{L} \boldsymbol{A}_l \lambda_l v_l \approx \frac{1}{\sqrt{m}} \tilde{\boldsymbol{Y}}^T \boldsymbol{u}$. The number of top principal components L is estimated by an iterative approach. Due to software licensing issues, we could not implement the FaST-LMM-EWASher directly into RnBeads. Instead, RnBeads can export preprocessed DNA methylation data in a format that can be directly loaded into FaST-LMM-EWASher. It is important to note that these reference-free methods model cell-mediated associations without explicitly knowing the concept of a cell type. Furthermore, they entail strong linearity assumptions. It is therefore hard to determine to what extent they capture actual cell-type contributions in contrast to other sources of variation and the results from the above analysis methods should be carefully checked for statistical as well as biological plausibility.

Furthermore, the process of aging is associated with characteristic DNA methylation signatures [Horvath 2013]. RnBeads can predict a sample's biological age using an elastic-net predictor that was inferred from a large panel of datasets [Scherer 2016]. The predicted age can be correlated with the annotated age (if available). Deviation from the annotated age can indicate potential sample mix-ups or a biologically relevant phenotype. For instance, methylation patterns indicative of accelerated aging have been associated with cancer [Horvath 2013]. Furthermore, the inferred age can be used as a covariate in exploratory and differential analyses in order to identify and adjust for systematic methylation that is due to aging. Optionally, a dataset-specific predictor can be trained and validated from the available data. The predictors have been extensively validated for 450K data and we have extended their applicability to genome-scale bisulfite sequencing data [Scherer 2016] (unpublished work together with Michael Scherer).

3.2.2 Implementation Details and Package Design

RnBeads and its companion data packages currently comprise a base of approximately 32,000 lines of R code, including more than 200 exported functions, classes and methods. To structure all functionality in a flexible and easily understandable way, RnBeads utilizes elements of object-oriented programming available in R. Specifically, all DNA methylation data is organized in an `S4` class hierarchy. Each analysis module is implemented as an independent unit operating on an `RnBSet` object. Module results are written to hypertext reports that employ XHTML and JavaScript to enable self-contained interactivity. The RnBeads reports include figures, which are collections of related plots spanning relevant parts of the parameter space. This setup enables users to dynamically explore the parameter space of each figure without the need to rerun the analysis. Dedicated R packages such as ggplot2 [Wickham 2009] are used to generate publication-grade plots, which are incorporated in the reports as bitmaps for quick visualization and as vector graphics for high-resolution printing and further processing.

All exported functions are documented and examples are provided. Detailed tutorials and executable examples can be found on the package website[8].

3.2.3 Scalability and Performance

RnBeads was designed to scale with large sample sizes. Parallel computation is implemented using the foreach and doParallel packages. Moreover, large R objects can be maintained directly on hard disk using the ff package. Small-scale analyses can be completed on a standard personal computer, whereas analyses of large datasets are recommended to be executed on a scientific computing cluster or on adequately powered cloud computing infrastructure. RnBeads implements convenience functionality to directly distribute analysis tasks to nodes of a scientific computing cluster. For users who prefer a web-based workflow or who lack access to a suitable infrastructure, a web server supporting analyses with up to 24 samples is available.

RnBeads has been tested successfully on Infinium data sets comprising thousands of samples and on RRBS and WGBS data sets with hundreds of samples. Table 3.1 lists runtime measurements of RnBeads for several large datasets. Although this benchmarking has been performed with an earlier version of RnBeads (version 0.99.15) and corresponding analyses complete much faster using a current version of the package (due to code optimization and maintenance that has been performed in the meantime) the numbers provide a good indication on how the runtime depends on analysis parameters such as the number of CpGs analyzed or the number of sample annotations provided. Not unexpectedly, increasing the number of categorical columns in the sample annotation table used for differential and exploratory analysis increases the number of comparisons and the number function calls and hence leads to an overall increase in runtime. Notably, this increase is more drastic than the increase incurred by higher numbers of CpGs. Time-critical steps that are not essential to every RnBeads analysis (e.g. plotting variability distributions or composite plots) can be disabled using package options resulting in significantly reduced running times. For instance, disabling site-specific exploratory and differential analyses and using a current version of RnBeads (version 1.5.0) a dataset of 478 WGBS profiles (cf. Chapter 4.4) could be processed in less than three days on the same computing cluster.

3.2.4 Methylome Resource

A methylome resource was established by applying RnBeads to some of the largest public datasets that are currently available for WGBS, for RRBS and for the 450K assay. This resource provides a reference for large-scale DNA methylation analyses that can be used in various ways. For example, researchers can browse through the reports online, explore biological hypotheses, and investigate relevant aspects of the data visually or through custom data analysis with R or other software tools. Furthermore, researchers can download the data and configuration files of the Methylome Resource, add their own DNA methylation data and then run RnBeads in order to analyze their data in the context of high-quality methylome data sets that span a broad set of tissue types. For the 450K assay, we downloaded raw intensity files for 4,034 primary tumor and normal control samples which have been collected by the The Cancer Genome Atlas (TCGA) consortium [Weisenberger 2014]. Additionally, we obtained RRBS DNA methylation profiles for 216 samples with coverage of 2,295,083 CpGs from the ENCODE project [ENCODE

[8] http://rnbeads.mpi-inf.mpg.de

Table 3.1: Performance benchmark for large DNA methylation analyses with RnBeads

Data type[a]	No. of samples[b]	No. of CpGs[c]	No. of annotations[d]	No. of comparisons[e]	Runtime (node)[f]	Runtime (cluster)[g]
450k	100	482,421	2	2	7h 5m	2h 24m
450k	500	482,421	6	6	1d 8h 14m	12h 2m
450k	1000	482,421	10	10	3d 2h 49m	1d 6h 23m
450k	4034[h]	482,421	4	18	25d 11h 3m	8d 23h 22m
RRBS	10	1,742,404	2	2	7h 47m	2h 11m
RRBS	50	2,162,686	6	6	1d 1h 52m	6h 56m
RRBS	100	2,221,889	10	10	1d 23h 9m	12h 17m
RRBS	216[h]	2,295,083	7	11	2d 15h 26m	22h 46m
WGBS	5	28,133,531	2	2	10d 4h 16m	2d 11h 53m
WGBS	10	28,150,019	6	6	32d 13h 26m	8d 6h 13m
WGBS	20	28,153,044	10	10	55d 14h 42m	13d 21h 33m
WGBS	41[h]	28,158,385	4	6	31d 21h 22m	7d 19h 35m

Data as of May 2014. RnBeads version 0.99.15 was used.

[a] Data from the following sources were included in the analysis: TCGA (450k), ENCODE (RRBS), Ziller *et al.* [2013] (WGBS)

[b] Subsets of the full datasets were generated by random sampling in order to assess the effect of sample size on runtime

[c] Number of sites represented in at least one sample

[d] Adding more columns to the sample annotation table increases the complexity and runtime of an analysis

[e] Including more pairwise comparisons for differential analysis of sample groups in the analysis strongly increases runtime but can be parallelized effectively

[f] Serial runtime measured on a computing cluster (16 nodes), summing up the runtime of all contributing nodes

[g] Parallel runtime/time to completion on a computing cluster (16 nodes)

[h] Full dataset available from the RnBeads methylome resource website

Project Consortium 2012], which comprises cell lines and primary samples of various normal and cancerous tissue types [Varley *et al.* 2013]. The resource also encompasses a dataset of 41 samples across a broad range of human cell types, sequenced using WGBS with total coverage of more than 28 million CpGs [Ziller *et al.* 2013]. Finally, 81 blood-related cell types were characterized using RnBeads in the context of the BLUEPRINT project and a detailed analysis of this dataset is provided in Section 4.2. All data was processed according to a standardized RnBeads workflow that could be completed in a few days on a scientific computing cluster (Table 3.1).

These types of resources are particularly valuable for researchers who have generated specialized DNA methylation datasets and want to assess data quality and/or biological relevance in context of a broad range of reference methylomes. The concept of preconfigured and rerunnable analyses of reference epigenome data also provides the means for making data from large-scale epigenome mapping projects more accessible to smaller-scale and mechanism-centered studies, thereby contributing to reproducibility, data sharing and the broader relevance of large-scale epigenome mapping projects.

3.2.5 Availability

RnBeads is available under the GPLv3 open source license and is part of Bioconductor [R. C. Gentleman *et al.* 2004]. The package vignette documents and tutorial sessions provide a detailed introduction into RnBeads and its modules and describe example analysis on basic as well as advanced levels. The RnBeads website[9] contains supplementary information, the package vignettes, tutorials, a web service for smaller-scale RnBeads analysis, an FAQ section, example analysis reports as well as the methylome resource of reports.

3.2.6 Use Case: Analysis of DNA Methylation During Adult Stem Cell Differentiation

To illustrate the practical use of RnBeads, we applied the software to datasets for which the underlying biology is relatively well understood. The example described here focuses on an RRBS dataset assessing the DNA methylation dynamics of blood and skin stem cell differentiation in mice [Bock *et al.* 2012]. This dataset comprises 13 blood and 6 skin cell populations at various stages of adult stem cell differentiation. DNA methylation of approximately two million CpGs in each sample was measured in biological replicates. The global distribution of DNA methylation is characteristically bimodal (Figure 3.4a). Discrete peaks at 33 %, 50 % and 67 % DNA methylation disappear after removing CpGs with low sequencing coverage in the preprocessing step. Exploratory analysis confirms that the difference between blood and skin cell types dominates the analysis (Figure 3.4b). DNA methylation levels are generally higher in blood cells than in skin cells when computing regional averages over all annotated genes, particular in genomic regions representing gene bodies (Figure 3.4c). Hierarchical clustering perfectly discriminates between blood and skin cell types (Figure 3.4d), confirming that DNA methylation patterns tend to be associated more strongly with cell lineage than with other properties such as cellular proliferation or differentiation state.

RnBeads also identifies DMRs that are statistically significant and exhibit pronounced DNA methylation differences between the two lineages. This analysis has been performed for single CpGs and also for sets of predefined genomic regions such as CpG islands, genes, promoters and genome-wide tiling regions. Such DMR analyses, which

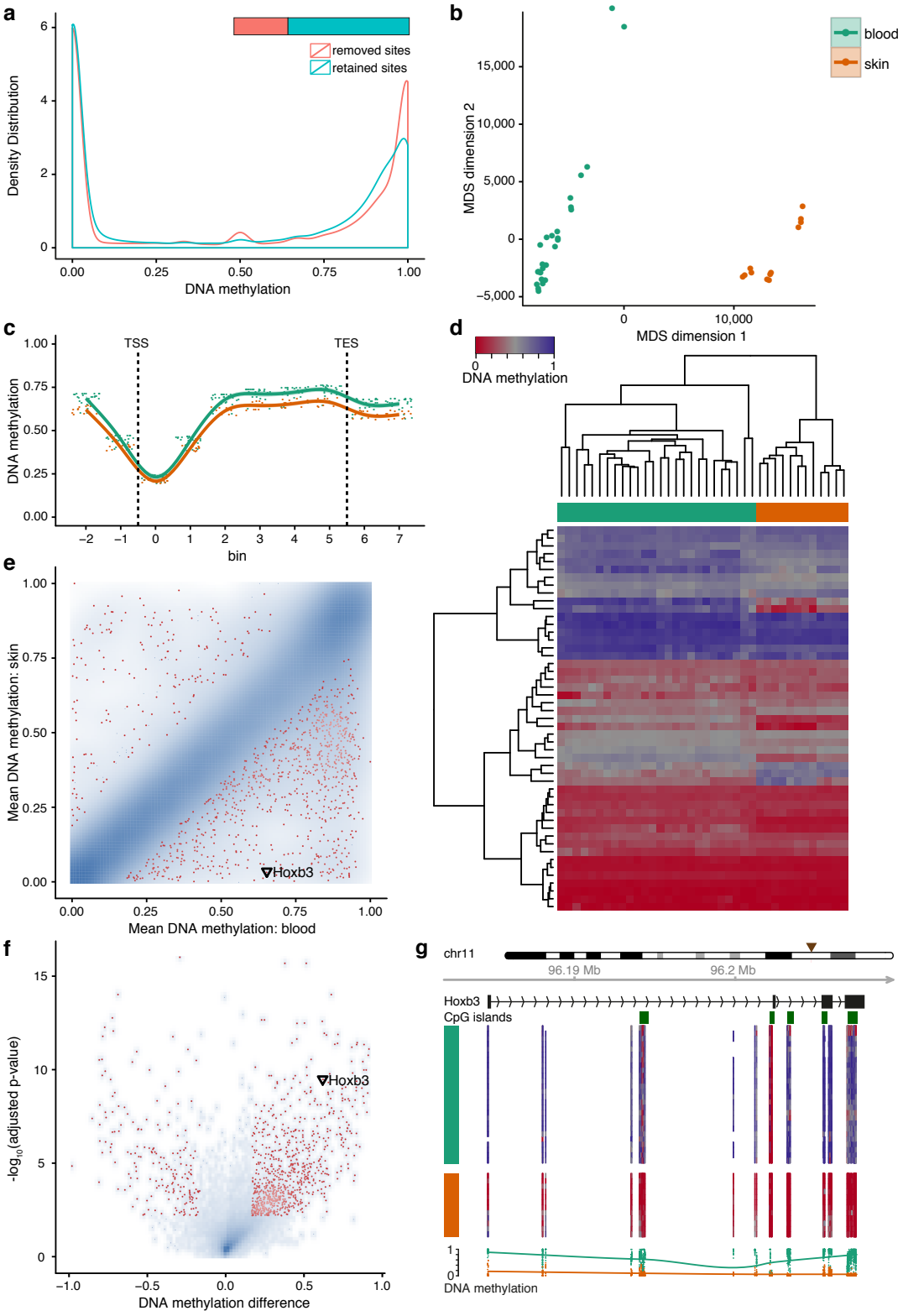
[9] http://rnbeads.mpi-inf.mpg.de

are based on regions of interest, provide an effective way of increasing the statistical power to detect differential DNA methylation, and they also increases the interpretability of identified DMRs [Bock 2012]. RnBeads' priority-ranked list of DMRs between blood and skin cell types contains many genes with established roles in blood and skin tissues, such as members of the homeobox and keratin gene families. Scatterplots illustrate the overall frequency of DMRs for a region type of interest (Figure 3.4e shows data for gene loci as regions of interest) and volcano plots provide a convenient way of visualizing the relationship between effect size and significance of the DMRs (Figure 3.4f). The *HOXB3* gene is highlighted as an example of blood-specific DNA methylation and Figure 3.4g illustrates how the `rnb.plot.locus.profile(...)` function of RnBeads can be employed to produce locus-based views of methylation. It thus provides an example of how RnBeads' utility functions and `RnBSet` objects, which are produced by the standard pipeline, can be used in custom R scripts for further data exploration.

3.2.7 Discussion

RnBeads provides an integrated framework for comprehensive DNA methylation analysis that is in accordance with the principle of reproducible research. Due to its modularized design and versatile analysis options, the package offers extensibility and flexibility for both first-time users and experienced researchers. The generated hypertext reports provide a convenient way of obtaining a general overview of any DNA methylation dataset on the basis of single CpGs or genomic regions of interest. Moreover, they facilitate browsing, sharing and archiving individual analyses and thus make RnBeads a suitable tool to be used by bioinformatics core facilities. Notable applications of the software include scenarios in which large datasets are common, such as the analysis of

Figure 3.4 : *(On the next page)* Analysis of DNA methylation during adult stem cell differentiation. RnBeads was used to reanalyze RRBS data comprising 19 cell types of the blood and skin lineages [Bock *et al.* 2012]. All plots have been generated with RnBeads, but have been reformatted. **(a)** Global distribution of DNA methylation levels among retained and removed CpGs after the preprocessing step. **(b)** Relative similarity and differences of DNA methylation profiles between cell types. Two maximally informative dimensions were calculated using MDS based on the matrix of average methylation levels in 5-kb tiling regions. **(c)** Composite plot of DNA methylation levels in blood and skin cell types averaged across all genes. Each gene was covered by six equally sized bins and by two flanking regions of the same size. Smoothing was done using cubic splines. **(d)** Heatmap and hierarchical clustering of DNA methylation levels among lineage marker genes that are specifically expressed in the blood lineage. Clustering employed average linkage and Manhattan distance. **(e)** Scatterplot of groupwise mean DNA methylation levels across genes. The 1,000 highest-ranking differentially methylated genes are highlighted in red. Point density is indicated by blue shading. **(f)** Volcano plot illustrating effect size and statistical significance across genes. Coloring as in (e). **(g)** DNA methylation profile of the *HOXB3* gene locus on chromosome 11 (triangle). Heatmaps show DNA methylation levels of single CpGs according to the color scheme in (d). Smoothing of DNA methylation levels (bottom) was done using cubic splines.


a
Density Distribution
removed sites
retained sites
DNA methylation
b
blood
skin
MDS dimension 2
MDS dimension 1
c
TSS
TES
DNA methylation
bin
d
DNA methylation
e
Mean DNA methylation: skin
Mean DNA methylation: blood
Hoxb3
f
-log10(adjusted p-value)
DNA methylation difference
Hoxb3
g
chr11
96.19 Mb
96.2 Mb
Hoxb3
CpG islands
DNA methylation

EWAS and epigenetic biomarker discovery in cancer cohorts. Finally, RnBeads is well-accepted by the scientific community: at the time of writing this thesis, the package is downloaded 500 to 600 times by approximately 200 to 300 unique users every month. According to Google Scholar, the corresponding article [Assenov *et al.* 2014] has been cited 73 times since its publication (status: November 2016).

3.3 Global Analysis of Epigenomic Marks in Repetitive Elements

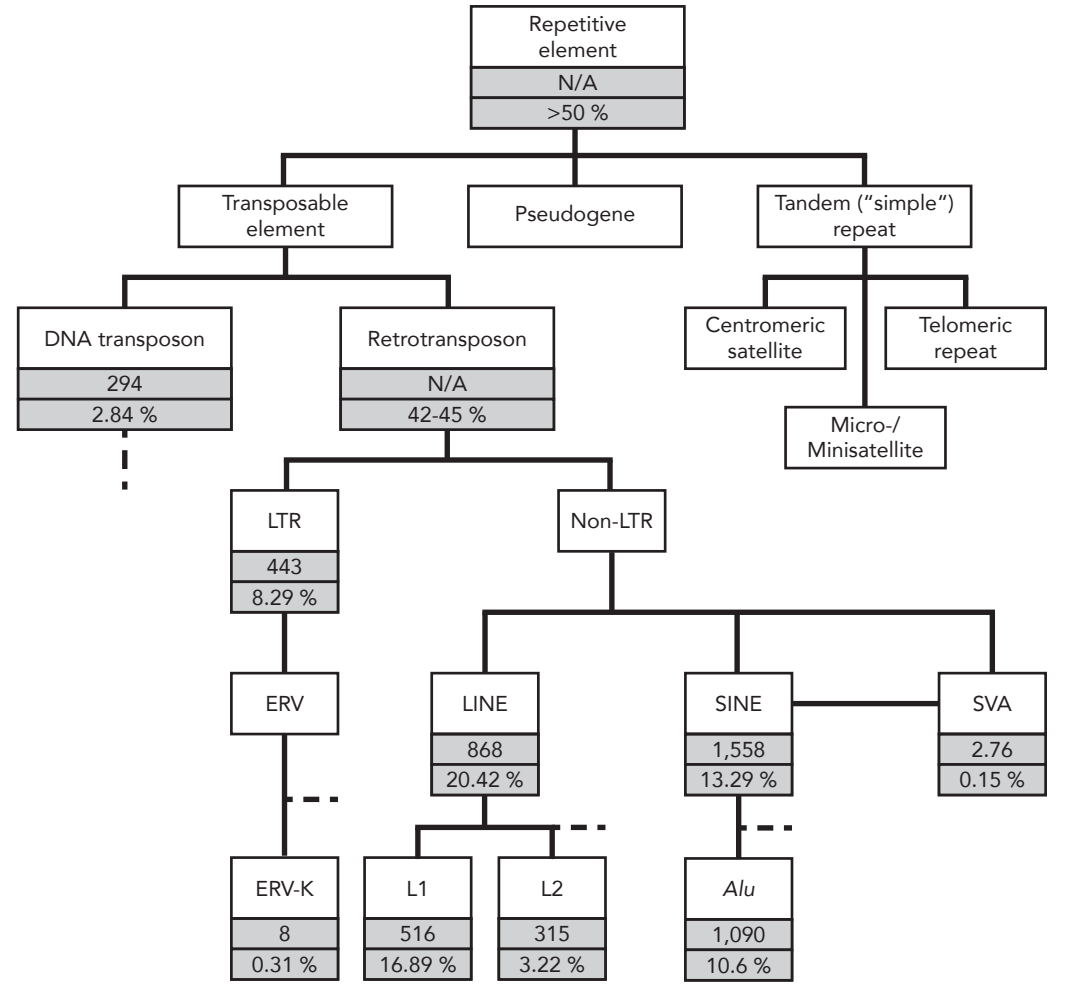
Repetitive elements are an integral part of the human genome [Lander *et al.* 2001]. In the past, they have been referred to as "junk DNA". However, recent evidence suggests that these elements play a vital role in structuring the genome and contribute to shaping genome architecture in the course of evolution. They are not only subject to epigenetic regulation, but are also themselves functionally involved in the epigenetic regulation of gene expression — a discovery that was acknowledged in 1983 when the Nobel Prize in Physiology or Medicine was awarded to Barbara McClintock for her seminal work on transposons and their role in phenotype regulation in maize plants [McClintock 1950].

In order to decipher the genome-wide regulatory patterns in repetitive elements we have devised computational methods for quantifying and analyzing epigenetic marks corresponding to the sequences of repeat subfamilies. These methods have been implemented in epiRepeatR, one of the first software packages for this type of analysis.

3.3.1 Repetitive Elements and Epigenetic Regulation

Repetitive elements in the human genome can be broadly grouped into **transposons**, **pseudogenes** and **simple repeats** [Jurka *et al.* 2011] (Figure 3.5). Transposons or Transposable Elements (TEs) are estimated to constitute approximately 50 % of the human genome [Lander *et al.* 2001; Mandal and Kazazian 2008]. A more recent study estimates that up to 69 % of our DNA could originate from repeats [Koning *et al.* 2011]. A large fraction of today's knowledge on TEs has been derived from plant studies [Fedoroff 2012]. However, this thesis focuses on repetitive elements in mammalian genomes. TEs possess the ability of changing their positions in the genome. When these insertions occur in the germline, TEs gradually amass copies of themselves during the course of evolution and have therefore have been described as "selfish DNA". The entirety of TEs in the genome has been referred to as the "mobilome" [Cowley and Oakey 2013; Burns and Boeke 2012]. Due to mutations and mechanisms of epigenetic silencing most of these elements have become degenerate and inactive in human. However, certain types of transposons are still mobile.

TEs can be classified by their structure and mode of transposition [Cowley and Oakey 2013; Z. Wang and Kunze 2015; Goodier and Kazazian 2008]. They are typically categorized into families and subfamilies based on their sequence [Jurka *et al.* 2005; Jurka *et al.* 2011]. **DNA transposons** change their position in the genome by a *cut-and-paste* mechanism and make up approximately three percent of human DNA. Their sequence length is typically in the range of one to several kilobases. There is currently no type of DNA transposon known to be active in the human genome. In contrast, **retrotransposons** are believed to be of retroviral origin and employ a *copy-and-paste* mechanism to replicate via an RNA intermediate and to reinsert DNA in a different locus. Thus, during the course of evolution they accumulated a plethora of copies in the genome. 42 % to 45 % of human DNA is estimated to be of this origin [Lander *et al.* 2001; Mandal and Kazazian 2008; Burns and Boeke 2012] (Figure 3.5). They can be further classified into **Long Terminal**



Active subfamilies	HERV-K	L1, L1HS	*AluY*	SVA-A to SVA-F
# specfic insertions	73	~2,000	~7,000	~1,000
Frequency of inserstion per birth		1/200 to 1/20	1/20	1/900

Figure 3.5: Hierarchy of repetitive elements in the human genome. Grey boxes specify the abundance of repeat categories in terms of number of copies (in thousands) and percent of the human genome covered. Dashed branches are not referred to in the main text and are truncated. The table at the bottom shows active transposon subfamilies along with their numbers of insertions specific to the human genome and frequency of germline insertions. The numbers were compiled from [Mandal and Kazazian 2008; Lander *et al.* 2001; Burns and Boeke 2012; Beck *et al.* 2011; Mills *et al.* 2007; Jurka *et al.* 2011].

Repeat (LTR) transposons and **non-LTR transposons**. The basis for this classification is the presence or absence of stretches of characteristic, repetitive DNA of 300 to 1,200 basepairs at the terminal ends of their sequences [Cowley and Oakey 2013]. The group of LTR transposons consists of **Endogenous Retroviruses (ERVs)**. These elements originate from ancient viral germline infections. Their sequences are typically seven to nine kilobases in length and they contain gene loci typical for viruses, such as *Gag*, *Pol* and *Env*. LTR transposons can occur in full length or as fragments ("solo" LTRs) and LTR-derived sequences account for about 8.3 % of the human genome. Recent transposition events have only been described for **human ERV (HERV)** subfamilies, in particular the HERV-K subfamily, while other LTR transposons are inactive in human. Notably, it has been reported that particularly those elements become active in the germline that exhibit high copy numbers throughout the genome [Thompson *et al.* 2016]. **Long Interspersed Nuclear Elements (LINEs)** are a group of non-LTR transposons. Their sequence (6 kb) contains two Open Reading Frames (ORFs) coding for the machinery required for their transposition and they are therefore considered autonomous elements. 868,000 LINE copies make up approximately 20.4 % of the human genome. While several LINE subfamilies exist, only members of the L1 subfamily are known to be active in human and germline insertions are found in 0.5 to 5 % of births [Beck *et al.* 2011]. With almost 1.6 million copies, **Short Interspersed Nuclear Elements (SINEs)** are the most numerous class of repeats in the human genome, covering approximately 13.3 % of it. They are also the most active: there is approximately one germline insertion of an *Alu* element in every 20 births. *Alu*s, the most prominent members of the SINE class, are short in sequence (~300 bp) and do not encode any proteins. For retrotransposition they use the machinery supplied by L1 elements and are therefore termed non-autonomous. LINEs and SINEs probably integrated into our ancestor's genome about 150 to 80 million years ago [Burns and Boeke 2012]. Furthermore, sequences of **SVA elements** constitute another class of non-LTR transposons. They are 700 bp to 4 kb in length and their acronym derives from the fact that these elements are composed of stretches of DNA originating from SINEs, a variable number of tandem repeats and *Alu*s. SVA elements are non-autonomous and utilize the L1 transposition machinery. Only a few thousand copies are present in the human genome. Nonetheless, there is about one insertion in every 900 births [Beck *et al.* 2011].

The group of **Pseudogenes** is characterized by sequence similarity to other genes in the genome. However, pseudogenes are often non-functional due to accumulation of mutations during the course of evolution. They can arise from gene duplication events. Complementary DNA (cDNA) copied from mRNA (and in some cases ncRNA) can also be incorporated into the L1 machinery and leads to the formation of **processed pseudogenes** [Beck *et al.* 2011; Z. Wang and Kunze 2015].

In contrast to transposons, **simple repeats** are short tandem repeats and typically do not possess the ability of translocation. Many such **satellite** repeats are located in the pericentromeric, heterochromatic regions of the genome where they exist in tens of thousands of copies and play a role in centromere assembly [Z. D. Smith and Meissner 2013]. Similarly, telomeric repeats can be found at the chromosomal ends and their length, which is maintained by telomerase enzymes, is an indicator for cellular age. Furthermore, short tandemly repeated stretches of one to six or 15 or more basepairs are can be found throughout the genome. They are referred to as **microsatellites** and **minisatellites**, respectively [Jurka *et al.* 2011].

In human, only 35 to 40 subfamilies of L1, *Alu*, SVA and potentially HERV elements are actively transposing [Mills *et al.* 2007]. Elements that only have been introduced into

our ancestors' genomes only recently and that are specific to the human species are particularly active [Burns and Boeke 2012]. It has been estimated that there are about 2,000 L1 insertions, 7,000 *Alu* insertions, 1,000 SVA insertions and 73 LTR insertions specific to human [Burns and Boeke 2012; Beck *et al.* 2011] (Figure 3.5). Comparative genomics approaches have identified several thousand polymorphic transposition events in human populations [Beck *et al.* 2011; Mills *et al.* 2007] and transposition has been associated with multiple diseases, such as hemophilia and cancer [Lee *et al.* 2012; Beck *et al.* 2011; Hancks and Kazazian 2012]. Somatic insertions under normal conditions have been reported in ESCs and neural progenitor cells [Goodier and Kazazian 2008].

While transposons have been primarily considered parasitic elements in the past, it is now assumed that they play a substantial role in driving genome evolution by exaptation [Bourque 2009; Fedoroff 2012; Xie *et al.* 2013]. Transposition can have structural and gene-regulatory effects: large-scale genome rearrangements, duplications and deletions can be the consequence of recombination events that can occur when homologous sequences, such as TE loci are in close spatial proximity (non-allelic homologous recombination) [Goodier and Kazazian 2008]. TE insertions can disrupt the ORFs of genes, introduce novel exons and lead to alternative splicing [Cowley and Oakey 2013; Goodier and Kazazian 2008; Beck *et al.* 2011]. Many transposition events also result in target site alterations, e.g. via deletion or insertion of short stretches of DNA [Goodier and Kazazian 2008; Beck *et al.* 2011; Z. Wang and Kunze 2015], and the machinery responsible for transposition of L1 and SVA elements can transduce genomic sequences flanking the locus of origin [Beck *et al.* 2011]. Furthermore, regulatory elements such as promoters and polyadenylation signals can be introduced into novel genomic contexts. Additionally, TEs frequently associate with DNase hypersensitive sites and TFBSs [Jacques *et al.* 2013]. Nearly a third of all binding sites for the ESR1, p53, OCT4, SOX2 and CTCF transcription factors are embedded in TE sequences [Bourque *et al.* 2008; T. Wang *et al.* 2007]. Particularly LTR elements contain regulatory elements such as TFBSs and can function as alternative promoters [Thompson *et al.* 2016]. It has been hypothesized that these transposons could be involved in the formation of tissue-specific enhancers. Finally, it has also been proposed that transcribed RNA pertaining to pseudogenes and other elements can regulate gene expression by competitive binding of miRNAs [Salmena *et al.* 2011].

Considering the potential consequences of transposition mentioned above, it is not surprising that TEs are subject to tight epigenetic regulation. In general DNA methylation and histone modifications indicative of closed chromatin have been associated with the silencing of these elements. In mammals, the majority of repetitive DNA is highly methylated and transposons are subject to *de novo* DNA methylation by DNMT3 enzymes in germline and in the early embryo [Yoder *et al.* 1997; Z. D. Smith and Meissner 2013]. If DNA methylation is depleted in ESCs derived from Dnmt1 knockout mice, Intracisternal A-particle (IAP) transposons become transcriptionally active [Burns and Boeke 2012], indicating epigenetic silencing. Particularly, satellite repeats in the mammalian pericentric heterochromatin exhibit high levels of DNA methylation and methylation of H3K9 [Z. D. Smith and Meissner 2013] and these modifications presumably prevent expression of these elements. RNAi represents another important mechanism for keeping transposition at bay. First discovered in *Drosophila*, but also observed in mammals, PIWI-Interacting RNAs (piRNAs) can direct *de novo* DNA methylation to TEs in order to silence them [Burns and Boeke 2012]. Still, in sperm cells and, to a lesser extent, also in ESCs blocks of CpG-dense DNA located in ERV, LINE, SINE and SVA elements exhibit low methylation levels [Molaro *et al.* 2011]. Particularly evolutionarily young

elements and full-length elements with high sequence similarity to the consensus sequence of the respective subfamily tend to have lower methylation levels, potentially indicating silencing escapees. Patterns of dynamic chromatin organization have also been found in ESCs, where LTR transposons are subject to trimethylation of H3 histone proteins, particularly H3K9me3 [Thompson *et al.* 2016]. These elements exhibit tissue specific expression and become active during embryonic pre-implantation and in the placenta. Similarly to their genic counterparts, epigenetic patterns of activation, such as H3K4me3 and chromatin accessibility, can be found in these elements. In addition, they are frequently associated with the transcription of lncRNAs. Epigenomic patterns can become deregulated in disease and hypomethylation in transposons has been associated with cancer onset and progression [Cowley and Oakey 2013; Sharma *et al.* 2010]. Here, derepression of transposition may contribute to altered genotypes in cancer cells which could result in tumor growth and malignancy.

Fedoroff [2012] argues that epigenetic silencing mechanisms might have contributed significantly to the expansion of TEs in eukaryotic genomes. Since certain repetitive elements have been shown to escape the global demethylation that occurs during epigenomic reprogramming in germline development, it has also been speculated that these elements can act as messenger vessels for trans-generational epigenetic inheritance [Cowley and Oakey 2013].

3.3.2 Computational Methods for the Analysis of Repeat Epigenomes

Despite the regulatory importance of epigenomic patterns in repeats only few computational methods exist for their analysis. In this work, we present one of the first bioinformatic approaches for quantifying DNA methylation levels and histone modifications in repetitive elements [Bock *et al.* 2010]. Since our initial implementation, a number of alternative approaches employing similar strategies to characterize repeat epigenomes have emerged.

Rosenfeld *et al.* [2009] mapped ChIP-seq reads to consensus sequences of centromeric and telomeric repeats. Histone modifications associated with heterochromatin, such as H3K20me3 and H3K9me3 were found to be prevalent in centromeric repeats and are presumably preventing gene expression in these regions. It is also speculated that certain histone modifications in centromeric regions could be associated with chromosomal replication. In contrast, the activating marks H2BK5me1 and H3K4me3 were associated with telomeric repeats and were consistent with findings of active transcription in telomeres. Day *et al.* [2010] also aggregated repetitive elements into consensus sequences. They quantified enrichment of histone modifications in repetitive elements by aligning ChIP-seq reads to these references. By also including flanking sequences of repeat instances, the number of aligned sequencing reads could be significantly increased, because reads mapping to the boundary regions could be mapped more accurately. They further employed a phylogenetic approach to combine repeat types into hierarchically organized groups. Using this approach more reads could be uniquely assigned to individual groups and histone mark enrichment could be quantified on the aggregate level rather than on the level of individual repeat types. They applied their approach to published histone modification data from mouse ESCs and found an enrichment of H3K9me3 and H4K20me3 marks in the ERV-K and ERV1 transposon subfamilies while ERV-L repeats enrich for H3K27me3. These repeat groups contain actively transposing elements in mouse and the authors explain these differences in histone modification patterns by alternate silencing mechanisms associated with these marks. Findings from their repeat-focused reanalysis of ChIP-seq in human $CD4^+$ T cells include an

enrichment of H3K4me1 in *Alu* repeats as well as H4K20me1 and H2BK5me1 in SVA elements. The co-localization of these elements with CG-rich regions of the genome could explain these patterns which are generally associated with active chromatin. Furthermore, consistent with Rosenfeld *et al.* [2009], the repressive marks H3K20me3 and H3K9me3 were found enriched in different satellite repeats. Xie *et al.* [2013] employed a similar, consensus-based strategy: a reference of the repetitive portion of the genome along with sequences flanking the instances of repeats was constructed. Reads which mapped to multiple positions in the genome were assigned to subfamilies of TEs if their best mapped positions only correspond to members of one subfamily. The authors quantified DNA methylation by MeDIP-seq and MRE-seq in 11 different human cell types. They identified tissue-specific hypomethylation in LTR elements and more specifically in ERV transposons. These elements co-localized with genes with cell-type specific functions and tissue-specific enhancer signatures such as H3K4me1 and p300 occupancy. Other approaches focus on the alignment problem itself and employ Gibbs sampling [J. Wang *et al.* 2010] or an iterative reweighting algorithm [D. Chung *et al.* 2011] in order to resolve mapping ambiguities induced by repetitive DNA.

However, to the best of our knowledge no integrated software framework exists that can readily be applied to large epigenome datasets originating from bisulfite sequencing as well as enrichment-based methods, such as ChIP-seq. We therefore decided to build on our initial analysis efforts and developed the epiRepeatR software package, which is described in the following sections.

3.3.3 A Pipeline for the Analysis of Epigenomic Marks in Repetitive Element Subfamilies

In order to obtain a global perspective of epigenetic patterns in the repetitive portion of the genome, we devised a computational pipeline that aggregates measurements obtained by bisulfite-sequencing as well as enrichment-based sequencing assays (e.g. ChIP-seq, MeDIP-seq, etc.) in repeat subfamilies (Figure 3.6). It provides methods for inspecting epigenetic variation across subfamilies and for comparing theses patterns between different groups of samples. The pipeline has been implemented in the epiRepeatR software package which implements an easy-to-use workflow for analyzing epigenetic patterns in repetitive elements across samples.

epiRepeatR employs one of two alternative approaches to quantify the signal for each epigenomic mark in each subfamily (details are provided below). Subsequently, the pipeline employs steps for quality control based on the coverage of repetitive elements. Heatmap plots visualizing DNA methylation levels and enrichment scores facilitate the exploration of epigenetic dynamics across repeat subfamilies and samples. Provided with corresponding annotation, differences in the epigenetic patterns between groups of samples in different repeat subfamilies can be visually inspected. Furthermore, unsupervised learning techniques, such as dimension reduction and sample clustering, provide the means of exploring inter-sample relationships based on the epigenomic profiles of repeats.

Quantification of Epigenomic Marks in Repeat Subfamilies

We implemented two alternative approaches for quantifying epigenomic marks in repetitive elements:

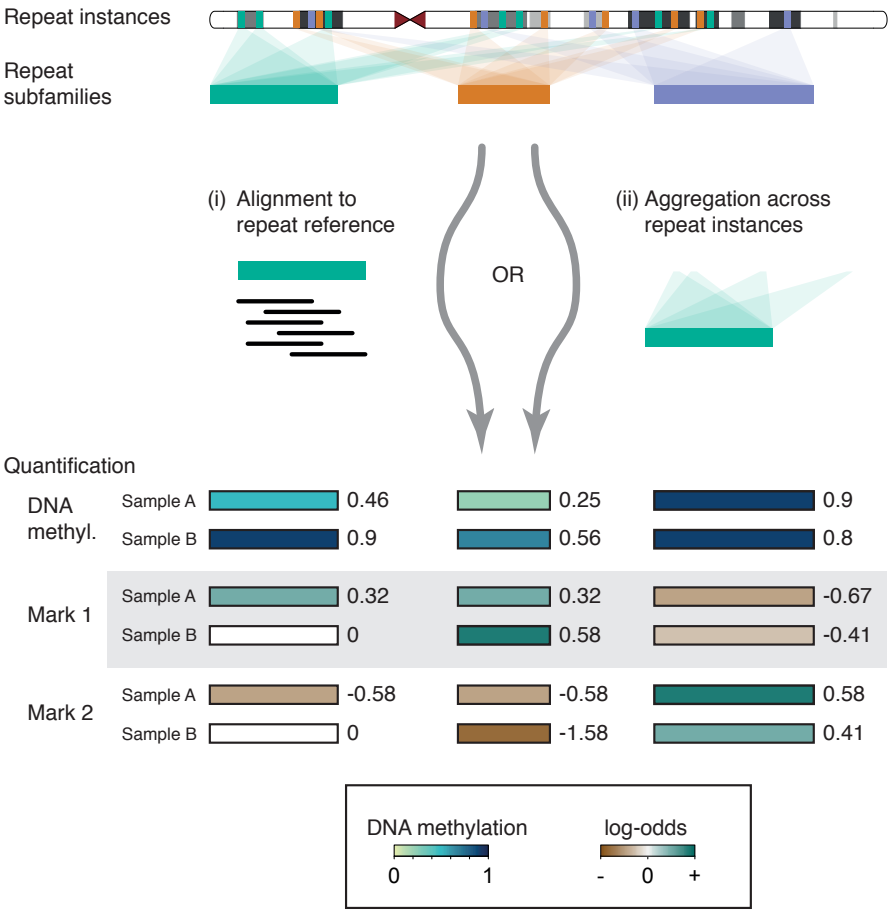


Figure 3.6: A pipeline for the analysis of epigenomic marks in subfamilies of repetitive elements. Each subfamily represents multiple repeat instances in the genome. Two alternative approaches are employed to quantify epigenetic marks in each subfamily. The schema illustrates hypothetical, constructed examples of three repeat subfamilies (left to right) assayed with bisulfite-sequencing and ChIP-seq for two different epigenetic marks. Quantitative measurements for methylation levels and ChIP-seq enrichment of the two marks are color coded and shown at the bottom of the figure. Top and bottom rows for each mark denote two epigenomes (Samples A and B).

(i) **Mapping to a reference of consensus sequences:** Sequencing reads are aligned to a reference of consensus sequences for subfamilies of repetitive elements. The epigenetic signal is quantified based on a single consensus sequence for each subfamily. Using bisulfite sequencing reads, methylation levels for each CpG in the reference are quantified and averaged for each subfamily. For enrichment-based data, log-odds scores are computed from the relative number of reads aligning to the consensus sequence.

(ii) **Aggregation across genome-wide instances of repetitive elements:** The epigenetic signal for each repeat subfamily is quantified from genome-wide alignment data. Using an annotation of repeat instances across a given reference genome, signals in individual instances are aggregated in order to obtain a single score for each subfamily. In the case of DNA methylation, the input to the pipeline consists of genome-wide methylation calls and the methylation levels for each subfamily is quantified by averaging across all CpGs in all repeat instances of that subfamily. For enrichment-based data, read counts for each subfamily are obtained by summing up the numbers of reads aligned to individual instances and a combined log-odds score is computed.

The approaches employ two opposing model assumptions. In approach (i) we assume that repeat instances are derived from ancestral sequences. For accurate signal quantification, individual repeat instances should share high sequence conservation within each subfamily. In contrast, approach (ii) requires sequences that have diverged from the consensus and allows for sequence variability across instances. If the instances of a subfamily exhibit high sequence similarity, it is likely that reads cannot be uniquely assigned to a specific instance, resulting in potential coverage biases between subfamilies.

The consensus-based approach (i) is highly dependent on a reference of consensus sequences that adequately represent each repeat subfamily. In epiRepeatR, sequencing reads are aligned to a reference obtained from the Repbase Update database [Jurka *et al.* 2005]. The BSMAP [Xi and W. Li 2009] and BWA [H. Li and Durbin 2009] mappers can be used for the alignment of bisulfite-sequencing and enrichment-sequencing data respectively. Other mapping software can be integrated into the pipeline with just a few lines of code. In contrast, the aggregation-based approach (ii) requires a reference genome along with a detailed annotation of repeat instances. epiRepeatR uses the annotation provided in the RepeatMasker[10] tracks of the UCSC database [Kent *et al.* 2002]. Metadata for repeat subfamilies, such as the repeat family and species specificity, is obtained from the Repbase Update database by matching of repeat names.

In both approaches, methylation levels and log-odds for enrichment are computed for each repeat subfamily. For bisulfite-sequencing assays, CpG methylation levels are averaged to obtain aggregate levels for each subfamily. Computing the log-odds for enrichment-based sequencing methods requires reads originating from both, the enrichment experiment (ChIP, MeDIP, etc.) and a genome-wide library (Input/WCE). A log-odds score is used to quantify the relative number of reads aligning to a given subfamily relative to the relative number of input reads. Taking the logarithm results in numerical stability compared to the ratio. We define the log-odds score as

$$\text{logOdds}\,(r) = \log_2 \frac{s_r/S}{b_r/B} \approx \log_2 \frac{s_r B + \epsilon}{b_r S + \epsilon}$$

Where s_r and b_r denote the number of signal and background reads aligning to repeat subfamily r. The total number of aligned signal and background reads is denoted by

[10] http://www.repeatmasker.org

S and B respectively. ϵ is a small constant to avoid division by 0. The input signal can either be specific to each sample or a shared background is used to normalize the signal for all samples. In an optional step, epiRepeatR aggregates reads from multiple input read files into a shared background.

Implementation Details

Provided with a tabular sample annotation file that contains file names, file types and sample metadata, analyses can be started using a single command in R or a shell script. For consensus-based quantification, the input to epiRepeatR consists of BAM files of pre-processed sequencing reads to be mapped and the annotation table. For the aggregation-based approach, sequencing reads which have been aligned to a corresponding reference genome and CpG methylation calls are used for computing enrichment scores and methylation levels respectively. Due to its modular software design, individual steps in the downstream analysis can be executed based on the binary object output of the preceding steps, thus providing flexibility to rerun parts of the pipeline in case of incomplete analyses or parameter changes. epiRepeatR can run in a multi-process environment or can be distributed across multiple nodes of a computing cluster, thereby enabling parallel processing of large datasets.

Availability

epiRepeatR is currently in an advanced development state and we are evaluating the software in the context of the DEEP and BLUEPRINT projects. The code is available from GitHub[11] under the GPL-3 license.

3.3.4 Applications

A prototype of the pipeline was previously developed in the Python and R programming languages and only supported consensus-based quantification (approach (i) above). It constituted one of the first approaches for the characterization of epigenomic patterns in repeats and was applied in a number of studies. In [Bock *et al.* 2010], we compared different sequencing-based methods for the genome-wide quantification of DNA methylation. Overall, DNA methylation profiles were similar across different technologies and we observed that CpG-rich repeat sequences that are highly abundant in the human genome were highly methylated compared to varying methylation levels in CpG poor and sparse elements.

Tobi *et al.* [2014] analyzed the effect of prenatal famine exposure on the DNA methylome in whole blood using RRBS and also used our prototype pipeline in oder to characterize DNA methylation in repetitive elements. Differentially methylated regions between exposed and unexposed individuals globally co-occurred with repetitive elements as quantified by an EpiGRAPH [Bock *et al.* 2009] analysis. However, we observed no interpretable famine-associated changes of DNA methylation on the level of individual repeat subfamilies.

Deplus *et al.* [2014] analyzed DNA methylation in the context of phosphorylation of *de novo* DNA methyltransferases. The article shows that the CK2 enzyme phosphorylates Dnmt3a and that inhibited phosphorylation results in reduced methylation activity in a mouse model. Moreover, genome-wide DNA methylation was quantified by

[11] https://github.com/MPIIComputationalEpigenetics/epiRepeatR

MeDIP-seq in human U2OS osteosarcoma cells with (i) normal expression of CK2 and with (ii) RNAi-reduced expression of CK2. While genome-wide methylation patterns were highly correlated between the two conditions in the non-repetitive portion of the genome, significant differences were observed in repetitive element using our repeat pipeline. LINE, LTR and satellite elements were hypomethylated in CK2-depleted cells compared to the control and SINEs, most notably *Alu*s, exhibited increased methylation. Hypermethylation in *Alu*s was confirmed experimentally by locus-specific bisulfite sequencing. Further immunofluorescence experiments showed that Dnmt3a was predominantly localized in the heterochromatic portion of the genome while this localization shifted when phosphorylation was inhibited. Taken together, the results indicate that CK2-mediated phosphorylation of Dnmt3a prevents DNA methylation in euchromatic regions of the genome and associated SINEs while heterochromatic regions are maintained in a methylated state.

3.3.5 Use Case: Epigenomic Signatures of Repetitive Elements in Human Blood Cells

Here, we provide a global perspective on the epigenomics of repetitive elements by applying epiRepeatR to the epigenomes of various human blood cell types. Our dataset comprises 11 samples derived from biological replicates of monocytes, macrophages and three $CD4^+$ T cell populations involved in immune memory formation [Wallner *et al.* 2016; Durek *et al.* 2016]. For each sample, epigenome profiles were generated by the DEEP consortium and include DNA methylomes assayed by WGBS and histone modification maps for six marks, profiled by ChIP-seq.

epiRepeatR was applied using both implemented quantification approaches. We first present the results of the aggregation-based approach and compare it to the consensus-based approach later in the section. ChIP-seq reads and genome-wide DNA methylation calls were obtained for three replicates of monocytes (Mono) and two replicates of macrophages (Mf), Central Memory T cells (TCMs), Effector Memory T cells (TEMs) and Naive T cells (TNs) and were used as input to epiRepeatR. The hg19 human genome assembly was used throughout the analysis. For each repeat subfamily, the mean DNA methylation level across all CpGs that were covered by at least five sequencing reads was computed. Enrichment over a joint background signal of ChIP-seq input libraries[12] was quantified. The analyses shown in this section are based on a filtered set of 331 repeat subfamilies that contained at least 50 CpGs across all instances in the genome, that contained a CpG covered by at least 50 sequencing reads in bisulfite experiments and that were covered by at least 200 reads in all ChIP-seq experiments and the joint input.

Cell types could be distinguished based on their epigenomic profiles in repetitive elements in unsupervised analyses (Figure 3.7). DNA methylation patterns were characteristic of blood cell types and in particular of different stages of T cell memory formation (Figure 3.7a). Interestingly, as monocytes and macrophages constitute highly related cell types, genetic differences due to donor origin dominate cell-type-specific differences in the biological replicates. Monocytes and macrophages could be distinguished from T cells based on histone modifications, such as H3K27ac (Figure 3.7b), but characteristic patterns within the T cell group were less pronounced. All samples were obtained in the context of DEEP and have been processed using standardized pipelines. However, it is important to note that the two groups of cells (monocytes and macrophages vs. T cells) were obtained from different sample providers and corresponding sequencing

[12] The input library of one sample (51_Hf03_BlTN_Ct) was excluded since it failed sequencing quality checks.

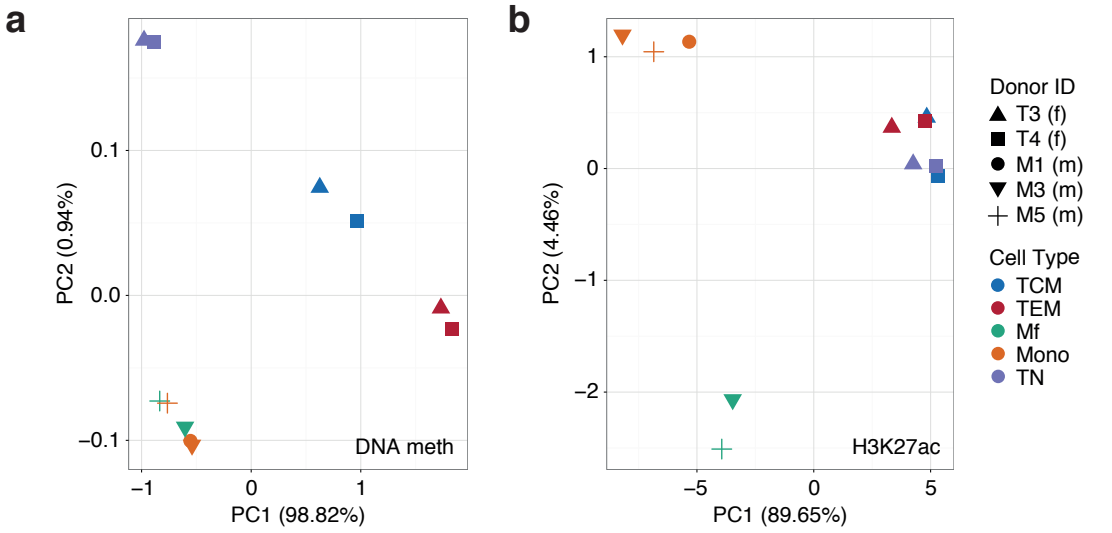


Figure 3.7: Principal component analysis based on the epigenetic signatures of repetitive elements in human blood cells. PCA based on **(a)** DNA methylation levels and **(b)** H3K27ac log-odds scores in 331 repeat subfamilies are shown for all 11 samples. Point colors indicate cell types and shapes denote donors. Numbers in parentheses indicate the percentage of variance explained by the first two principal components.

experiments have been conducted at different institutes. Furthermore, monocytes and macrophages were obtained from the blood of male donors while T cell profiles were derived from pools of female donors. We therefore cannot exclude the possibility that the observed differences between the two groups could be the result of technical variability and confounding rather than true biological variability.

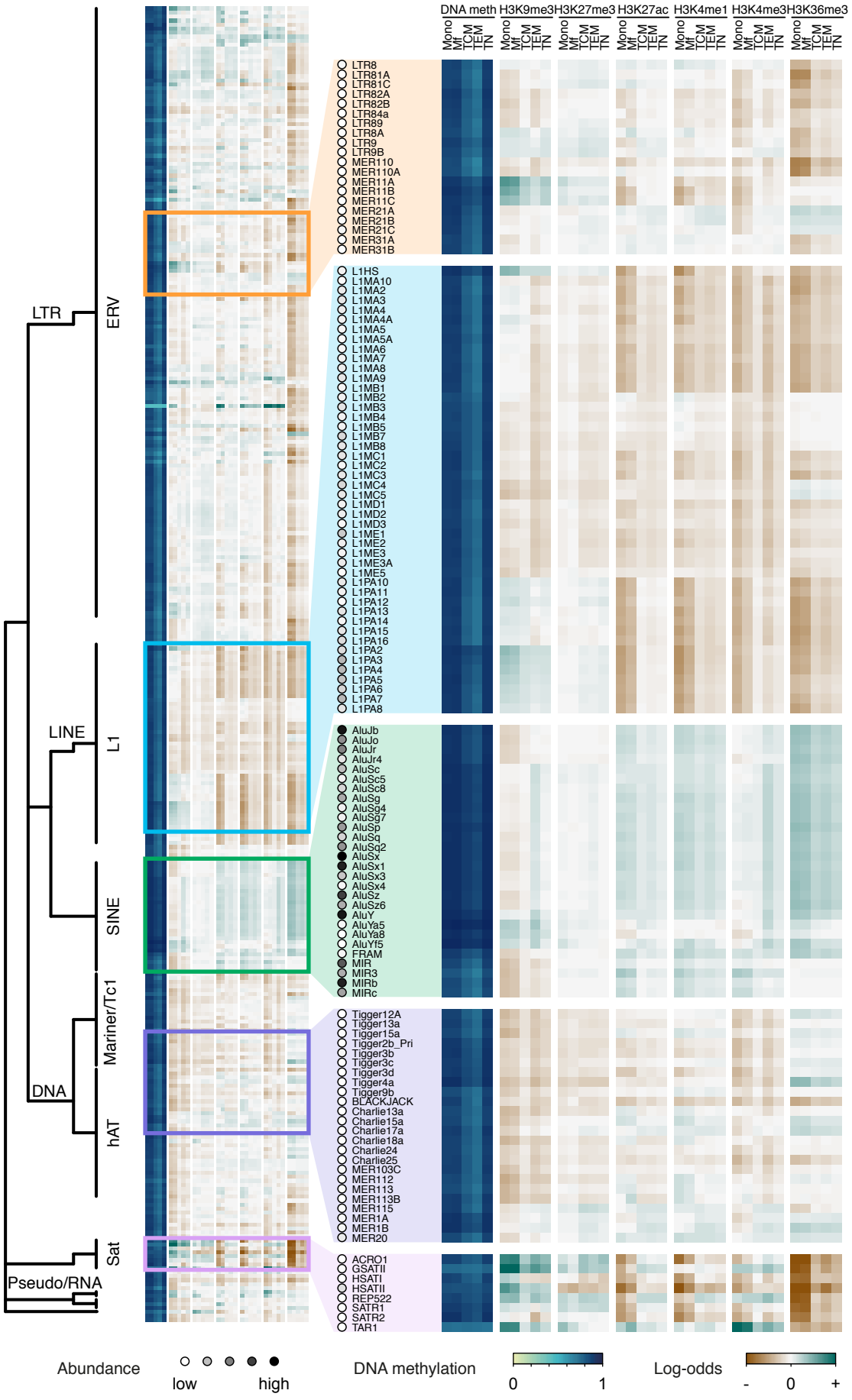
epiRepeatR enables exploratory analysis by providing hierarchical heatmap views of the epigenetic patterns in repeat families and subfamilies across all groups of samples in a dataset (Figure 3.8). Repeat subfamilies can be grouped hierarchically by annotating the rows of the heatmap with dendrograms. These dendrograms can be constructed based on hierarchical clustering or based on a preannotated subfamilies hierarchy of repetitive elements. In human blood cells we observe consistently high DNA methylation levels across all repeat subfamilies but also a certain degree of variability between different repeats and cell types. For instance, *Alu* elements, which are highly abundant in the genome, appear particularly highly methylated compared to other repeats. Furthermore, the memory states of T cells exhibit lower DNA methylation levels compared to the naive state across subfamilies. Consistent with previous studies, the repressive H3K9me3 was particularly enriched in subfamilies of satellite repeats and also present in ERVs. We also observe high levels of H3K9me3 in the primate-specific L1PA subfamilies of LINE elements, which have been shown to escape repression by DNA methylation during the preimplantation stage of human embryos [Z. D. Smith *et al.* 2014]. *Alu* elements show a similar pattern: *AluY* elements which are evolutionarily young and actively transposing in the human genome contain higher levels of H3K9me3 than other *Alu*s. In contrast, DNA transposons are relatively depleted in H3K9me3. Furthermore, the repressive mark H3K27me3 is found in the majority of ERV elements. Histone marks

characteristic of enhancers (H3K4me1 and H3K27ac) are generally absent in LINE elements, but frequently occur in *Alu* repeats — a finding which is also in accordance with the results of Day *et al.* [2010]. H3K4me3, which is typically found in active promoter regions, exhibits a similar, but less pronounced pattern. H3K36me3 is strongly depleted in satellite repeats and to a lesser extent in LINE and ERV elements. In contrast, *Alu* elements and certain DNA transposons are generally enriched for this histone mark, which is frequently located in gene bodies and has been associated with transcriptional elongation.

We further characterized the association of epigenomic marks with attributes of repeat subfamilies (Table 3.2). These attributes include measures of genomic abundance, species-specificity and sequence features. DNA methylation levels vary across the taxonomy levels with which individual repeat subfamilies are annotated and which provide an indicator of how specific a repeat subfamily is to the human species (Figure 3.9a). Generally, DNA methylation levels in repeat subfamilies tend to increase with increased genomic abundance and CpG content of the repeats (Figure 3.9b, 3.9d and Table 3.2). In contrast, the number of thymines in repeat instances exhibits the inverse behavior (Figure 3.9c).

In addition to the aggregation-based approach (approach (ii)), we also quantified repeat epigenomes using the consensus-based approach (approach (i)). A median of 6.2 % of bisulfite reads aligned to the Repbase Update consensus sequences. ChIP-seq alignment rates range from a median of 4.5 % for H3K4me3 to 8.7 % for H3K9me3. Compared to the aggregation-based approach, the consensus-based approach resulted in a higher dynamic range of DNA methylation levels across repeat subfamilies (Figure 3.10a). Evolutionarily young and highly abundant elements such as active L1, *AluY* and SVA elements generally were highly methylated while older, lowly abundant elements such as DNA transposons, *AluJ* and inactive L1 were unmethylated (data not shown). ERVs exhibited variable methylation patterns. In contrast, the repeat epigenomes quantified by the two approaches were concordant when considering histone modifications such as H3K9me3 (Figure 3.10b). In order to elucidate the apparent disagreement in DNA methylation levels between the two approaches, we characterized the sequence composition of the RepeatMasker instances and the Repbase Update consensus sequences (Figure 3.11). Repbase Update consensus sequences generally contain more cytosines and CpG dinucleotides. In agreement with this observation, it is important to note that the authors of Repbase Update employ special adjustments for the preservation of CpGs in the assembly of consensus sequences [Bao *et al.* 2015] (and personal communication).

Figure 3.8 : *(On the next page)* Epigenetic signatures of repetitive elements in blood cells. The average epigenetic signals for three replicates of monocytes and two replicates of macrophages, TCMs, TEMs and TNs are shown (columns in each block of the heatmap). Blocks in the heatmap denote different epigenomic marks. Group mean DNA methylation levels and log-odds scores are depicted according to corresponding color scales. The 331 filtered repeat subfamilies are shown. Selected portions of the heatmap are shown in detail on the right. Each subfamily (row) is annotated with genomic abundance (filled circles), which is quantified as the mean relative number of sequencing reads aligning to all instances. Subfamilies are named according to RepeatMasker annotation. The dendrogram on the left corresponds to a preannotated hierarchy of repetitive elements.


DNA meth
H3K9me3
H3K27me3
H3K27ac
H3K4me1
H3K4me3
H3K36me3
Mono
Mf
TCM
TEM
TN
LTR
ERV
LINE
L1
SINE
DNA
Mariner/Tc1
hAT
Sat
Pseudo/RNA
LTR8
LTR81A
LTR81C
LTR82A
LTR82B
LTR84a
LTR89
LTR8A
LTR9
LTR9B
MER110
MER110A
MER11A
MER11B
MER11C
MER21A
MER21B
MER21C
MER31A
MER31B
L1HS
L1MA10
L1MA2
L1MA3
L1MA4
L1MA4A
L1MA5
L1MA5A
L1MA6
L1MA7
L1MA8
L1MA9
L1MB1
L1MB2
L1MB3
L1MB4
L1MB5
L1MB7
L1MB8
L1MC1
L1MC2
L1MC3
L1MC4
L1MC5
L1MD1
L1MD2
L1MD3
L1ME1
L1ME2
L1ME3
L1ME3A
L1ME5
L1PA10
L1PA11
L1PA12
L1PA13
L1PA14
L1PA15
L1PA16
L1PA2
L1PA3
L1PA4
L1PA5
L1PA6
L1PA7
L1PA8
AluJb
AluJo
AluJr
AluJr4
AluSc
AluSc5
AluSc8
AluSg
AluSg4
AluSg7
AluSp
AluSq
AluSq2
AluSx
AluSx1
AluSx3
AluSx4
AluSz
AluSz6
AluY
AluYa5
AluYa8
AluYf5
FRAM
MIR
MIR3
MIRb
MIRc
Tigger12A
Tigger13a
Tigger15a
Tigger2b_Pri
Tigger3b
Tigger3c
Tigger3d
Tigger4a
Tigger9b
BLACKJACK
Charlie13a
Charlie15a
Charlie17a
Charlie18a
Charlie24
Charlie25
MER103C
MER112
MER113
MER113B
MER115
MER1A
MER1B
MER20
ACRO1
GSATII
HSATI
HSATII
REP522
SATR1
SATR2
TAR1
Abundance
low
high
DNA methylation
0
1
Log-odds
-
0
+

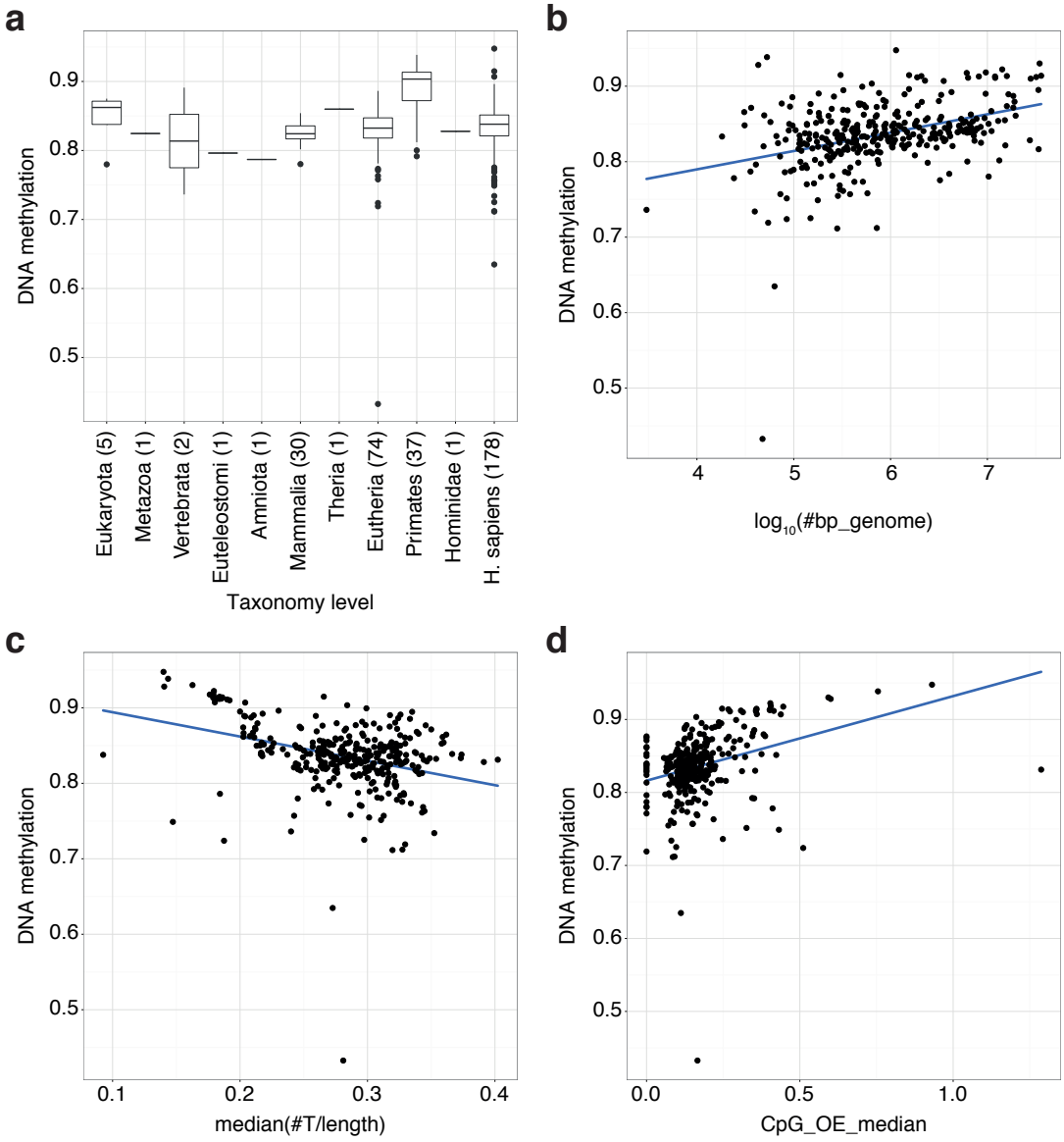
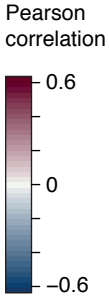


Figure 3.9: Association of DNA methylation with features of repeat subfamilies. Mean DNA methylation levels across all samples in the dataset are shown. **(a)** Boxplot showing the distribution of DNA methylation levels across annotated taxonomy levels. Numbers in parentheses indicate numbers of subfamilies annotated with a given species level. **(b)**, **(c)**, **(d)** Scatterplots of DNA methylation levels compared to the logarithm of the number of basepairs covered, median relative number of thymines and median CpG observed/expected ratio across all RepeatMasker instances of a given repeat subfamily. Lines depict linear regression estimates. Corresponding correlation coefficients can be found in Table 3.2.

Table 3.2: Correlation of repeat subfamily features and epigenomic marks

	DNAmeth	H3K9me3	H3K27me3	H3K27ac	H3K4me1	H3K4me3	H3K36me3
$\log_{10}$(#instances)	0.36	−0.26	−0.16	0	0.04	−0.15	0.36
$\log_{10}$(#bp_genome)	0.39	−0.03	−0.1	−0.15	−0.14	−0.22	0.21
median(#A/length)	0.18	−0.32	−0.54	−0.52	−0.42	−0.46	−0.06
median(#C/length)	−0.08	0.48	0.57	0.41	0.27	0.46	−0.01
median(#G/length)	0.3	0.3	0.41	0.37	0.34	0.34	0.37
median(#T/length)	−0.34	−0.38	−0.34	−0.14	−0.09	−0.24	−0.22
median(#AA/length)	0.14	−0.23	−0.45	−0.36	−0.28	−0.32	0
median(#CA/length)	0.13	0.18	0.12	−0.11	−0.15	0	0.01
median(#GA/length)	0.16	0.01	−0.04	−0.1	−0.11	−0.12	0.02
median(#TA/length)	−0.03	−0.46	−0.56	−0.49	−0.36	−0.48	−0.18
median(#AC/length)	0.08	0.1	−0.01	−0.14	−0.17	−0.03	−0.07
median(#CC/length)	−0.09	0.48	0.57	0.39	0.26	0.44	−0.05
median(#GC/length)	0.19	0.42	0.58	0.56	0.5	0.55	0.36
median(#TC/length)	−0.28	0.21	0.29	0.19	0.1	0.18	−0.18
median(#AG/length)	0.18	0.05	0.21	0.19	0.18	0.08	0.2
median(#CG/length)	0.31	0.53	0.31	0.26	0.2	0.43	0.22
median(#GG/length)	0.31	0.38	0.46	0.34	0.3	0.37	0.32
median(#TG/length)	0.03	−0.01	0.11	0.17	0.17	0.04	0.19
median(#AT/length)	0	−0.4	−0.61	−0.64	−0.52	−0.56	−0.24
median(#CT/length)	−0.28	0.21	0.44	0.43	0.33	0.34	−0.07
median(#GT/length)	0.02	−0.12	−0.01	0.13	0.18	−0.02	0.26
median(#TT/length)	−0.34	−0.35	−0.35	−0.06	−0.03	−0.18	−0.19
CpG_OE_median	0.32	0.37	0.02	0.04	0	0.19	0.13



Association is quantified by Pearson correlation coefficients of the features and DNA methylation levels or ChIP-seq log-odds scores. Features for each subfamily include the logarithm of the number of instances and the number of basepairs covered as well as median counts of nucleotides and dinucleotides normalized by instance length. The CpG observed/expected ratio was quantified based on median instance length, base and CpG dinucleotide counts. Instances were defined according to RepeatMasker annotation.

Moreover, the difference in DNA methylation levels between the two approaches is correlated with differences in sequence composition: we observe larger methylation differences in repeat subfamilies which exhibit fewer cytosines and CpGs in RepeatMasker compared to Repbase Update (Figure 3.11). These findings are consistent with the hypothesis that the Repbase Update consensus sequences are representative of ancestral sequences from which the RepeatMasker instances derive. During the course of evolution, repeat instances degenerate due to deamination and other mutations and thus are depleted in cytosines and CpGs. Alignment of bisulfite reads does not discriminate between cytosines and thymines. Consequently, the methylation patterns of reads aligned to the consensus might not actually reflect the methylation state of specific instances, but rather the degree to which they diverge from the consensus. This sequence bias is particularly pronounced for older repeats. Sequences of evolutionarily young elements exhibit a higher similarity to the consensus and their methylation levels are therefore more consistent between the two approaches. This reasoning raises the question whether the CpGs in the Repbase Update consensus sequences represent suitable reference loci for bisulfite-based quantification of all instances of a given repeat subfamily.

3.3.6 Discussion

Sequencing technology and processing pipelines have improved significantly in recent years, the high sequence similarity of instances of repetitive elements still pose a challenge to the processing of high-throughput sequencing data. Although overall read

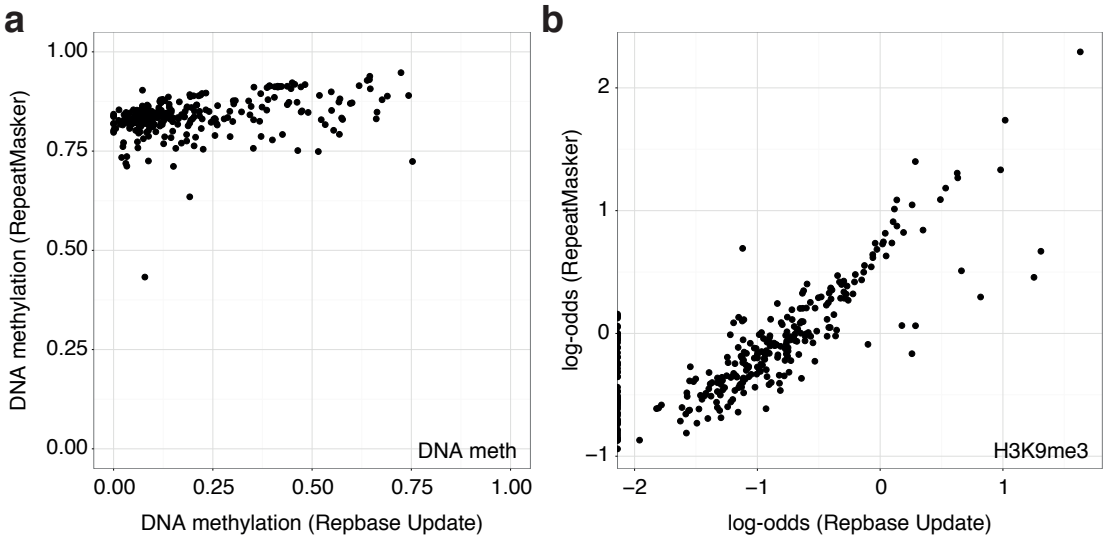


Figure 3.10: Comparison of epigenetic signals in aggregation and consensus-based approaches. Scatterplots show **(a)** DNA methylation levels and **(b)** log-odds for H3K9me3 according to the consensus (Repbase Update) and aggregation-based (RepeatMasker) approaches.

alignment rates are high, repeats can accumulate a large percentage of reads that cannot be uniquely assigned to a single position in the genome. Processing pipelines typically employ one of two strategies to deal with these multi-mapped reads: (i) ambiguously mapping reads are discarded from further processing, leading to potential coverage biases and discarding potentially useful sequence information or (ii) they are randomly assigned to one of the possible genomic positions, leading to lower mapping qualities and diluting the information content at the respective coordinates. The effects are particularly strong for short sequencing read strategies. Our pipeline addresses the issue of multi-mapped reads by adopting a genome-wide view on repetitive elements: rather than requiring accurate, unique alignments to individual repeat instances, our approach aggregates instances of repeats into subfamilies and provides an increased robustness when instances are highly similar. However, it should be noted that this type of analysis does not focus on individual insertion loci of repetitive elements. It is therefore possible that the identified epigenomic patterns of a particular repeat subfamily are only based on a few instances in the genome. Additionally, between-instance variability is currently not taken into account.

The application of our pipeline to a dataset of human blood cells results in observations that argue for epigenetic regulation of repetitive elements. Particularly *Alu* elements seem to co-localize with regulatory regions such as promoters and enhancers, which are marked by distinct epigenetic patterns. It remains to be investigated whether the distinct epigenetic marks are merely placed into these elements due to the co-localization or whether the elements themselves exert or are subject to epigenetic regulation. Furthermore, evolutionarily young elements (e.g. *AluY* and L1PA) appear methylated and marked by H3K9 methylation. These elements are particularly rich in CpG, since they represent relatively novel additions to our genomes and therefore did not diverge from the consensus. Notably, in contrast to other CpG-rich regions in the genome,

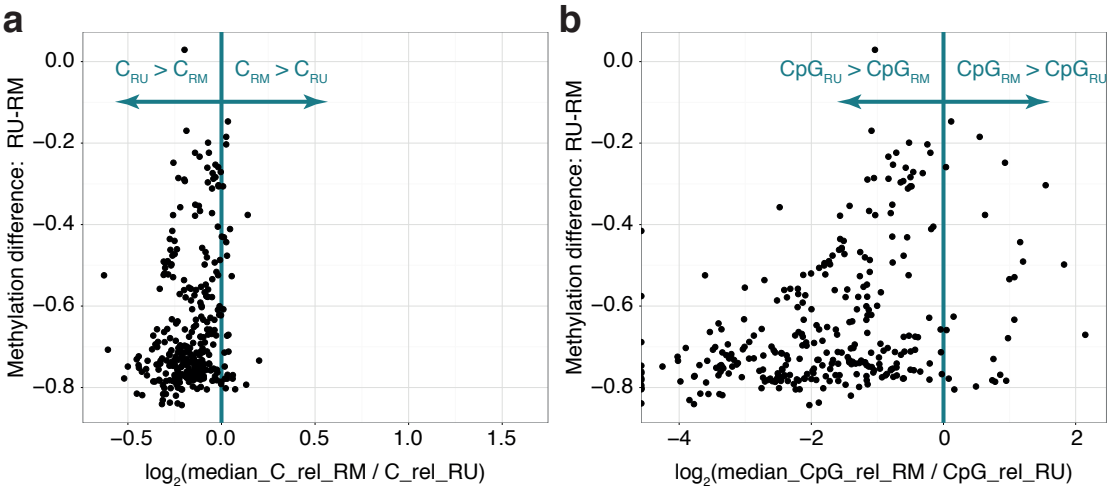


Figure 3.11: Deviations in sequence composition explain DNA methylation differences between the aggregation and consensus-based approaches. Scatterplots compare the relative difference in sequence composition between RepeatMasker (RM) instances and the Repbase Update (RU) consensus against the difference in methylation levels between the consensus-based and aggregation-based approaches. The difference in sequence composition is quantified as the logarithm of the ratio of the median relative sequence content across RepeatMasker instances and the relative sequence content of the Repbase Update consensus. The sequence content is shown for the relative number of **(a)** cytosines and **(b)** CpG dinucleotides.

DNA methylation is high in these elements. Epigenetic downregulation potentially provides an important mechanism for keeping these elements at bay, protecting structural genomic integrity.

Interestingly, DNA methylation patterns discriminate between different populations of human T cells, with memory T cells exhibiting globally demethylated repeats compared to naive T cells. These findings are further discussed in Section 4.3 which provides a characterization of genome-wide DNA methylation during T cell memory formation.

The apparent disagreement of the two outlined approaches for quantifying DNA methylation levels raises the question whether the Repbase Update consensus sequences provide a suitable reference for bisulfite-based analyses. CpGs in the consensus sequences could reflect ancestral or even artificial states rather than corresponding to actual CpGs in repeat instances. Therefore, the analysis of repetitive elements in [Bock *et al.* 2010] and [Tobi *et al.* 2014] should be interpreted with caution. Furthermore, in our use case, only a median of 6.2 % of bisulfite reads aligned to consensus sequences, although repetitive elements have been postulated to cover more than half of the human genome. It remains to be evaluated whether a more suitable reference of consensus sequences for bisulfite sequencing can be derived. Subdividing repeat subfamilies or including flanking sequences could provide more granularity [Day *et al.* 2010; Xie *et al.* 2013], but might also lead to a signal that is diluted across many reference sequences. Other directions of future development include the analysis of expression data quantified by RNA-seq as well as enabling quantitative differential and comparative analysis of repeat epigenomes between groups of samples.

In summary, the described pipeline facilitates the characterization of epigenomic patterns in repetitive elements. Future applications in the context of embryonic development and diseases like cancer could shed light on the patterns that regulate the expression of repetitive DNA elements and could contribute to elucidating how repeats themselves represent regulatory elements.

4

Charting the Epigenomic Landscape of Hematopoiesis

The work described in this chapter contributed to the BLUEPRINT [Adams et al. 2012] and DEEP projects.
The first analytic part of the chapter consists of a descriptive analysis of methylome data of differentiated hematopoietic cell types from the BLUEPRINT project, employing RnBeads. Cells were obtained by multiple collaboration partners in the project and sequencing library preparation as well as primary data processing was performed by the group of Simon Heath at the Centro Nacional de Análisis Genómico (Barcelona, Spain).
Second, epigenome maps for human T helper cells during the process of immune memory formation have been generated in the context of DEEP [Durek et al. 2016]. The study was led by Pawel Durek, Karl Nordström, Gilles Gasparoni, Jörn Walter, Alf Hamann and Julia Polansky. I contributed descriptive analyses and methods for the processing of DNA methylation data.
Third, within the context of BLUEPRINT, we analyzed methylomes of hematopoietic progenitor cells and a corresponding article has been published recently [Farlik et al. 2016]. Data analysis was performed in collaboration with Matthias Farlik, Florian Halbritter, Peter Ebert and Johanna Klughammer. Bisulfite sequencing libraries were prepared by Matthias Farlik. Elisa Laurenti, Thomas Lengauer, Mattia Frontini and Christoph Bock supervised the project (cf. [Farlik et al. 2016] for a details on author contributions). I conducted analyses comprising data preprocessing, the general characterization of the dataset and also derived and interpreted statistical learning classifiers for cell type prediction. Furthermore, I had a leading role in the planning of the analysis, method development, writing the manuscript and designing figures and supplementary material. Corresponding text and figures in Section 4.4 has been adapted from the manuscript.

Approximately one trillion (10^{12}) blood cells are formed in the human bone marrow every day in a process called hematopoiesis [Doulatov *et al.* 2012]. Blood cells and their progenitors are organized in a differentiation hierarchy that is tightly regulated by epigenetic mechanisms (Figure 4.1). Employing the methods described in Chapter 3, this chapter presents detailed analyses of DNA methylomes at different levels in the hematopoietic hierarchy (cf. color coding in Figure 4.1). Section 4.1 provides the biological background for the remainder of the chapter. It introduces relevant blood cell types and discusses current models for hematopoietic differentiation. Furthermore, known epigenetic cues regulating blood-cell differentiation are outlined. Section 4.2 takes on

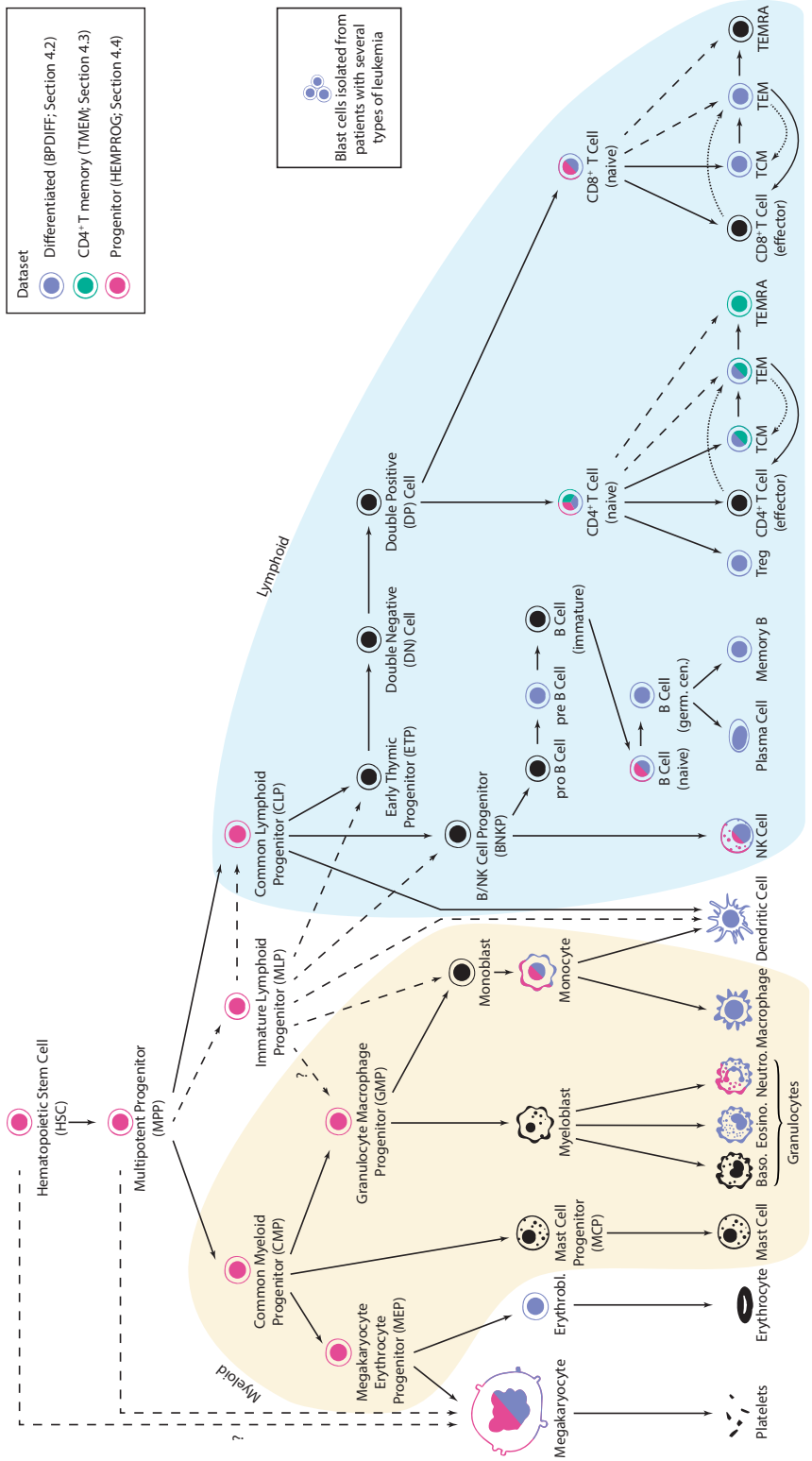
a broad perspective on the hierarchy and focuses on the terminal cell types. It contains a descriptive analysis of DNA methylation in differentiated cells assayed by the BLUEPRINT project. For clarity, the dataset analyzed in this section will be referred to as BPDIFF. Rather than providing an in-depth characterization of hematopoiesis, this introductory section aims at providing a birds-eye view on between-cell-type heterogeneity in human blood. Section 4.3 focuses on one branch of the hierarchy and spotlights the DNA methylation dynamics in human T cell memory formation (TMEM dataset). It highlights the relationship between epigenetic patterns and immunity-related cell function. Finally, the stem of the hierarchy is explored in Section 4.4 which constitutes the core of the chapter and contains a detailed characterization of hematopoietic progenitor cell types and their paths of differentiation. Within the context of BLUEPRINT, a whole-genome DNA methylation dataset (HEMPROG) was generated using low-input bisulfite sequencing. We provide a detailed analysis of this dataset and exploit statistical methods for modeling the *in vivo* DNA methylation landscape of blood stem cell differentiation.

4.1 Epigenetic Regulation of Hematopoiesis

The main constituents of blood are red blood cells (**erythrocytes**), which are responsible for the transport of oxygen. **Leukocytes** or white blood cells are cells of the immune system that defend the organism against pathogens and clear dysfunctional cells. **Platelets**, sometimes referred to as thrombocytes, contribute to the repair of blood vessels and blood clotting. Organs important for hematopoiesis include the bone marrow, where the bulk of blood cells are formed, and the lymphoid system and its associated tissues such as lymph nodes, spleen and thymus, where many immune cells mature. Many immune cells eventually reside in peripheral tissues as well as in blood.

The blood is one of the best characterized developmental systems in mammals [Doulatov *et al.* 2012]. Human hematopoietic cells are typically obtained from peripheral blood, which is widely available, or from bone marrow, whose acquisition employs more invasive methods. A multitude of different blood cell types exist, which arise during hematopoiesis from proliferating progenitor cells that become increasingly restricted in their potential to give rise to different cell lineages in a stepwise fashion. Dozens of hematopoietic cell types have been defined based on morphology, pluripotency and

Figure 4.1 : *(On the next page)* Hierarchy of the hematopoietic system. Based on literature review, the differentiation trajectories in the hematopoietic system are shown. Solid lines represent the canonical model of hematopoiesis. MLPs which are attributed with myelo-lymphoid differentiation have been incorporated into the tree using dashed lines [Doulatov *et al.* 2012]. In addition, a possible direct branching of megakaryocytes from HSCs or MPPs has been proposed [Notta *et al.* 2016; Woolthuis and C. Y. Park 2016] and is also depicted using dashed lines. Alternative models of T cell memory formation are indicated in dashed and dotted arrows and are further explained in Section 4.3. Orange and blue background colors indicate the myeloid and lymphoid branches of the hierarchy, respectively. The color coding of cell pictograms indicates the availability of methylation data in the BLUEPRINT and/or DEEP projects and corresponds to the respective dataset described in this chapter.


Dataset
Differentiated (BPDIFF; Section 4.2)
CD4+ T memory (TMEM; Section 4.3)
Progenitor (HEMPROG; Section 4.4)
Blast cells isolated from patients with several types of leukemia
Hematopoietic Stem Cell (HSC)
Multipotent Progenitor (MPP)
Immature Lymphoid Progenitor (MLP)
Common Lymphoid Progenitor (CLP)
Early Thymic Progenitor (ETP)
Double Negative (DN) Cell
Double Positive (DP) Cell
B/NK Cell Progenitor (BNKP)
pro B Cell
pre B Cell
B Cell (immature)
B Cell (naive)
B Cell (germ. cen.)
Memory B
Plasma Cell
NK Cell
CD4+ T Cell (naive)
CD4+ T Cell (effector)
Treg
TCM
TEM
TEMRA
CD8+ T Cell (naive)
CD8+ T Cell (effector)
Lymphoid
Common Myeloid Progenitor (CMP)
Granulocyte Macrophage Progenitor (GMP)
Megakaryocyte Erythrocyte Progenitor (MEP)
Monoblast
Monocyte
Dendritic Cell
Myeloblast
Baso. Eosino. Neutro.
Granulocytes
Macrophage
Mast Cell Progenitor (MCP)
Mast Cell
Erythrobl.
Erythrocyte
Megakaryocyte
Platelets
Myeloid

the presence of surface markers. Differentiation steps are triggered by distinct signaling molecules called cytokines and are subject to tight epigenetic regulation. Figure 4.1 depicts the canonical view of differentiation events based on current literature and highlights selected aspects that are currently debated (dashed arrows).

Hematopoietic Stem Cells (HSCs) are placed at the apex of the hematopoietic hierarchy. They are defined based on the property to be able to give rise to all constituent cell types of blood and are capable of long-term self-renewal. HSCs are a very rare cell type, represented by only one in about 10,000 cells in the bone barrow [Alberts *et al.* 2008]. They can either divide symmetrically producing further HSCs or asymmetrically spawning a progeny of **Multipotent Progenitors (MPPs)**. MPPs can still give rise to all hematopoietic cell types but are shorter-lived than HSCs and restricted in their potential to self-renew. This observation is reflected in the classical model of hematopoiesis in which the loss of capability of self-renewal precedes lineage commitment [Doulatov *et al.* 2012]. A number of more lineage-committed progenitor cell populations derive from MPPs. They typically undergo a limited number of cell divisions, thereby amplifying through a cascade of differentiating progeny cells that become increasingly restricted in their differentiation potential. According to the classical model, the differentiation path of MPPs bifurcates into the two common progenitor cell types of the myeloid and lymphoid lineages.

Common Myeloid Progenitors (CMPs) can give rise to all myeloid cells. These are primarily associated with the innate immune response, which represents an organism's first line of defense against pathogens. Cells involved in this defense include cells derived from **Granulocyte Macrophage Progenitors (GMPs)**: **granulocytes**, which comprise **basophils**, **eosinophils** and **neutrophils**, utilize lysosomes and other vesicles in their cytosol to neutralize pathogens. In basophils and eosinophils these vesicles contain toxins that can be released for killing pathogens of external origin. Neutrophils ingest pathogens, which are subsequently destroyed using the vesicles' contents. **Monocytes** differentiate into **macrophages** which are capable of disabling invaders through phagocytosis and of eliciting inflammatory immune responses through the secretion of cytokines. CMPs have also been attributed the potential to differentiate into progenitors of megakaryocytes and the erythroid lineage (**Megakaryocyte Erythrocyte Progenitors (MEPs)**). **Megakaryocytes** are large cells which can become highly polyploidic during maturation when they lose their ability to divide. They finally give rise to platelets.

There are approximately 2×10^{12} **lymphocytes** in the human body [Alberts *et al.* 2008]. According to the canonical model of hematopoiesis all of them derive from **Common Lymphoid Progenitors (CLPs)**. However, recently, the existence of **Immature Lymphoid Progenitors (MLPs)** populations has been postulated. These cells represent human lymphoid progenitors that potentially also differentiate into myeloid cells, but are incapable of giving rise to erythroid cells or megakaryocytes (dashed arrows in Figure 4.1) [Doulatov *et al.* 2012]. In mouse, **Lymphoid-primed Multipotent Progenitors (LMPPs)** have been attributed with a similar role. Undergoing further steps of lineage restriction, CLPs and MLPs eventually differentiate into lymphocytes, including **Natural Killer cells (NK cells)**, **B cells** and **T cells**. NK cells are part of the innate immune system and kill tumor and infected cells by secreting cytokines capable of inducing apoptosis or lysis. B cells and T cells form the basis of the adaptive immune system, which is only found in vertebrates. They originally exist in a naive state and upon exposure to a pathogen become clonally activated, proliferate and enter an effector state responsible for a specific immune response. A subset of cells also enters a memory state in which,

upon subsequent exposure to the same antigen, they can rapidly generate further effector and memory cells specific to the antigen and produce an increased amount of cytokines or antibodies for immune signaling [Alberts *et al.* 2008]. B lymphocytes carry out their immune function by producing and secreting antibodies specific to the encountered pathogen, which can render pathogens such as viruses inactive or mark them for destruction. T cells exert different immune functions: **cytotoxic T cells ($CD8^+$ T cells)** kill infected cells via recognition of antigens presented on the cell surface while **T helper cells ($CD4^+$ T cells)** are capable of activating other immune cells such as macrophages, granulocytes, dendritic cells, B cells and cytotoxic T cells. **Regulatory T cells (Tregs)** provide cues for the inhibition of other T cells and dendritic cells and thus arrange for a controlled immune response. Notably, their control is frequently brought out of balance in autoimmune diseases. **Dendritic cells** can derive from a myeloid or lymphoid origin. They activate other adaptive immune cells by presenting antigens which they acquire by ingesting foreign particles [DeFranco *et al.* 2007].

Our current knowledge of hematopoiesis is largely based on functional studies in mice combined with fewer, more limited studies in human cells [Doulatov *et al.* 2012]. Nonetheless, focused functional assays have succeeded in characterizing certain aspects of human hematopoiesis and represent the main contributions to our current understanding of the blood-cell-type-hierarchy (Figure 4.1). *In vitro*, colony formation assays are employed that can characterize the outgrowth of cell populations from selected clones. *In vivo*, selected murine and human hematopoietic cell populations can be (xeno)transplanted in order to gauge their capability of repopulating the blood system of mice, whose native blood-forming cells have been destroyed by irradiation [Doulatov *et al.* 2012].

While the depicted canonical model of differentiation has proven useful for our understanding of hematopoiesis, several of its aspects remain controversial. Recently, it has been postulated that megakaryocytes could be derived from multipotent progenitors rather than share common origin with myeloid cells (dashed arrows in Figure 4.1) [Notta *et al.* 2016; Paul *et al.* 2015; Woolthuis and C. Y. Park 2016]. Direct derivation of lymphoid and myeloid progenitors from HSC and MPP populations that are more lineage-restricted than CMPs and CLPs has also been proposed [Perié *et al.* 2015; Notta *et al.* 2016; Cabezas-Wallscheid *et al.* 2014]. The existence of MLPs in human and LMPPs in mice represents a further deviation from a strictly tree-like model. Moreover, it is important to note that the definition of many hematopoietic cell types is mainly based on the expression of surface markers which can be used for separating cells using Fluorescence-Activated Cell Sorting (FACS). This particularly holds true for progenitor cell populations. However, it is currently debated to what extent surface marker expression corresponds to molecular function and heterogeneity within the cell population [Notta *et al.* 2016]: a pool of sorted cells expressing an identical set of surface markers might still contain distinct classes of cells that possess heterogeneous functional potential. Concretely, CMPs might represent a heterogeneous population of cells that contain multiple lineage-restricted progenitor cells rather than a homogeneous population of universal myeloid-erythroid-megakaryocyte progenitors [Paul *et al.* 2015; Perié *et al.* 2015; Notta *et al.* 2016]. Further examples are provided by MPPs which exhibit different potentials for proliferation and repopulation [Cabezas-Wallscheid *et al.* 2014] and by the possible existence of HSC subpopulations primed for myeloid, lymphoid and megakaryocyte development [Muller-Sieburg *et al.* 2012; Woolthuis and C. Y. Park 2016].

Single-cell protocols provide the means for a more accurate definition of cell type and differentiation potential by enabling an in-depth molecular characterization of population heterogeneity. Paul *et al.* [2015] profiled the transcriptomes of myeloid progenitors in mouse using single-cell RNA-seq. They identified clusters of cell subpopulations that partially deviated in their expression of surface markers from commonly employed gating strategies in FACS. Lineage-specific TF-mediated regulation was linked to these clusters.

The process of hematopoiesis can be steered by a relatively small set of transcription factors [Doulatov *et al.* 2012; Álvarez-Errico *et al.* 2015; Novershtern *et al.* 2011; Rosenbauer and Tenen 2007] which are in turn regulated by epigenetic signatures. Using defined transcription factors, it is possible to transdifferentiate between cell types of different blood lineages [Álvarez-Errico *et al.* 2015; Orkin and Zon 2008]. Moreover, epigenomic dysregulation is linked to disease. Mutations in proteins catalyzing epigenetic changes are associated with malignancies [Shih *et al.* 2012; Álvarez-Errico *et al.* 2015]. For instance, *TET2*, an important component responsible for DNA demethylation, is mutated in several myeloid malignancies [Álvarez-Errico *et al.* 2015] and TF dysregulation in the myeloid lineage could contribute to the onset and progression of myeloid leukemia [Rosenbauer and Tenen 2007].

Distinct DNA methylation patterns resemble hallmarks of hematopoietic cell identity: Ji *et al.* [2010] profiled DNA methylation in murine blood progenitors using a microarray-based approach. They found DNA methylation patterns which are markedly different between lymphoid and myeloid lineages, with myeloid cells generally exhibiting lower methylation levels in specific regulatory regions. They concluded that lymphoid development depends on suppressing myeloerythroid regulators via DNA methylation. This is in line with findings which show that mice with reduced activity of *Dnmt1* exhibit differentiation skewed towards the myeloid lineage [Bröske *et al.* 2009] and that active demethylation plays a key role in myeloid development [Álvarez-Errico *et al.* 2015]. Along with evidence that the lymphoid cells emerged later in evolution, these findings indicate that myeloid differentiation could constitute a default hematopoietic pathway and that lymphoid development represents a superimposed alternative [Álvarez-Errico *et al.* 2015]. Bock *et al.* [2012] employed RRBS and expression microarrays to dissect stem cell differentiation in the blood and skin of mice. They identified loss of methylation in lineage-specific regulatory elements and gain in regulatory elements of other lineages during differentiation towards the myeloid or lymphoid lineage. In particular promoters of myeloid TFs gained methylation during differentiation towards lymphoid lineage. Moreover, a hierarchy of cell differentiation was inferred using an approach that assigns ranks of differentiation and proliferation and uses this ordering to connect cell types in a two-dimensional space inferred from averaging methylation and expression-based distances between samples. Cabezas-Wallscheid *et al.* [2014] provided an integrative view on the proteome, transcriptome and methylome of murine HSCs and MPPs populations. In their study, different MPP subpopulations exhibited different rates of proliferation and repopulation functionality.

Chromatin-based regulation represents another avenue of steering blood formation and chromatin-modifying enzymes are attributed with central roles in myeloid differentiation and immune cell activation [Álvarez-Errico *et al.* 2015]. The dynamics of chromatin marks characteristic of enhancers have been profiled by Lara-Astiaso *et al.* [2014] using ChIP-seq. The authors concluded that some enhancers are already active in HSCs and are maintained only in the respective lineage while others are initially inactive and only gain active enhancer marks in certain lineages.

A number of(epi)genome-scale datasets describing human hematopoiesis have been published, but only a few of them include hematopoietic stem and progenitor cells. Novershtern *et al.* [2011] have profiled transcription during human hematopoiesis and discovered that the degree of differential expression between hematopoietic cell types is on a scale comparable to that of the overall heterogeneity across human tissues. Furthermore, they determined modules of coexpressed genes and TFs likely to be involved in the regulation of hematopoietic differentiation. Another recent study described cell-type-specific expression and splicing events in human progenitors and two myeloid precursor cell types [L. Chen *et al.* 2014]. M. R. Corces *et al.* [2016] generated maps of chromatin accessibility in hematopoitic progenitors and mature cell types using ATAC-seq. Moreover, they identified regulatory signatures in these cell types which were able to characterize cell type contributions in leukemia cells. WGBS has been employed to profile DNA methylation in HSCs, neutrophils and in B cells [Hodges *et al.* 2011; Kulis *et al.* 2015]. The latter study revealed a gradual loss of methylation during B cell differentiation that primarily occurs in enhancer regions in the early stages of differentiation.

International initiatives provide essential contributions to the quest for uncovering the epigenetic basis of hematopoietic regulation. The BLUEPRINT project is specifically dedicated to characterizing epigenomes of the hematopoietic system in health and malignancy. The DEEP consortium places a particular focus on inflammatory and metabolic diseases and charts epigenome maps of important immune cell types.

4.2 DNA Methylation BLUEPRINTs of Differentiated Hematopoietic Cells

In order to dissect the epigenomic landscape of the human blood system in breadth, we analyzed whole-genome DNA methylation data of 81 blood-related cell types profiled by the BLUEPRINT consortium. This rich resource encompasses a broad spectrum of differentiated blood cells (Figures 4.1, 4.2). Here, we provide a general characterization of inter-cell-type variability within the hematopoietic system and, focusing on the comparison between monocytes and neutrophils, investigate cell-type-specific DNA methylation patterns. This section also highlights the utility of the RnBeads pipeline, which was employed throughout for the assessment of data quality and for a global characterization of the dataset. The results presented here suggest that hematopoietic cell identity is reflected in DNA methylation signatures. In particular, methylation patterns in regulatory elements such as enhancers and promoters can differentiate between cell types and could be associated with roles in regulating cell-type-specific gene expression.

4.2.1 Methods

Cell Isolation and Whole Genome Bisulfite Sequencing

Blood cell populations were collected and sorted by the BLUEPRINT consortium. Detailed experimental and bioinformatic protocols can be found on the project's data coordination website[1]. WGBS library preparation and primary data processing was performed by the group of Simon Heath at the Centro Nacional de Análisis Genómico (Barcelona, Spain). Primary analysis steps included mapping of bisulfite reads to the genome using the GEM3 aligner and calling of methylation levels at individual cytosines using an in-house pipeline [Kulis *et al.* 2015].

[1] http://dcc.blueprint-epigenome.eu/#/md/methods

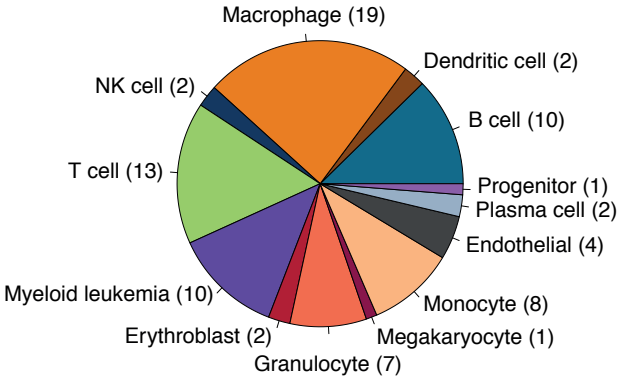


Figure 4.2: Cell types included in the BPDIFF dataset. Sample numbers for each cell type are denoted in parentheses. The dataset contains 81 samples in total.

Description of the Dataset

The initial dataset comprised the 82 methylomes that were part of the seventh BLUEPRINT data release (September 2015). One B cell sample covered a significantly lower number of CpGs and was removed from subsequent analyses. Therefore, the BPDIFF dataset analyzed here contains 81 blood-related samples (Figure 4.2).

The dataset contains a heterogeneous panel of blood-related cell types extracted and sorted from different tissues[2] (Figure 4.2). Myeloid cells isolated from cord or peripheral blood of healthy individuals constitute a large portion of the dataset: monocytes and derived macrophages of three different activation states contribute 27 samples. The group of granulocytes comprises one eosinophil and six neutrophil samples. Lymphocytes in the dataset include multiple stages of B cell differentiation [Kulis *et al.* 2015] and T cell types. Concretely, populations of pre-B cells, naive B cells, memory B cells and plasma cells were isolated from either bone marrow, cord blood or peripheral blood. Germinal center B cells were extracted from tonsil. $CD4^+$ and $CD8^+$ T cells were sorted from peripheral blood, with the exception of two samples which were derived from cord blood. One sample of $CD4^+$ regulatory T cells, two NK cell populations and two conventional dendritic cell samples were also obtained. In addition, the dataset includes one hematopoietic progenitor cell sample. Notably, four endothelial samples from umbilical vein are also included. Finally, ten acute myeloid or promyeloid leukemia samples from bone marrow and peripheral blood represent myeloid malignancy in the dataset. Four of these were subject to treatment with specific drug compounds.

Data Processing and Analysis

The human genome assembly version GRCh38 was used throughout the analysis. RnBeads analysis reports were created using RnBeads version 1.1.8 (Section 3.2) and further plots were generated using custom R scripts. Methylation levels were computed based an individual CpGs. In addition average methylation levels were computed for gene promotors, putative regulatory regions and genome-tiling regions of size 5 kb. Promoters were defined as regions ranging from -1,500 bp to +500 bp around the TSS of Gencode22 annotated genes [Harrow *et al.* 2012]. Putative regulatory regions were obtained

[2] For the sake of simplicity, the term "tissue" is used here to refer to solid tissues as well as body fluids, in particular blood of different origins.

from the BLUEPRINT edition of the Ensembl regulatory build [Zerbino *et al.* 2015] (seventh data release[3]; September 2015).

Data Availability

Bisulfite sequencing data are available as part of the seventh BLUEPRINT data release (September 2015). Links to the corresponding data repositories are provided on the BLUEPRINT website[4]. Full RnBeads analysis reports are available from the RnBeads methylome resource website[5].

4.2.2 Results

Quality Control and Filtering of CpGs

The dataset contains DNA methylation levels for a total of 28,571,135 CpGs annotated in the human genome. Individual samples covered between 23 and 28 million CpGs at different read depths (Figure 4.3), with a median of 31 reads covering a CpG. It is important to note that monocyte and neutrophil samples that were sequenced relatively early in the project generally had higher read coverages than other samples and that two monocyte samples were sequenced at significantly higher read depths. Running the RnBeads pipeline removed CpGs which (i) overlapped with annotated SNPs in the genome, which (ii) were covered by less than five sequencing reads in more than 40 samples (50 % of the dataset), (iii) represented high-coverage outliers exceeding 50 times the 95th percentile of read coverage or (iv) were located on sex chromosomes. This filtering procedure resulted in 23,607,313 CpGs that were used in subsequent analysis steps.

Characterization of Hematopoietic Methylomes

Unsupervised statistical learning methods were employed in order to explore between-cell-type heterogeneity. Reducing the high-dimensional space of methylation levels to two dimensions using PCA or MDS revealed patterns of inter-sample relationships (Figure 4.4a). We characterized the dataset in terms of mean methylation levels in candidate regulatory regions which were defined based on a consensus chromatin state segmentations and include putative enhancers and transcriptionally active regions [Zerbino *et al.* 2015]. Overall, patterning in DNA methylation pertaining to different cell types is clearly visible and dominates other effects and biases incurred by factors such as donor sex and tissue origin. Myeloid cells are more similar to each other than to other blood cells and form a relatively tight cluster while lymphoid cells appear more diverse. Particularly, B cells contribute to the between-sample differences on which the dimension reduction is based. Not surprisingly, myeloid leukemia cells also exhibit highly heterogeneous DNA methylation patterns. These observations are also confirmed by hierarchical clustering analysis (Figure 4.4b). Most of the regulatory regions with highest DNA methylation variability across the entire dataset are markedly hypomethylated in myeloid cells compared to other cell types. Interestingly, the groups B and T lymphocytes are each divided into two subgroups (Figure 4.4b): one cluster contained predominantly naive B and T cells while effector and memory cells were exclusivity assigned to

[3] ftp://ftp.ebi.ac.uk/pub/databases/blueprint/
[4] http://www.blueprint-epigenome.eu
[5] http://rnbeads.mpi-inf.mpg.de/methylomes.php

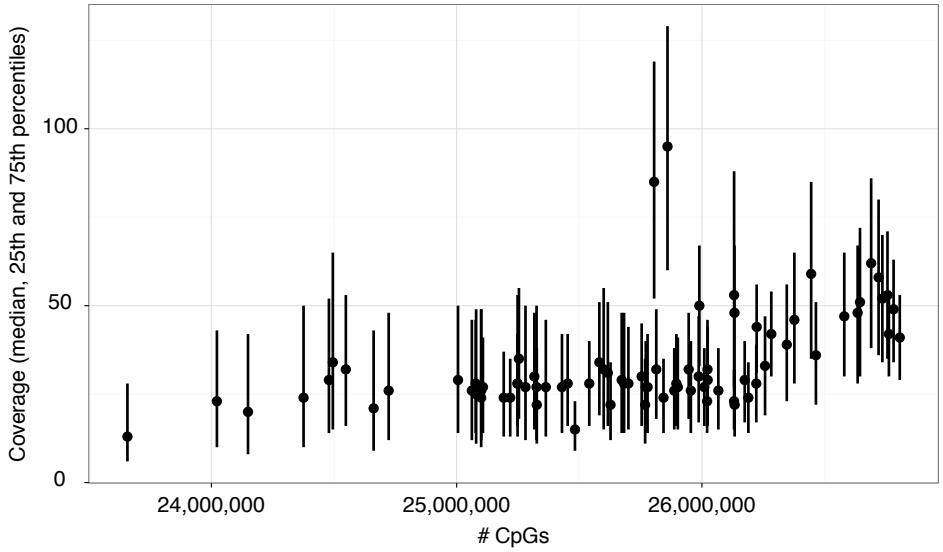


Figure 4.3: CpG read coverage distribution in the BPDIFF dataset. The number of covered CpGs and median coverages are shown for each sample. Vertical bars depict inter-quartile ranges.

the other cluster. Notably, the naive cluster exhibited higher similarity with the myeloid samples than with the effector lymphoid cluster. The four endothelial samples and erythroblasts group together with effector lymphocytes in their respective own, smaller subclusters. Leukemia samples form a single cluster with heterogeneous patterns of DNA methylation.

Differential Methylation between Monocytes and Neutrophils

Monocytes and neutrophils represent two of the most common nucleated myeloid cell types in mammalian blood. Neutrophils, which contribute 50 % to 60 % of leukocytes, are especially abundant. Both cell types derive from common progenitors and carry out important effector functions in innate immunity. They play essential roles as first-line defenders against foreign pathogens: both are phagocytic cell types and can trigger or repress inflammatory responses, yet they differ in their morphology and capacity to release inflammatory cytokines [Dale *et al.* 2008]. Neutrophils are recruited to the site of infection in order to neutralize pathogens by ingesting them or releasing cytotoxins. They are short-lived, while active monocytes can proliferate and differentiate into macrophages or dendritic cells and thus also play a role in adaptive immunity.

Here, differences in DNA methylation signatures between monocytes and neutrophils are elucidated. The methylomes of eight monocytes (six extracted from adult peripheral blood and two from cord blood of newborns) and six neutrophils (four peripheral blood and two cord blood) were compared. Differentially Methylated Regions (DMRs) were identified using the rank-based approach employed by RnBeads (cf. Chapter 3.2), adjusting for donor sex and tissue origin in the computation of p-values for differential methylation. On the genome-wide level, DMRs tend to be hypomethylated in neutrophils as compared to monocytes (Figure 4.5a). Furthermore, a more gene-centric view revealed a set of promoters with lower methylation levels in neutrophils (Figure 4.5b). Enrichment analyses for GO terms of the genes associated with these

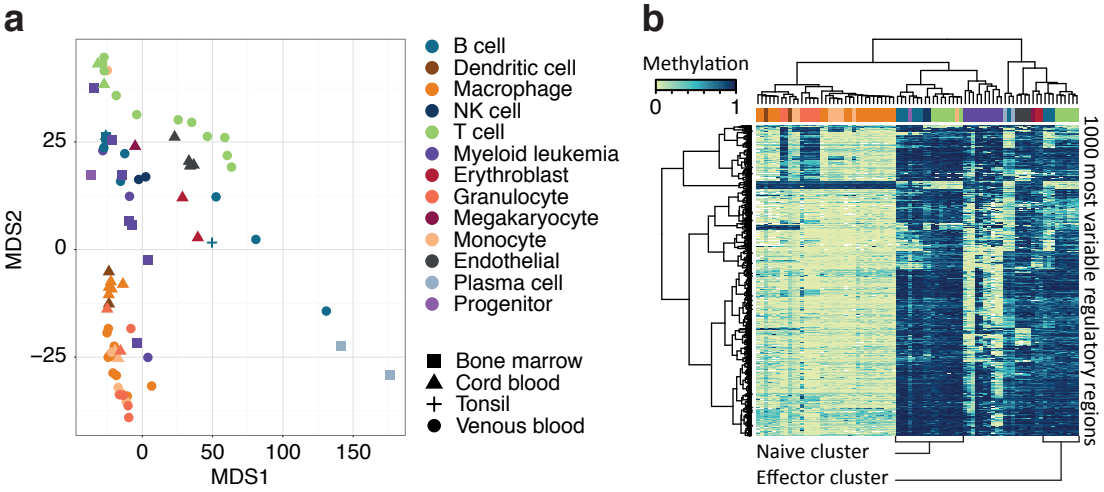


Figure 4.4: Unsupervised analysis of the BPDIFF dataset. **(a)** MDS plot of the dataset. **(b)** Heatmap and hierarchical clustering (using Ward's linkage method; as implemented by the method parameter `ward.D` in the `hclust` function of the STATS R package). In both cases Euclidean distances have been computed from mean methylation levels in putative regulatory regions. Colors and point shapes denote cell types and tissues of origin. Red/orange and blue/green colors denote cells of the myeloid and lymphoid lineages, respectively.

promoters confirmed neutrophil-specific activity to be highly represented (Table 4.1). A large portion of significantly enriched terms were associated with immune response and epigenetic regulation.

The promoter region of the *DEFA4* gene was found among the highest-ranking differentially methylated promoters (Figure 4.5b). DEFA4 belongs to the defensin family of cytotoxic peptides which is involved in antimicrobial defense. Inspection of basepair-resolution DNA methylation levels revealed reduced methylation across the gene locus and almost complete loss around the TSS in neutrophils (Figure 4.6), indicating epigenetic regulation of gene expression.

4.2.3 Discussion

We analyzed of one of the largest collections of whole-methylome data for blood cells. The respective dataset spans the breadth of the hematopoietic hierarchy and captures epigenetic variability across differentiated cell types. Our analysis shows that hematopoietic cell identity is inscribed in DNA methylation signatures. We observed DNA methylation variability which were characteristic of cell type in putative regulatory regions. Methylation variability in cells of the lymphoid lineage was higher when compared to myeloid cells, which could be indicative of an increased regulatory plasticity in cells involved in the adaptive immune response. Furthermore, most variably methylated regulatory regions exhibited marked hypomethylation in myeloid cells, potentially indicating a default differentiation trajectory of hematopoietic progenitors towards the myeloid lineage, that is suppressed by DNA methylation in the lymphoid lineage [Ji *et al.* 2010; Bröske *et al.* 2009]. DNA methylation in regulatory regions was also indicative of the activation state of lymphoid cells. Section 4.3, which characterizes

Table 4.1: GO terms enriched in promoters hypomethylated in neutrophils compared to monocytes

GO ID	P-value	Odds-ratio	GO term
GO:0050832	0	122.248	defense response to fungus
GO:0051707	0	13.6664	response to other organism
GO:0009607	0	13.0245	response to biotic stimulus
GO:0045087	0	10.7557	innate immune response
GO:0006954	0	12.9736	inflammatory response
GO:0016045	2.00E-04	127.425	detection of bacterium
GO:0098581	3.00E-04	90.9821	detection of external biotic stimulus
GO:0002376	4.00E-04	5.8884	immune system process
GO:0031640	5.00E-04	70.7361	killing of cells of other organism
GO:0019731	6.00E-04	65.8491	antibacterial humoral response
GO:0010043	9.00E-04	53.0208	response to zinc ion
GO:0051716	0.0011	5.4018	cellular response to stimulus
GO:0009605	0.0011	5.3563	response to external stimulus
GO:0010963	0.0012	NaN	regulation of L-arginine import
GO:0044356	0.0012	NaN	clearance of foreign intracellular DNA by conversion of DNA cytidine to uridine
GO:0035872	0.0014	41.4674	nucleotide-binding domain, leucine rich repeat receptor signaling pathway
GO:0070301	0.0017	37.3897	cellular response to hydrogen peroxide
GO:0070988	0.0021	34.0402	demethylation
GO:0070488	0.0023	900.2941	neutrophil aggregation
GO:0071557	0.0023	900.2941	histone H3-K27 demethylation
GO:0050830	0.0025	30.7339	defense response to Gram-positive bacterium
GO:0001816	0.0033	7.7384	cytokine production
GO:0090467	0.0035	450.1176	arginine import
GO:1902023	0.0035	450.1176	L-arginine transport
GO:0031349	0.0043	10.5789	positive regulation of defense response
GO:0035821	0.0045	22.6518	modification of morphology or physiology of other organism
GO:0032827	0.0047	300.0588	negative regulation of natural killer cell differentiation involved in immune response
GO:0070269	0.0047	300.0588	pyroptosis
GO:0006919	0.0047	22.1221	activation of cysteine-type endopeptidase activity involved in apoptotic process
GO:0006950	0.0047	4.0482	response to stress
GO:1901565	0.0054	9.739	organonitrogen compound catabolic process
GO:0042742	0.0058	20.0517	defense response to bacterium
GO:0010269	0.0059	225.0294	response to selenium ion
GO:0070383	0.0059	225.0294	DNA cytosine deamination
GO:0030307	0.0069	18.0964	positive regulation of cell growth
GO:0032815	0.007	180.0118	negative regulation of natural killer cell activation
GO:0031638	0.008	16.6579	zymogen activation
GO:0009253	0.0082	150	peptidoglycan catabolic process
GO:0051092	0.0083	16.3685	positive regulation of NF-kappaB transcription factor activity
GO:2001056	0.0093	15.4299	positive regulation of cysteine-type endopeptidase activity
GO:0009635	0.0094	128.563	response to herbicide
GO:0032119	0.0094	128.563	sequestering of zinc ion
GO:0051597	0.0094	128.563	response to methylmercury

Enrichment in biological process ontology terms was computed for the 100 promoters most hypomethylated in neutrophils as compared to monocytes according to the combined rank. Terms with a p-value less than 0.01 are shown.

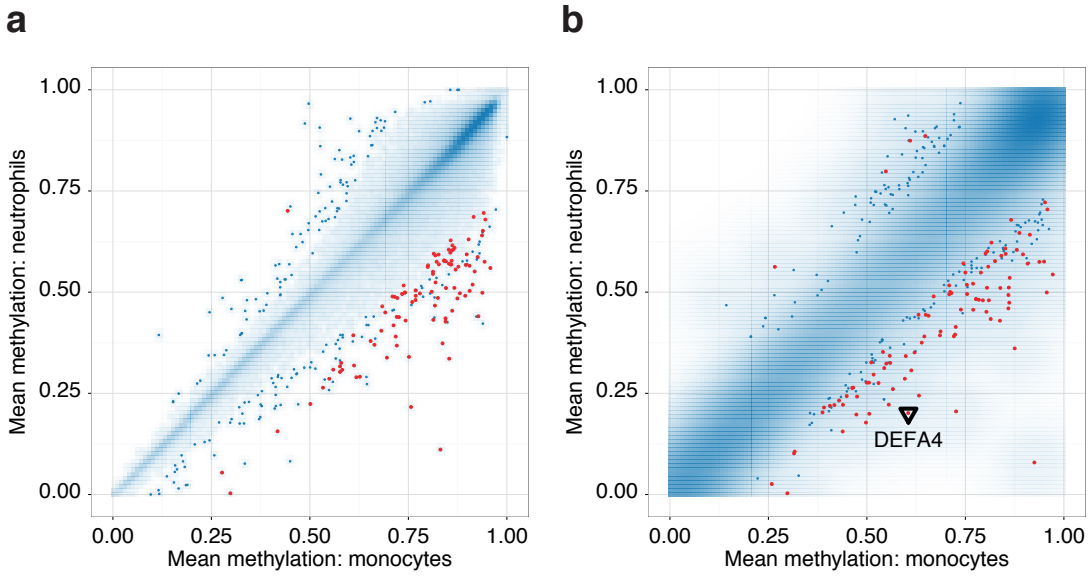


Figure 4.5: Differential DNA methylation between monocytes and neutrophils. Scatterplots show the mean DNA methylation levels in monocytes and neutrophils for **(a)** 5-kb tiling windows and **(b)** promoters. Point density is denoted by blue shading and the 100 highest ranking differentially methylated regions have been highlighted in red in each panel. The *DEFA4* promoter is marked by a triangle and Figure 4.6 shows a detailed view of DNA methylation patterns in that locus.

the DNA methylation dynamics during T cell memory formation, explores this issue further.

In addition to genome-wide variability in regulatory elements, we also find characteristic changes in promoter methylation of genes with cell-type-specific functions. We compared monocytes and neutrophils as two common myeloid cell types and observed high agreement in their genome-wide DNA methylation profiles. However, we also found neutrophil-specific activity of genes whose promoters were hypomethylated in neutrophils, indicating epigenetic gene regulation and providing an example of specific immune function reflected in DNA methylation.

Our descriptive analyses provide a starting point for a more detailed characterization of the DNA methylation dynamics in blood. Following up on our findings in differentiated blood cells, in Section 4.4, we take on a progenitor-focused view on the hematopoietic system and employ computational modeling of cellular identity using DNA methylation signatures in regulatory elements.

Importantly, our analyses were facilitated by the RnBeads software (Chapter 3.2), illustrating the utility of analysis pipelines that provide an integrated view on large-scale datasets and also enable the user to address specific questions in-depth. The RnBeads analysis reports are part of the BLUEPRINT secondary data release and provide a resource for further integrative data analysis projects. The reference methylome maps analyzed here capture variability across blood cell types and could therefore be used as a point of reference for methylation-based cell-type deconvolution on the genome-scale. Furthermore, DNA methylation profiles obtained from samples of malignant cells could

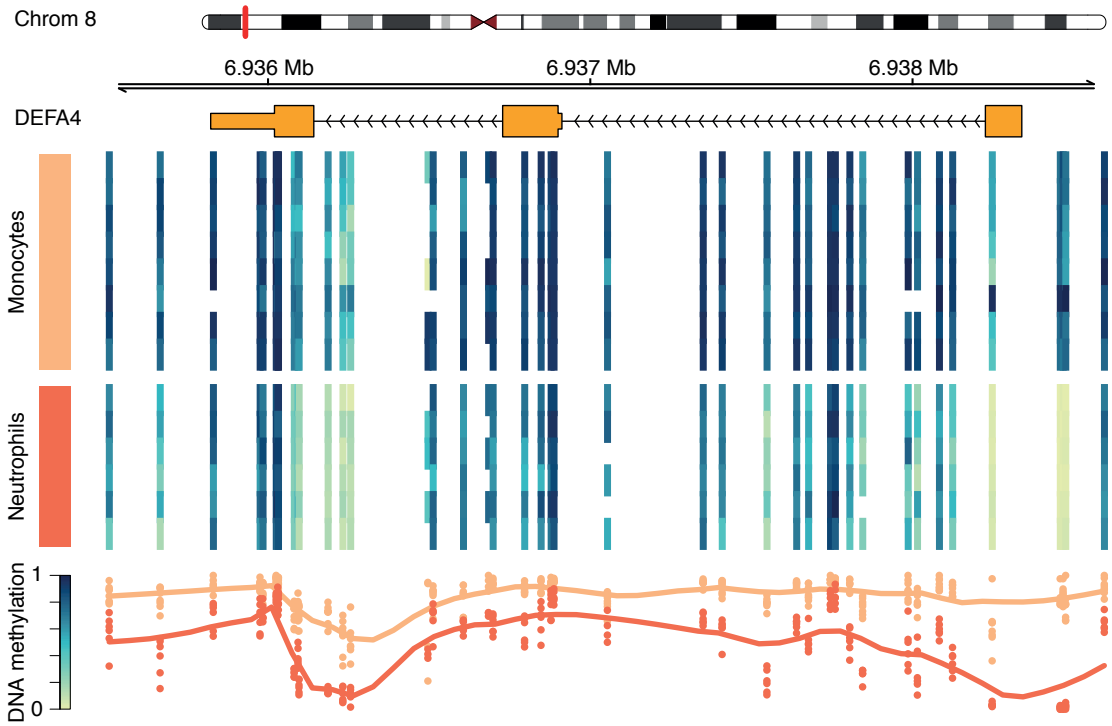


Figure 4.6: Neutrophils are hypomethylated compared to monocytes at the *DEFA4* gene locus. The *DEFA4* gene model (exons interspersed by introns) is annotated in orange. The heatmap shows DNA methylation levels at individual covered CpGs (columns of the heatmap). Each row represents a monocyte or neutrophil sample in the dataset. Local Regression (LOESS) [Cleveland 1979] on the methylation levels was employed in order to obtain smoothed estimates for monocytes and neutrophils respectively (bottom).

be compared to the reference data in order to dissect tumor heterogeneity based the epigenomic resemblance to healthy cells.

In summary, our observations are consistent with a role of DNA methylation patterns located in enhancers and other regulatory elements as key contributors to cell identity. Here, we provide a mostly descriptive analysis that entails implications for the involvement of DNA methylation dynamics in regulating cell identity in the immune-response, hematopoietic differentiation and blood-related malignancies.

4.3 Epigenome Reprogramming in Human T Cells

T cells represent a key component of adaptive immunity and are therefore considered critical players in disease. So far, research focused on vaccine development has been directed at eliciting immune memory by antibody formation in the B cell lineage. However, T cell memory also plays an important role in the organism's arsenal of defense. Both, $CD4^+$ T helper cells and $CD8^+$ cytotoxic T lymphocytes derive from double positive cells ($CD4^+$, $CD8^+$) located in the thymus (Figure 4.1) but carry out different effector functions: while $CD8^+$ T cells are capable of killing infected cells and clearing pathogens, $CD4^+$ T lymphocytes play a stimulatory role by activating other immune cells. When an organism is exposed to a pathogen, **Naive T cells (TNs)** become activated and specialize when they recognize antigens presented on the surface of other cells by the major histocompatibility complex (MHC) via their T-cell receptor (TCR) for the first time. This specialization occurs for each cell individually and cells specifically adapt to a given antigen. Upon encountering a pathogen, the majority of cells become effector cells, but a subset of cells enters a stage of immune memory in which they are capable of rapidly responding to subsequent exposure to the same pathogen. During the adaptive immune response, the specialized T cells rapidly expand until the pathogen is cleared. A contraction phase follows in which up to 90 % of effector cells undergo apoptosis. However, a pool of **memory T cells** persists [Kaech and Cui 2012]. Memory T cells are the most abundant lymphoid cell type in the adult human body [Farber *et al.* 2014]. They are characterized by a significantly lowered activation threshold and enhanced cytokine production upon a subsequent re-encounter of a pathogen. Memory T cell clones can persist for years to decades in human [Farber *et al.* 2014] and are mostly quiescent or exhibit intermittent proliferative activity [Pepper and Jenkins 2011]. Several subpopulations of memory T cells have been identified and they differ in their surface marker expression, response to and production of cytokines as well as tissue localization. **Central Memory T cells (TCMs)** show the closest resemblance to the naive phenotype. They are located in lymphoid tissues, but also circulate through blood and possess an increased proliferative capacity [Farber *et al.* 2014]. In contrast, **Effector Memory T cells (TEMs)** preferentially locate in peripheral tissue and exhibit only limited proliferation. They carry out certain effector functions such as the production of cytotoxins [Pepper and Jenkins 2011; Farber *et al.* 2014]. Furthermore, **$CD45RA^+$ Memory T cells (TEMRAs)** have recently been characterized [Henson *et al.* 2012]. They also possess reduced proliferative capacity, carry out effector functions and play a role in chronic inflammation.

Despite intensive research in the area, the precise ontogeny of human T cell memory formation is still under debate and several alternative models for the relationships between effector and memory cell populations exist [Ahmed *et al.* 2009; Kaech and Cui 2012; Restifo and Gattinoni 2013; Pepper and Jenkins 2011; Farber *et al.* 2014]. The first model states that during pathogen exposure, TEMs emerge directly from effector cell

populations and give rise to TCMs as a generalized, long-lived pool for immune memory (dotted arrows in the T cell branches in Figure 4.1). According to another model, TCMs and TEMs are both derived directly from TNs (dashed arrows in Figure 4.1). In contrast, a progressive differentiation model more closely resembles the process of fate determination in stem cell (solid arrows in Figure 4.1). In this model, TNs resemble the most general state and upon activation can give rise to TCMs which are still long-lived and capable of self-renewal. Populations of memory stem cells represent potential intermediate precursors in this line of differentiation [Farber *et al.* 2014]. TCMs in turn are direct ancestors to TEMs which already exert effector functions. Effector T cells as the most differentiated and short-lived state emerge from them.

Given the distinct identities of memory T cell populations, it is logical to assume that their epigenomes play a vital role in regulating immune memory formation. So far, studies investigating T cell epigenetics during human memory formation on a genome scale have been lacking. Related efforts include the work by Russ *et al.* [2014] who profiled H3K4me3 and H3K27me3 histone modifications in CD8 positive naive, effector and memory T cells of mice that had been exposed to an influenza virus strain. Crompton *et al.* [2015] assayed the same histone marks in naive cells, T memory stem cells, central memory cells and effector memory cells. Hashimoto *et al.* [2013] used a methylation-sensitive restriction approach to assay DNA methylation at a subset of murine CpGs in $CD4^+$ T memory cells. Another study describes differential DNA methylation between human $CD4^+$ naive and memory T cells in selected CpGs in using locus-specific bisulfite sequencing and correlated those patterns to gene expression [Komori *et al.* 2015]. Kulis *et al.* [2015] investigated the DNA methylation dynamics during human B cell maturation using WGBS and 450K and their dataset also included different memory stages.

We sought to identify and characterize DNA methylation signatures that contribute to human $CD4^+$ T cell memory formation. In an effort of the DEEP consortium, epigenomes of TN, TCM, TEM and TEMRA cells were profiled. Utilizing this resource and the computational pipelines established in DEEP, we explore the genome-wide DNA methylation patterns at various stages of memory formation and relate their dynamics to hypothesized differentiation pathways. Our results reveal stepwise losses of DNA methylation at the genome scale and provide further evidence for a progressive model of human T cell differentiation.

4.3.1 Methods

Cell Isolation and Sequencing Library Preparation

Genome-wide DNA methylation in human $CD4^+$ TN, TCM, TEM and TEMRA cells was assessed within the context of DEEP. TN, TCM and TEM were obtained from peripheral blood by the groups of Alf Hamann and Julia Polansky at the Deutsches Rheuma-Forschungszentrum (Berlin, Germany). TEMRA cells were sorted in the group of Birgit Sawitzki at the Institute of Medical Immunology (Charité University Medicin, Berlin, Germany). Samples from 3 to 10 female donors were pooled to obtain sufficient material for sequencing and to mitigate inter-individual sample variance. Sequencing libraries for WGBS and NOMe-seq were prepared for two replicate pools of TN, TCM and TEM cells. The dataset contains only one pool of TEMRA cells which was assayed by NOMe-seq. Detailed methods for cell isolation and sequencing library preparation can be found in [Durek *et al.* 2016].

Processing of Bisulfite Sequencing Reads

DNA methylation data obtained from WGBS and NOMe-seq was processed using uniform processing pipelines established in the DEEP project. Primary data processing was performed by project partners at the Deutsches Krebsforschungszentrum (DKFZ; Heidelberg, Germany). In brief, the SeqPrep tool[6] was used to trim adapter sequences and reads were mapped to an *in silico* bisulfite converted version of the human reference genome (hg19/GRCh37d5 assembly) using MethylCtools [Hovestadt *et al.* 2014]. BWA (version 0.6.2-tpx) was employed using default parameters with the exception of setting the quality trimming threshold (`-q`) to 20 and disabling Smith-Waterman alignment for the unmapped mate read (`-s`). Duplicate reads were removed using Picard[7] and reads with an alignment score of less than one were discarded.

A standardized pipeline for determining DNA methylation levels at individual CpGs was developed in collaboration with Karl Nordström (Saarland University, Saarbrücken, Germany). The pipeline is based on the Bis-SNP tool [Liu *et al.* 2012] and also employs Picard, samtools [H. Li *et al.* 2009], BamUtil[8] and UCSC tools [Kent *et al.* 2010]. In the quantification of CpG methylation from NOMe-seq data, only methylation levels for cytosines in HCG contexts (where H represents any base other than guanine) were retained in order to avoid confounding by the artificial methylation step. The DEEP data analysis center provides detailed process descriptions[9] of the employed computational pipelines in a standardized XML format [Ebert *et al.* 2015].

Analysis reports were created using RnBeads version 1.1.4 and further plots were generated using custom R scripts. In addition to the default regions of RnBeads, methylation levels were averaged in genomic tiling windows of size 1 kb and putative regulatory regions (BLUEPRINT edition of the Ensembl regulatory build [Zerbino *et al.* 2015]). RnBeads filtering steps removed CpGs that overlapped with SNPs or that were covered by fewer than five sequencing reads in more than seven samples (60 % of the dataset). Employing these criteria, 26,117,504 CpGs were retained for further analysis. Differential methylation between different cell types was quantified using RnBeads' ranking approach (cf. Chapter 3.2). Enrichment analysis in regions of interest were conducted using LOLA [Sheffield and Bock 2016] with its core and extended databases of genomic and epigenomic features (using the November 2015 version).

Data Availability

Sequencing read data have been deposited to European Genome-phenome Archive (EGA) under accession number `EGAS00001001624`. Full RnBeads analysis reports are available from the RnBeads methylome resource website[10].

4.3.2 Results

In order to provide a global perspective on inter-sample variability of DNA methylation, unsupervised statistical learning methods were employed. Projecting single-CpG-resolution methylation levels onto the first two principal components revealed that DNA methylation patterns are characteristic of the samples' cell types (Figure 4.7). These

[6] http://github.com/jstjohn/SeqPrep
[7] http://broadinstitute.github.io/picard/
[8] http://genome.sph.umich.edu/wiki/BamUtil
[9] http://doi.org/10.17617/1.2W
[10] http://rnbeads.mpi-inf.mpg.de/methylomes.php

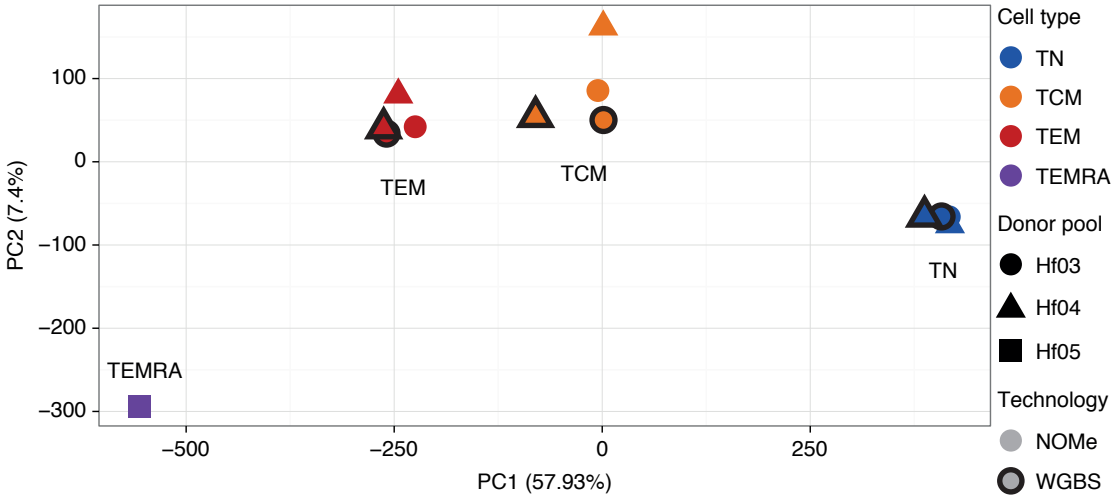


Figure 4.7: Stages of T cell memory formation are distinguishable by their DNA methylation signatures. PCA was used to project T cell samples onto two dimensions from DNA methylation measured at individual CpGs. Percentages annotate the variance explained by the first and second principal component. Point colors and shapes denote cell types and donor pools. Outlined and plain points represent WGBS and NOMe-seq experiments respectively.

cell-type-specific differences dominate variability introduced by other sources such as donor pool and experimental protocol. Although not statistically significant due to the relatively small sample size, particularly the first principal component, which explained 57.9 % of the dataset variance, is visibly associated with a sample's annotated cell type (p-value < 0.12, Kruskal-Wallis test; Figure 4.7). Hierarchical clustering of samples confirmed the presence of cell-type-specific DNA methylation signatures (Figure 4.8). The memory cell states (TCMs and TEMs) exhibited more similar DNA methylation patterns to each other than to naive T cells.

Genome-wide inspection of DNA methylation levels exposed global hypomethylation in the memory cell stages when compared to the naive stage (Figure 4.8). To obtain an overview of methylation dynamics across cell types, genome-wide tiling regions were grouped according to their methylation patterns using k-means clustering (Figure 4.9). Evaluating different numbers of clusters by visual inspection, we selected $k = 5$. Fewer clusters did not fully capture all modes of cell-type-dependent differences and increasing k led to multiple clusters containing highly similar patterns (data not shown). The five resulting region clusters represent modes of cell-type-associated changes. Two of these clusters represent regions exhibiting consistently low and high methylation levels across the dataset (clusters 1 and 4 in Figure 4.9 respectively). The regions in the remaining three clusters are characterized by intermediate to high methylation levels in TNs, hypomethylation in TCMs and TEMs and a further decrease in TEMRAs. Generally, TCMs more closely resemble TNs than TEMs in terms of genome-wide DNA methylation levels. Taken together, in their most simple interpretation, these findings argue for a stage-wise decrease in DNA methylation levels in the order TN→TCM→TEM→TEMRA and are in line with the progressive differentiation model

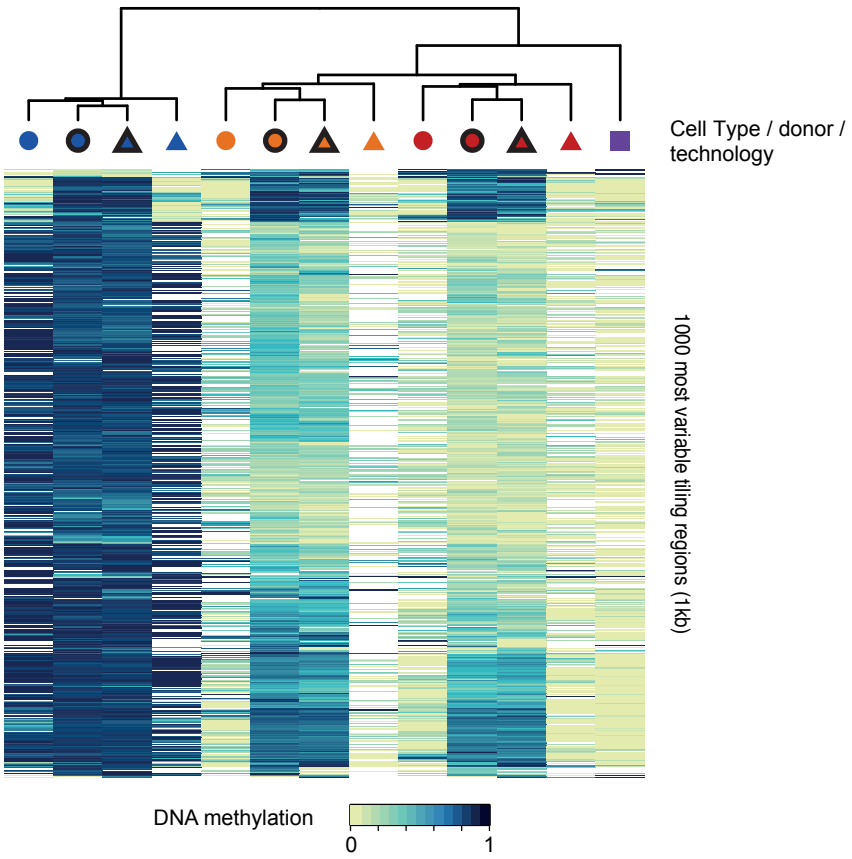


Figure 4.8: T cells become increasingly hypomethylated during memory formation. The heatmap shows the 1,000 most variable tiling regions (size: 1 kb). Hierarchical clustering employing Manhattan distance and complete linkage has been used to group samples. The same shape and color coding as in Figure 4.7 is applied for sample annotation.

of T cell memory formation. This is also reflected in the two-dimensional view of principal components depicted in Figure 4.7. Differences linked to memory formation were most pronounced in cluster 3 which is characterized by intermediate methylation levels in naive T cells. This is consistent with previous observations stating that widespread changes in DNA methylation occur in Partially Methylated Domains (PMDs) [Lister *et al.* 2009; Berman *et al.* 2012; Hon *et al.* 2012]. These PMDs frequently overlap with enhancers and other regulatory elements.

Focusing on the comparison of TCMs to TNs, RnBeads' rank-based method was employed in order to identify genomic regions differentially methylated during memory formation. The highest-ranking differentially methylated promoters comprise multiple genes related to immunity, cell signaling and hematopoiesis such as the chemokine receptor *CCR5* (Figure 4.10). Putative regulatory regions were generally hypomethylated in TCMs compared to TNs (Figure 4.11a). Enrichment analysis for genomic and epigenomic features revealed associations with DNase hypersensitive regions that are specific to blood-related cells (Figure 4.11b). Moreover, ChIP-seq peaks in lymphoblastoid cell lines for transcription factors involved in the regulation of hematopoiesis such as BATF, RUNX3, IRF4 and NF-κB significantly overlap with differentially methylated regulatory regions.

In the dataset analyzed here, CpG methylation was quantified by WGBS and NOMe-seq. NOMe-seq samples were generally sequenced at a lower sequencing depth compared to WGBS (approximately 10 to 50 % fewer reads per sample). Additionally, CpGs that could potentially be biased by the artificial methylation step (CpGs in GCG context) were discarded. This protocol resulted in a median of 19,345,622 assayed CpGs per NOMe-seq sample compared to 25,838,238.5 in case of WGBS. The median number of sequencing reads covering a CpG was 13 and 23.5 respectively. Despite these differences in read coverage (as exemplified by white gaps in Figure 4.8), overall, the quantified methylation levels are consistent between these two bisulfite sequencing protocols: Pearson correlations coefficients between 0.91 and 0.95 were observed when comparing CpG methylation between technical replicates of the same cell-type and donor-pool combination while biological replicates comparing the two donor pools in the same cell type using WGBS yielded coefficients ranging from 0.93 to 0.96. Furthermore, the unsupervised learning approaches discussed above (Figures 4.7, 4.8) confirmed that technical variation is clearly dominated by biological, cell-type dependent variation.

4.3.3 Discussion

Genome-wide DNA methylation patterns are capable of distinguishing between different naive and memory T cell types. The observed differences in DNA methylation support the hypothesis of a stepwise loss of methylation along the TN→TCM→TEM→TEMRA axis and are thus consistent with the progressive differentiation model of T cell memory formation. These results are further substantiated by experimental evidence on decreasing telomere lengths and differentiation potential along this axis [Restifo and Gattinoni 2013; Farber *et al.* 2014]. Additional datasets, potentially on the level of single cells, could help to shed further light on the precise dynamics during memory formation.

Dynamic DNA methylation is preferentially located in genomic regions with intermediate to high levels of methylation in naive T cells (PMDs). These regions overlap with regulatory elements that exhibit epigenomic marks indicative of open chromatin and TFBSs associated with hematopoiesis. Furthermore, PMDs have been linked to LADs

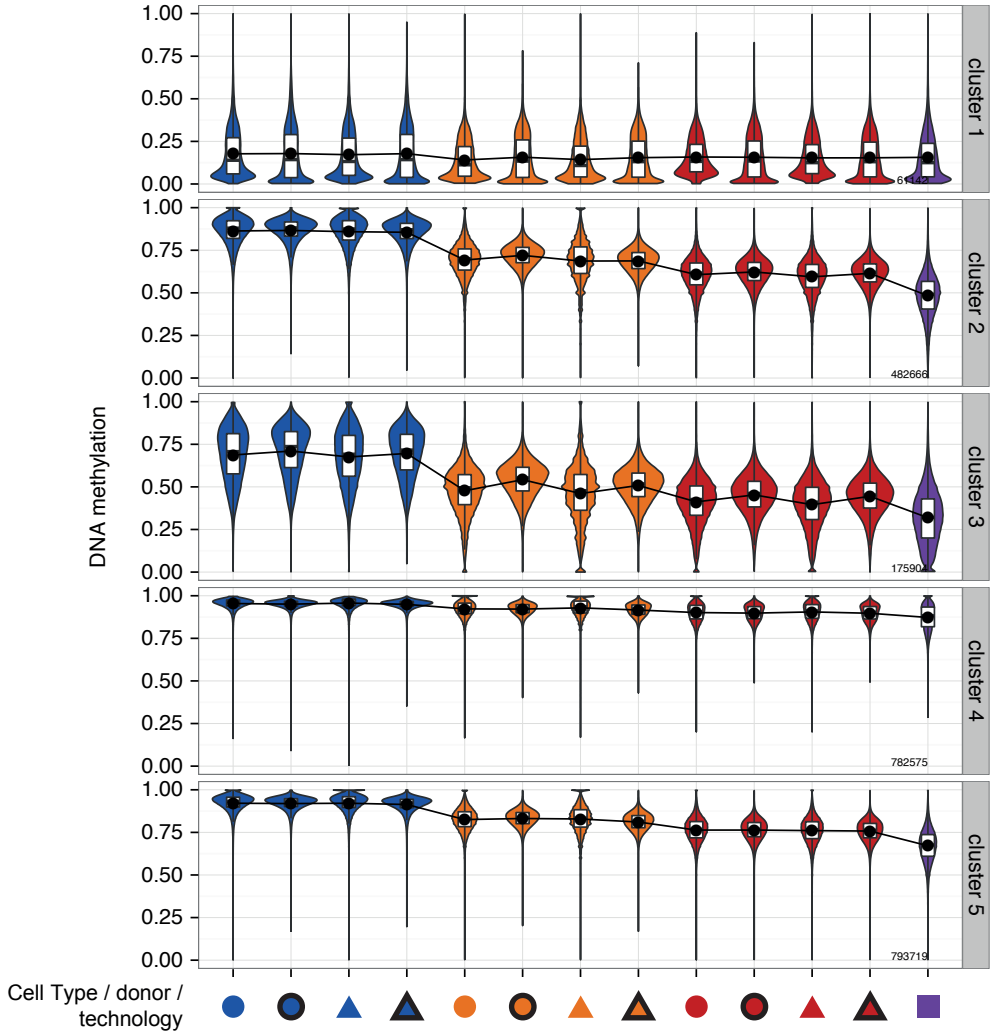


Figure 4.9: Clusters of genomic regions exhibit varying degrees of hypomethylation during memory formation. Violin and boxplots show the distribution of DNA methylation levels in five clusters of 1-kb genomic tiling regions. These clusters were obtained from a k-means clustering ($k = 5$) based on mean methylation levels in those regions. Cluster mean methylation levels are denoted by dots that are connected by lines. Numbers in the lower right corners denote the number of regions in each cluster. The sample annotation uses the same shape and color coding as in Figure 4.7.

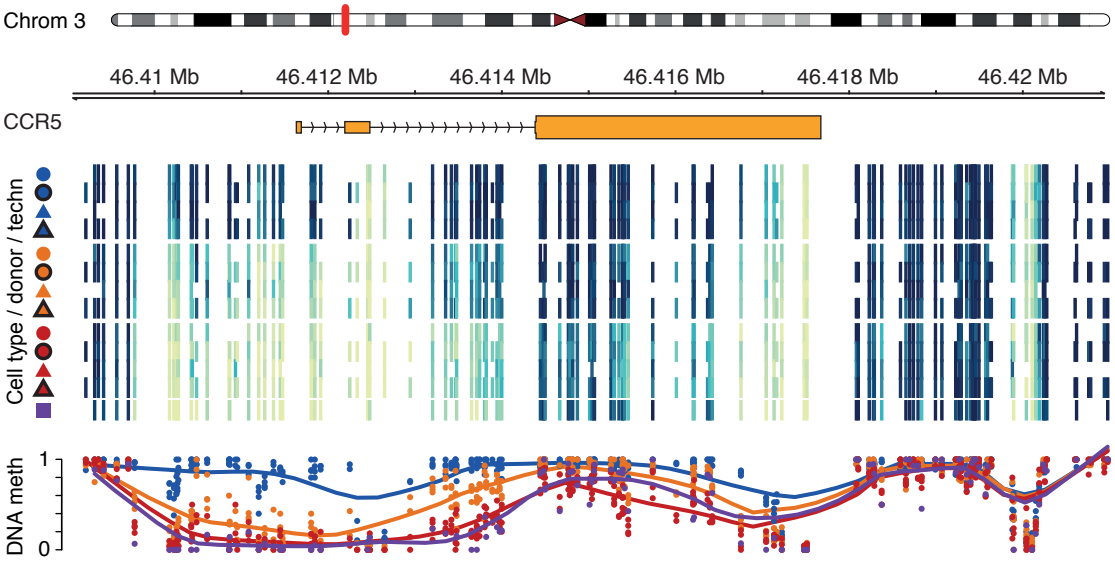


Figure 4.10: The *CCR5* gene locus is hypomethylated during memory formation. The heatmap shows DNA methylation levels at individual CpGs (columns of the heatmap). Each row represents a sample in the dataset. LOESS on the methylation levels was employed in order to obtain smoothed estimates for each cell type (bottom). The same color and shape encoding as in Figure 4.7 was used.

and regions of late replication timing in the context of cell proliferation [Aran *et al.* 2011; Berman *et al.* 2012]. This raises the hypothesis that DNA methylation could reflect a cell's proliferative history: according to the progressive differentiation model of T cell formation, memory cells rapidly divide after encountering an antigen, giving rise to a pool of TCMs, which are mostly resting. Upon reactivation, TCMs reenter a proliferative stage and give rise to TEMs, characterized by low proliferative activity. Along this path of differentiation, DNA methylation is lost in a stepwise manner.

Interestingly, inspecting genes that lose DNA methylation along this path revealed that methylation changes frequently occurred in loci representing Small Nucleolar RNAs (snoRNAs). snoRNAs have previously been attributed with roles in mRNA splicing and miRNA regulation [Williams and Farzaneh 2012]. It remains to be tested whether they are indeed regulated by DNA methylation or whether the observations represent spurious associations due to the fact that snoRNAs are primarily derived from body regions of coding genes which could overlap with other regulatory elements.

In addition to DNA methylation, the DEEP consortium has profiled chromatin accessibility, six different histone modifications as well as regulatory and messenger RNA in the discussed cell populations, thereby producing full epigenomes according to the IHEC definition. Durek *et al.* [2016] provide a more detailed discussion on these marks. In brief, the progressive differentiation model is also supported by other data types, particularly by expression patterns of ncRNA, which are consistent with DNA methylation changes. Further integrative analyses revealed that changes in DNA methylation in regions of accessible chromatin are associated with changes in gene expression and confirm the enrichment of potential transcription factor binding events in differentially methylated regions.

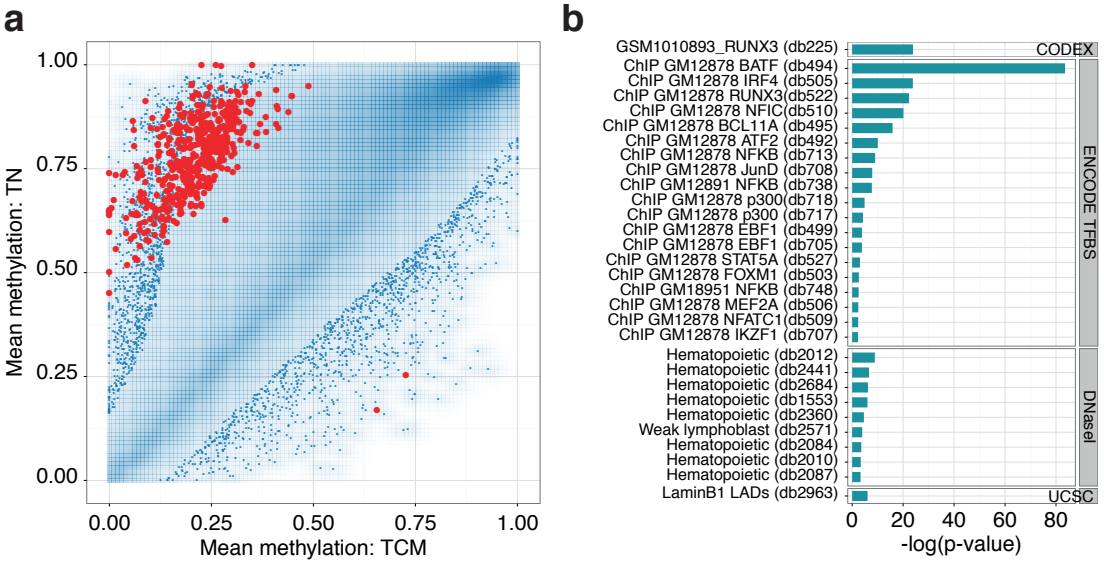


Figure 4.11: Regulatory regions are hypomethylated in TCMs compared to TNs. **(a)** Scatterplot showing the mean DNA methylation levels of putatitive regulatory regions in TCMs and TN. Point density is denoted by blue shading and the 500 highest ranking differentially methylated regions have been highlighted in red. **(b)** *p*-values for selected terms obtained from a LOLA enrichment analysis of the 500 highest ranking differentially methylated regulatory regions. These terms include ChIP-seq peak regions identified in the ENCODE and CODEX projects and DNaseI-seq experiments and are contained in the LOLA database. GM12878 denotes a lymphoblastoid cell line profiled in the ENCODE project.

Little is currently known about the role of TEMRA cells. The methylation data presented here suggests a placement of TEMRA downstream of TEMs in the trajectory of memory formation. However, with only one TEMRA sample at hand, it was not possible to reach robust conclusions and DEEP is currently generating further profiles for this interesting cell type. In addition, this study focuses on T helper cells ($CD4^+$) in human. It remains to be investigated to what extent the presented findings generalize to memory formation in $CD8^+$ cytotoxic T cells and to other vertebrates.

In summary, T cell memory formation plays an important role in adaptive immunity. Our results suggest that the differentiation stage of memory cells is reflected in their DNA methylation signatures. These signatures might therefore provide useful tools and indicators in the context of immunity-related diseases, allergies as well as vaccine development. Understanding the (epigenomic) identity of immune cells could also result valuable for the design and application of future cell therapies.

4.4 DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation

The base of the hematopoietic system consists of stem and progenitor cells that can give rise to all blood cell types (Figure 4.1). Epigenome remodeling plays a crucial part in the process of lineage commitment. However, while differentiated blood cells are readily available for sampling and epigenomic profiling, only small amounts of cell material can be obtained from hematopoietic stem and progenitor cells. It is therefore not surprising that the epigenomes of the early stages of hematopoietic differentiation on the genome-scale in human have been poorly characterized, so far. Within the context of BLUEPRINT, we applied protocols for low-input and single-cell whole genome bisulfite sequencing (µWGBS) to human HSCs, MPPs and to various lineage-committed progenitor cell types purified by FACS (Figure 4.1; Table 4.2). The resulting dataset constitutes a valuable resource for reassessing current models of human hematopoiesis from an epigenomic perspective.

Here, we provide a detailed account on the DNA methylation dynamics during hematopoietic stem and progenitor cell differentiation in regulatory regions of the genome. We found notable differences in the methylation patterns between stem cells, myeloid progenitors and cells of the lymphoid lineage. Lineage-specific changes in DNA methylation appear to be linked to cell-type-specific chromatin accessibility. Moreover, statistical learning models were able to accurately infer cell types from DNA methylation signatures and could be used for data-driven reconstruction of the human hematopoietic system. Our observations illustrate the power of DNA methylation analysis for the *in vivo* dissection of differentiation landscapes as a complementary approach to lineage tracing and *in vitro* differentiation assays.

4.4.1 Methods

Cell Purification and Sequencing Library Preparation

Suitable sorting schemes were devised for different hematopoietic stem and progenitor cell types (Table 4.2). HSCs and MPPs were sorted from peripheral blood, fetal liver, cord blood and bone marrow. Eight additional progenitor cell types were sorted from peripheral blood. For each of these cell types low-input bisulfite libraries were prepared using the µWGBS protocol [Farlik *et al.* 2015], which were sequenced by the Biomedical

Table 4.2: Surface markers used for sorting progenitor cell types by FACS

	HSC	MPP	CMP	MEP	GMP	CLP	MLP0	MLP1	MLP2	MLP3
Lin	-	-	-	-	-	-	-	-	-	-
CD34	+	+	+	+	+	+	+	+	+	+
CD38	-	-	+	+	+	+	-	-	-	-
CD90	+	-					-	-	-	-
CD45RA	-	-	-	-	+	+	+	+	+	+
CD49f	+	-								
CD123			dim	-	+					
FLT3				-						
CD36				-						
CD110				+						
CD41				-						
CD7						-	-	-	+	+
CD10					-	+	-	+	-	+

Cell type colors were selected such that early progenitors are represented by purple colors, lymphoid progenitors are depicted in orange, red or yellow and lymphoid cell types are assigned shades of blue and green. By courtesy of Matthias Farlik.

Sequencing Facility at the CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences (Vienna, Austria) using a 2×75 bp paired-end setup on the Illumina HiSeq 3000/4000 platform. In order to avoid high PCR duplication rates, libraries were sequenced at a relatively low coverage. In total, 639 bisulfite sequencing libraries passed quality control, and 3.1 terabases of sequencing data were produced.

Among the profiled cell types were four populations of MLPs which exhibited distinct differentiation capabilities and DNA methylation patterns (MLP0 to MLP3). Additionally, seven mature blood cell types were profiled (Figure 4.1). Each cell type was obtained from three healthy donors to account for inter-individual heterogeneity. Progenitor cells were profiled using a pooling strategy comprising different nominal cell numbers. Specifically, for each cell type and donor, eight pools of ten cells, two pools of 50 cells and one pool of 1,000 cells were processed individually. Detailed methods for sample collection, cell isolation, cell culture and sequencing library preparation can be found in [Farlik *et al.* 2016].

DNA Methylation Sequencing Data Processing

Adapter sequences were trimmed using Trimmomatic version 0.32 [Bolger *et al.* 2014], and the trimmed reads were aligned using Bismark version 0.12.2 employing Bowtie2 version 2.2.4 [Krueger and Andrews 2011; Langmead and Salzberg 2012] with parameters `--minins 0 --maxins 6000 --bowtie2`. The GRCh38 assembly of the human reference genome was used throughout the study. Duplicate reads mapping to identical genomic coordinates were removed as potential PCR artifacts. Reads with a bisulfite conversion rate below 90 % or with fewer than three cytosines outside a CpG context were discarded as potential post-bisulfite contamination. The Bismark methylation extractor was used to estimate CpG methylation levels. Replicates of different nominal cell numbers which belonged to the same donor and cell type were merged by summing up the total number of methylated and unmethylated reads per CpG across all

replicates. Merged and unmerged datasets were further processed using RnBeads version 1.5 [Assenov *et al.* 2014]. This generated standard reports for data exploration and quality control. DNA methylation values of individual CpGs were aggregated based on predefined genomic regions such as genomic tiling regions (5 kb) and regulatory regions annotated by the BLUEPRINT edition of the Ensembl regulatory build [Zerbino *et al.* 2015] (August 2015 data release). The aggregate values produced by RnBeads were used for further data analysis using custom R scripts.

Integration of DNA Methylation and Chromatin Accessibility Data

We downloaded publicly available peak regions and fragment count data from ATAC-seq experiments [M. R. Corces *et al.* 2016] (GEO accession `GSE74912`) and transformed the peak coordinates to genome assembly GRCh38 using the UCSC liftOver tool[11]. Mean DNA methylation levels for all merged samples were computed in all ATAC-seq peaks. We used a one-sided Wilcoxon test to identify cell-type-specific regions of open chromatin. Specifically, for each cell type in the ATAC-seq dataset we selected those peak regions in which samples of that cell type exhibited a significantly higher ATAC-seq fragment count than samples not belonging to that cell types (FDR-adjusted p-value less than 0.05).

DNA-Methylation-Based Prediction of Cell Types

Regularized general linear models, as implemented in the glmnet R package [Friedman *et al.* 2010; Krishnapuram *et al.* 2005], were used to classify samples into ten progenitor cell types based on their DNA methylation profiles across regulatory regions. The annotation of regulatory regions was obtained from the BLUEPRINT edition of the Ensembl regulatory build [Zerbino *et al.* 2015]. The classifiers were trained on the DNA methylation levels in a total of 319 10-cell, 50-cell, and 1,000-cell replicates. In order to avoid biases incurred by the different tissue origins of samples, models were trained exclusively on DNA methylation data from peripheral blood samples. Due to the relatively low sequencing coverage the replicates contained an average of 56 % missing values of all covered CpGs. These were imputed with the impute R package[12] using 5-nearest neighbor averaging on the entire dataset of replicates (including non-peripheral-blood samples and differentiated cell types). Elastic net regularization was applied to multinomial logistic regression classifiers. Concretely, these models infer model parameters by employing methods for coordinate descent in order to maximize the penalized log-likelihood:

$$\max_{\beta_{0l},\beta_l} \Bigg[\underbrace{\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{l=1}^{K} y_{il}\left(\beta_{0l} + x_i^T\beta_l\right) - \log\left(\sum_{l=1}^{K} e^{\beta_{0l}+x_i^T\beta_l}\right)\right)}_{\text{multinomial logistic regression log-likelihood}} - \underbrace{\lambda\left((1-\alpha)||\beta||_F^2/2 + \alpha\sum_{j=1}^{p}||\beta_{j\cdot}||_2\right)}_{\text{elastic-net penalty}} \Bigg]$$

with N observations (here: samples) in the training set and measurements (methylation levels) $x_i \in \mathbb{R}^p$ for p features (regions) that are classified to one of K classes (cell

[11] https://genome.ucsc.edu/cgi-bin/hgLiftOver
[12] https://bioconductor.org/packages/release/bioc/html/impute.html

types). $y_{il} \in \{0,1\}$ indicates whether observation i of the training set belongs to class l. The index 0 denotes parameters for the intercept. The result of the coordinate descent algorithm is a $p \times K$ parameter matrix β with column vectors $\beta_l \in \mathbb{R}^p$ that contain parameter values for each class (cell type) and row vectors $\beta_{j\cdot}$. The penalty term ensures that model parameters are regularized consistently across classes [Friedman *et al.* 2010]. It is described by a linear combination of the Frobenius matrix norm for regularizing all coefficients across all features and classes (penalty as in ridge regression: $||\beta||_F^2 = \sum_{l=1}^{K}\sum_{j=1}^{p}|\beta_{lj}|^2$) and the Euclidean norm for each individual feature across all classes (penalty as in grouped-lasso regression). In this study, the regularization parameter λ was obtained by nested 10-fold cross-validation as the value resulting in the most regularized model whose error is still within one standard error of the minimum. α was set to 0.5 to reflect equal mixing of the ridge and lasso penalty terms.

Class importance was defined in terms of model parameters as the Euclidean norm aggregate of per-class coefficients in the model ($\overline{\beta}_{j\cdot} = \sqrt{\sum_{l=1}^{K}\beta_{lj}^2}$). We defined signature regions which were able to discriminate between cell types as those regions whose class importance value was non-zero ($\overline{\beta}_{j\cdot} \neq 0$) in a model trained on the entire dataset of 319 replicates.

Class probabilities were defined as fitted probabilities from the logistic regression model: given a feature vector x, the probability of class l is defined by

$$\Pr(G = l|x) = \frac{e^{\beta_{0l}+x^T\beta_l}}{\sum_{k=1}^{K} e^{\beta_{0k}+x^T\beta_k}}$$

For assessing model quality, 10-fold cross validation was performed and misclassification rates (per class and overall) were averaged in the cross-validation test sets. Receiver Operating Characteristic curves (ROC curves) and Area Under the Curve (AUC) values were obtained by evaluating the class probabilities in the one-vs-all setting for each class. In brief, the score resulting from the difference of the class probability of the assigned class and the largest class probability excluding that class was computed for each sample. ROC curves are then defined by applying all possible thresholds on this score by computing false positive and true positive rates for every threshold value and AUC values were derived.

For assigning cross-class probabilities of individual progenitor cell types, the samples of one class were excluded from the model, the model was trained based on the data for all other classes and applied to cross-predict the data points of the class that was excluded from training (leave-one-class-out classifiers).

Inference of Cell-Type-Similarity Graph

In the cell-type-similarity graph (Figure 4.23), nodes represent cell types and directed edges represent the probabilities of predicting one cell type to another according to the corresponding leave-one-class-out classifier. Specifically, for each pair of source and target cell type, the edge weight corresponds to the average class probability assigned by the leave-one-class-out classifier for the target cell type to all peripheral blood samples of the source cell type. The graph in Figure 4.23 shows the directed edge pairs for each pair of nodes as trapezoids in which the widths at the target and source node correspond to weights of the directed edges (e.g. the predictor which did not include HSC samples assigned a higher probability to classify HSC samples as MPP than the probabilities the predictor which did not include MPP samples assigned to predicting MPP samples as

HSC). Differentiated cell types (circles) were predicted based on the classifier trained on all 319 stem/progenitor samples from peripheral blood. Only edges which correspond to an average prediction probability exceeding 0.1 are shown.

Genomic Region Enrichment Analysis

We used LOLA [Sheffield and Bock 2016] to identify significant overlaps (adjusted p-value < 0.05) with genomic region sets associated with transcription factor binding. These regions were previously defined from ChIP-seq experiments by ENCODE [Harrow *et al.* 2012] and from the data contained in the CODEX database [Sánchez-Castillo *et al.* 2015]. In brief, provided with genomic regions of interest and a background set of genomic regions (universe), LOLA quantifies the overlap of the regions of interest with predefined sets of genomic regions. For each of these sets, a p-value is calculated using Fisher's Exact Test, which is adjusted for multiple testing by controlling of the FDR [Benjamini and Yekutieli 2001]. Additionally, log-odds scores and the number of overlapping regions are computed. To facilitate visualization and interpretation, we manually curated the LOLA database annotations to group the cell types on which the region sets in the database are based into broader categories. Plots showing the enrichment of TFBSs that were significant in at least one comparison (Figure 4.20b) were generated using custom R scripts provided by Johanna Klughammer.

Data Availability and Supplementary Website

A supplemental website with genome browser tracks, diagrams and tables as well as direct links to the data sources is available[13]. The website also contains the code for the DNA-methylation-based prediction of cell types. Processed methylation calls can be obtained from GEO under accession number `GSE87197` and raw sequencing data are deposited in EGA under controlled access (study accession number `EGAS00001002070`). The dataset is included in the epigenome registry[14] of IHEC (accession numbers `IHECRE00002734` to `IHECRE00002810`) and track data is also accessible through the IHEC data portal[15] and the DeepBlue Epigenomic Data Server [Albrecht *et al.* 2016].

4.4.2 Results

DNA Methylation Maps of Human Hematopoietic Stem and Progenitor Cells

We generated reference methylome maps for ten different hematopoietic progenitor cell types and seven differentiated cell types obtained from peripheral blood of healthy donors (Figure 4.1; Table 4.2). Additionally, HSCs and MPPs sorted from fetal liver, cord blood and bone marrow were profiled. However, in order to not be confounded by the tissue of origin, we focus on analysis of DNA methylation in peripheral blood cells in this section. Progenitor cells were profiled in replicate pools of different nominal cell numbers (cf. Section 4.4.1). For most analysis steps, we pursued a composite approach to methylation mapping by combining the bisulfite sequencing reads from these replicates for the same individual and cell type. Combining these profiles results in high-resolution methylome maps that intrinsically account for epigenetic variability within and between cell types.

[13] `http://blueprint-methylomes.computational-epigenetics.org`
[14] `http://www.ebi.ac.uk/vg/epirr`
[15] `http://epigenomesportal.ca/ihec/`

Globally, DNA methylation levels were similarly high across progenitor cell types (Figure 4.12a). We observed slightly decreased genome-wide methylation levels in differentiated cells of the myeloid lineage. The distribution of DNA methylation across the genome followed the expected patterns: DNA methylation in CpG islands and promoter regions was reduced and exhibited a bimodal distribution and the vast majority of 5-kb tiling regions were highly methylated (Figure 4.12b). Putative enhancer elements exhibited more variable, intermediate to high levels of DNA methylation. To provide a robust and biologically meaningful basis for analyzing DNA methylation differences between cell types, we aggregated DNA methylation levels across genomic regions defined by the BLUEPRINT version of the Ensembl regulatory build [Zerbino *et al.* 2015] which integrates epigenome data across many cell types into a catalog of six types of putative regulatory regions. In our dataset, these regions exhibited broadly varying DNA methylation levels (Figure 4.12c).

Figure 4.13 shows selected examples of DNA methylation variation in regulatory regions. At the *KCNH2* gene locus (a key factor for erythroid development), two CTCF sites and a distal element showed decreased DNA methylation in the myeloid lineage and also increased expression in CMP and GMP cells (data not shown). A putative enhancer of the myeloid-linked *TREML1* gene displays decreased DNA methylation in HSCs, MPPs and myeloid progenitors. In contrast, CTCF sites in the lymphoid-linked *SUSD3* gene exhibit lower DNA methylation in lymphoid progenitors. Furthermore, promoter-associated regulatory regions in the *EXOC6* locus displayed lower methylation levels in HSCs and MPPs that were not linked to changes in gene expression (data not shown).

DNA methylation levels in regulatory elements could also be used to discriminate between lineage in unsupervised analyses based on individual replicates. Dimension reduction using MDS revealed two compact clusters comprising lymphoid and myeloid cells, while HSC and MPP profiles were more dispersed (Figure 4.14). This lineage-specific clustering indicates that neither technical biases arising from the different cell numbers in the replicates nor inter-individual variation between donors had a strong influence on our investigation of cell-type-specific DNA methylation patterns.

Robustness Analysis

We obtained DNA methylation profiles from replicates with relatively low cell numbers and coverage. In order to confirm that our observations are not merely the result of technical variation due to differences in nominal cell numbers or sequencing depth, we quantified the similarities between replicates of different pool sizes in putative regulatory regions (Figure 4.15). We observed moderate correlation coefficients between individual 10-cell and 50-cell replicates. The agreement was higher when 50-cell replicates were compared to 1,000-cell replicates. We also aggregated the methylation calls of five 10-cell replicates into virtual 50-cell replicates *in silico*. The correlation coefficients obtained for comparisons between physical 50-cell replicates and these virtual 50-cell replicates closely reflected those of physical 50-cell replicates among each other. Furthermore, lineage-specific differences in DNA methylation patterns outweigh variation due to pool sizes in unsupervised analyses (Figure 4.14). Together, these results argue against the presence of strong biases induced by technical differences between replicates of different cell numbers and therefore confirm the evaluation of the µWGBS protocol conducted by Farlik *et al.* [2015]. However, we also observed only moderate

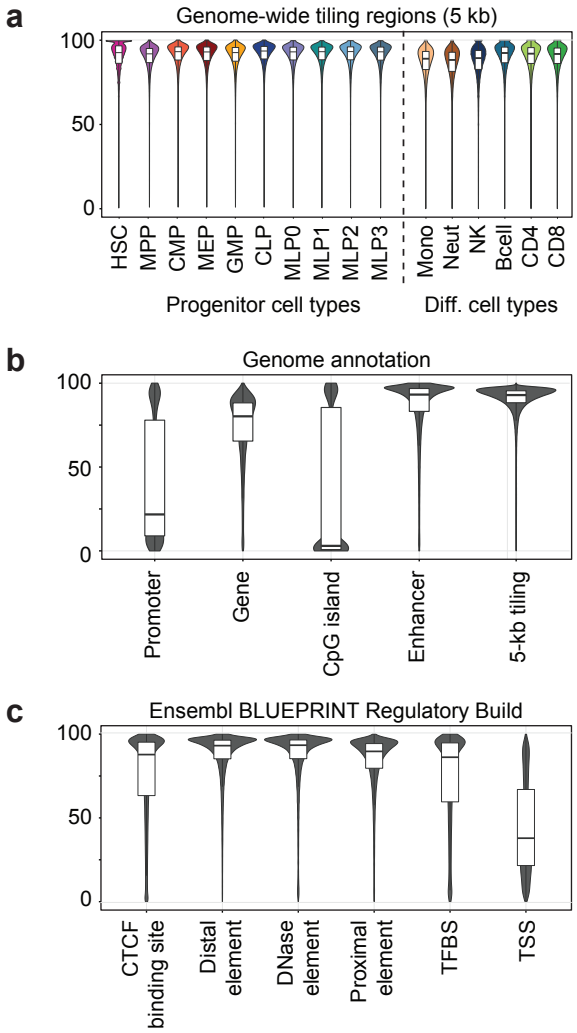


Figure 4.12: Distribution of methylation levels in hematopoietic cells. Violin and box plots show mean DNA methylation levels in **(a)** genome-wide (5-kb) tiling regions for each cell type, **(b)** summarized across cell types for various types of annotated genomic regions (Ensembl genes and promoters, CpG islands, putative enhancers and 5-kb tiling regions) and **(c)** for elements contained in the BLUEPRINT edition of the Ensembl regulatory build. Distributions in **(b)** and **(c)** are based on merged progenitor cell samples obtained from peripheral blood.

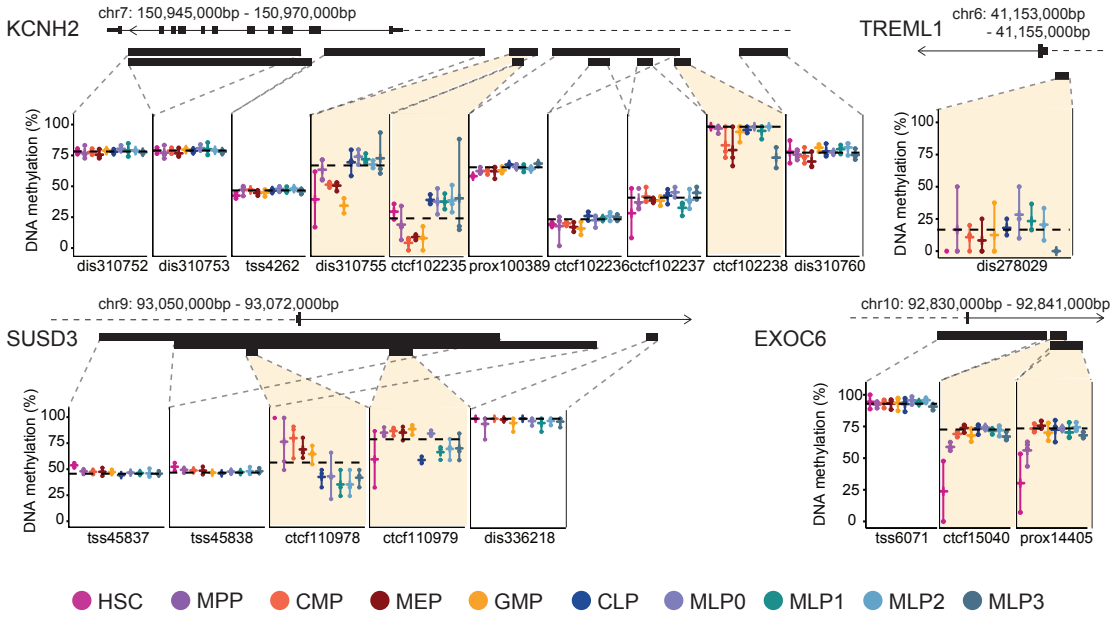


Figure 4.13: DNA methylation levels in regulatory regions in four gene loci. Black bars denote the positions of regulatory regions according to the BLUEPRINT Ensembl Regulatory Build. Horizontal, dashed, black lines indicate the sample medians for the respective regions. Colored vertical bars connect the highest and lowest observed merged methylation values in samples obtained from peripheral blood. By courtesy of Matthias Farlik.

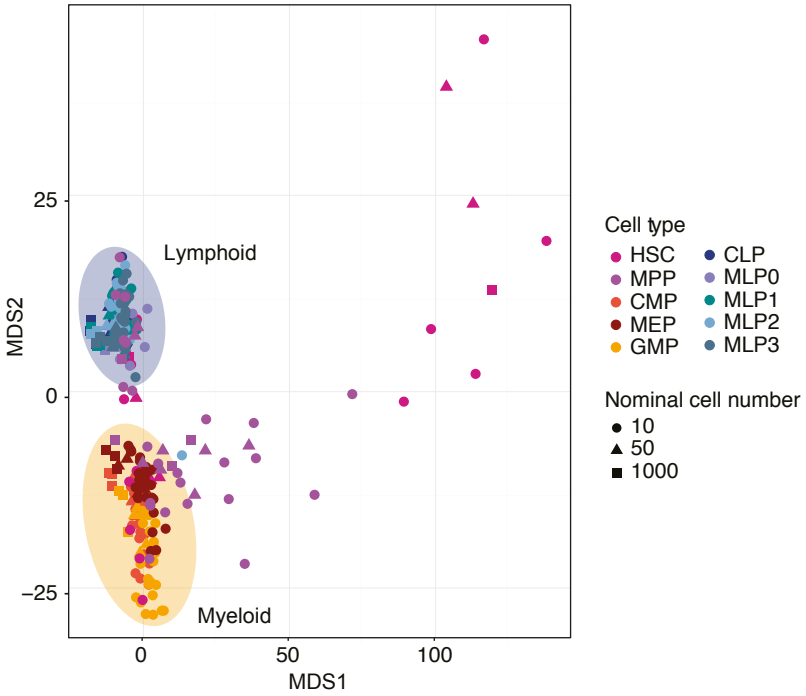


Figure 4.14: Unsupervised analysis of individual replicates of hematopoietic samples based on DNA methylation in putative regulatory regions. MDS coordinates of DNA methylation profiles for all 10-cell, 50-cell, and 1,000-cell replicates of hematopoietic stem/progenitor cell types are shown.

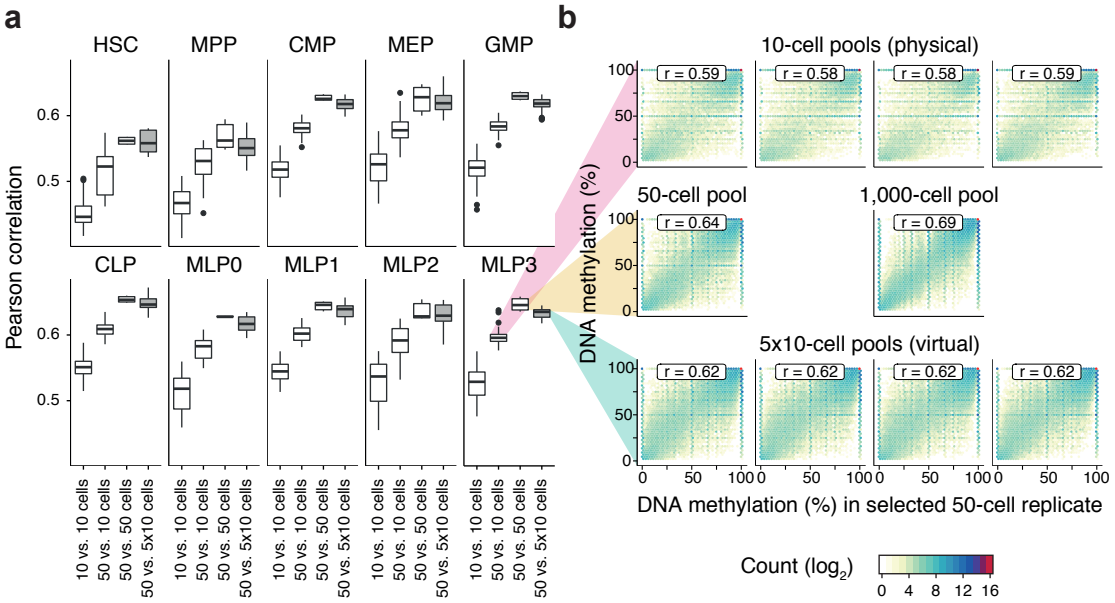


Figure 4.15: Correlation in DNA methylation profiles in replicates of different pool sizes. **(a)** Box plots summarizing the Pearson correlation coefficients of DNA methylation levels in putative regulatory regions between pairs of 10-cell replicates, mixed pairs of one 10-cell and one 50-cell replicates, pairs of 50-cell samples and mixed pairs of 50-cell samples with virtual 50-cell samples, which were derived by merging five 10-cell samples. The analysis was performed for each cell type separately and only datasets from the same donor were compared to exclude effects of variability between donors. **(b)** Density scatterplots of average DNA methylation levels showing the agreement of one 50-cell replicate (MLP3_50_D1_2) with selected physical 10-cell, 50-cell and 1000-cell samples of the same cell type and donor (top two rows). The bottom row shows the agreement of the 50-cell replicate with four selected virtual combinations of five 10-cell samples. r denotes Pearson correlation coefficients for each comparison. By courtesy of Florian Halbritter.

overall correlation between replicates of low cell numbers, which is likely due to the low sequencing coverage of individual replicates.

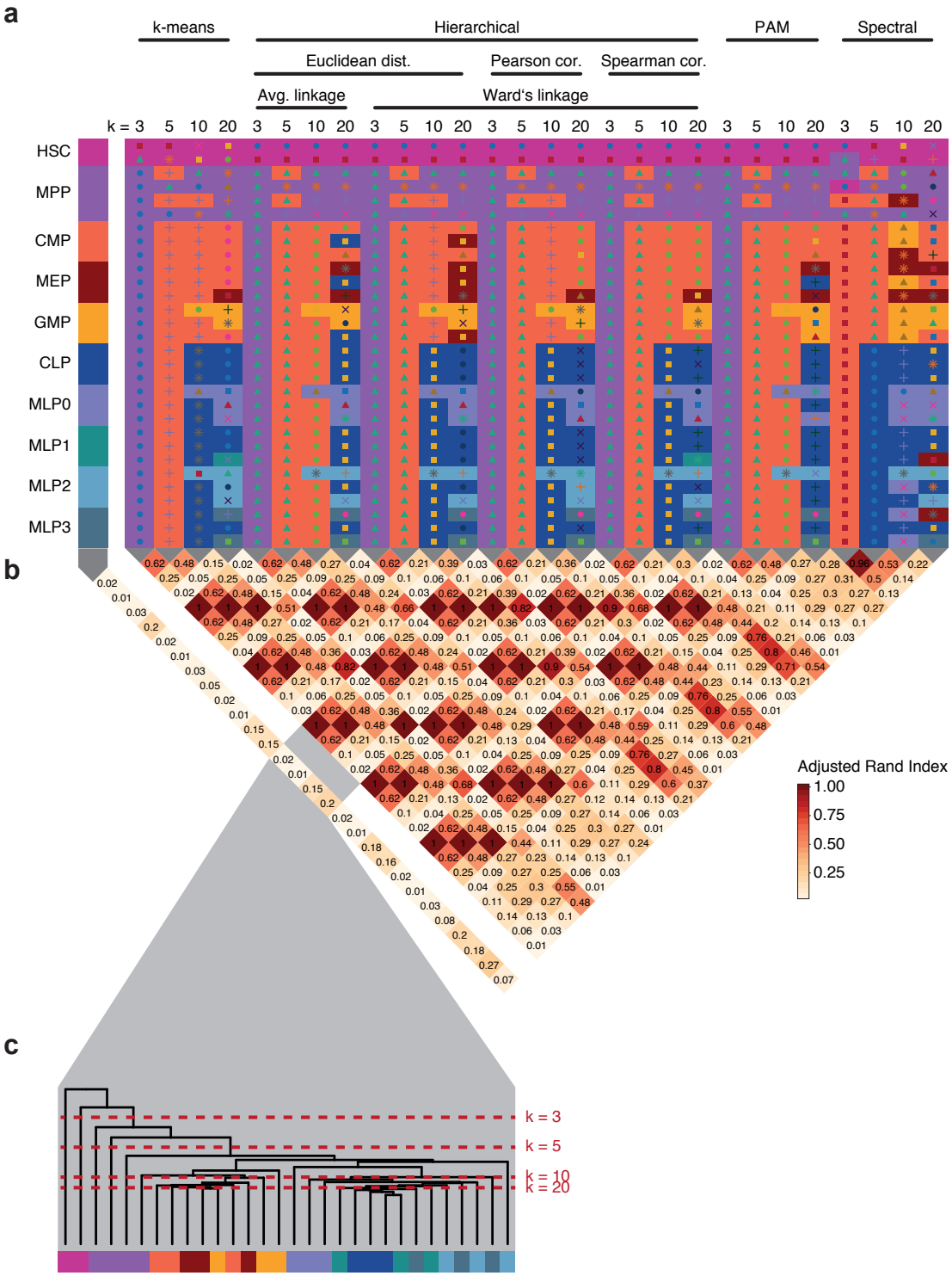
In order to overcome these limitations of relatively low sequencing coverage we employed a composite approach in which we combined methylation levels of all 10-cell, 50-cell, and 1,000-cell replicates for each cell type, tissue source and donor. We verified that these composite methylomes robustly captured biological variation. To this end, we employed different clustering methods and parameter settings to DNA methylation data in putative regulatory regions and quantified the agreement of cluster assignments (Figure 4.16). We chose different numbers of clusters k in this evaluation. For each cluster in each method and parameter setting, we determined the cell type associated with that cluster (Figure 4.16a). Across several parameter settings, the HSC and MPP samples were assigned separately to different clusters of size 1 and thus dominate the clustering results for small values of k. This observation is consistent with the previously observed heterogeneity in early progenitors. As we increased k, the clustering and associated cell types indicated differences between myeloid and lymphoid samples and heterogeneous

HSC/MPP samples (Figures 4.16a, 4.16c). This observation was stable across clustering methods and parameter settings which were in agreement when similar values for k were employed (Figure 4.16b). We therefore conclude that the biological variation in DNA methylation levels is robustly captured across all applied clustering methods and parameter settings.

DNA Methylation in Cell-Type-Specific Regions of Open Chromatin

We characterized DNA methylation in regions of accessible chromatin (Figure 4.17). To that end, we took advantage of publicly available ATAC-seq data [M. R. Corces *et al.* 2016] and identified regions that exhibited open chromatin specifically in different hematopoietic cell types (cf. Section 4.4.1). We observed variable DNA methylation patterns across regions and cell types. Regions specifically accessible in HSCs generally exhibited low overall methylation levels across all cell types. Regions with open chromatin in differentiated blood cells were highly methylated in most progenitors and lost methylation only in the respective differentiated cell types. Strikingly, regions with accessible chromatin in myeloid and lymphoid progenitor cell types were markedly hypomethylated in differentiated cells of the respective lineage. In contrast to regions which are accessible in lymphoid progenitors, peaks specifically open in CMPs already exhibit reduced methylation at the level of myeloid progenitor cells. These results could indicate that lineage-specific opening of chromatin precedes demethylation in hematopoiesis.

Figure 4.16 : *(On the next page)* Agreement of cluster assignments by selected clustering methods. **(a)** Cluster and cell type assignments by selected clustering methods and parameter settings. Rows in the heatmap correspond to methylation profiles in putative regulatory regions in progenitor samples from peripheral blood and columns correspond to different clustering methods and parameter settings. Points of different colors and shapes denote assignments to different clusters for each method. Colored boxes indicate the cell type most associated with each cluster, i.e. the cell type that had the maximum Jaccard Index with the respective cluster. The annotated cell types of all samples are shown on the left. The applied clustering methods include k-means clustering, hierarchical clustering with different parameter settings for the employed distance method (Euclidean distance and 1-correlation) and linkage method (average linkage and Ward's linkage), Partitioning Around Medoids (PAM) and spectral clustering. For each method the number of clusters k was varied ($k \in \{3, 5, 10, 20\}$). **(b)** Heatmap of pairwise agreement between clustering methods. The adjusted Rand index [Hubert and Arabie 1985] was used to assess the agreement of clusterings among each other and with the annotated cell type. This measure quantifies the number of agreements in relation to the number of disagreements between clusters and additionally accounts for agreements due to chance. **(c)** Dendrogram for hierarchical clustering using 1-Pearson correlation as distance metric and Ward's linkage method (as implemented by the method parameter `ward.D` in the `hclust` function of the STATS R package). Colors denote annotated cell types.


a
k-means
Hierarchical
PAM
Spectral
Euclidean dist.
Pearson cor.
Spearman cor.
Avg. linkage
Ward's linkage
k = 3 5 10 20 3 5 10 20 3 5 10 20 3 5 10 20 3 5 10 20 3 5 10 20 3 5 10 20
HSC
MPP
CMP
MEP
GMP
CLP
MLP0
MLP1
MLP2
MLP3
b
Adjusted Rand Index
1.00
0.75
0.50
0.25
c
k = 3
k = 5
k = 10
k = 20

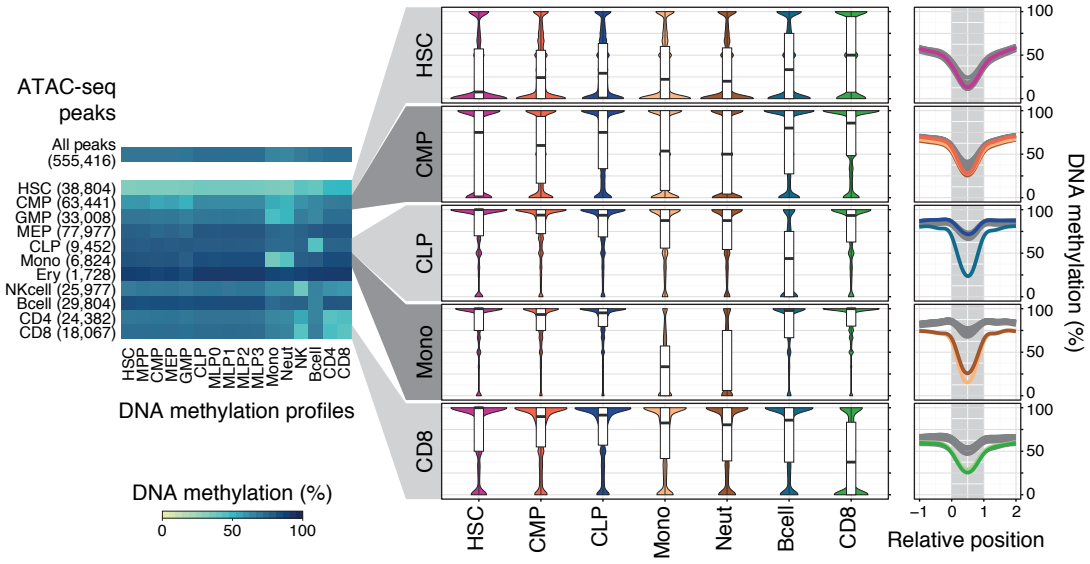


Figure 4.17: DNA methylation in cell-type-specific regions of open chromatin. The heatmap on the left shows average DNA methylation levels in cell-type-specific regions of open chromatin (rows) across hematopoietic progenitors and differentiated blood cell types (columns). Numbers in parentheses denote the number of accessible regions specific for a cell type. The label "All peaks" refers to all regions of open chromatin in the dataset. The violin and box plots in the middle show the distributions of DNA methylation levels for cell type-specific open-chromatin regions in selected hematopoietic cell types. The composite plots on the right show DNA methylation levels locally averaged across regions. In these plots, CpGs in the neighborhood of the ATAC-seq peak regions were annotated with coordinates relative to the start and end of the regions (x-axis). CpGs with a relative coordinate of 0 and 1 are located at the start and end of a peak and coordinates -1 and 2 correspond to one peak length upstream and downstream of the peak. The curves represent cubic spline smoothers of DNA methylation levels for each cell type.

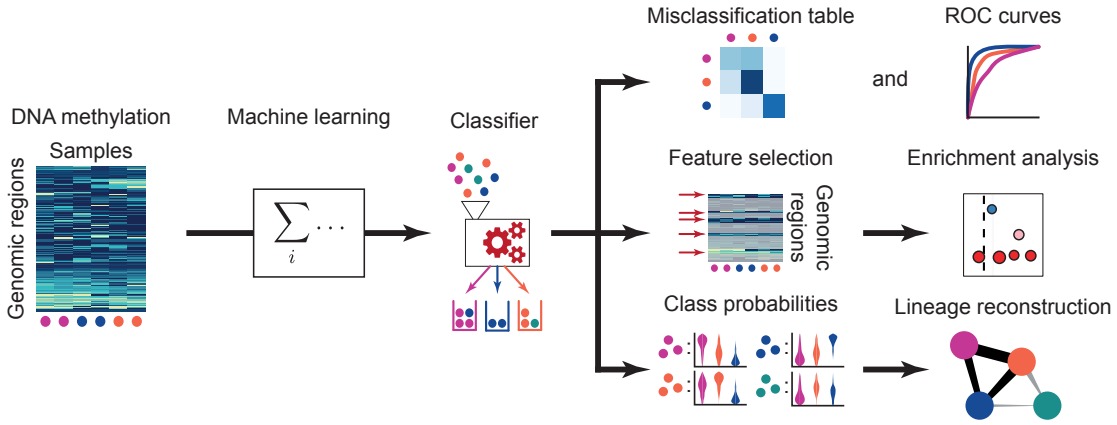


Figure 4.18: Overview of the statistical learning approach for cell type prediction. After statistical validation, the described models were used to infer the cell type based on DNA methylation levels in regulatory regions, to identify and characterize signature regions and to infer similarities between cell types based on prediction probabilities.

Statistical Modeling Identifies Epigenetic Signatures Distinctive of Cell Lineage

Given that DNA methylation patterns could discriminate between myeloid, lymphoid, and multipotent cells in unsupervised analyses, we investigated whether DNA methylation maps could also identify individual stem/progenitor cell types in a supervised learning setting. For this purpose, we inferred classifiers that were able to accurately predict cell types based on DNA methylation signatures of regulatory regions of the genome (Figure 4.18).

Specifically, we trained and evaluated elastic net-regularized general linear models [Friedman *et al.* 2010; Krishnapuram *et al.* 2005] for predicting the cell type of each sample. These classifiers were trained on the DNA methylation levels of all BLUEPRINT Regulatory Build regions in each of the 319 10-cell, 50-cell, and 1,000-cell replicates from peripheral blood. Model performance was assessed based on test set results in a 10-fold cross-validation setting (Table 4.3). Misclassifications occurred most frequently between cells of the same lineage in the hematopoietic hierarchy (myeloid, lymphoid and HSC/MPP) indicating high similarity in the DNA methylation profiles. Overall, we observed high prediction accuracy for all cell types with mean area under the ROC curve (AUC) values between 0.85 and 1.0 (Figure 4.19). The highest accuracy was observed for myeloid progenitor cell types (GMP, CMP, MEP) and for the MLP0 population. Consistent with high similarity in methylation profiles, lymphoid progenitors (CLP, MLP1, MLP2, MLP3) were generally more difficult to distinguish, but still received high AUC values. Lower AUC values indicative of higher variation in DNA methylation patterns were observed for the HSC and MLP2 cell populations, which were frequently confused with MPPs and CLPs, respectively (Table 4.3). In order to assess the robustness of the statistical learning method approach, we also employed alternative models using Support Vector Machines (SVMs) with a linear kernel [Cortes and Vapnik 1995] and random forests [Breiman 2001]. These models achieved comparable predictive performances (data not shown).

Additionally, we employed our classifier to single-cell methylome data of six cell types (HSC, MPP, CMP, GMP, CLP and MLP0) obtained using the same µWGBS protocol.

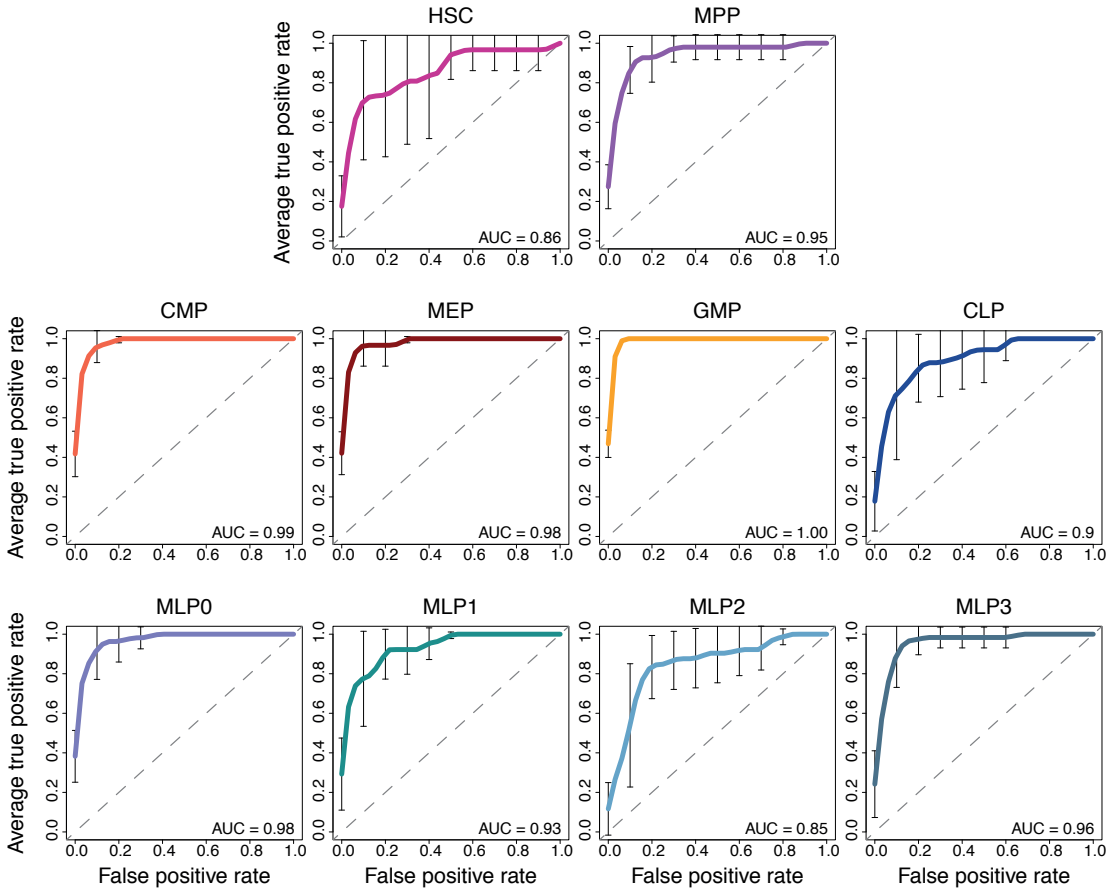
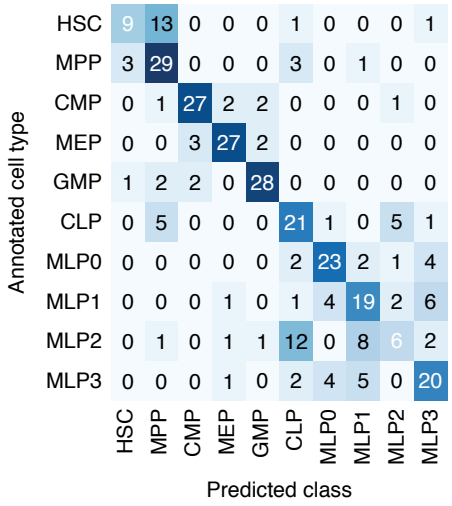


Figure 4.19: Performance of methylation-based classifiers for cell type prediction. ROC curves and mean AUC values corresponding to 10-fold cross-validation are shown separately for each cell type. Prediction was based on DNA methylation levels at regulatory regions. The ROC curves show the prediction performance in a one-vs-all setting for each class, i.e. by sliding a threshold along a value calculated as the difference of the class probability and the largest class probability excluding that class. Vertical bars denote standard deviations across 10-fold cross-validation folds. Diagonal dashed lines correspond to the expected performance of random guessing (AUC = 0.5).

Table 4.3: Confusion matrix for progenitor cell classification

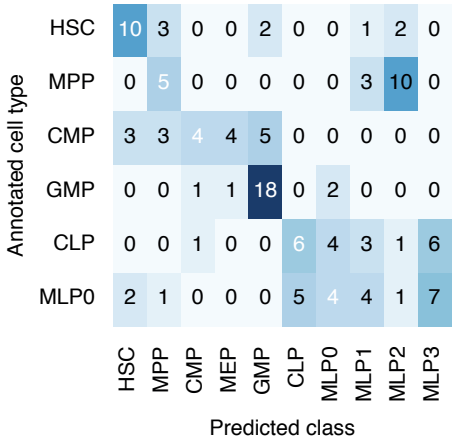


The table is based on 10-fold cross-validation of the cell type classifiers trained and evaluated on 10-cell, 50-cell, and 1,000-cell replicates sorted from peripheral blood.

Due to the low amounts of input material and sequencing coverage, a large portion of regions contained missing values that were computationally imputed before prediction and the resulting predictive performance was low: all replicates were classified as HSC, which corresponds to the cell type that had overall the lowest sequencing coverage and thus also contained the largest fraction of imputed values in our training dataset. To mitigate these coverage-related artifacts, we trained an additional classifier on an alternative feature set of aggregate methylation values for sets of genomic regions. These values were obtained by averaging the region methylation values for different categories defined in the LOLA core database [Sheffield and Bock 2016]. This aggregation method has previously been shown to result in robust results in the evaluation of the µWGBS assay [Farlik *et al.* 2015]. Overall, the resulting classifier assigned classes corresponding to the lineage of the annotated myeloid and lymphoid cell type (Table 4.4). A number of cells labeled as CMP and MLP0 were classified as HSC or MPP, potentially indicating epigenetic memory of differentiation in these cells. In contrast, a considerable fraction HSCs and MPPs were assigned classes corresponding to myeloid or lymphoid progenitors. In particular, MPPs exhibited similarity to MLP2s. These results could be reflective of lineage-primed cell states in early progenitors.

Having established, that the classifiers were able to accurately infer cell types, we took advantage of the regularization employed by our elastic net models. Using the built-in feature selection, we identified 1,234 regulatory regions whose DNA methylation levels collectively distinguished hematopoietic cell types with high accuracy and robustness (Figure 4.20a). Although we found select regions that were hypomethylated in specific cell types of the myeloid or lymphoid lineage, cell-type-specific methylation was not directly apparent for the majority of signature regions. These results suggest that the inferred classifier captures more complex signatures of methylation variability which are indicative of cell type. LOLA enrichment analysis on the 1,234 signature regions identified transcription factors involved in myeloid and lymphoid differentiation, such as *GATA1*, *TAL1* and *MYB* (Figure 4.20b). Furthermore, the signature regions were able to discriminate cell lineages in unsupervised analyses (Figure 4.21). Myeloid and

Table 4.4: Confusion matrix for single-cell progenitor cell classification



The table is based on predicting the cell type of single cells using a classifier trained on 10-cell, 50-cell, and 1,000-cell replicates sorted from peripheral blood, using aggregate methylation values across region sets contained in the LOLA core database [Sheffield and Bock 2016].

lymphoid progenitor samples separated in the first principal component but no clear clustering within each group was apparent. Differentiated cell types of the myeloid and lymphoid lineage formed separate clusters in the vicinity of their corresponding progenitors. HSCs and MPPs were more dispersed due to their heterogeneity in DNA methylation patterns.

Finally, we inferred a data-driven model of human hematopoiesis from our classifiers. To that end, we derived separate classifiers that were trained on all peripheral blood progenitor replicates, excluding one cell type (leave-one-class-out models). These models were then used to classify replicates of the excluded cell type into the remaining classes. We used the class probabilities assigned by the respective classifiers to derive a network reflecting similarities between the cell types. Here, the predicted class probabilities (Figure 4.22) served as a measure of cell type similarity. The resulting network (Figure 4.23) showed high intra-lineage similarity between cell types and captured cell type relationships that were in accordance with the current models of hematopoietic differentiation. For instance, HSCs were predicted as MPP with high probability and vice versa. MPPs possessed moderate probabilities to be classified as cells from the myeloid and lymphoid lineage, possibly indicating primed samples in our dataset. Cells within the myeloid and lymphoid lineages exhibited high similarity and cross-prediction patterns appear nearly symmetric. Notably, MEPs exhibited a limited tendency to be classified as MLP0, which is in line with a possible myeloid potential of MLPs. Mature cell types were frequently classified within their respective lineage. Interestingly, T cells and NK cells also exhibited limited probabilities to be placed into myeloid classes.

4.4.3 Discussion

The epigenetic basis of regulatory mechanisms that direct hematopoietic differentiation are just beginning to be understood. While many epigenetic signatures have been charted genome-wide across the hematopoietic hierarchy in mice [Paul *et al.* 2015; Cabezas-Wallscheid *et al.* 2014; Lara-Astiaso *et al.* 2014; Bock *et al.* 2012] comprehensive profiling of epigenomic marks in human blood progenitor cells has been limited

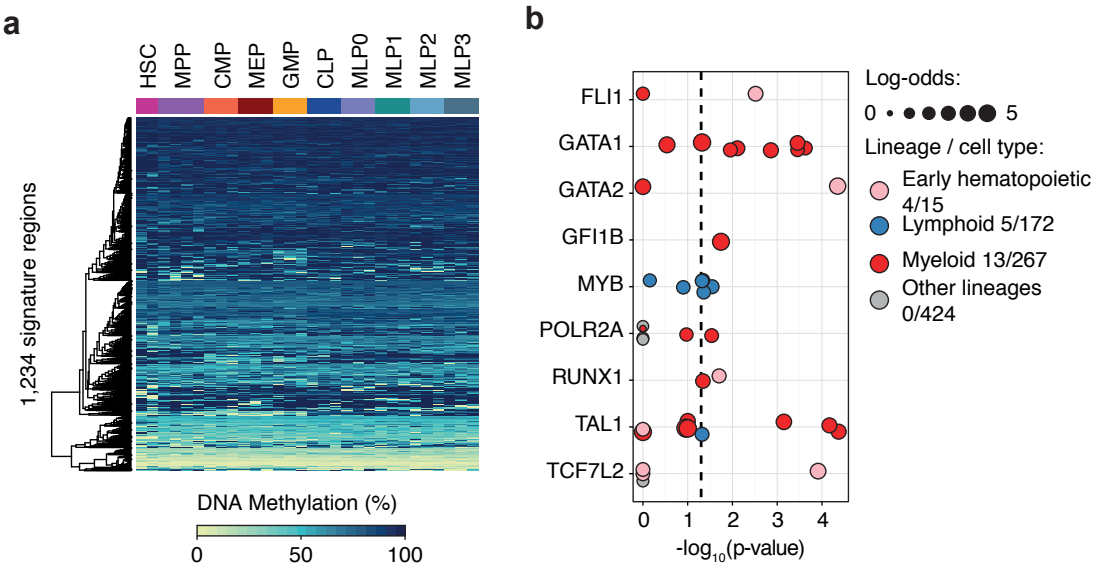


Figure 4.20: Characterization of progenitor cell type signature regions. **(a)** Heatmap showing average DNA methylation levels of merged replicates (one column for each cell type and donor) for the 1,234 signature regions extracted from the trained classifier. Regions (rows) were arranged using hierarchical clustering with Euclidean distance and complete linkage. **(b)** LOLA enrichment analysis [Sheffield and Bock 2016] of selected TFBS in regulatory regions associated with the signature regions. Colored dots represent ChIP-seq experiments in the indicated cell type or lineage. The size of the dots represents log-odds ratios computed using Fisher's exact test. Panel **(b)** by courtesy of Johanna Klughammer.

to RNA expression [Novershtern *et al.* 2011; L. Chen *et al.* 2014] and maps of chromatin accessibility [M. R. Corces *et al.* 2016]. Here, we complement this view by establishing genome-wide profiles of DNA methylation dynamics in human hematopoietic differentiation. The generated methylome maps provide a comprehensive resource for studying epigenetic regulation of cell differentiation and blood-related diseases.

A key result of our study is the identification of DNA methylation signatures that were not only definitive of cell lineage, but that could also discriminate between individual cell types. In our approach, we specifically selected elastic net-regularized general linear models due to their high predictive performance and interpretability through build-in feature weighting and regularization. After ascertaining that predictive performance was high in our models, we exploited the interpretability of derived model parameters. We used feature selection by regularization to identify epigenetic signatures that were characteristic of cell type and could be partially explained by lineage-specific transcription factor binding. The majority of individual signature regions exhibited only small DNA methylation differences between cell types, but collectively these regions supported highly accurate cell type prediction. We showed that prediction based on DNA methylation in regulatory regions was able to place individual cell populations into their developmental context. Using epigenome-based cell type prediction as a quantitative measure of relatedness between cell types, we inferred a data-driven model of

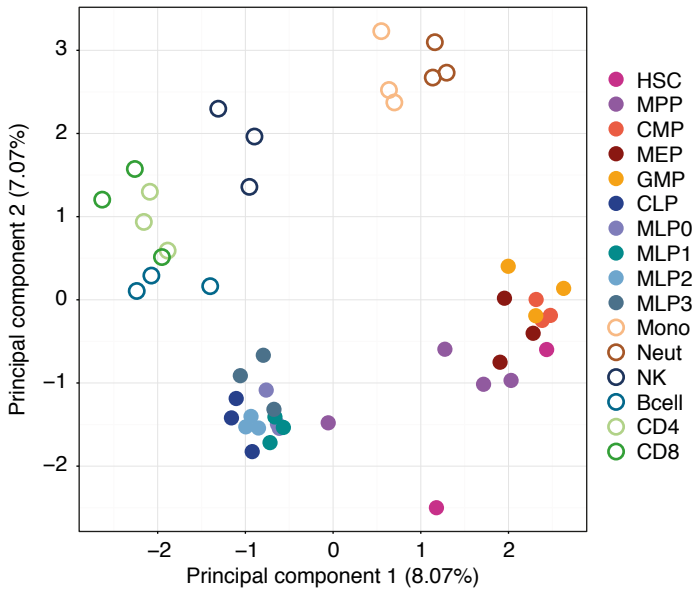


Figure 4.21: Principal component analysis of hematopoietic samples based on DNA methylation levels in signature regions. The first two principal components are shown and the numbers in parentheses indicate the percent variance explained. Point colors indicate cell types of merged samples (one point for each cell type and donor). Progenitor cell types are shown with a solid fill and differentiated cell types have no fill.

human hematopoiesis directly from the DNA methylation maps. Our predictive modeling approach thereby provides an avenue that can lead to a robust, quantitative definition of cell type which is complementary to methods relying purely on cell surface marker expression. States of individual replicates or even single-cells can be defined based on their similarity to reference cell types which can be derived from prediction probabilities.

In addition to peripheral blood samples, we also profiled DNA methylation in HSCs and MPPs from bone marrow, cord blood and fetal liver and observed differences in the methylation patterns between different tissue sources [Farlik *et al.* 2016]. In particular, we identified regulatory regions that exhibited markedly lower DNA methylation levels in peripheral blood HSCs compared to other sources and that were associated with binding sites of the CTCF insulator and cohesine complex proteins. These hypomethylation events could reflect changes in chromatin architecture and three-dimensional structure that ultimately control gene expression. Peripheral blood is readily accessible and therefore highly relevant for clinical diagnosis, while bone marrow, cord blood and fetal liver are also commonly used in basic research. Understanding these epigenetic differences and related changes in gene regulation in hematopoietic cells of different sources is therefore of high importance for reproducible and transferable research.

A large portion of the variation in our dataset originates from differential methylation between the myeloid and lymphoid lineage. In [Farlik *et al.* 2016], we investigated these differences in detail and observed an asymmetric pattern: while regulatory regions showing low methylation levels in myeloid cells and high methylation levels in lymphoid cells were enriched for binding sites of transcription factors associated with

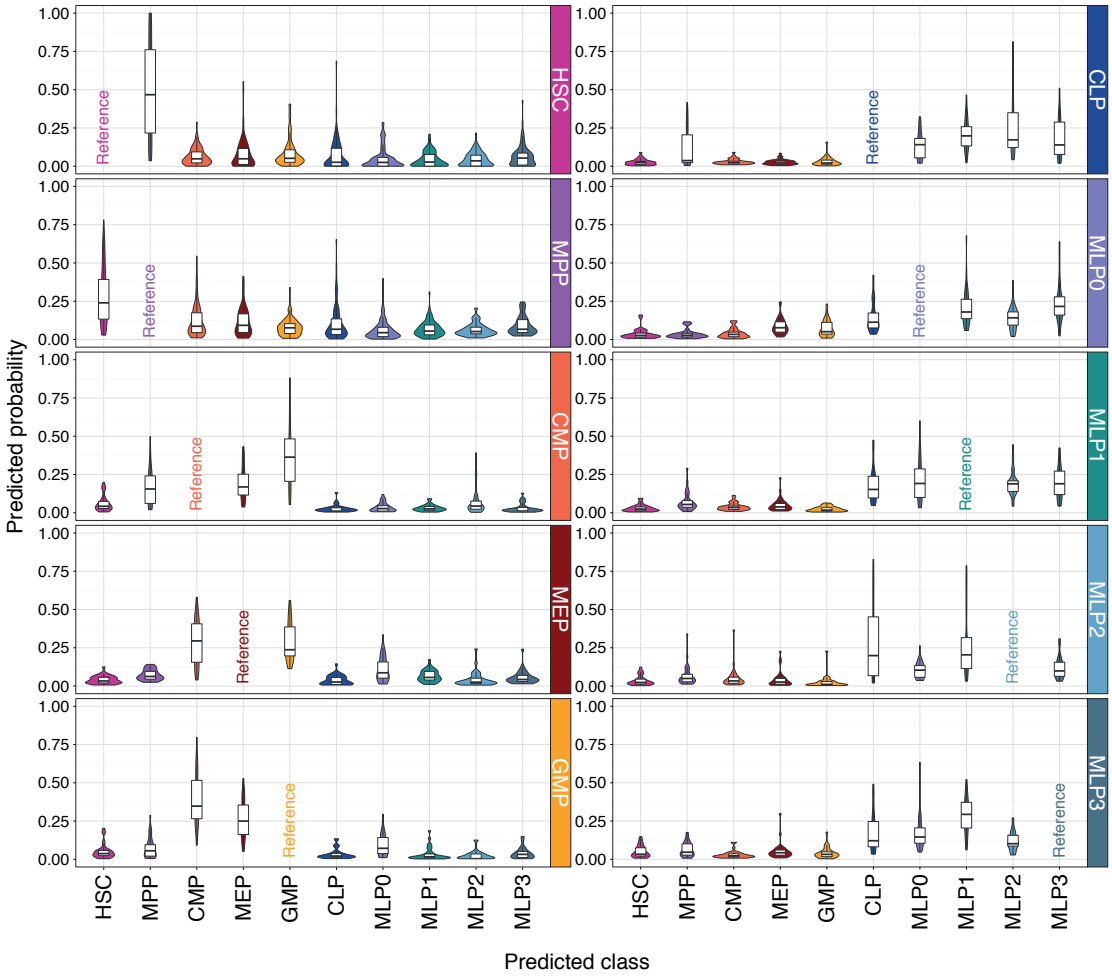


Figure 4.22: Distributions of cross-class probability for progenitor cell types. Violin plots and boxplots show the distribution of class probabilities across all samples of a given cell type according to the leave-one-class-out classifiers (cf. Section 4.4.1).

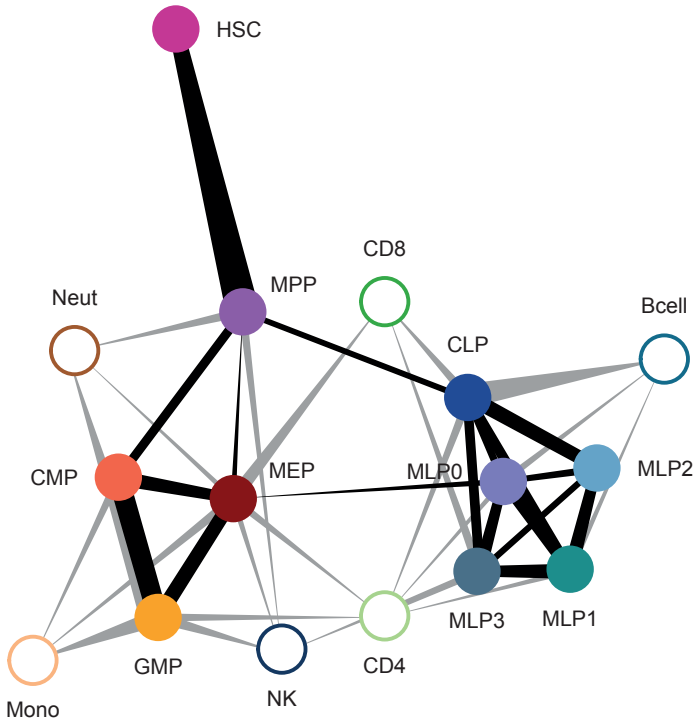


Figure 4.23: Similarity graph of hematopoietic cell types based on prediction probabilities. Nodes represent cell types and edges are weighted by cross-prediction class probabilities (cf. Section 4.4.1). Edges between progenitors are shown in black and gray edges show connections between differentiated and progenitor cell types. The graph layout was automatically generated using the Fruchterman-Reingold algorithm implemented in the IGRAPH R package [Csardi and Nepusz 2006].

hematopoietic differentiation, lymphoid malignancy and myeloid differentiation, regions that were hypermethylated in the myeloid lineage compared to the lymphoid lineage did not exhibit such characteristic transcription factor binding. Related studies have found similar patterns in murine hematopoiesis [Bock *et al.* 2012]. Together with experiments that showed only limited lymphoid differentiation when DNA methylation levels were reduced by knocking out methyltransferases in mice [Bröske *et al.* 2009], the data support the view that myeloid lineage commitment might represent a default differentiation pathway which is suppressed by DNA methylation in the lymphoid lineage.

In order to dissect early lymphoid differentiation, four different populations of MLP were characterized. Using *in vitro* assays, we observed multi-lineage differentiation potential of early lymphoid progenitors [Farlik *et al.* 2016]. MLP0 cells exhibited the highest potential to give rise to cell colonies of lymphoid as well myeloid cells in these assays. This trait was in accordance with distinctive DNA methylation patterns in sets of regions which were defined based on transcription factor binding profiles. These patterns could be reflective of cellular plasticity and could help to shed light on the regulatory basis of multi-lineage potential that has recently been discussed [Notta *et al.* 2016; Paul *et al.* 2015].

Additionally, we supplemented our dataset with RNA-expression profiles for hematopoietic progenitor cell types. However, when we related differential gene expression between cells of the myeloid and lymphoid lineages to differential methylation, we observed only moderate association between expression and DNA methylation in promoter regions and only found a small number of myeloid and lymphoid regulators with concordant differences [Farlik *et al.* 2016]. In contrast, integration of our methylation data with profiles of chromatin accessibility, which were published recently [M. R. Corces *et al.* 2016], revealed myeloid-specific and lymphoid-specific regions of open chromatin in progenitors that exhibit low methylation levels in mature cell types of the corresponding lineage. Lineage-specific opening of chromatin seems to precede DNA demethylation events in mature cell types. Chromatin opening could thus prime regulatory regions for stable activation through loss of methylation in a cell-type-specific manner. Together these results support the view that epigenetic regulation of hematopoietic differentiation involves mechanisms that go far beyond promoter-driven regulation. For instance, the involvement of distal elements such as enhancers and insulators and associated epigenetic marks remains to be elucidated.

A key technical challenge of our analysis derives from the relatively low sequencing coverage obtained for replicates with low cell numbers. We addressed this issue by assessing the reproducibility of DNA methylation profiles in individual replicates and employed a composite approach in which we merged the methylation levels of all 10-cell, 50-cell and 1,000-cell replicates for each cell type, tissue source and donor. We also evaluated a panel of unsupervised analysis methods with different parameter settings and concluded that the biological variation in the data could be captured across methods. Rather than studying DNA methylation on the level of single CpGs, in which coverage artifacts are extensive, we focused our analyses on the aggregate methylation profiles in genomic regions, such as putative regulatory elements. This approach not only mitigated biases due to coverage but also reduced the statistical complexity inherent to large feature numbers. When analyzing single-cell methylomes (data shown in [Farlik *et al.* 2016]), further aggregation across a panel of several hundred sets of such regions is advisable and has proven highly effective in low-input DNA methylation profiling [Farlik *et al.* 2015].

In summary, we identified signatures of human hematopoietic differentiation inscribed in DNA methylation. Understanding these signatures is not only important for understanding and tackling diseases such as leukemia and immune defects, but also as a model of how complex and dynamic tissues are formed and maintained by stem cell differentiation and lineage commitment. The described approach (low-input DNA methylation sequencing combined with predictive modeling) is directly transferable to dissecting the *in vivo* dynamics of other tissues such as the gut, skin and brain. Moreover, given the increasing interest and technical feasibility of DNA methylation biomarkers in clinical applications [Bock, Halbritter, *et al.* 2016], unraveling patterns of methylation involved in immunity-related and cardiovascular diseases, blood-cell malignancy as well as vaccine development could represent an important step for advancing precision medicine.

5

Analyzing and Manipulating DNA Methylation Patterns in Leukemia

The work described in this chapter has been conducted in collaboration with Giovanni Amabile and Annalisa Di Ruscio and has been published in [Amabile et al. 2015]. In the project, I was responsible for the bioinformatic analysis of DNA methylation data obtained by Reduced Representation Bisulfite Sequencing (RRBS) and designed corresponding figures. The benchwork was conducted by Giovanni Amabile, Annalisa Di Ruscio, Robert S. Welner, Alexander K. Ebralidze, Hong Zhang, Lihua Qi, Michelle M. Le Beau and Elena Levantini. The manuscript was written by Giovanni Amabile, Annalisa Di Ruscio, Christoph Bock and Daniel G. Tennen. Figures in this chapter have been adapted from the publication.

This chapter expands on the characterization of the DNA methylation landscape in human hematopoiesis by providing a perspective on the methylation dynamics that occur in malignant blood cells. We investigated how DNA methylation patterns are altered in models for human leukemia and how these patterns can be reverted using cellular reprogramming.

Chronic Myeloid Leukemia (CML) is characterized by a genetic translocation of the q34.1 region of chromosome 9 to the q11.2 region of chromosome 22 which gives rise to the *BCR-ABL* fusion gene and the so-called Philadelphia chromosome. Through its tyrosine kinase activity, the protein product of *BCR-ABL* can facilitate the increased proliferation of hematopoietic stem and progenitor cells [Machova Polakova *et al.* 2013]. Therefore, kinase inhibitors such as imatinib represent frequent treatment options for CML patients. Furthermore, aberrant DNA methylation in human leukemia is functionally involved in the onset and progression of cancer [Feinberg *et al.* 2006; Baylin and Jones 2011; Jones 2012]. In CML, aberrant methylation patterns in regulator genes have been implicated in cell proliferation and potential resistance to kinase inhibitor treatment. These findings are indicative of a complex interplay between genetic and epigenetic aberrations during leukemia progression. Using cellular reprogramming, it is possible to erase most tissue-specific epigenetic patterns and to establish a pluripotent cell state that resembles that of embryonic stem cells [Mikkelsen *et al.* 2008]. Induced Pluripotent Stem Cells (iPSCs) can be generated from a number of cell types, including malignant cells [Carette *et al.* 2010; Miyoshi *et al.* 2010; Kumano *et al.* 2012; Stricker *et al.* 2013]. Therefore, reprogramming represents a suitable tool for unraveling the relationships between epigenetic markup, proliferative activity, pluripotency and malignancy.

Employing the methods and pipelines described in Chapter 3, we characterized DNA methylation profiles in human leukemia cells and iPSC clones derived from these cells. We show that cellular reprogramming is able to reset cancer-related DNA methylation

signatures to those typical for pluripotent cells. Moreover, when transplanted into immunocompromised mice, reprogrammed leukemia cells exhibited reduced oncogenic potential. Finally, we describe aberrant methylation patterns that were induced when *BCR-ABL* was activated in transgenic mouse models.

5.1 Methods

Cell Cultures and Transgenic Mouse Models

K562 and KBM7 human leukemia cell lines, which harbor the *BCR-ABL* fusion gene, were cultured as described in [Amabile *et al.* 2015]. Primary CML cells were obtained from the bone marrow of two donors. Additionally, $CD34^+$ blood progenitor cells were profiled in the study. All cell types were reprogrammed using retroviral vectors carrying the transcription factor genes *OCT4*, *SOX2*, *KLF4* and *c-MYC* and were cultured as described in [Amabile *et al.* 2015]. Two iPSC clones reprogrammed from each of the K562, KBM7 cell lines and from primary CML cells were picked for further analysis. One clone was picked from the reprogrammed $CD34^+$ cell populations. For primary CML, iPSCs were generated from a single donor. In the following, iPSC lines derived from leukemic cells are referred to as Leukemia Induced Pluripotent Stem Cell (LiPSC). KBM7 cells and derived LiPSCs were provided by the Brummelkamp laboratory [Carette *et al.* 2010].

Furthermore, hematopoietic progenitor cell populations (Lin^-, $Sca1^+$, Kit^+ (LSK)), were selected from bone marrow mononuclear cells extracted from a transgenic mouse model in which the expression of BCR-ABL is subject to the control of an enhancer of the *Scl* (Stem Cell Leukemia) gene, thus allowing for the mostly specific expression of BCR-ABL in progenitor cells. This expression can be conditionally induced in the cells by withdrawal of tetracycline from the drinking water of the mice [Koschmieder *et al.* 2005] and can be reversed (rescued) when the drinking water is re-substituted with tetracycline.

Generation and Analysis of RRBS Data

RRBS was conducted on genomic DNA isolated from K562, KBM7 and primary CML cells as well as their reprogrammed counterparts and $CD34^+$-iPSCs, as described in [Amabile *et al.* 2015]. Sequencing libraries were prepared from two replicates corresponding to different passage numbers for K562, KBM7 and LiPSCs and in triplicate for $CD34^+$-iPSCs. Sequencing was performed on two libraries containing size-selected fractions of 40-120 bp and 120-220 bp respectively, using an Illumina GA *IIx* machine with a read length of 36 bp and an Illumina HiSeq with read length of 76 bp ($CD34^+$-iPSCs). Analogously, RRBS profiles were generated for two replicates of the non-induced (control), induced (leukemic) and rescued state of bone marrow cells from the transgenic mouse model. Sequencing, primary read processing and alignment was conducted at the Cancer Science Institute (National University of Singapore) and, for $CD34^+$-iPSCs, at the CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences (Vienna, Austria). Sequencing reads were mapped to the human genome assembly NCBI37/hg19 and mm9 for human and mouse data, respectively, using RRBSMAP [Xi *et al.* 2012], allowing for two mismatches. $CD34^+$-iPSC reads were mapped using BSMAP [Xi and W. Li 2009]. RRBS alignment data for human H1 ESC line passage 37 and 38 and methylation call data for $CD34^+$ cells were obtained from the NIH Roadmap Epigenomics Mapping Consortium [Roadmap Epigenomics Consortium *et al.*

2015]. Methylation levels were called using the Bis-SNP software [Liu *et al.* 2012]. Further, integrative processing of methylation data was performed using RnBeads version 0.99.17 [Assenov *et al.* 2014] (cf. Section 3.2). After filtering of CpGs with low coverage across more than 50 % of the samples and of CpGs on the sex chromosomes, the integrative analysis was based on 1,009,592 CpGs. Promoter methylation was defined as the mean methylation level for CpGs within a window of 1,500 bp to 500 bp of the TSS of Ensembl-annotated genes. Differentially methylated sites and regions were identified using the ranking-based approach described in Section 3.2.

Data Availability

Sequencing reads and DNA methylation level data have been deposited to the GEO under accession number `GSE50456`. Full RnBeads analysis reports, including tables characterizing differential methylation, can be accessed from the supplementary website[1].

5.2 Results

In order to dissect the DNA methylation dynamics during the reprogramming of human leukemia cells to a pluripotent state, LiPSC lines were generated from K562 and KBM7 CML cell lines as well as from primary bone marrow cells from a BCR-ABL positive donor and $CD34^+$ hematopoietic progenitor cells. Analysis of genotyping data assayed by SNP arrays confirmed that the reprogrammed clones contained the same mutational patterns as their parental cells [Amabile *et al.* 2015]. We obtained high-resolution DNA methylation profiles for leukemia cells and all iPSCs using RRBS. Additionally, RRBS data for human ESCs (H1 cell line) and $CD34^+$ assayed by the REMC were included in the analysis.

Leukemia-Specific Methylation Patterns are Reprogrammed During Induction of Pluripotency

Cellular reprogramming of leukemia cells led to widespread changes in methylation profiles. Globally, LiPSCs exhibited genome-wide hypomethylation compared to their leukemic counterparts. The effect was more pronounced during the reprogramming of cell lines than in the primary CML samples, when considering variation across all samples (Figure 5.1). Although the overall methylation patterns of reprogrammed cells were highly similar to those of ESCs, hierarchical clustering generally grouped reprogrammed cells along with the cells from which they were derived, thereby indicating the retention of epigenetic memory during the generation of iPSCs [Kim *et al.* 2010; Lister *et al.* 2011; Doi *et al.* 2009; Bock *et al.* 2011]. Notably, K562 cells and their derived LiPSCs exhibited marked hypermethylation in the promoters that varied most in the dataset (Figure 5.1). These patterns clearly distinguished them from other cells included in the analysis and thus represent cell-line-specific methylation signatures.

Focusing on regions differentially methylated between leukemia cells and LiPSCs, reprogramming-induced hypomethylation was observed in CpG islands, promoters and most common types of repetitive genomic elements (Figure 5.2). This effect was most pronounced in K562 and KBM7 cell lines, but was also observable in primary CML cells. Notably, differences between $CD34^+$ progenitor cells and their reprogrammed

[1] `http://reprogramming-leukemia.computational-epigenetics.org`

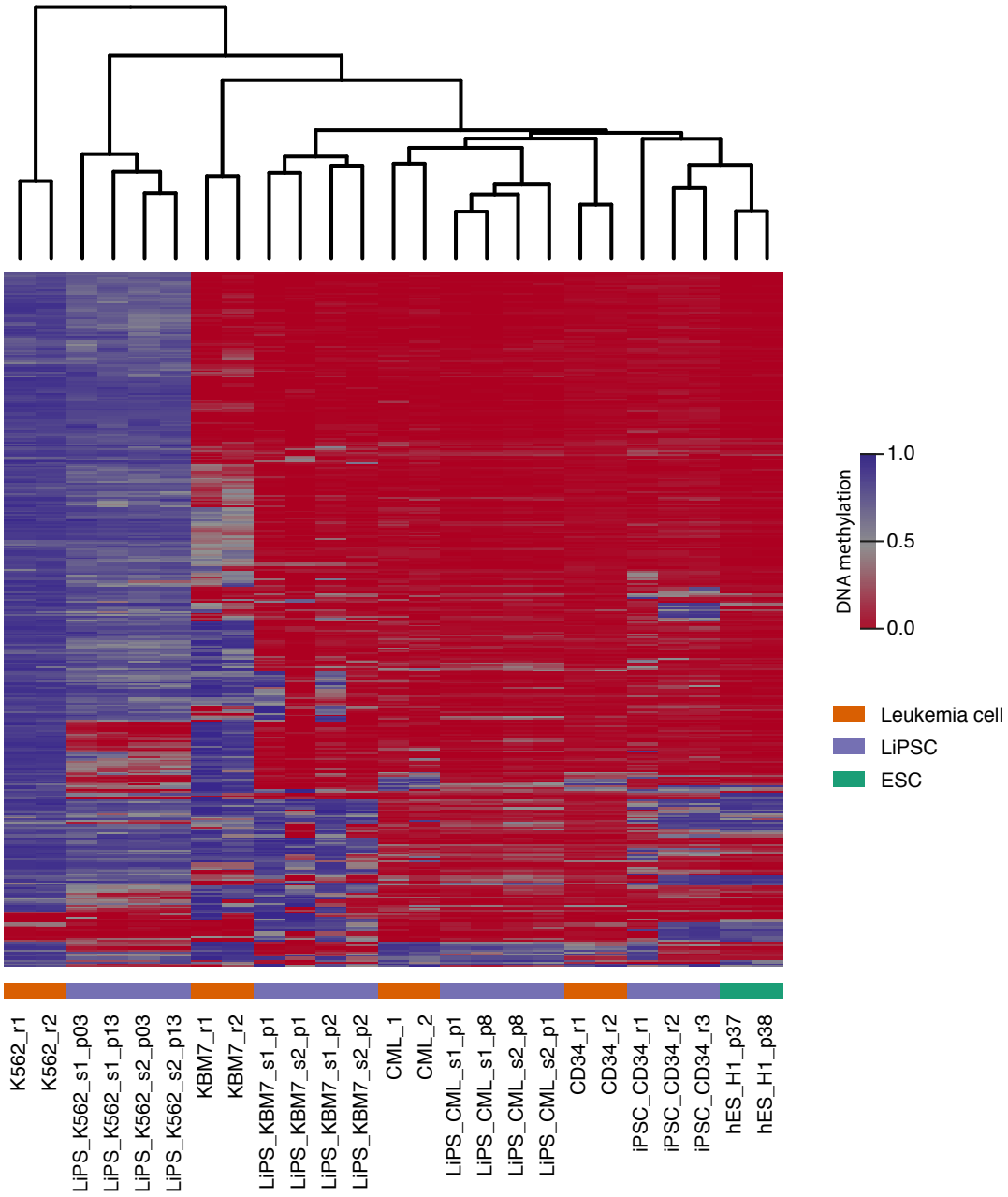


Figure 5.1: DNA methylation patterns in promoters are reprogrammed during the generation of iPSC. The heatmap shows mean methylation levels in the 402 (2 %) most variable promoters (rows) across all samples (columns). The dendrogram above shows the result of hierarchical clustering according to all surveyed CpGs using Manhattan distance and average linkage. The sample coloring below indicates state of reprogramming and pluripotency.

counterparts were low in those regions compared to the changes associated with reprogramming cancer cells. Examples of gene promoters hypomethylated in reprogrammed leukemia cells include developmental and pluripotency-associated transcription factors such as *SALL4* and *HOXA5* as well as candidate tumor suppressor genes such as *BRCA1* (Figure 5.3).

Employing the RnBeads' ranking-based approach, we identified those promoters that exhibited most evidence for differential methylation between all replicates of LiPSC and their parental leukemia cells (Figure 5.4). The majority of differentially methylated promoters lost methylation in LiPSCs. These differences were associated with development, differentiation and cell signaling when conducting a GO enrichment analysis on the biological process ontology (Figure 5.4b). In contrast, the promoters exhibiting higher DNA methylation in the parental, leukemic cells were enriched for hematopoiesis-related categories such as lymphocyte activation and immune response. To obtain a more unbiased view on genome-wide changes in methylation during the reprogramming of leukemia cells, we also characterized the most differentially methylated genome-wide tiling regions using EpiExplorer [Halachev *et al.* 2012] (Figure 5.5). We qualitatively identified enrichment of differential methylation in regulatory regions associated with CpG islands, promoters, enhancers as well as PcG protein repression. The latter is consistent with previous studies which associated hypermethylation in cancer with Polycomb targeting in ESCs [Ohm *et al.* 2007; Schlesinger *et al.* 2007; Widschwendter *et al.* 2007]. Moreover, in line with previous findings on aberrant methylation in cancer [Hon *et al.* 2012; Berman *et al.* 2012], we also observed differential methylation in putative enhancer regions (marked by H3K4me1 and H3K27ac and respective chromatin states).

Taken together, our results suggest the erasure of cancer-associated DNA methylation patterns during reprogramming. Pluripotency and developmental pathways become activated due to the demethylation of respective promoters and regulatory elements while a comparatively small number of hematopoiesis-associated genes are inactivated through targeted hypermethylation.

DNA Methylation Patterns in a BCR-ABL Inducible Mouse Model

In addition to the methylomes of human cell lines, single-CpG resolution DNA methylation maps of hematopoietic progenitor cells obtained from a CML mouse model were generated. In this model, the expression of BCR-ABL can be induced by the withdrawal of tetracycline from the drinking water thereby leading to a chronic myoproliferative disorder closely resembling CML. This expression can be reverted when tetracycline is re-added. We used RRBS to profile methylation in (i) control mice, (ii) mice in which BCR-ABL was induced and (iii) mice in which the expression was reduced following induction (rescued). However these validation experiments were limited in terms of genome-wide coverage and sample replicates: after filtering for CpGs located on the sex chromosomes and CpGs with low read coverage, DNA methylation measurements for 788,407 sites were retained. Additionally, only two replicates were available for each condition. Therefore, we were only able make qualitative statements and our conclusions need to be subjected to more rigorous validation. Globally, due to the low number of replicates per condition, we detected modest, statistically insignificant differences in DNA methylation patterns between conditions. Nonetheless, we observed patterns of aberrant methylation upon BCR-ABL expression, when focusing our analysis on regions gaining methylation in induced mice compared to control mice (Figure 5.6). Compared to control mice, leukemic mice exhibited a moderate, global increase in methylation

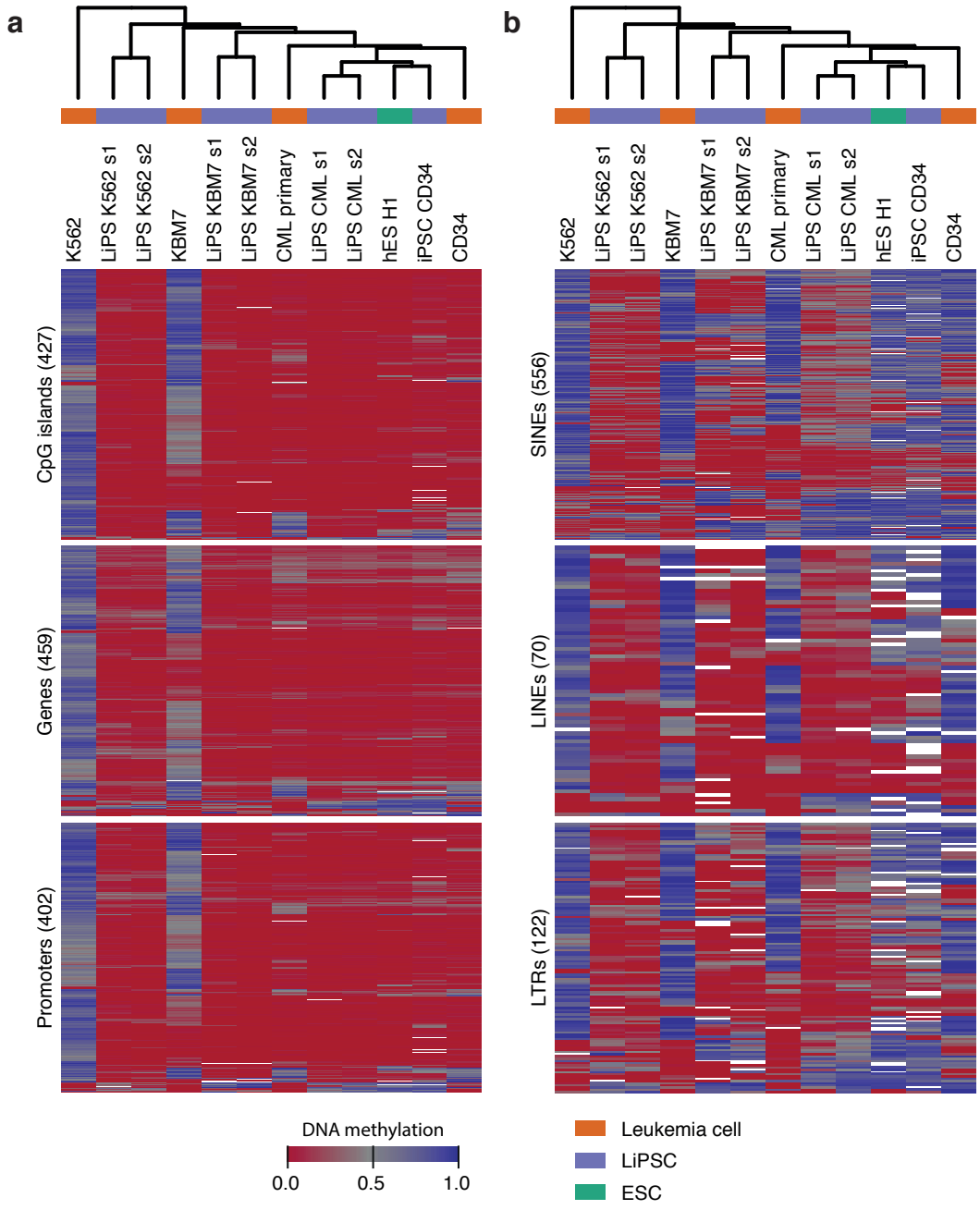


Figure 5.2: Reprogramming erases aberrant DNA methylation patterns in LiPSCs. Heatmaps visualize average DNA methylation levels in **(a)** selected region types and **(b)** selected families of repetitive elements. For each region type, the 2 % of regions (numbers are indicated in parentheses) that scored highest according to a ranking score in differential analysis comparing the group of K562, KBM7 and primary CML samples to the group including all derived iPS cell lines are shown. The dendrograms above show the result of hierarchical clustering according to all surveyed CpGs using Manhattan distance and average linkage. Methylation levels for replicates and for the two primary CML donors are averaged.

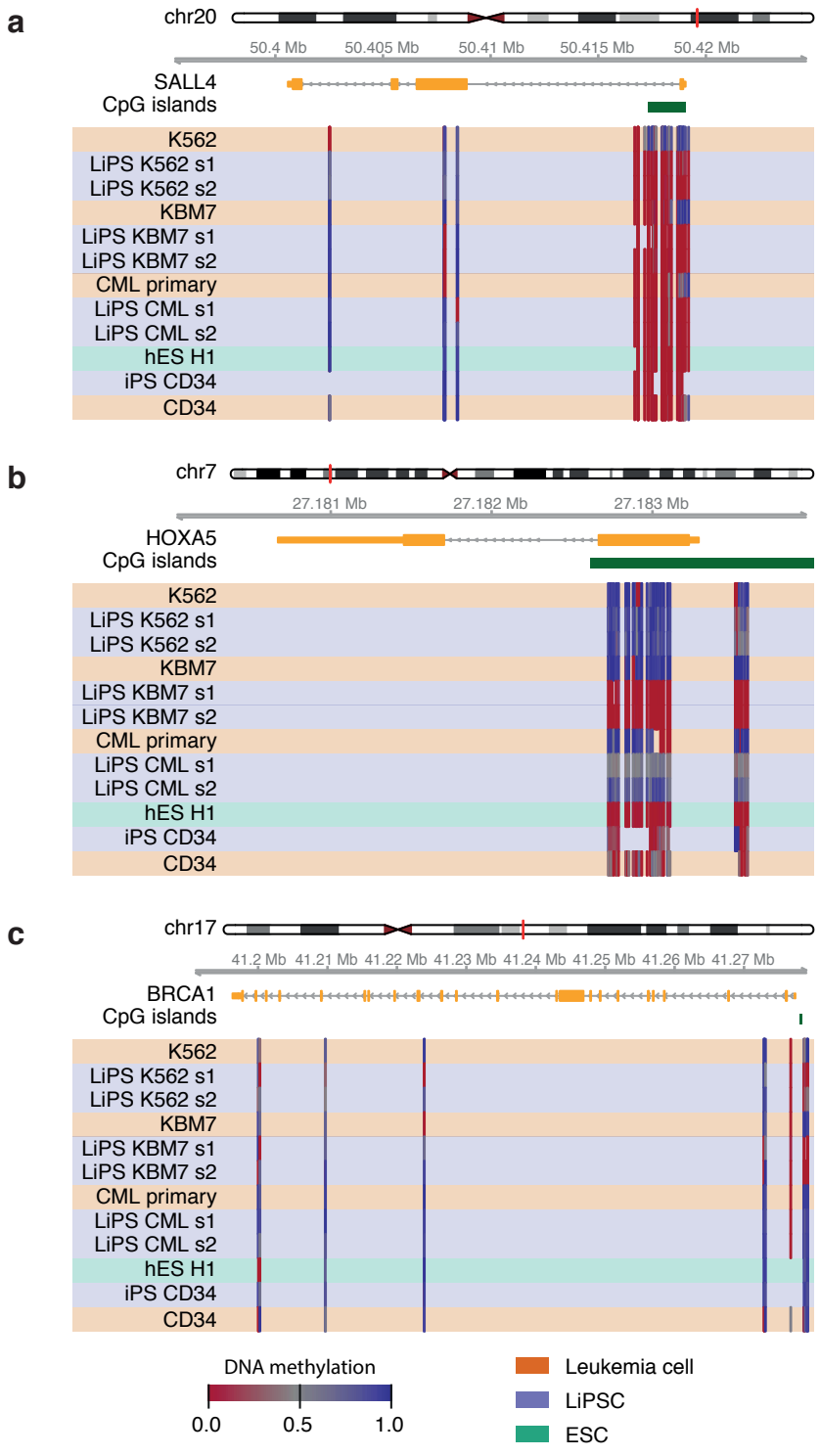


Figure 5.3: Leukemia-specific methylation patterns are reset in regulatory gene loci. The **(a)** *SALL4*, **(b)** *HOXA5* and **(c)** *BRCA1* loci are shown. Coordinates for Ensembl genes and CpG islands are represented by yellow and green bars respectively. A heatmap in which each column depicts the methylation levels of a single CpG, averaged across replicates is shown below.

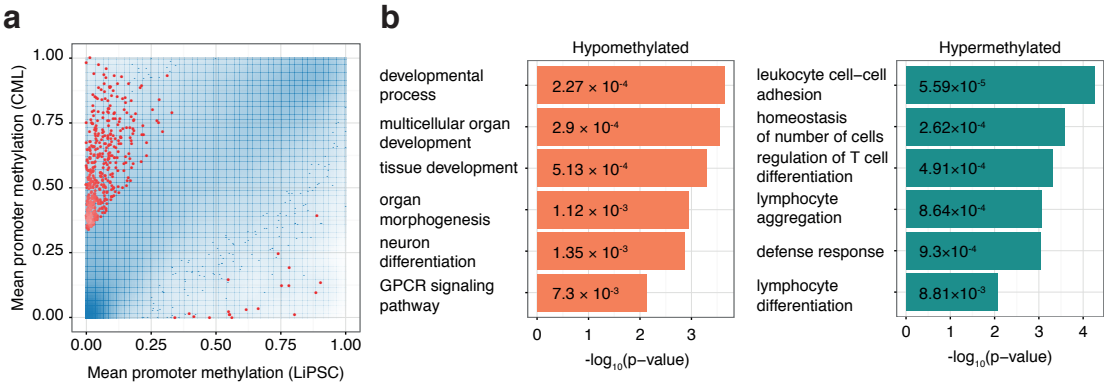


Figure 5.4: Differentially methylated promoters in LiPSCs are associated with pluripotency and hematopoiesis. **(a)** Scatterplot showing mean promoter methylation levels of LiPSCs and leukemia cells. The 500 promoters with the highest extent of differential methylation (according to the ranking-based approach described in Section 3.2) are plotted in red. Point density is shown as blue shading. **(b)** *p*-values for selected GO terms enriched in the 500 most hypermethylated and hypomethylated promoters in LiPSCs compared to leukemia cells, respectively. *p*-values were computed using RnBeads and the GOstats R package [Falcon and R. Gentleman 2007].

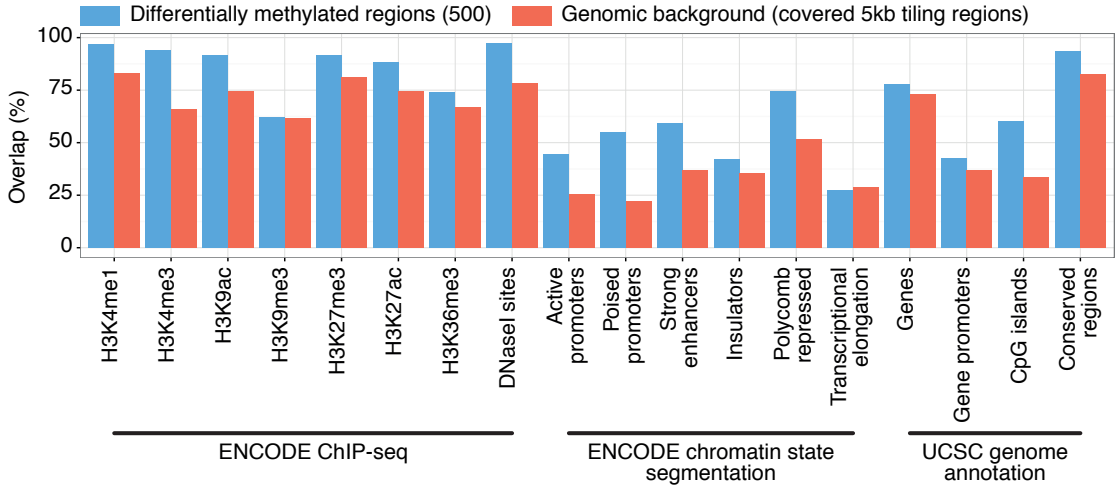


Figure 5.5: Differentially methylated regions in LiPSCs are associated with regulatory regions in the genome. Bars plots indicate overlaps with genomic and epigenomic features obtained by an EpiExplorer analysis [Halachev *et al.* 2012]. The 500 highest-ranking differentially methylated tiling regions (blue) are compared with a background of 72,469 of 5-kb tiling regions covered in the dataset (red). Features include ChIP-seq peaks and chromatin state segmentations obtained from the ENCODE project [ENCODE Project Consortium 2012] and annotations obtained from the UCSC Genome Browser [Kent *et al.* 2002]. Overlaps of at least 1 bp are shown as percentage of all annotated regions.

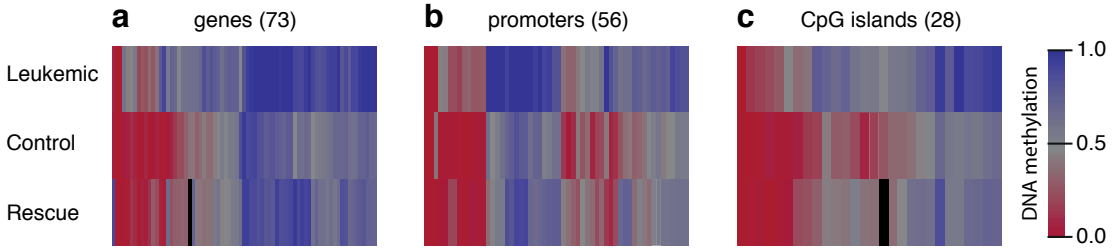


Figure 5.6: BCR-ABL expression induces aberrant methylation in mouse hematopoietic progenitor cells. Heatmaps show the average methylation levels in genes **(a)**, promoters **(b)** and CpG islands **(c)**, which exhibit evidence of hypermethylation in BCR-ABL induced mice compared to control mice (difference in mean methylation levels > 0.2). Methylation levels of the two replicates are averaged for each condition.

levels associated with CpG islands and promoters. Upon BCR-ABL induction, we observed hypermethylation events at select gene promoters linked to development, such as *Hoxb1* (data not shown). Globally, the rescued state appeared to be more similar to control state than to the induced state. It remains to be elucidated how these altered DNA methylation patterns are mechanistically linked to the expression of BCR-ABL.

5.3 Experimental Evidence and Discussion

Using experimental methods, we also tried to answer the question of whether DNA methylation changes inherent to induced pluripotency result in resetting the cells to a "normal" epigenomic state associated with reduced malignancy. Experimental evidence in [Amabile *et al.* 2015] indeed confirmed that reprogramming leads to a reduced oncogenic potential in mouse models: LiPSC-derived cells from primary CML and K562 could be differentiated towards myeloid and erythroid lineages (as measured by surface marker expression) in spite of the genetic alteration of *BCR-ABL*. In contrast, leukemic cells appeared to be locked in an undifferentiated cell state. Interestingly, when DNA methylation was depleted by treating leukemic cells with the demethylating agent 5-Azacytidine (AZA) they exhibited differentiation potential towards cells of the myeloid lineage, indicating that aberrant DNA methylation could disrupt differentiation. Moreover, transplantation of leukemia cells into immunocompromised mice led to a malignant phenotype indicated by infiltration of the spleen by these cells and lower survival rates. In contrast, mice transplanted with LiPSCs-derived cells showed no evidence of disease onset.

Furthermore, mice in which the expression of BCR-ABL was induced developed leukemia and had low survival rates. In contrast, mice in which either the genetic cause of leukemia was inhibited by blocking BCR-ABL using the tyrosine kinase inhibitor imatinib or in which DNA methylation patterns were reverted by use of AZA exhibited a relatively normal phenotype. In addition, survival of immunocompromised mice transplanted with BCR-ABL-induced leukemic bone marrow was significantly reduced compared to survival of mice transplanted with imatinib or AZA-treated, BCR-ABL-induced bone marrow. These results suggest that malignancy is associated with widespread changes in DNA methylation linked to BCR-ABL expression. When these changes are

artificially reverted through AZA treatment or through blocking of BCR-ABL via imatinib, a relatively normal phenotype can be re-established.

In contrast to the reduced oncogenic potential we observe in LiPSCs, a related study reported that neural progenitor cells derived from reprogrammed glioblastoma cells retained their malignancy [Stricker *et al.* 2013]. This indicates that different tumor types are differentially driven by genetic and epigenetic events. In leukemia, genetic lesions and epigenetic aberrations seem to contribute to disease development in concert while glioblastoma are apparently more driven by alterations in DNA sequence.

In conclusion, our results emphasize the close interconnection of genetic and epigenetic contributions to cancer development. In CML, the induction of BCR-ABL can trigger substantial changes in DNA methylation patterns. However, the molecular mechanisms by which the genetic lesions are responsible for these characteristic methylation signatures are not yet well understood. Importantly, the changes in DNA methylation can, to a certain extent, be reset by means of cellular reprogramming. Finally, our findings lend further support to the relevance of studying the potential of DNA-demethylating drugs for anti-cancer therapy.

6 Perspective

More than 200 canonical cell types of the human body have been labeled [Alberts *et al.* 2008]. These cells carry largely identical DNA sequences and yet they fulfill many different functions and exhibit markedly different phenotypes and varying degrees of specialization. These attributes are governed by regulatory mechanisms and their inherent epigenetic profiles. Furthermore, cells of the same canonical type can display a high degree of heterogeneity, indicating the need for more detailed descriptions of cell state. This thesis provides a detailed account on the analysis of DNA methylation signatures of cell identity. Methods and computational tools for the identification, characterization and comparison of these signatures have been developed in the context of this work. This chapter concludes the thesis by summarizing its key developments and results and provides a broad perspective on their role in the interpretation of epigenome maps.

6.1 Conclusion

A multitude of software packages for processing of epigenome data are available in public repositories such as GitHub[1], SourceForge[2] and Bioconductor [R. C. Gentleman *et al.* 2004]. These packages provide access to established methods in the field as well as to novel approaches for specific tasks and thus represent the building blocks for computational pipelines. Emerging guidelines and standards in DNA methylation analysis have recently been outlined [Bock 2012] and are starting to be standardized as the result of expert discussions [Michels *et al.* 2013]. Chapter 3 presents software tools such as RnBeads and epiRepeatR which implement these standards that are now widely used in epigenome analysis. Furthermore, we developed a pipeline for the quantification of DNA methylation levels from aligned bisulfite sequencing reads that has been applied and validated in two large-scale methylome studies [Ziller *et al.* 2011; Ziller *et al.* 2013] and continues to be employed. In recent years, the development of user-friendly software tools that can be generally applied when interpreting epigenome data has shifted into focus. Prime examples include Galaxy [Giardine *et al.* 2005] and EpiExplorer [Halachev *et al.* 2012]. The availability of such tools enables researchers with little background in computer science or bioinformatics to conduct basic (epi)genome-wide analyses. In line with this development, we have implemented RnBeads, a pipeline for the comprehensive analysis of DNA methylation in large datasets. RnBeads supports virtually all experimental platforms providing single-base-resolution DNA methylation measurements. The tool facilitates start-to-finish analysis by specifying only a few lines of R code, while its modular design enables highly configurable workflows at the same time. With 200 to 300 unique downloads from the Bioconductor repository monthly (status:

[1] https://github.com
[2] https://sourceforge.net

November 2016), the tool is well-accepted by the epigenomics community. Quantifying differential patterns between cell populations is important for identifying and understanding development- and disease-related changes in epigenetic regulation. In the DNA methylation field, a plethora of methods and implementations is available for this task (cf. Appendix B). In this work, we devised a robust and interpretable approach that quantifies differential methylation based on predefined genomic regions using a ranking-based method that takes statistical significance into account in addition to absolute and relative differences in methylation.

Although repetitive DNA elements are highly abundant and subject to epigenetic regulation in the human genome, only limited software tools exists for the analysis of epigenomic marks in repeats on the genome scale. We therefore developed epiRepeatR, a pipeline for the global epigenomic characterization of repeat subfamilies across samples in large datasets originating from bisulfite and enrichment-based sequencing. Applying our pipeline to a dataset of human blood cells we confirmed, that particularly young, CpG rich elements are marked by epigenomic signatures which are characteristic of repressed chromatin, such as DNA methylation, H3K27me3 and H3K9me3. In contrast, some *Alu* elements co-localize with epigenomic marks indicative of enhancers and promoters which potentially indicates a gene-regulatory role of repetitive element sequences.

Chapter 4, we dissected the *in vivo* DNA methylation dynamics in the human hematopoietic system, using the tools described in Chapter 3. We identified epigenetic signatures characteristic of cell state and explored how these signatures are altered during cell differentiation, immune memory formation, cellular reprogramming and in leukemia. The epigenome-wide analysis of one of the largest methylome datasets available revealed characteristic variability across blood cell types. Overall, related cell types exhibited highly similar signatures and methylation patterns closely resembled their respective cell lineage. Dynamic patterns were found in regulatory regions of the genome that were previously annotated with cell-type-specific immune activity. Focusing on immune memory formation in T helper cells, we discovered genome-wide, gradual loss of methylation in regions that showed overall intermediate to high levels of DNA methylation. The identified changes could reflect the cells' respective proliferative history and support a model of progressive differentiation during memory formation.

With the goal to provide a better understanding and characterization of epigenetic changes in the early steps of the formation of blood cell identity, we obtained low-input, whole-methylome data for hematopoietic stem and progenitor cells. Statistical modeling identified signature regions indicative of cell lineage which coincided with cell-type-specific regulatory patterns. These signatures are in agreement with a model of hematopoietic differentiation in which a cell type is defined by a population of cells exhibiting common epigenetic signatures, but also cell-to-cell heterogeneity (see next section). In line with this model, we explored variation of DNA methylation in putative regulatory regions of the genome, identified DNA methylation signatures of hematopoietic differentiation and established relationships in the epigenomic patterns of different blood cell types.

Perturbations of epigenetic signatures and their interaction with regulators that catalyze transcriptional and epigenetic changes are frequently associated with disease. Identifying and understanding disease-linked epigenomic changes is fundamental for the discovery of biomarkers and the development of drugs that alter epigenetic signatures. They could thus serve as important diagnostic tools and therapy instruments in precision medicine [Bock and Lengauer 2012]. Furthermore, epigenetic regulation plays

a role in the adaptation of immune cells to a pathogen. The methods and software packages described in this work provide versatile tools for characterizing disease-associated changes in DNA methylation patterns. Specifically, our RnBeads software can be efficiently employed for the interpretation of EWAS data and it is particularly useful for the discovery of DNA methylation biomarkers. Methods for validating and testing these biomarkers have recently been evaluated [Bock, Halbritter, *et al.* 2016] and their clinical implementation could have considerable impact on diagnostics and the development of personalized therapies.

Furthermore, the study described in Chapter 5 analyzes characteristic changes in DNA methylation induced by the reprogramming of leukemia cells to a pluripotent state and thus provides another perspective on cancer treatment. Although some epigenetic memory of the cell of origin was retained upon the induction of pluripotency, reprogrammed cells lost DNA methylation patterns indicative of leukemia and were highly reminiscent of embryonic stem cells. Our mouse models suggest that genetic lesions can lead to genome-wide changes in DNA methylation and that the resetting of epigenome state induced by reprogramming of leukemia cells is associated with reduced oncogenic potential. Therefore, assessing the efficacy of epigenome-altering drugs could be of relevance for future anti-cancer therapies.

6.2 Probabilistic Interpretation of the Epigenomic Landscape

The inscription of a cell's identity in its epigenetic signatures and the consequent localization in the epigenomic landscape relative to other cells are recurring themes in this thesis. Huang [2012] interpreted Waddington's epigenetic landscape [Waddington 1957] from a gene regulatory network perspective. He introduced the concept of an "epigenetic, quasi-potential landscape" which is shaped by regulatory interactions of genes. Gene interactions determine the stability of a cell's state, which is characterized by its gene expression levels. Stable cell states are described as attractors in the quasi-potential landscape and give rise to developmental trajectories and distinct phenotypes. The interactions, in turn, are determined by (epi-)genetic factors and are assumed to be hard-wired for a given genome. Figure 6.1 outlines a complementary interpretation of the epigenetic landscape that extends this view and emphasizes the regulatory role of epigenomic factors. In this probabilistic perspective, each cell can be observed in a state whose description comprises the entirety of DNA sequence, RNA expression, protein and metabolite levels, the structural organization of the DNA inside the nucleus as well as regulatory parameters imprinted in patterns of DNA methylation, histone modifications and other epigenomic marks. Thus, a cell can be placed in a hypothetical, high-dimensional probability space and a corresponding probability distribution captures the likelihood of observing a cell in a given configuration. Due to its extremely high dimension, describing this probability space in its entirety is unrealistic and modeling approaches resort to characterizing subspaces of it. Importantly, as outlined in Section 2.1, epigenetic factors greatly influence genomic architecture and gene expression. It is therefore crucial to include these factors in the parametrization of the probability space. The (unknown) probability distribution itself is determined by the physical and chemical properties of the factors that contribute to the cell state as well as their interactions. Spatio-temporal factors, in particular the environment in which the cell is placed, influence these properties. Figure 6.1 shows a two-dimensional abstraction of the high-dimensional probability space. Cell configurations profiled by various

"omics"-techniques are sampled from the probability distribution and represent phenotypically distinct cell types, as for instance defined by the expression of cell surface markers (colored dots in Figure 6.1). This sampling approximates the local maxima of the probability distribution (triangles in Figure 6.1). The probability (or height in the landscape) corresponds to the prevalence of cell states in an organism and therefore influences sampling. Moreover, the variance within each cell type can be attributed to cell-to-cell heterogeneity that is due to cell-intrinsic factors (e.g. the stage in the cell cycle), but partly results from stochastic variation. Additionally, depending on the profiling techniques employed, this variance also includes contributions due to measurement errors. Certain cell types can be relatively well defined (e.g. the light blue cell type in Figure 6.1) or exhibit high variability (e.g. the turquoise cell type). High variability leads to potentially ambiguous definitions of phenotypically-defined cell types. For instance, in practice, different sorting criteria are employed by different research groups in order to define highly similar canonical cell types. Models that describe cell types on a multidimensional, continuous scale could therefore complement the discrete definition based on surface markers and lead to more statistical notion of cell type. These models can be used to place cells into the epigenomic landscape based on their respective representation and allow for assessing similarities and differences between cells. They could thus be particularly useful for characterizing populations of malignant cells which exhibit high heterogeneity.

Furthermore, currently available epigenome data are measured on the basis of cell populations which can result in a reduced signal-to-noise ratio depending on the heterogeneity of the population. Particularly in the context of DNA methylation analyses in EWAS settings, accounting for intra-sample heterogeneity is of major importance [Jaffe and Irizarry 2014] and computational methods have been devised for dealing with multiple cell types represented in a sample [Houseman *et al.* 2012; Houseman *et al.* 2014; Zou *et al.* 2014]. As can be seen from the illustration above, cell-to-cell variability is present even in populations that were subject to rigorous cell-sorting procedures. Emerging single-cell technologies will continue to provide increasing amounts of data and combining the information from multiple levels of of "omics" data will likely result in valuable insights into the molecular basis of this variability [Bock, Farlik, *et al.* 2016]. Devising corresponding models constitutes an important challenge in computational epigenomics. Additionally, as stated above, the environment contributes to shaping the topology of the epigenomic landscape and therefore also intra- and inter-cell-type heterogeneity. Currently, environmental and temporal covariates are just beginning to be understood and are only infrequently incorporated into respective models.

Biological processes such as cell differentiation during development and cellular reprogramming entail transitions between cell states. In the probabilistic view outlined here, these transitions exhibit different degrees of stochasticity and can be interpreted as directed random walks from one mode to another and they are guided by the topology of the underlying probability distribution (dashed lines in Figure 6.1). Employing a geographical metaphor, this corresponds to traversing descending and ascending slopes from one summit in the landscape to another. If these slopes are fairly broad, the walk can take several alternative routes at random and multiple paths may be equally likely whereas the path is more or less predetermined when following a narrow ridge.

In the classical view of the epigenetic landscape, elevation is associated with differentiation potential and its quantification poses an important aspect in understanding cellular plasticity. For instance, Banerji *et al.* [2013] proposed a measure of network

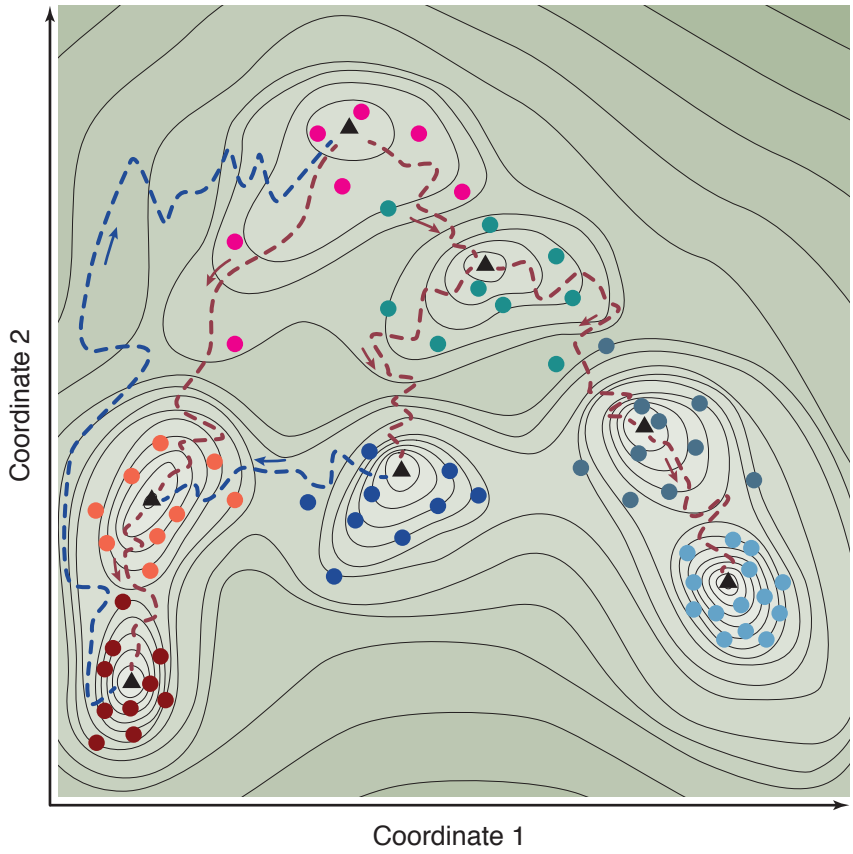


Figure 6.1: Probabilistic interpretation of the epigenomic landscape. The two coordinates depicted represent a simplified, high-dimensional epigenomic space that is defined by cell characteristics that can be quantified by various experimental procedures (e.g. DNA sequence, RNA expression levels, genome-wide DNA methylation levels, histone modifications, etc.). The out-of-the-plain axis represents the probability of observing a cell in a given epigenomic configuration, and the local probability density is indicated by contour lines. Triangles denote local maxima of the underlying probability distribution. Phenotypically-defined cell types are illustrated by different colors. Dashed lines show paths of reprogramming during cell differentiation (red) as well as cellular reprogramming to pluripotency and transdifferentiation (blue). The figure is based on hypothetical data.

entropy, which was quantified by integrating gene expression levels over a protein interaction network, as a surrogate for the energy potential (elevation) in the differentiation landscape. However, different interpretations for the property of cellular plasticity exist and it is unlikely that it can be reduced to a single quantity. Pluripotent progenitor cells generally constitute relatively rare populations in an organisms. In the probabilistic landscape interpretation, they are therefore represented by less prevalent states, which exhibit lower altitude and are more dispersed compared to differentiated states (exemplified by the pink and turquoise cell types in Figure 6.1). Notably, computational methods have been devised in order to gauge the differentiation potentials of a variety of populations of pluripotent cells [Bock *et al.* 2011]. Interestingly, this study also showed that cells often retain signatures characteristic of their differentiation history, which is reflected by their proximity in the epigenomic landscape.

Obviously, the dimensionality and parametrization of the complete probability space is too large and complex to be exhaustively captured by current modeling techniques. Therefore, models resort to parsimonious representations of cell state, in which important aspects of the overall distribution are conserved and which thus provide an adequate approximation of the epigenomic landscape[3]. Multiple "omics" technologies provide the means of characterizing subspaces of the cell-state space. Each of these subspaces is high-dimensional in itself and parametrized by up to millions of features. Further reducing the number of features can mitigate common statistical issues associated with high-dimensional data such as model complexity and resultant overtraining and biases incurred by multiple-testing.

This work employs DNA-methylation-based representations of the cell-state space that focus on predefined genomic regions such as promoters, repetitive genomic elements or putative regulatory regions. Considering whole-epigenome data from multiple blood-related cell states we started to chart the territories in the methylation landscape occupied by distinct cell types. We employed computational models in order to quantify similarities and differences between cell populations associated with hematopoietic development and malignancy. The resulting descriptions complement the phenotypically-defined notion of cell type and can be used to characterize cell-to-cell heterogeneity.

6.3 Outlook

This work describes a comprehensive ensemble of methods and tools for interpreting DNA methylation patterns. Nevertheless, models integrating multiple layers of the regulatory code become increasingly important in order to capture the complex interplay of regulatory factors. Two approaches are generally employed: (i) derive separate models for different feature types and assess the congruency of those models, or (ii) incorporate interactions of factors in a joint model. Typically, the first approach results in relatively interpretable models for single feature types, and complexity through data integration is introduced by the quantification of agreement across the different layers. The integrative analysis parts in this work generally employ this approach. Models derived from the latter approach are generally more complex, more susceptible to overtraining and

[3] This problem is well-known in the statistical learning field and addressed by dimension reduction techniques. Notably, the dimension can always be reduced to the number of observed cell states without loss of information.

harder to interpret, but have the potential to directly or indirectly capture the interactions involved in epigenetic regulation. They thus represent promising directions of expanding the methylation-focused view of this thesis. An interesting question to be asked here is what sets of epigenetic features are required to address specific problems and how much information can be gained by incorporating additional features. Statistical models for assessing the associations between different epigenomic marks on a genome-wide scale are still underrepresented in the literature. However, the focus of computational epigenomics is increasingly shifting towards integrative analyses. For instance, various predictive approaches model the associations between DNA sequence, epigenomic marks and gene expression: epigenomic profiles could be inferred from DNA sequence [Whitaker *et al.* 2015] or from other epigenomic patterns profiled in similar cell types [Ernst and Kellis 2015], and epigenomic marks were shown to be predictive of gene expression [Karlić *et al.* 2010]. Other prime examples include the identification of chromatin states [Ernst and Kellis 2010; Ernst *et al.* 2011] and the quantification of co-occurrence of chromatin marks and their modifiers by means of partial correlation coefficients [Lasserre *et al.* 2013; Perner *et al.* 2014]. The remodeling potential of epigenomic marks can be assessed through profiles of transcription factor binding [Ziller *et al.* 2015] and TF binding can be predicted from open chromatin data [Schmidt *et al.* 2016].

In the concrete case of DNA methylation and its association with other regulatory factors, the mindset has shifted from a promoter-centric view assuming that DNA methylation levels are strictly anti-correlated with gene expression to a more regulatory perspective in which the binding of transcription factors, potentially at distal sites, can be governed by complex patterns of methylation that act in concert with other epigenetic signals [Schübeler 2015]. Changes in DNA methylation associated with development and disease occur genome-wide, in particular in regions exhibiting intermediate to high overall methylation levels and putative enhancers [Lister *et al.* 2009; Berman *et al.* 2012; Hon *et al.* 2012; Stadler *et al.* 2011]. However, many interesting questions remain: Exactly how does DNA methylation affect transcription factor binding, transcription initiation and elongation? What roles does the interaction with other epigenomic marks play in the definition of enhancers and insulators? How are the epigenetic patterns in turn regulated by signaling and DNA-binding factors? Do epigenetic signatures and associated enhancer activities affect different subpopulations of cells in different ways?

On the genome-scale, regulatory associations are attributed mostly on the basis of correlative analyses and only few exceptions exist (e.g. [Yu *et al.* 2008]). Although bystander patterns can be exploited for an in-depth description of cell state, identifying signatures that are causally associated with a given phenotype represents a major challenge of epigenomic data analysis. Dependencies in epigenetic profiles could potentially be exploited in graphical models like Bayesian networks and structural equation models [Bollen 1989] in order to infer a statistical notion of causality [Pearl 2009]. Nonetheless, true causality is unlikely to be derivable from observational data alone and perturbation experiments are needed. Therefore, the formation of testable hypothesis from genome-wide data is of high importance. Such hypotheses could be derived from the interpretable models described in this work. Experimental tools are rapidly advancing and technologies like high-throughput RNAi screening and CRISPR/Cas9 editing (reviewed in [Sander and Joung 2014]) enable highly specific targeting of regulators as well as the deposition of epigenetic marks. These methods could thus hold the key to interventional validation of model-derived hypotheses. Those validation experiments could in turn provide the basis for refined computational models and new hypotheses.

The resolution of the epigenome atlas increases as more and more datasets become available. Large-scale mapping efforts undertaken by national and international research consortia such as ENCODE, REMC, BLUEPRINT and DEEP are the main contributors to globally characterizing the landscape spanned by the human epigenome. The umbrella IHEC project aims at charting more than 1,000 complete epigenomes by the year 2020 and the generated reference maps cover a plethora of cell phenotypes. Diseases like neurological, inflammatory and metabolic disorders as well as cancer are in the focus of associated research projects and the efforts are complemented by numerous EWAS and global initiatives focusing on malignancy, such as the International Cancer Genome Consortium (ICGC) and TCGA. The sheer data deluge thus produced by these efforts calls for efficient algorithms and computational solutions. These consortia also further the establishment of good standard practices and comparable metrics for quality control as well as the documentation of experimental procedures and employed computational pipelines, which are essential to reproducible research [Ebert *et al.* 2015] and are actively discussed by the scientific community. Providing standardized access to the enormous amounts of generated data is of prime importance for the efficient use of the generated reference maps — an issue that is starting to be addressed by tools like the ENCODE data portal[4] and DeepBlue [Albrecht *et al.* 2016]. Furthermore, novel technologies are established and thus enable the mapping of novel subspaces of the epigenomic landscape. Examples of these emerging methods include charting the three-dimensional structure of the DNA inside a cell's nucleus and characterizing cell-to-cell heterogeneity by mapping the epigenomes of single cells.

In conclusion, these are exciting times in which a deluge of epigenome-wide data becomes available and the regulatory roles of the epigenome and its influence on cell identity are just beginning to be unraveled.

[4] https://www.encodeproject.org

A

Glossaries

List of Abbreviations

27K	Illumina Infinium HumanMethylation27 BeadChip
3C	Chromosome Conformation Capture
450K	Illumina Infinium HumanMethylation450 BeadChip
4C	Circularized Chromosome Conformation Capture
5C	Carbon-Copy Chromosome Conformation Capture
5caC	5-Carboxylcytosine
5fC	5-Formylcytosine
5hmC	5-Hydroxymethylcytosine
5mC	5-Methylcytosine
AML	Acute Myeloid Leukemia
ATAC-seq	Assay for Transposase-Accessible Chromatin using Sequencing
AUC	Area Under the Curve
AZA	5-Azacytidine
bp	Basepair
cDNA	Complementary DNA
ceRNA	Competing Endogenous RNA
CGI	CpG Island
ChIA-PET	Chromatin Interaction Analysis by Paired-End Tag Sequencing
ChIP-seq	Chromatin Immunoprecipitation-Sequencing
chr	Chromosome
CIMP	CpG Island Methylator Phenotype
circRNA	Circular RNA
CLP	Common Lymphoid Progenitor
CML	Chronic Myeloid Leukemia
CMP	Common Myeloid Progenitor
DEEP	Deutsches Epigenom Programm
DMR	Differentially Methylated Region
DNaseI-seq	DNaseI-Sequencing

DNMT	DNA Methyltransferase
EGA	European Genome-phenome Archive
ENCODE	Encyclopedia of DNA Elements
EPIC	Illumina Infinium MethylationEPIC BeadChip
eRNA	Enhancer RNA
ERV	Endogenous Retrovirus
ESC	Embryonic Stem Cell
EWAS	Epigenome-Wide Association Study
FACS	Fluorescence-Activated Cell Sorting
FDA	US Food and Drug Administration
FDR	False Discovery Rate
GEO	Gene Expression Omnibus
GMP	Granulocyte Macrophage Progenitor
GO	Gene Ontology
HAT	Histone Acetyltransferase
HDAC	Histone Deacetylase
HDM	Histone Demethylase
HIV	Human Immunodeficiency Virus
HMM	Hidden Markov Model
HMT	Histone Methyltransferase
HSC	Hematopoietic Stem Cell
IAP	Intracisternal A-particle
ICGC	International Cancer Genome Consortium
IHEC	International Human Epigenome Consortium
iPSC	Induced Pluripotent Stem Cell
LAD	Lamina-Associated Domain
LINE	Long Interspersed Nuclear Element
LiPSC	Leukemia Induced Pluripotent Stem Cell
LMPP	Lymphoid-primed Multipotent Progenitor
LMR	Low-Methylated Region
lncRNA	Long Non-Coding RNA
LOESS	Local Regression
LTR	Long Terminal Repeat
MBD	Methyl-CpG Binding Domain
MDS	Multidimensional Scaling
MeDIP-seq	Methylated DNA Immunoprecipitation Sequencing
MEP	Megakaryocyte Erythrocyte Progenitor
miRNA	Micro RNA
MLP	Immature Lymphoid Progenitor
MNase-seq	MNase-Sequencing
MPP	Multipotent Progenitor

MRE-seq	Methylation-Sensitive Restriction Enzyme Sequencing
mRNA	Messenger RNA
µWGBS	Low-input Whole Genome Bisulfite Sequencing
ncRNA	Non-Coding RNA
NDR	Nucleosome-Depleted Region
NGS	Next Generation Sequencing
NK cell	Natural Killer cell
NOMe-seq	Nucleosome Occupancy and Methylome-Sequencing
nt	Nucleotide
ORF	Open Reading Frame
PAM	Partitioning Around Medoids
PBMC	Peripheral Blood Mononuclear Cell
PCA	Principal Component Analysis
PcG	Polycomb Group
PCR	Polymerase Chain Reaction
piRNA	PIWI-Interacting RNA
PMD	Partially Methylated Domain
QC	Quality Control
REMC	NIH Roadmap Epigenomics Mapping Consortium
RNA-seq	RNA-Sequencing
RNAi	RNA Interference
ROC curve	Receiver Operating Characteristic curve
RRBS	Reduced Representation Bisulfite Sequencing
scWGBS	Single-Cell Whole Genome Bisulfite Sequencing
SINE	Short Interspersed Nuclear Element
siRNA	Small Interfering RNA
snoRNA	Small Nucleolar RNA
SNP	Single-Nucleotide Polymorphism
SNV	Single-Nucleotide Variation
SVA	Surrogate Variable Analysis
SVA element	SINE/Variable number tandem repeat/Alu element
SVM	Support Vector Machine
TAD	Topological Associated Domain
TCGA	The Cancer Genome Atlas
TCM	Central Memory T cell
TE	Transposable Element
TEM	Effector Memory T cell
TEMRA	$CD45RA^+$ Memory T cell
TF	Transcription Factor
TFBS	Transcription Factor Binding Site

TN	Naive T cell
Treg	Regulatory T cell
TrxG	Trithorax Group
TSS	Transcription Start Site
UMR	Unmethylated Region
WCE	Whole-Cell Extract
WGBS	Whole Genome Bisulfite Sequencing

List of Genes, Transcripts, Proteins and Complexes

ABL	Abelson murine leukemia viral oncogene homolog 1
BATF	basic leucine zipper transcription factor, ATF-like
BCR	breakpoint cluster region protein
BRCA1	breast cancer 1, early onset
c-MYC	MYC
CCR5	C-C chemokine receptor type 5
CD4	cluster of differentiation 4
CD8	cluster of differentiation 8
CD10	cluster of differentiation 10
CK2	casein kinase 2
CTCF	CCCTC-binding factor
DEFA4	defensin alpha 4
DNMT1	DNA methyltransferase 1
DNMT3A	DNA methyltransferase 3α
DNMT3B	DNA methyltransferase 3β
DNMT3L	DNA methyltransferase 3-Like
Env	envelope (viral origin)
ESR1	estrogen receptor 1
EXOC6	exocyst complex component 6
EZH2	enhancer of zeste homolog 2
Gag	group-specific antigen (viral origin)
GATA1	GATA binding protein 1
H1	histone H1 (linker histone)
H2A	histone H2A
H2A.Z	histone H2A.Z
H2B	histone H2B
H3	histone H3
H3.3	histone H3.3
H4	histone H4
HOXA5	homeobox A5
HOXB3	homeobox B3
IDH1	isocitrate dehydrogenase 1
IDH2	isocitrate dehydrogenase 2
IRF4	interferon regulatory factor 4
ISWI	ISWI chromatin remodeling complex
KCNH2	potassium voltage-gated channel subfamily H member 2
KDM4A	lysine-specific demethylase 4A
KLF4	Kruppel-like factor 4

MGMT	O(6)-methylguanine-DNA methyltransferase
MHC	major histocompatibility complex
MYB	Myb proto-oncogene
NF-κB	nuclear factor kappa-light-chain-enhancer of activated B cells
OCT4	octamer-binding transcription factor 4
p300	p300 coactivator
p53	tumor protein p53
PIWI	P-element induced wimpy testis
Pol	DNA polymerase (viral origin)
PRC2	Polycomb repressive complex 2
RNAPII	RNA Polymerase II
RUNX3	Runt-related transcription factor 3
SALL4	Spalt-like transcription factor 4
SOX2	SRY-related high mobility group box 2
SOX17	SRY-related high mobility group box 17
SUSD3	sushi domain containing 3
SWI/SNF	Iswitch/sucrose nonfermentable chromatin remodeling complex
TAL1	T-cell acute lymphocytic leukemia protein 1
TCR	T-cell receptor
TDG	thymine-DNA glycosylase
TET	ten-eleven translocation
TET2	ten-eleven translocation 2
TREML1	trem-like transcript 1
XIST	X-inactive specific transcript

Glossary

active DNA demethylation
: Demethylation of cytosines involving the conversion of 5mC to 5hmC and to its derivates 5fC and 5caC (catalyzed by enzymes of the TET family). Subsequently, an unmethylated state is the result of either passive dilution via DNA replication or active restoration by means of base excision involving TDG enzymes and repair [Kohli and Zhang 2013].

adaptive immunity
: Pathogen-specific defense mechanism facilitated by specialized cells of the lymphoid lineage. After an initial exposure to a given pathogen an immune memory is formed that is specific to that pathogen and can trigger a highly accelerated response upon subsequent exposures. Adaptive immunity is exclusive to vertebrates.

batch effect
: Source of variation in biological data due to study design, sample handling or technical artifacts. Examples include the manufacturing batch of chips and reagents, the scientist conducting the experiments, the date on which the experiment was run or imbalanced distribution of biological traits across groups of phenotypes. Computational methods for correcting for batch effects include empirical Bayes methods [Johnson *et al.* 2007], SVA and considering corresponding features in differential analysis [Leek *et al.* 2010].

β-value
: An estimate for the methylation level of a given cytosine. For Illumina methylation arrays it is defined as $\beta = \frac{\max(M,0)}{\max(M,0)+\max(U,0)+\epsilon}$, where M and U correspond to the methylated and unmethylated intensity signal respectively and ϵ is a constant (typically set to 100).

biomarker
: Measurable (molecular) indicator for a given (disease) condition or state.

bivalent (chromatin) domain
: Region of chromatin marked by methylation of both lysine 4 and lysine 27 on histone H3 [Bernstein *et al.* 2006].

cellular reprogramming
: Inducing a pluripotent cell state by nuclear transfer, cell fusion or the introduction of exogenous transcription factors [Yamanaka and Blau 2010].

chromatin
: Macromolecule complex consisting of DNA (and RNA) and packaging proteins (histones).

CpG
: DNA dinucleotide consisting of cytosine followed by guanine, linked by phosphate.

CpG island
: A genomic region particularly rich in CpG dinucleotides. Multiple alternative definitions and algorithms for defining them exist. The most widely-used definition was introduced by Gardiner-Garden and Frommer [1987].

cytokine
Peptides or small proteins involved in cell signaling. Among other functions, they direct the immune response and provide differentiation queues to cells of the hematopoietic system.

DNaseI hypersensitive site
Accessible genomic region frequently cleaved by DNaseI nucleases. Typically indicates open chromatin and low nucleosome occupancy.

enhancer
Region containing TFBSs located distal or proximal to gene promoters involved in transcriptional activation. Active enhancers are typically characterized by the presence of H3K4me1, H3K4me2, H3K27ac, DNaseI hypersensitivity and p300 occupancy [Ong and V. G. Corces 2011].

epigenetic/epigenomic mark
Umbrella term for molecular manifestations of epigenetic regulation, such as DNA methylation, histone modifications, etc..

epigenome
The collective of all "potentially stable and, ideally, heritable changes in gene expression or cellular phenotype that occurs without changes in Watson-Crick base-pairing of DNA" [Goldberg *et al.* 2007].

euchromatin
Loosly packed chromatin structure.

exaptation
Shift of function of a genomic unit during the course of evolution.

genomic imprinting
Epigenetically regulated, parent-of-origin dependent expression of genes.

hematopoiesis
Formation of blood cell types from progenitor cells.

heterochromatin
Tightly packed chromatin structure.

hypermethylated
Higher DNA methylation level in a given condition compared to a reference condition.

hypomethylated
Lower DNA methylation level in a given condition compared to a reference condition.

Induced Pluripotent Stem Cell (iPSC)
Cells reprogrammed to a pluripotent state by means of exogenous transcription factors.

innate immunity
Non-specific immune response representing an organisms first line of defense against pathogens. Involved mechanisms include biological barriers and specialized cell types of the myeloid lineage which can disable pathogens using toxins and phagocytosis. The innate immune response is typically fast and can also recruit the adaptive immune response.

insulator
(Distal) gene regulatory element representing a barrier for the spreading of of signals to other gene regions. Generally marked by CTCF.

Long Non-Coding RNA (lncRNA)
RNA longer than 200 nucleotides, not coding for a protein. lncRNAs are often polyadenylated and can occur in the nucleus as well as in the cytoplasm.

lymphoid
Arising from progenitor cells of the lymphoid lineage of the hematopoietic system. Lymphoid cell types include natural killer cells, B cells and T cells.

M-value
An estimate for the methylation level of a given cytosine. For Illumina methylation arrays it is defined as $M = \log_2 \frac{\max(M,0)+\alpha}{\max(U,0)+\alpha}$, where M and U correspond to the methylated and unmethylated intensity signal respectively and α is a constant (typically set to 1). It corresponds to the logit transformation of the β-value or methylation level: $M = \log_2 \frac{\beta}{1-\beta}$

methylation level
For bisulfite sequencing based methods, this is the proportion of methylated read-cytosines among all reads covering a particular reference cytosine. For methylation arrays it corresponds to the β-value.

methylome
The collective of all DNA methylation events in a given entity (such as a single cell, cell type or organism).

multi-mapped reads
Reads originating from high-throughput sequences that align to multiple positions in the reference genome.

myeloid
Arising from progenitor cells of the myeloid lineage of the hematopoietic system. Myeloid cell types include monocytes, macrophages and granulocytes.

Next Generation Sequencing (NGS)
Umbrella term for technologies capable of high-throughput sequencing of millions of short DNA reads. The most commonly employed NGS technology is Illumina sequencing using the contemporary *HiSeq* machine model or the older *Genome Analyzer*.

nucleosome
Histone octamer comprising the H3, H4, H2A and H2B units with 147 bp of DNA wrapped around it.

passive DNA demethylation
Demethylation of cytosines involving dilution via DNA replication.

phagocytosis
Process in which a cell ingests a particle and potentially decomposes it.

Philadelphia chromosome
Chromosomal abnormality associated with the translocation of the q34.1 region of chromosome 9 to the q11.2 region of chromosome 22 (termed t(9;22)(q34;q11.2)) which gives rise to the *BCR-ABL* fusion gene in CML.

promoter
Genomic region located near the TSS to which the transcriptional machinery is recruited. Typically defined to extend from a few kilobases upstream of the TSS to a few hundred bases downstream of the TSS.

RNA Interference (RNAi)
Post-transcriptional or transcriptional reduction in mRNA levels mediated by small, non-coding RNAs.

silencer
Distal gene regulatory element associated with transcriptional silencing.

transcription factory
Nuclear subcompartment containing multiple promoters and enhancers, associated with open chromatin and high transcriptional activity.

transdifferentiation
Natural or artificial conversion of one cell type to another without going through a pluripotent cell state.

transposon/Transposable Element (TE)
A stretch of DNA capable of changing its position in the genome by transposition. Transposons can be further classified by their structure and mode of transposition.

B

Supplementary Material

Methods for Identifying Differential DNA Methylation

Table B.1 contains a curated list of publications and methods for the detection of differential DNA methylation between (groups of) samples. This list is based on the manual identification of relevant publications from the titles and abstracts of weekly PubMed search queries for the terms “DNA” and “methylation” or for the combined term “DNA methylation” between February 2014 and November 2016. Additional references were added based on customized queries for differential analysis methods.

Table B.1: Methods for identifying differential DNA methylation

#	Author, year	PubMed ID	Supports 450K	Supports BS-seq	DM sites	DM regions	Region detection	Covariate adjustment	Software	Methods
1	Akalin *et al.*, 2012	23034086	✓	✓	✓	✓	P	✓	methylKit	Logistic regression or Fisher's Exact Test
2	Akman *et al.*, 2014	24618468	(X)	✓	X	✓	H	X	BEAT	Bayesian, beta-binomial mixture models; pooling of successive CpGs into regions
3	Almeida *et al.*, 2016	27794558	✓	X	✓	✓	H	(✓)	DiMmeR	t-test for binary outcomes, regression models for continuous outcomes; empirical p-value computation; adjustment for cell composition using minfi
4	Aryee *et al.*, 2014	24478339	✓	X	✓	✓	✓	(X)	minfi	Linear models, F-test, bump hunting; aggregation of DMRs by block finding
5	Assenov *et al.*, 2014	25262207	✓	✓	✓	✓	P	✓	RnBeads	Combined ranking of absolute and relative methylation difference and limma p-values
6	Bacalini *et al.*, 2015	25701668	✓	(X)	✓	✓	X	✓	-	MANOVA; mixture of single probes and groups of probes
7	Barfield *et al.*, 2012	22451269	✓	X	✓	X	X	✓	CpGassoc	ANOVA
8	Baumann *et al.*, 2014	24589664	(X)	✓	✓	✓	P	X	-	Fisher's Exact Test (single sample) or logistic regression (replicates)
9	Butcher *et al.*, 2014	25461817	✓	(X)	✓	✓	✓	(X)	ChAMP	limma, probe lasso
10	Chen *et al.*, 2012	22368244	✓	(✓)	✓	X	X	(✓)	-	One-sided t-test; combine p-values from different age groups; adjustment for age
11	Chen *et al.*, 2013	23452721	✓	X	X	✓	H	✓	methylPCA	Association tests based on linear models using PCA features; region detection based on covariance
12	Chen *et al.*, 2013	23369576	✓	(✓)	✓	X	X	(✓)	-	Kruskal-Wallis test on different age groups. Then combine p-values; adjustment for age
13	Chen *et al.*, 2014	24884464	✓	(✓)	✓	X	X	(✓)	-	Cuzik test for obtaining a combined p-value for ordinal data; combine p-values from different age groups; adjustment for age
14	Dolzhenko *et al.*, 2014	24962134	X	✓	✓	✓	✓	✓	RADMeth	Beta-binomial regression; sliding combination of p-values for defining DMRs; Z-test for region detection
15	Feng *et al.*, 2014	24561809	X	✓	✓	✓	H	X	DSS	Bayesian hierarchical model and Wald test for modeling read counts in conditions
16	Gao *et al.*, 2015	26140213	(✓)	✓	X	✓	X	X	SMAP	Pearson's chi-square test, t-test
17	Hansen *et al.*, 2012	23034175	(X)	✓	✓	✓	✓	X	Bsmooth	Smoothing, t-test
18	Hebestreit *et al.*, 2013	23658421	(✓)	✓	✓	✓	H	✓	BiSeq	Smoothing, beta regression, hierarchical testing

#	Author, year	PubMed ID	Supports 450K	Supports BS-seq	DM sites	DM regions	Region detection	Covariate adjustment	Software	Methods
19	Houseman *et al.*, 2012	22568884	✓	(X)	-	-	-	-	-	Identification of cell population heterogeneity by surrogate variable analysis, given reference cell types
20	Houseman *et al.*, 2014	24451622	✓	(✓)	✓	X	X	✓	RefFreeEWAS	Reference-free identification of cell population heterogeneity using singular value decomposition
21	Huang *et al.*, 2013	23497201	✓	(✓)	✓	X	X	(✓)	-	Baumgartner-Weiß-Schindler (BWS) on different age groups; subsequently combine p-values; adjustment for age
22	Jaffe *et al.*, 2012	22422453	✓	(✓)	✓	✓	H	✓	minfi	Regression in genomic regions using surrogate variable analysis for batch effect removal (linear); LOESS smoothing; threshold-based identification of bumps representing contiguous significant regions; permutation tests and Bayesian model
23	Jühling *et al.*, 2016	26631489	✓	✓	X	✓	✓	X	metilene	Segmentation; non-parametric 2D Kolmogorov-Smirnoff Test
24	Kilaru *et al.*, 2012	22430798	✓	X	✓	X	X	✓	MethLAB	Linear mixed effect model
25	Lea *et al.*, 2015	26599596	X	✓	✓	X	X	✓	MACAU	Binomial mixed models inferred by MCMC sampling; takes into account population structure
26	Li *et al.*, 2013	23735126	(X)	✓	X	✓	H	X	eDMR	Gaussian mixture model; distance-based region identification
27	Liang *et al.*, 2014	24497972	(X)	✓	X	✓	H	X	WBSA	detecting differential methylation between single samples; Wilcoxon test
28	Morris *et al.*, 2013	24336642	✓	(X)	✓	✓	✓	(X)	ChAMP	limma, probe lasso
29	Park *et al.*, 2014	24836530	X	✓	✓	✓	X	X	methylSig	Beta-binomial model; approximation algorithm to estimate parameters; likelihood ratio test to test for difference in means; can incorporate locality via a kernel to adapt for proximity of CpGs
30	Park *et al.*, 2016	26819470	X	✓	✓	X	X	X	DSS	Beta-binomial model with arcsine link function; Wald test to obtain p-values
31	Peters *et al.*, 2015	25972926	✓	(X)	✓	✓	✓	✓	DMRcate	Smoothing/kernel density modeling; DMR identification by kernel smoothing
32	Preussner *et al.*, 2015	26628921	✓	X	✓	✓	P	X	ADMIRE	Mann-Whitney U test, p-value joining using comb-p
33	Raineri *et al.*, 2014	24824426	X	✓	✓	X	X	X	diff_methyl	detecting differential methylation between single samples; Beta distribution model
34	Rijlaarsdam *et al.*, 2014	26889969	✓	(✓)	X	✓	H	X	DMRforPairs	detecting differential methylation between single samples; thresholding; Wilcoxon and Kruskal-Wallis test
35	Saito *et al.*, 2014	24423865	(X)	✓	✓	✓	✓	X	Bisulfighter	detecting differential methylation between single samples; HMM

#	Author, year	PubMed ID	Supports 450K	Supports BS-seq	DM sites	DM regions	Region detection	Covariate adjustment	Software	Methods
36	Smyth *et al.*, 2004	16646809	(✓)	(✓)	✓	(✓)	X	✓	limma	Hierarchical linear models; empirical Bayes
37	Sofer *et al.*, 2013	23990415	✓	(✓)	X	✓	✓	✓	Aclust	Clustering of adjacent CpGs (Aclust); Generalized Estimating Equation (GEE)
38	Song *et al.*, 2013	24324667	(✓)	✓	X	✓	✓	X	MethPipe	detecting differential methylation between single samples; HMM for detecting low methylation regions
39	Stockwell *et al.*, 2014	24608764	(X)	✓	X	✓	P	X	DMAP	Fisher's Exact test or ANOVA on regions; DMRs based on windows/RRBS-fragments
40	Su *et al.*, 2012	22941633	X	✓	✓	✓	H	X	CpG_MPs	Thresholds on methylation levels or Fishers Exact test; DMR identification based on threshold extension or sliding windows
41	Sun *et al.*, 2014	24565500	(✓)	✓	✓	✓	✓	X	MOABS	Hierarchical model (beta binomial) fit by emprirical Bayes; confidence interval on difference statistic; DMR identification using HMM
42	Sun *et al.*, 2016	26854292	(✓)	✓	✓	✓	H	X	HMM-Fisher	Segmentation by HMM followed by Fisher's Exact Test
43	Teng *et al.*, 2012	22956892	X	X	✓	X	X	X	-	Empiral Bayes using Gamma distribution assumption
44	Valavanis *et al.*	24808224	✓	X	✓	X	X	X	-	Control sample based filtering of t-test identified DMPs
45	Wahl *et al.*, 2014	24994026	✓	(X)	✓	X	X	✓	gamlss,mgcv	Generalised Additive Models for Location Scale and Shape (GAMLSS)
46	Wang *et al.*, 2011	21818777	✓	(✓)	✓	X	X	X	-	Mixture model of uniform and truncated normal distribution
47	Wang *et al.*, 2012	22253290	✓	(X)	✓	✓	X	✓	IMA	Wilcoxon rank-sum test, t-test, limma
48	Wang *et al.*, 2015	26176536	X	✓	X	✓	H	X	swDMR	Hypothesis tests and thresholding on sliding windows (t-test,Wilcoxon, Fisher, ANOVA, Kruskal-Wallis, Chi-Square) and window merging
49	Warden *et al.*, 2013	23598999	✓	✓	X	✓	X	X	COHCAP	ANOVA, t-test, Fisher's Exact Test
50	Wessely *et al.*, 2012	22936948	✓	(X)	✓	X	X	X	NIMBL	Differential methylation scores based single-linkage differences and median difference; can be combined with statistical tests
51	Wu *et al.*, 2013	24040221	✓	X	✓	✓	H	(X)	FastDMA	ANCOVA
52	Wu *et al.*, 2015	26184873	X	✓	✓	✓	H	X	DSS	detecting differential methylation between single samples; extend the DSS software to estimate within-group variance by spatially close CpGs; Smoothing, empirical Bayes
53	Xu *et al.*, 2013	23554163	X	✓	✓	X	X	X	-	Test statistic taking group variances into acount; binomial methylation model

#	Author, year	PubMed ID	Supports 450K	Supports BS-seq	DM sites	DM regions	Region detection	Covariate adjustment	Software	Methods
54	Yu *et al.*, 2016	26887041	(X)	✓	X	✓	✓	X	HMM-DM	DMR detection by HMM; software available as script
55	Zhang *et al.*, 2013	24283878	✓	(✓)	✓	X	X	X	-	Kullback-Leibler Divergence/Entropy; only applicable to matched samples
56	Zhang *et al.*, 2015	25865601	✓	(✓)	✓	✓	H	X	-	Distance discriminant analysis (DDA), integration with expression data; DMR detection by windowing
57	Zou *et al.*, 2014	24464286	✓	(✓)	✓	(X)	X	✓	FaST-LMM-EWASher	Integration of Linear Mixed Models and PCA to estimate cell composition

✓: supported, X: not supported, (✓): not explicitely supported, but potentially extensible, (X): not supported by design, requires major modifications to be supported, H: region detection by simple heuristics, P: based on predefined genomic regions

C

List of Publications

Joint-first Author Publications

Amabile*, G., Di Ruscio*, A., **Müller*, F.**, Welner, R. S., Yang, H., Ebralidze, A. K., Zhang, H., Levantini, E., Qi, L., Martinelli, G., Brummelkamp, T., Le Beau, M. M., Figueroa, M. E., Bock, C., and Tenen, D. G. (2015). "Dissecting the role of aberrant DNA methylation in human leukaemia." *Nature Communications* 6.7091. DOI: 10.1038/ncomms8091.

Assenov*, Y., **Müller*, F.**, Lutsik*, P., Walter, J., Lengauer, T., and Bock, C. (2014). "Comprehensive analysis of DNA methylation data with RnBeads." *Nature Methods* 11.11, pp. 1138–1140. DOI: 10.1038/nmeth.3115.

Farlik, M., Halbritter, F., **Müller*, F.**, Choudry, F. A., Ebert, P., Klughammer, J., Farrow, S., Santoro, A., Ciaurro, V., Mathur, A., Uppal, R., Stunnenberg, H. G., Ouwehand, W. H., Laurenti, E., Lengauer, T., Frontini, M., and Bock, C. (2016). "DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation." *Cell Stem Cell* 19.6, pp. 808–822. DOI: 10.1016/j.stem.2016.10.019.

Ziller*, M. J., **Müller*, F.**, Liao, J., Zhang, Y., Gu, H., Bock, C., Boyle, P., Epstein, C. B., Bernstein, B. E., Lengauer, T., Gnirke, A., and Meissner, A. (2011). "Genomic Distribution and Inter-Sample Variation of Non-CpG Methylation across Human Cell Types." *PLoS Genetics* 7.12, e1002389. DOI: 10.1371/journal.pgen.1002389.

Contributing Author Publications

Bock, C., Tomazou, E. M., Brinkman, A. B., **Müller, F.**, Simmer, F., Gu, H., Jäger, N., Gnirke, A., Stunnenberg, H. G., and Meissner, A. (2010). "Quantitative comparison of genome-wide DNA methylation mapping technologies." *Nature Biotechnology* 28.10, pp. 1106–1114. DOI: 10.1038/nbt.1681.

Deplus, R., Blanchon, L., Rajavelu, A., Boukaba, A., Defrance, M., Luciani, J., Rothé, F., Dedeurwaerder, S., Denis, H., Brinkman, A. B., Simmer, F., **Müller, F.**, Bertin, B., Berdasco, M., Putmans, P., Calonne, E., Litchfield, D. W., Launoit, Y. de, Jurkowski, T. P., Stunnenberg, H. G., Bock, C., Sotiriou, C., Fraga, M. F., Esteller, M., Jeltsch, A., and Fuks, F. (2014). "Regulation of DNA methylation patterns by CK2-mediated phosphorylation of Dnmt3a." *Cell Reports* 8.3, pp. 743–753. DOI: 10.1016/j.celrep.2014.06.048.

Durek, P. *et al.* (2016). "Epigenomic Profiling of Human CD4+ T Cells Supports a Linear Differentiation Model and Highlights Molecular Regulators of Memory Development." *Immunity* 45.5, pp. 1148–1161. DOI: 10.1016/j.immuni.2016.10.022.

Ebert, P., **Müller, F.**, Nordström, K., Lengauer, T., and Schulz, M. H. (2015). "A general concept for consistent documentation of computational analyses." *Database* 2015. DOI: 10.1093/database/bav050.

Feuerbach, L., Halachev, K., Assenov, Y., **Müller, F.**, Bock, C., and Lengauer, T. (2012). "Analyzing epigenome data in context of genome evolution and human diseases." *Evolutionary Genomics : Statistical and Computational Methods, Volume 2*. Ed. by M. Anisimova. Vol. 856. Methods in Molecular Biology. New York, NY: Springer. Chap. 18, pp. 431–467. ISBN: 978-1-61779-584-8. DOI: 10.1007/978-1-61779-585-5_18.

Martínez-Cardús, A., Moran, S., Musulen, E., Moutinho, C., Manzano, J. L., Martinez-Balibrea, E., Tierno, M., Élez, E., Landolfi, S., Lorden, P., Arribas, C., **Müller, F.**, Bock, C., Tabernero, J., and Esteller, M. (2016). "Epigenetic Homogeneity Within Colorectal Tumors Predicts Shorter Relapse-free and Overall Survival Times for Patients With Locoregional Cancer." *Gastroenterology*. DOI: 10.1053/j.gastro.2016.08.001.

Planello, A. C., Ji, J., Sharma, V., Singhania, R., Mbabaali, F., **Müller, F.**, Alfaro, J. A., Bock, C., De Carvalho, D. D., and Batada, N. N. (2014). "Aberrant DNA methylation reprogramming during induced pluripotent stem cell generation is dependent on the choice of reprogramming factors." *Cell Regeneration* 3.4. DOI: 10.1186/2045-9769-3-4.

Sandoval, J. *et al.* (2013). "A prognostic DNA methylation signature for stage I non-small-cell lung cancer." *Journal of Clinical Oncology* 31.32, pp. 4140–4147. DOI: 10.1200/JCO.2012.48.5516.

Schneider, E., El Hajj, N., **Müller, F.**, Navarro, B., and Haaf, T. (2015). "Epigenetic Dysregulation in the Prefrontal Cortex of Suicide Completers." *Cytogenetic and Genome Research* 146.1, pp. 19–27. DOI: 10.1159/000435778.

Tobi, E. W., Goeman, J. J., Monajemi, R., Gu, H., Putter, H., Zhang, Y., Slieker, R. C., Stok, A. P., Thijssen, P. E., **Müller, F.**, Zwet, E. W. van, Bock, C., Meissner, A., Lumey, L. H., Eline Slagboom, P., and Heijmans, B. T. (2014). "DNA methylation signatures link prenatal famine exposure to growth and metabolism." *Nature Communications* 5, p. 5592. DOI: 10.1038/ncomms6592.

Wallner, S., Schröder, C., Leitão, E., Berulava, T., Haak, C., Beisser, D., Rahmann, S., Richter, A. S., Manke, T., Bönisch, U., Arrigoni, L., Fröhler, S., Klironomos, F., Chen, W., Rajewsky, N., **Müller, F.**, Ebert, P., Lengauer, T., Barann, M., Rosenstiel, P., Gasparoni, G., Nordström, K., Walter, J., Brors, B., Zipprich, G., Felder, B., Klein-Hitpass, L., Attenberger, C., Schmitz, G., and Horsthemke, B. (2016). "Epigenetic dynamics of monocyte-to-macrophage differentiation." *Epigenetics & Chromatin* 9.1, p. 33. DOI: 10.1186/s13072-016-0079-z.

Xi, Y., Bock, C., **Müller, F.**, Sun, D., Meissner, A., and Li, W. (2012). "RRBSMAP: a fast, accurate and user-friendly alignment tool for reduced representation bisulfite sequencing." *Bioinformatics (Oxford, England)* 28.3, pp. 430–432. DOI: 10.1093/bioinformatics/btr668.

Ziller, M. J., Gu, H., **Müller, F.**, Donaghey, J., Tsai, L. T. Y., Kohlbacher, O., De Jager, P. L., Rosen, E. D., Bennett, D. A., Bernstein, B. E., Gnirke, A., and Meissner, A. (2013). "Charting a dynamic DNA methylation landscape of the human genome." *Nature* 500.7463, pp. 477–481. DOI: 10.1038/nature12433.

References

Adams, D. *et al.* (2012). "BLUEPRINT to decode the epigenetic signature written in blood." *Nature Biotechnology* 30.3, pp. 224–226. DOI: 10.1038/nbt.2153.

Ahmed, R., Bevan, M. J., Reiner, S. L., and Fearon, D. T. (2009). "The precursors of memory: models and controversies." *Nature Reviews Immunology* 9.9, pp. 662–668. DOI: 10.1038/nri2619.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2008). *Molecular Biology of the Cell*. 5th Edition. Garland Science. ISBN: 9780815341062.

Albrecht, F., List, M., Bock, C., and Lengauer, T. (2016). "DeepBlue epigenomic data server: programmatic data retrieval and analysis of epigenome region sets." *Nucleic Acids Research*. DOI: 10.1093/nar/gkw211.

Allis, C. D., Jenuwein, T., and Reinberg, D., eds. (2007). *Epigenetics*. Cold Spring Harbour Laboratory Press, Cold Spring Harbour, New York. ISBN: 9780879697242.

Álvarez-Errico, D., Vento-Tormo, R., Sieweke, M., and Ballestar, E. (2015). "Epigenetic control of myeloid cell differentiation, identity and function." *Nature Reviews Immunology* 15.1, pp. 7–17. DOI: 10.1038/nri3777.

Amabile, G., Di Ruscio, A., Müller, F., Welner, R. S., Yang, H., Ebralidze, A. K., Zhang, H., Levantini, E., Qi, L., Martinelli, G., Brummelkamp, T., Le Beau, M. M., Figueroa, M. E., Bock, C., and Tenen, D. G. (2015). "Dissecting the role of aberrant DNA methylation in human leukaemia." *Nature Communications* 6.7091. DOI: 10.1038/ncomms8091.

Aran, D., Toperoff, G., Rosenberg, M., and Hellman, A. (2011). "Replication timing-related and gene body-specific methylation of active human genes." *Human Molecular Genetics* 20.4, pp. 670–680. DOI: 10.1093/hmg/ddq513.

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., and Irizarry, R. A. (2014). "Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays." *Bioinformatics (Oxford, England)* 30.10, pp. 1363–1369. DOI: 10.1093/bioinformatics/btu049.

Assenov, Y., Müller, F., Lutsik, P., Walter, J., Lengauer, T., and Bock, C. (2014). "Comprehensive analysis of DNA methylation data with RnBeads." *Nature Methods* 11.11, pp. 1138–1140. DOI: 10.1038/nmeth.3115.

Banerji, C. R. S., Miranda-Saavedra, D., Severini, S., Widschwendter, M., Enver, T., Zhou, J. X., and Teschendorff, A. E. (2013). "Cellular network entropy as the energy potential in Waddington's differentiation landscape." *Scientific Reports* 3, p. 3039. DOI: 10.1038/srep03039.

Bao, W., Kojima, K. K., and Kohany, O. (2015). "Repbase Update, a database of repetitive elements in eukaryotic genomes." *Mobile DNA* 6, p. 11. DOI: 10.1186/s13100-015-0041-9.

Baylin, S. B. and Jones, P. A. (2011). "A decade of exploring the cancer epigenome - biological and translational implications." *Nature Reviews Cancer* 11.10, pp. 726–734. DOI: 10.1038/nrc3130.

Beck, C. R., Garcia-Perez, J. L., Badge, R. M., and Moran, J. V. (2011). "LINE-1 Elements in Structural Variation and Disease." *Annual Review of Genomics and Human Genetics* 12, pp. 187–215. DOI: 10.1146/annurev-genom-082509-141802.

Bell, O., Tiwari, V. K., Thomä, N. H., and Schübeler, D. (2011). "Determinants and dynamics of genome accessibility." *Nature Reviews Genetics* 12.8, pp. 554–564. DOI: 10.1038/nrg3017.

Benjamini, Y. and Yekutieli, D. (2001). "The Control of the False Discovery Rate in Multiple Testing under Dependency." *Annals of Statistics* 29, pp. 1165–1188.

Benjamini, Y. and Hochberg, Y. (1995). "Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society Series B* 57, pp. 289–300.

Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., Jaenisch, R., Wagschal, A., Feil, R., Schreiber, S. L., and Lander, E. S. (2006). "A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells." *Cell* 125.2, pp. 315–326. DOI: 10.1016/j.cell.2006.02.041.

Berman, B. P., Weisenberger, D. J., Aman, J. F., Hinoue, T., Ramjan, Z., Liu, Y., Noushmehr, H., Lange, C. P. E., Dijk, C. M. van, Tollenaar, R. A. E. M., Berg, D. van den, and Laird, P. W. (2012). "Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains." *Nature Genetics* 44.1, pp. 40–46. DOI: 10.1038/ng.969.

Bergman, Y. and Cedar, H. (2013). "DNA methylation dynamics in health and disease." *Nature Structural & Molecular Biology* 20.3, pp. 274–281. DOI: 10.1038/nsmb.2518.

Bestor, T. H. (2000). "The DNA methyltransferases of mammals." *Human Molecular Genetics* 9.16, pp. 2395–2402.

Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., Vitale, L., Pelleri, M. C., Tassani, S., Piva, F., Perez-Amodio, S., Strippoli, P., and Canaider, S. (2013). "An estimation of the number of cells in the human body." *Annals of Human Biology* 40.6, pp. 463–471. DOI: 10.3109/03014460.2013.807878.

Bibikova, M., Le, J., Barnes, B., Saedinia-Melnyk, S., Zhou, L., Shen, R., and Gunderson, K. L. (2009). "Genome-wide DNA methylation profiling using Infinium® assay." *Epigenomics* 1.1, pp. 177–200. DOI: 10.2217/epi.09.14.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., Delano, D., Zhang, L., Schroth, G. P., Gunderson, K. L., Fan, J.-B., and Shen, R. (2011). "High density DNA methylation array with single CpG site resolution." *Genomics* 98.4, pp. 288–295. DOI: 10.1016/j.ygeno.2011.07.007.

Bird, A. (2002). "DNA methylation patterns and epigenetic memory." *Genes & Development* 16.1, pp. 6–21. DOI: 10.1101/gad.947102.

Bird, A. (1986). "CpG-rich islands and the function of DNA methylation." *Nature* 321.6067, pp. 209–213. DOI: 10.1038/321209a0.

Bock, C., Reither, S., Mikeska, T., Paulsen, M., Walter, J., and Lengauer, T. (2005). "BiQ Analyzer: visualization and quality control for DNA methylation data from bisulfite sequencing." *Bioinformatics* 21.21, pp. 4067–4068. DOI: 10.1093/bioinformatics/bti652.

Bock, C., Paulsen, M., Tierling, S., Mikeska, T., Lengauer, T., and Walter, J. (2006). "CpG Island Methylation in Human Lymphocytes is Highly Correlated with DNA sequence, Repeats, and Predicted DNA Structure." *PLoS Genetics* 2.3, e26. DOI: 10.1371/journal.pgen.0020026.

Bock, C., Halachev, K., Büch, J., and Lengauer, T. (2009). "EpiGRAPH: user-friendly software for statistical analysis and prediction of (epi)genomic data." *Genome Biology* 10.2, R14. DOI: 10.1186/gb-2009-10-2-r14.

Bock, C., Tomazou, E. M., Brinkman, A. B., Müller, F., Simmer, F., Gu, H., Jäger, N., Gnirke, A., Stunnenberg, H. G., and Meissner, A. (2010). "Quantitative comparison of genome-wide DNA methylation mapping technologies." *Nature Biotechnology* 28.10, pp. 1106–1114. DOI: 10.1038/nbt.1681.

Bock, C., Kiskinis, E., Verstappen, G., Gu, H., Boulting, G., Smith, Z. D., Ziller, M., Croft, G. F., Amoroso, M. W., Oakley, D. H., Gnirke, A., Eggan, K., and Meissner, A. (2011). "Reference Maps of Human ES and iPS Cell Variation Enable High-Throughput Characterization of Pluripotent Cell Lines." *Cell* 144.3, pp. 439–452. DOI: 10.1016/j.cell.2010.12.032.

Bock, C., Beerman, I., Lien, W.-H., Smith, Z. D., Gu, H., Boyle, P., Gnirke, A., Fuchs, E., Rossi, D. J., and Meissner, A. (2012). "DNA Methylation Dynamics during In Vivo Differentiation of Blood and Skin Stem Cells." *Molecular Cell* 47.4, pp. 633–647. DOI: 10.1016/j.molcel.2012.06.019.

Bock, C. and Lengauer, T. (2012). "Managing drug resistance in cancer: lessons from HIV therapy." *Nature Reviews Cancer* 12.7, pp. 494–501. DOI: 10.1038/nrc3297.

Bock, C., Farlik, M., and Sheffield, N. C. (2016). "Multi-Omics of Single Cells: Strategies and Applications." *Trends in biotechnology* 34.8, pp. 605–608. DOI: 10.1016/j.tibtech.2016.04.004.

Bock, C., Halbritter, F., *et al.* (2016). "Quantitative comparison of DNA methylation assays for biomarker development and clinical applications." *Nature Biotechnology* 34.7, pp. 726–737. DOI: 10.1038/nbt.3605.

Bock, C. (2012). "Analysing and interpreting DNA methylation data." *Nature Reviews Genetics* 13.10, pp. 705–719. DOI: 10.1038/nrg3273.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data." *Bioinformatics (Oxford, England)* 30.15, pp. 2114–2120. DOI: 10.1093/bioinformatics/btu170.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. Wiley-Interscience. ISBN: 9780471011712.

Booth, M. J., Branco, M. R., Ficz, G., Oxley, D., Krueger, F., Reik, W., and Balasubramanian, S. (2012). "Quantitative Sequencing of 5-Methylcytosine and 5-Hydroxymethylcytosine at Single-Base Resolution." *Science (New York, NY)* 336.6083, pp. 934–937. DOI: 10.1126/science.1220671.

Bourque, G., Leong, B., Vega, V. B., Chen, X., Lee, Y. L., Srinivasan, K. G., Chew, J.-L., Ruan, Y., Wei, C.-L., Ng, H.-H., and Liu, E. T. (2008). "Evolution of the mammalian transcription factor binding repertoire via transposable elements." *Genome Research* 18.11, pp. 1752–1762. DOI: 10.1101/gr.080663.108.

Bourque, G. (2009). "Transposable elements in gene regulation and in the evolution of vertebrate genomes." *Current Opinion in Genetics & Development* 19.6, pp. 607–612. DOI: 10.1016/j.gde.2009.10.013.

Breiman, L. (2001). "Random Forests." *Machine Learning* 45.1, pp. 5–32. DOI: 10.1023/A:1010933404324.

Bröske, A.-M., Vockentanz, L., Kharazi, S., Huska, M. R., Mancini, E., Scheller, M., Kuhl, C., Enns, A., Prinz, M., Jaenisch, R., Nerlov, C., Leutz, A., Andrade-Navarro, M. A., Jacobsen, S. E. W., and Rosenbauer, F. (2009). "DNA methylation protects hematopoietic stem cell multipotency from myeloerythroid restriction." *Nature Genetics* 41.11, pp. 1207–1215. DOI: 10.1038/ng.463.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position." *Nature Methods* 10.12, pp. 1213–1218. DOI: 10.1038/nmeth.2688.

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., and Greenleaf, W. J. (2015). "Single-cell chromatin accessibility reveals principles of regulatory variation." *Nature* 523.7561, pp. 486–490. DOI: 10.1038/nature14590.

Burns, K. H. and Boeke, J. D. (2012). "Human Transposon Tectonics." *Cell* 149.4, pp. 740–752. DOI: 10.1016/j.cell.2012.04.019.

Burger, L., Gaidatzis, D., Schübeler, D., and Stadler, M. B. (2013). "Identification of active regulatory regions from DNA methylation data." *Nucleic Acids Research* 41.16, e155–e155. DOI: 10.1093/nar/gkt599.

Cabezas-Wallscheid, N., Klimmeck, D., Hansson, J., Lipka, D. B., Reyes, A., Wang, Q., Weichenhan, D., Lier, A., Paleske, L. von, Renders, S., Wünsche, P., Zeisberger, P., Brocks, D., Gu, L., Herrmann, C., Haas, S., Essers, M. A. G., Brors, B., Eils, R., Huber, W., Milsom, M. D., Plass, C., Krijgsveld, J., and Trumpp, A. (2014). "Identification of Regulatory Networks in HSCs and Their Immediate Progeny via Integrated Proteome, Transcriptome, and DNA Methylome Analysis." *Cell Stem Cell* 15.4, pp. 507–522. DOI: 10.1016/j.stem.2014.07.005.

Carette, J. E., Pruszak, J., Varadarajan, M., Blomen, V. A., Gokhale, S., Camargo, F. D., Wernig, M., Jaenisch, R., and Brummelkamp, T. R. (2010). "Generation of iPSCs from cultured human malignant cells." *Blood* 115.20, pp. 4039–4042. DOI: 10.1182/blood-2009-07-231845.

Celniker, S. E., Dillon, L. A. L., Gerstein, M. B., Gunsalus, K. C., Henikoff, S., Karpen, G. H., Kellis, M., Lai, E. C., Lieb, J. D., MacAlpine, D. M., Micklem, G., Piano, F., Snyder, M., Stein, L., White, K. P., Waterston, R. H., and modENCODE Consortium (2009). "Unlocking the secrets of the genome." *Nature* 459.7249, pp. 927–930. DOI: 10.1038/459927a.

Chavez, L., Jozefczuk, J., Grimm, C., Dietrich, J., Timmermann, B., Lehrach, H., Herwig, R., and Adjaye, J. (2010). "Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage." *Genome Research* 20.10, pp. 1441–1450. DOI: 10.1101/gr.110114.110.

Chen, P.-Y., Cokus, S. J., and Pellegrini, M. (2010). "BS Seeker: precise mapping for bisulfite sequencing." *BMC Bioinformatics* 11.1, p. 203. DOI: 10.1186/1471-2105-11-203.

Chen, L. *et al.* (2014). "Transcriptional diversity during lineage commitment of human blood progenitors." *Science (New York, NY)* 345.6204, p. 1251033. DOI: 10.1126/science.1251033.

Chin, M. H., Mason, M. J., Xie, W., Volinia, S., Singer, M., Peterson, C., Ambartsumyan, G., Aimiuwu, O., Richter, L., Zhang, J., Khvorostov, I., Ott, V., Grunstein, M., Lavon, N., Benvenisty, N., Croce, C. M., Clark, A. T., Baxter, T., Pyle, A. D., Teitell, M. A., Pelegrini, M., Plath, K., and Lowry, W. E. (2009). "Induced Pluripotent Stem Cells and Embryonic Stem Cells are Distinguished by Gene Expression Signatures." *Cell Stem Cell* 5.1, pp. 111–123. DOI: 10.1016/j.stem.2009.06.008.

Chung, D., Kuan, P. F., Li, B., Sanalkumar, R., Liang, K., Bresnick, E. H., Dewey, C., and Keleş, S. (2011). "Discovering Transcription Factor Binding Sites in Highly Repetitive Regions of Genomes with Multi-Read Analysis of ChIP-Seq Data." *PLoS Computational Biology* 7.7, e1002111. DOI: 10.1371/journal.pcbi.1002111.

Cleveland, W. S. (1979). "Robust Locally Weighted Regression and Smoothing Scatterplots." *Journal of the American Statistical Association* 74.368, pp. 829–836. DOI: 10.1080/01621459.1979.10481038.

Coarfa, C., Yu, F., Miller, C. A., Chen, Z., Harris, R. A., and Milosavljevic, A. (2010). "Pash 3.0: A versatile software package for read mapping and integrative analysis of genomic and epigenomic variation using massively parallel DNA sequencing." *BMC Bioinformatics* 11.1, p. 572. DOI: 10.1186/1471-2105-11-572.

Corces, M. R., Buenrostro, J. D., Wu, B., Greenside, P. G., Chan, S. M., Koenig, J. L., Snyder, M. P., Pritchard, J. K., Kundaje, A., Greenleaf, W. J., Majeti, R., and Chang, H. Y. (2016). "Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution." *Nature Genetics*. DOI: 10.1038/ng.3646.

Cortes, C. and Vapnik, V. (1995). "Support-Vector Networks." *Machine Learning* 20.3, pp. 273–297. DOI: 10.1007/BF00994018.

Cowley, M. and Oakey, R. J. (2013). "Transposable Elements Re-Wire and Fine-Tune the Transcriptome." *PLoS Genetics* 9.1, e1003234. DOI: 10.1371/journal.pgen.1003234.

Crompton, J. G., Narayanan, M., Cuddapah, S., Roychoudhuri, R., Ji, Y., Yang, W., Patel, S. J., Sukumar, M., Palmer, D. C., Peng, W., Wang, E., Marincola, F. M., Klebanoff, C. A., Zhao, K., Tsang, J. S., Gattinoni, L., and Restifo, N. P. (2015). "Lineage relationship of CD8+ T cell subsets is revealed by progressive changes in the epigenetic landscape." *Cellular & molecular immunology*. DOI: 10.1038/cmi.2015.032.

Csardi, G. and Nepusz, T. (2006). "The igraph software package for complex network research." *InterJournal* Complex Systems, p. 1695. URL: http://igraph.org.

Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., Steemers, F. J., Trapnell, C., and Shendure, J. (2015). "Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing." *Science (New York, NY)* 348.6237, pp. 910–914. DOI: 10.1126/science.aab1601.

Dale, D. C., Boxer, L., and Liles, W. C. (2008). "The phagocytes: neutrophils and monocytes." *Blood* 112.4, pp. 935–945. DOI: 10.1182/blood-2007-12-077917.

Day, D. S., Luquette, L. J., Park, P. J., and Kharchenko, P. V. (2010). "Estimating enrichment of repetitive elements from high-throughput sequence data." *Genome Biology* 11.6, R69. DOI: 10.1186/gb-2010-11-6-r69.

DeFranco, A., Locksley, R. M., and Robertson, M. (2007). *Immunity: The Immune Response in Infectious and Inflammatory Disease*. New Science Press. ISBN: 0199206147.

Dekker, J., Marti-Renom, M. A., and Mirny, L. A. (2013). "Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data." *Nature Reviews Genetics* 14.6, pp. 390–403. DOI: 10.1038/nrg3454.

Deplus, R., Blanchon, L., Rajavelu, A., Boukaba, A., Defrance, M., Luciani, J., Rothé, F., Dedeurwaerder, S., Denis, H., Brinkman, A. B., Simmer, F., Müller, F., Bertin, B., Berdasco, M., Putmans, P., Calonne, E., Litchfield, D. W., Launoit, Y. de, Jurkowski, T. P., Stunnenberg, H. G., Bock, C., Sotiriou, C., Fraga, M. F., Esteller, M., Jeltsch, A., and Fuks, F. (2014). "Regulation of DNA methylation patterns by CK2-mediated phosphorylation of Dnmt3a." *Cell Reports* 8.3, pp. 743–753. DOI: 10.1016/j.celrep.2014.06.048.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J. B., Lipovich, L., Gonzalez, J. M., Thomas, M., Davis, C. A., Shiekhattar, R., Gingeras, T. R., Hubbard, T. J., Notredame, C., Harrow, J., and Guigó, R. (2012). "The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression." *Genome Research* 22.9, pp. 1775–1789. DOI: 10.1101/gr.132159.111.

Doi, A., Park, I.-H., Wen, B., Murakami, P., Aryee, M. J., Irizarry, R., Herb, B., Ladd-Acosta, C., Rho, J., Loewer, S., Miller, J., Schlaeger, T., Daley, G. Q., and Feinberg, A. P. (2009). "Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts." *Nature Genetics* 41.12, pp. 1350–1353. DOI: 10.1038/ng.471.

Doulatov, S., Notta, F., Laurenti, E., and Dick, J. E. (2012). "Hematopoiesis: A Human Perspective." *Cell Stem Cell* 10.2, pp. 120–136. DOI: 10.1016/j.stem.2012.01.006.

Durek, P. *et al.* (2016). "Epigenomic Profiling of Human CD4+ T Cells Supports a Linear Differentiation Model and Highlights Molecular Regulators of Memory Development." *Immunity* 45.5, pp. 1148–1161. DOI: 10.1016/j.immuni.2016.10.022.

Ebert, P., Müller, F., Nordström, K., Lengauer, T., and Schulz, M. H. (2015). "A general concept for consistent documentation of computational analyses." *Database* 2015. DOI: 10.1093/database/bav050.

ENCODE Project Consortium (2004). "The ENCODE (ENCyclopedia Of DNA Elements) Project." *Science (New York, NY)* 306.5696, pp. 636–640. DOI: 10.1126/science.1105136.

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

Ernst, J. and Kellis, M. (2010). "Discovery and characterization of chromatin states for systematic annotation of the human genome." *Nature Biotechnology* 28.8, pp. 817–825. DOI: 10.1038/nbt.1662.

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., and Bernstein, B. E. (2011). "Mapping and analysis of chromatin state dynamics in nine human cell types." *Nature* 473.7345, pp. 43–49. DOI: 10.1038/nature09906.

Ernst, J. and Kellis, M. (2015). "Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues." *Nature Biotechnology* 33.4, pp. 364–376. DOI: 10.1038/nbt.3157.

Falcon, S. and Gentleman, R. (2007). "Using GOstats to test gene lists for GO term association." *Bioinformatics (Oxford, England)* 23.2, pp. 257–258. DOI: 10.1093/bioinformatics/btl567.

Farber, D. L., Yudanin, N. A., and Restifo, N. P. (2014). "Human memory T cells: generation, compartmentalization and homeostasis." *Nature reviews Immunology* 14.1, pp. 24–35. DOI: 10.1038/nri3567.

Farlik, M., Sheffield, N. C., Nuzzo, A., Datlinger, P., Schönegger, A., Klughammer, J., and Bock, C. (2015). "Single-cell DNA Methylome Sequencing and Bioinformatic Inference of Epigenomic Cell-State Dynamics." *Cell Reports* 10.8, pp. 1386–1397. DOI: 10.1016/j.celrep.2015.02.001.

Farlik, M., Halbritter, F., Müller, F., Choudry, F. A., Ebert, P., Klughammer, J., Farrow, S., Santoro, A., Ciaurro, V., Mathur, A., Uppal, R., Stunnenberg, H. G., Ouwehand, W. H., Laurenti, E., Lengauer, T., Frontini, M., and Bock, C. (2016). "DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation." *Cell Stem Cell* 19.6, pp. 808–822. DOI: 10.1016/j.stem.2016.10.019.

Fedoroff, N. V. (2012). "Transposable elements, Epigenetics, and Genome Evolution." *Science (New York, NY)* 338.6108, pp. 758–767. DOI: 10.1126/science.338.6108.758.

Feinberg, A. P., Ohlsson, R., and Henikoff, S. (2006). "The epigenetic progenitor origin of human cancer." *Nature Reviews Genetics* 7.1, pp. 21–33. DOI: 10.1038/nrg1748.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33.1, pp. 1–22.

Frith, M. C., Mori, R., and Asai, K. (2012). "A mostly traditional approach improves alignment of bisulfite-converted DNA." *Nucleic Acids Research* 40.13, e100–e100. DOI: 10.1093/nar/gks275.

Fullwood, M. J. and Ruan, Y. (2009). "ChIP-Based Methods for the Identification of Long-Range Chromatin Interactions." *Journal of Cellular Biochemistry* 107.1, pp. 30–39. DOI: 10.1002/jcb.22116.

Garber, M., Grabherr, M. G., Guttman, M., and Trapnell, C. (2011). "Computational methods for transcriptome annotation and quantification using RNA-seq." *Nature Methods* 8.6, pp. 469–477. DOI: doi:10.1038/nmeth.1613.

Gardiner-Garden, M. and Frommer, M. (1987). "CpG Islands in Vertebrate Genomes." *Journal of molecular biology* 196.2, pp. 261–282. DOI: 10.1016/0022-2836(87)90689-9.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. (2004). "Bioconductor: open software development for computational biology and bioinformatics." *Genome Biology* 5.10, R80. DOI: 10.1186/gb-2004-5-10-r80.

Gentleman, R. C. (2005). "Reproducible Research: A Bioinformatics Case Study." *Statistical Applications in Genetics and Molecular Biology* 4.1, Article2. DOI: 10.2202/1544-6115.1034.

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W. J., and Nekrutenko, A. (2005). "Galaxy: A platform for interactive large-scale genome analysis." *Genome Research* 15.10, pp. 1451–1455. DOI: 10.1101/gr.4086505.

Gilbert, S. F. (2014). *Developmental Biology*. 10th Edition. Sinhauer Associates, Inc. ISBN: 1605351733.

Goldberg, A. D., Allis, C. D., and Bernstein, E. (2007). "Epigenetics: A Landscape Takes Shape." *Cell* 128.4, pp. 635–638. DOI: 10.1016/j.cell.2007.02.006.

Goodier, J. L. and Kazazian, H. H. (2008). "Retrotransposons Revisited: The Restraint and Rehabilitation of Parasites." *Cell* 135.1, pp. 23–35. DOI: 10.1016/j.cell.2008.09.022.

Grossniklaus, U., Kelli, W. G., Ferguson-Smith, A. C., Pembrey, M., and Lindquist, S. (2013). "Transgenerational epigenetic inheritance: how important is it?" *Nature Reviews Genetics* 14, pp. 228–235. DOI: 10.1038/nrg3435.

Gu, H., Bock, C., Mikkelsen, T. S., Jäger, N., Smith, Z. D., Tomazou, E., Gnirke, A., Lander, E. S., and Meissner, A. (2010). "Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution." *Nature Methods* 7.2, pp. 133–136. DOI: 10.1038/nmeth.1414.

Guenther, M. G., Frampton, G. M., Soldner, F., Hockemeyer, D., Mitalipova, M., Jaenisch, R., and Young, R. A. (2010). "Chromatin Structure and Gene Expression Programs of Human Embryonic and Induced Pluripotent Stem Cells." *Cell Stem Cell* 7.2, pp. 249–257. DOI: 10.1016/j.stem.2010.06.015.

Guo, H., Zhu, P., Wu, X., Li, X., Wen, L., and Tang, F. (2013). "Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing." *Genome Research* 23.12, pp. 2126–2135. DOI: 10.1101/gr.161679.113.

Hahne, F., Durinck, S., Ivanek, R., Mueller, A., Lianoglou, S., Tan, G., and Parsons, L. (2012). *Gviz: Plotting data and annotation information along genomic coordinates*. R package version 1.13.2.

Halachev, K., Bast, H., Albrecht, F., Lengauer, T., and Bock, C. (2012). "EpiExplorer: live exploration and global analysis of large epigenomic datasets." *Genome Biology* 13.10, R96. DOI: 10.1186/gb-2012-13-10-r96.

Hancks, D. C. and Kazazian, H. H. (2012). "Active human retrotransposons: variation and disease." *Current Opinion in Genetics & Development* 22.3, pp. 191–203. DOI: 10.1016/j.gde.2012.02.006.

Hansen, T. B., Jensen, T. I., Clausen, B. H., Bramsen, J. B., Finsen, B., Damgaard, C. K., and Kjems, J. (2013). "Natural RNA circles function as efficient microRNA sponges." *Nature* 495.7441, pp. 384–388. DOI: 10.1038/nature11993.

Harris, E. Y., Ponts, N., Levchuk, A., Roch, K. L., and Lonardi, S. (2010). "BRAT: bisulfite-treated reads analysis tool." *Bioinformatics (Oxford, England)* 26.4, pp. 572–573. DOI: 10.1093/bioinformatics/btp706.

Harris, R. A. *et al.* (2010). "Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications." *Nature Biotechnology* 28.10, pp. 1097–1105. DOI: 10.1038/nbt.1682.

Harrow, J. *et al.* (2012). "GENCODE: the reference human genome annotation for The ENCODE Project." *Genome Research* 22.9, pp. 1760–1774. DOI: 10.1101/gr.135350.111.

Hashimoto, S.-i., Ogoshi, K., Sasaki, A., Abe, J., Qu, W., Nakatani, Y., Ahsan, B., Oshima, K., Shand, F. H. W., Ametani, A., Suzuki, Y., Kaneko, S., Wada, T., Hattori, M., Sugano, S., Morishita, S., and Matsushima, K. (2013). "Coordinated changes in DNA methylation in antigen-specific memory CD4 T cells." *Journal of Immunology* 190.8, pp. 4076–4091. DOI: 10.4049/jimmunol.1202267.

Heard, E. and Martienssen, R. A. (2014). "Transgenerational Epigenetic Inheritance: Myths and Mechanisms." *Cell* 157.1, pp. 95–109. DOI: 10.1016/j.cell.2014.02.045.

Hegi, M. E., Diserens, A.-C., Gorlia, T., Hamou, M.-F., Tribolet, N. de, Weller, M., Kros, J. M., Hainfellner, J. A., Mason, W., Mariani, L., Bromberg, J. E. C., Hau, P., Mirimanoff, R. O., Cairncross, J. G., Janzer, R. C., and Stupp, R. (2005). "MGMT Gene Silencing and Benefit from Temozolomide in Glioblastoma." *The New England journal of medicine* 352.10, pp. 997–1003. DOI: 10.1056/NEJMoa043331.

Heijmans, B. T., Tobi, E. W., Stein, A. D., Putter, H., Blauw, G. J., Susser, E. S., Slagboom, P. E., and Lumey, L. H. (2008). "Persistent epigenetic differences associated with prenatal exposure to famine in humans." *Proceedings of the National Academy of Sciences* 105.44, pp. 17046–17049. DOI: 10.1073/pnas.0806560105.

Henson, S. M., Riddell, N. E., and Akbar, A. N. (2012). "Properties of end-stage human T cells defined by CD45RA re-expression." *Current Opinion in Immunology* 24.4, pp. 476–481. DOI: 10.1016/j.coi.2012.04.001.

Heyn, H. and Esteller, M. (2012). "DNA methylation profiling in the clinic: applications and challenges." *Nature Reviews Genetics* 13.10, pp. 679–692. DOI: 10.1038/nrg3270.

Heyn, H., Li, N., Ferreira, H. J., Moran, S., Pisano, D. G., Gomez, A., Diez, J., Sanchez-Mut, J. V., Setien, F., Carmona, F. J., Puca, A. A., Sayols, S., Pujana, M. A., Serra-Musach, J., Iglesias-Platas, I., Formiga, F., Fernandez, A. F., Fraga, M. F., Heath, S. C., Valencia, A., Gut, I. G., Wang, J., and Esteller, M. (2012). "Distinct DNA methylomes of newborns and centenarians." *Proceedings of the National Academy of Sciences* 109.26, pp. 10522–10527. DOI: 10.1073/pnas.1120658109.

Hodges, E., Molaro, A., Dos Santos, C. O., Thekkat, P., Song, Q., Uren, P. J., Park, J., Butler, J., Rafii, S., McCombie, W. R., Smith, A. D., and Hannon, G. J. (2011). "Directional DNA Methylation Changes and Complex Intermediate States Accompany Lineage Specificity in the Adult Hematopoietic Compartment." *Molecular Cell* 44.1, pp. 17–28. DOI: 10.1016/j.molcel.2011.08.026.

Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A., and Noble, W. S. (2012). "Unsupervised pattern discovery in human chromatin structure through genomic segmentation." *Nature Methods* 9.5, pp. 473–476. DOI: 10.1038/nmeth.1937.

Holliday, R. and Pugh, J. E. (1975). "DNA Modification Mechanisms and Gene Activity during Development." *Science (New York, NY)* 187.4173, pp. 226–232.

Hon, G. C., Hawkins, R. D., Caballero, O. L., Lo, C., Lister, R., Pelizzola, M., Valsesia, A., Ye, Z., Kuan, S., Edsall, L. E., Camargo, A. A., Stevenson, B. J., Ecker, J. R., Bafna, V., Strausberg, R. L., Simpson, A. J., and Ren, B. (2012). "Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer." *Genome Research* 22.2, pp. 246–258. DOI: 10.1101/gr.125872.111.

Horvath, S. (2013). "DNA methylation age of human tissues and cell types." *Genome Biology* 14.10, R115. DOI: 10.1186/gb-2013-14-10-r115.

Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., Wiencke, J. K., and Kelsey, K. T. (2012). "DNA methylation arrays as surrogate measures of cell mixture distribution." *BMC Bioinformatics* 13, p. 86. DOI: 10.1186/1471-2105-13-86.

Houseman, E. A., Molitor, J., and Marsit, C. J. (2014). "Reference-Free Cell Mixture Adjustments in Analysis of DNA Methylation Data." *Bioinformatics (Oxford, England)* 30.10, pp. 1431–1439. DOI: 10.1093/bioinformatics/btu029.

Hovestadt, V. *et al.* (2014). "Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing." *Nature* 510.7506, pp. 537–541. DOI: 10.1038/nature13268.

Huang, S. (2012). "The molecular and mathematical basis of Waddington's epigenetic landscape: a framework for post-Darwinian biology?" *BioEssays* 34.2, pp. 149–157. DOI: 10.1002/bies.201100031.

Hubert, L. and Arabie, P. (1985). "Comparing Partitions." *Journal of classification* 2.1, pp. 193–218. DOI: 10.1007/BF01908075.

Jacques, P.-É., Jeyakani, J., and Bourque, G. (2013). "The Majority of Primate-Specific Regulatory Sequences Are Derived from Transposable Elements." *PLoS Genetics* 9.5, e1003504–12. DOI: 10.1371/journal.pgen.1003504.

Jaffe, A. E. and Irizarry, R. A. (2014). "Accounting for cellular heterogeneity is critical in epigenome-wide association studies." *Genome Biology* 15.2, R31. DOI: 10.1186/gb-2014-15-2-r31.

Jenuwein, T. and Allis, C. D. (2001). "Translating the Histone Code." *Science (New York, NY)* 293.5532, pp. 1074–1080. DOI: 10.1126/science.1063127.

Ji, H., Ehrlich, L. I. R., Seita, J., Murakami, P., Doi, A., Lindau, P., Lee, H., Aryee, M. J., Irizarry, R. A., Kim, K., Rossi, D. J., Inlay, M. A., Serwold, T., Karsunky, H., Ho, L., Daley, G. Q., Weissman, I. L., and Feinberg, A. P. (2010). "Comprehensive methylome map of lineage commitment from haematopoietic progenitors." *Nature* 467.7313, pp. 338–342. DOI: 10.1038/nature09367.

Johnson, W. E., Li, C., and Rabinovic, A. (2007). "Adjusting batch effects in microarray expression data using empirical Bayes methods." *Biostatistics (Oxford, England)* 8.1, pp. 118–127. DOI: 10.1093/biostatistics/kxj037.

Jones, P. A. (2012). "Functions of DNA methylation: islands, start sites, gene bodies and beyond." *Nature Reviews Genetics* 13, pp. 484–492. DOI: 10.1038/nrg3230.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). "Repbase Update, a database of eukaryotic repetitive elements." *Cytogenetic and Genome Research* 110.1-4, pp. 462–467. DOI: 10.1159/000084979.

Jurka, J., Bao, W., Kojima, K., and Kapitonov, V. V. (2011). "Repetitive Elements: Bioinformatic Identification, Classification and Analysis." *eLS*. DOI: 10.1002/9780470015902.a0005270.pub2.

Kaati, G., Bygren, L. O., and Edvinsson, S. (2002). "Cardiovascular and diabetes mortality determined by nutrition during parents' and grandparents' slow growth period." *European Journal of Human Genetics* 10.11, pp. 682–688. DOI: 10.1038/sj.ejhg.5200859.

Kaech, S. M. and Cui, W. (2012). "Transcriptional control of effector and memory CD8+ T cell differentiation." *Nature Reviews Immunology* 12.11, pp. 749–761. DOI: 10.1038/nri3307.

Karlić, R., Chung, H.-R., Lasserre, J., Vlahoviček, K., and Vingron, M. (2010). "Histone modification levels are predictive for gene expression." *Proceedings of the National Academy of Sciences* 107.7, pp. 2926–2931. DOI: 10.1073/pnas.0909344107.

Kelly, T. K., Liu, Y., Lay, F. D., Liang, G., Berman, B. P., and Jones, P. A. (2012). "Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules." *Genome Research* 22.12, pp. 2497–2506. DOI: 10.1101/gr.143008.112.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). "The Human Genome Browser at UCSC." *Genome Research* 12.6, pp. 996–1006. DOI: 10.1101/gr.229102.

Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S., and Karolchik, D. (2010). "BigWig and BigBed: enabling browsing of large distributed datasets." *Bioinformatics (Oxford, England)* 26.17, pp. 2204–2207. DOI: 10.1093/bioinformatics/btq351.

Kim, K., Doi, A., Wen, B., Ng, K., Zhao, R., Cahan, P., Kim, J., Aryee, M. J., Ji, H., Ehrlich, L. I. R., Yabuuchi, A., Takeuchi, A., Cunniff, K. C., Hongguang, H., Mckinney-Freeman, S., Naveiras, O., Yoon, T. J., Irizarry, R. A., Jung, N., Seita, J., Hanna, J., Murakami, P., Jaenisch, R., Weissleder, R., Orkin, S. H., Weissman, I. L., Feinberg, A. P., and Daley, G. Q. (2010). "Epigenetic memory in induced pluripotent stem cells." *Nature* 467.7313, pp. 285–290. DOI: 10.1038/nature09342.

Klughammer, J., Datlinger, P., Printz, D., Sheffield, N. C., Farlik, M., Hadler, J., Fritsch, G., and Bock, C. (2015). "Differential DNA Methylation Analysis without a Reference Genome." *Cell Reports* 13.11, pp. 2621–2633. DOI: 10.1016/j.celrep.2015.11.024.

Kohli, R. M. and Zhang, Y. (2013). "TET enzymes, TDG and the dynamics of DNA demethylation." *Nature* 502.7472, pp. 472–479. DOI: 10.1038/nature12750.

Komori, H. K., Hart, T., LaMere, S. A., Chew, P. V., and Salomon, D. R. (2015). "Defining CD4 T Cell Memory by the Epigenetic Landscape of CpG DNA Methylation." *The Journal of Immunology* 194.4, pp. 1565–1579. DOI: 10.4049/jimmunol.1401162.

Koning, A. P. J. de, Gu, W., Castoe, T. A., Batzer, M. A., and Pollock, D. D. (2011). "Repetitive Elements May Comprise Over Two-Thirds of the Human Genome." *PLoS Genetics* 7.12, e1002384. DOI: 10.1371/journal.pgen.1002384.

Koohy, H., Down, T. A., Spivakov, M., and Hubbard, T. (2014). "A Comparison of Peak Callers Used for DNase-Seq Data." *PLoS ONE* 9.5, e96303. DOI: 10.1371/journal.pone.0096303.

Koschmieder, S., Gottgens, B., Zhang, P., Iwasaki-Arai, J., Akashi, K., Kutok, J. L., Dayaram, T., Geary, K., Green, A. R., Tenen, D. G., and Huettner, C. S. (2005). "Inducible chronic phase of myeloid leukemia with expansion of hematopoietic stem cells in a transgenic model of BCR-ABL leukemogenesis." *Blood* 105.1, pp. 324–334. DOI: 10.1182/blood-2003-12-4369.

Kouzarides, T. (2007). "Chromatin Modifications and Their Function." *Cell* 128.4, pp. 693–705. DOI: 10.1016/j.cell.2007.02.005.

Krishnapuram, B., Carin, L., Figueiredo, M. A. T., and Hartemink, A. J. (2005). "Sparse Multinomial Logistic Regression: Fast Algorithms and Generalization Bounds." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 27.6, pp. 957–968. DOI: 10.1109/TPAMI.2005.127.

Krueger, F. and Andrews, S. R. (2011). "Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications." *Bioinformatics (Oxford, England)* 27.11, pp. 1571–1572. DOI: 10.1093/bioinformatics/btr167.

Krueger, F., Kreck, B., Franke, A., and Andrews, S. R. (2012). "DNA methylome analysis using short bisulfite sequencing data." *Nature Methods* 9.2, pp. 145–151. DOI: 10.1038/nmeth.1828.

Kulis, M. *et al.* (2015). "Whole-genome fingerprint of the DNA methylome during human B cell differentiation." *Nature Genetics* 47.7, pp. 746–756. DOI: 10.1038/ng.3291.

Kumano, K., Arai, S., Hosoi, M., Taoka, K., Takayama, N., Otsu, M., Nagae, G., Ueda, K., Nakazaki, K., Kamikubo, Y., Eto, K., Aburatani, H., Nakauchi, H., and Kurokawa, M. (2012). "Generation of induced pluripotent stem cells from primary chronic myelogenous leukemia patient samples." *Blood* 119.26, pp. 6234–6242. DOI: 10.1182/blood-2011-07-367441.

Kunde-Ramamoorthy, G., Coarfa, C., Laritsky, E., Kessler, N. J., Harris, R. A., Xu, M., Chen, R., Shen, L., Milosavljevic, A., and Waterland, R. A. (2014). "Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing." *Nucleic Acids Research* 42.6, e43–e43. DOI: 10.1093/nar/gkt1325.

Laird, P. W. (2010). "Principles and challenges of genomewide DNA methylation analysis." *Nature Reviews Genetics* 11.3, pp. 191–203. DOI: 10.1038/nrg2732.

Lander, E. S. *et al.* (2001). "Initial sequencing and analysis of the human genome." *Nature* 409.6822, pp. 860–921. DOI: 10.1038/35057062.

Landt, S. G. *et al.* (2012). "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia." *Genome Research* 22.9, pp. 1813–1831. DOI: 10.1101/gr.136184.111.

Langmead, B. and Salzberg, S. L. (2012). "Fast gapped-read alignment with Bowtie 2." *Nature Methods* 9.4, pp. 357–359. DOI: 10.1038/nmeth.1923.

Lara-Astiaso, D., Weiner, A., Lorenzo-Vivas, E., Zaretsky, I., Jaitin, D. A., David, E., Keren-Shaul, H., Mildner, A., Winter, D., Jung, S., Friedman, N., and Amit, I. (2014). "Chromatin state dynamics during blood formation." *Science (New York, NY)* 345.6199, pp. 943–949. DOI: 10.1126/science.1256271.

Lasserre, J., Chung, H.-R., and Vingron, M. (2013). "Finding Associations among Histone Modifications Using Sparse Partial Correlation Networks." *PLoS Computational Biology* 9.9, e1003168. DOI: 10.1371/journal.pcbi.1003168.s001.

Leek, J. T. and Storey, J. D. (2007). "Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis." *PLoS Genetics* 3.9, pp. 1724–1735. DOI: 10.1371/journal.pgen.0030161.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. A. (2010). "Tackling the widespread and critical impact of batch effects in high-throughput data." *Nature Reviews Genetics* 11.10, pp. 733–739. DOI: 10.1038/nrg2825.

Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L. J., Lohr, J. G., Harris, C. C., Ding, L., Wilson, R. K., Wheeler, D. A., Gibbs, R. A., Kucherlapati, R., Lee, C., Kharchenko, P. V., Park, P. J., and Cancer Genome Atlas Research Network (2012). "Landscape of Somatic Retrotransposition in Human Cancers." *Science (New York, NY)* 337.6097, pp. 967–971. DOI: 10.1126/science.1222077.

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2012). "The sva package for removing batch effects and other unwanted variation in high-throughput experiments." *Bioinformatics (Oxford, England)* 28.6, pp. 882–883. DOI: 10.1093/bioinformatics/bts034.

Li, H., Ruan, J., and Durbin, R. (2008). "Mapping short DNA sequencing reads and calling variants using mapping quality scores." *Genome Research* 18.11, pp. 1851–1858. DOI: 10.1101/gr.078212.108.

Li, H. and Durbin, R. (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform." *Bioinformatics (Oxford, England)* 25.14, pp. 1754–1760. DOI: 10.1093/bioinformatics/btp324.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, 1. G. P. D. P. (2009). "The Sequence Alignment/Map format and SAMtools." *Bioinformatics (Oxford, England)* 25.16, pp. 2078–2079. DOI: 10.1093/bioinformatics/btp352.

Li, E., Bestor, T. H., and Jaenisch, R. (1992). "Targeted Mutation of the DNA Methyltransferase Gene Results in Embryonic Lethality." *Cell* 69.6, pp. 915–926.

Lieberman-Aiden, E., Berkum, N. L. van, Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J. (2009). "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." *Science (New York, NY)* 326.5950, pp. 289–293. DOI: 10.1126/science.1181369.

Lienert, F., Wirbelauer, C., Som, I., Dean, A., Mohn, F., and Schübeler, D. (2011). "Identification of genetic elements that autonomously determine DNA methylation states." *Nature Genetics* 43.11, pp. 1091–1097. DOI: 10.1038/ng.946.

Lim, J.-Q., Tennakoon, C., Li, G., Wong, E., Ruan, Y., Wei, C.-L., and Sung, W.-K. (2012). "BatMeth: improved mapper for bisulfite sequencing reads on DNA methylation." *Genome Biology* 13.10, R82. DOI: 10.1186/gb-2012-13-10-r82.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B., and Ecker, J. R. (2009). "Human DNA methylomes at base resolution show widespread epigenomic differences." *Nature* 462.7271, pp. 315–322. DOI: 10.1038/nature08514.

Lister, R., Pelizzola, M., Kida, Y. S., Hawkins, R. D., and Nery, J. R. (2011). "Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells." *Nature* 471.7336, pp. 68–73. DOI: 10.1038/nature09798.

Lister, R., Mukamel, E. A., Nery, J. R., Urich, M., Puddifoot, C. A., Johnson, N. D., Lucero, J., Huang, Y., Dwork, A. J., Schultz, M. D., Yu, M., Tonti-Filippini, J., Heyn, H., Hu, S., Wu, J. C., Rao, A., Esteller, M., He, C., Haghighi, F. G., Sejnowski, T. J., Behrens, M. M., and Ecker, J. R. (2013). "Global Epigenomic Reconfiguration during Mammalian Brain Development." *Science (New York, NY)* 341.6146, p. 1237905. DOI: 10.1126/science.1237905.

Liu, Y., Siegmund, K. D., Laird, P. W., and Berman, B. P. (2012). "Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data." *Genome Biology* 13.7, R61. DOI: 10.1186/gb-2012-13-7-r61.

Luger, K., Mäder, A. W., Richmond, R. K., and Sargent, D. F. (1997). "Crystal structure of the nucleosome core particle at 2.8 Å resolution." *Nature* 389.6648, pp. 251–260. DOI: 10.1038/38444.

Lutsik, P., Feuerbach, L., Arand, J., Lengauer, T., Walter, J., and Bock, C. (2011). "BiQ Analyzer HT: locus-specific analysis of DNA methylation by high-throughput bisulfite sequencing." *Nucleic Acids Research* 39.Web Server issue, W551–6. DOI: 10.1093/nar/gkr312.

Machova Polakova, K., Koblihova, J., and Stopka, T. (2013). "Role of Epigenetics in Chronic Myeloid Leukemia." *Current hematologic malignancy reports* 8.1, pp. 28–36. DOI: 10.1007/s11899-012-0152-z.

Maksimovic, J., Gordon, L., and Oshlack, A. (2012). "SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips." *Genome Biology* 13.6, R44. DOI: 10.1186/gb-2012-13-6-r44.

Makambi, K. (2003). "Weighted inverse chi-square method for correlated significance tests." *Journal of Applied Statistics* 30, pp. 225–234. DOI: 10.1080/0266476022000023767.

Mammana, A. and Chung, H.-R. (2015). "Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome." *Genome Biology* 16.1, p. 151. DOI: 10.1186/s13059-015-0708-z.

Mandal, P. K. and Kazazian, H. H. (2008). "SnapShot: Vertebrate Transposons." *Cell* 135.1, 192–192.e1. DOI: 10.1016/j.cell.2008.09.028.

McClintock, B. (1950). "The origin and behavior of mutable loci in maize." *Proceedings of the National Academy of Sciences of the United States of America* 36.6, pp. 344–355. DOI: 10.1073/pnas.36.6.344.

Meissner, A., Mikkelsen, T. S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B. E., Nusbaum, C., Jaffe, D. B., Gnirke, A., Jaenisch, R., and Lander, E. S. (2008). "Genome-scale DNA methylation maps of pluripotent and differentiated cells." *Nature* 454.7205, pp. 766–770. DOI: 10.1038/nature07107.

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., Maier, L., Mackowiak, S. D., Gregersen, L. H., Munschauer, M., Loewer, A., Ziebold, U., Landthaler, M., Kocks, C., Noble, F. le, and Rajewsky, N. (2013). "Circular RNAs are a large class of animal RNAs with regulatory potency." *Nature* 495.7441, pp. 333–338. DOI: 10.1038/nature11928.

Metzker, M. L. (2010). "Sequencing technologies - the next generation." *Nature Reviews Genetics* 11.1, pp. 31–46. DOI: 10.1038/nrg2626.

Michels, K. B., Binder, A. M., Dedeurwaerder, S., Epstein, C. B., Greally, J. M., Gut, I., Houseman, E. A., Izzi, B., Kelsey, K. T., Meissner, A., Milosavljevic, A., Siegmund, K. D., Bock, C., and Irizarry, R. A. (2013). "Recommendations for the design and analysis of epigenome-wide association studies." *Nature Methods* 10.10, pp. 949–955. DOI: 10.1038/nmeth.2632.

Mifsud, B., Tavares-Cadete, F., Young, A. N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S. W., Andrews, S., Grey, W., Ewels, P. A., Herman, B., Happe, S., Higgs, A., LeProust, E., Follows, G. A., Fraser, P., Luscombe, N. M., and Osborne, C. S. (2015). "Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C." *Nature Genetics* 47.6, pp. 598–606. DOI: 10.1038/ng.3286.

Mikeska, T., Bock, C., El-Maarri, O., Hübner, A., Ehrentraut, D., Schramm, J., Felsberg, J., Kahl, P., Büttner, R., Pietsch, T., and Waha, A. (2007). "Optimization of Quantitative MGMT Promoter Methylation Analysis using Pyrosequencing and Combined Bisulfite Restriction Analysis." *Journal of Molecular Diagnostics* 9.3, pp. 368–381. DOI: 10.2353/jmoldx.2007.060167.

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R. P., Lee, W., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E. S., and Bernstein, B. E. (2007). "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells." *Nature* 448.7153, pp. 553–560. DOI: 10.1038/nature06008.

Mikkelsen, T. S., Hanna, J., Zhang, X., Ku, M., Wernig, M., Schorderet, P., Bernstein, B. E., Jaenisch, R., Lander, E. S., and Meissner, A. (2008). "Dissecting direct reprogramming through integrative genomic analysis." *Nature* 454.7200, pp. 49–55. DOI: 10.1038/nature07056.

Mills, R. E., Bennett, E. A., Iskow, R. C., and Devine, S. E. (2007). "Which transposable elements are active in the human genome?" *Trends in Genetics* 23.4, pp. 183–191. DOI: 10.1016/j.tig.2007.02.006.

Miura, F., Enomoto, Y., Dairiki, R., and Ito, T. (2012). "Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging." *Nucleic Acids Research* 40.17, e136. DOI: 10.1093/nar/gks454.

Miyoshi, N., Ishii, H., Nagai, K.-i., Hoshino, H., Mimori, K., Tanaka, F., Nagano, H., Sekimoto, M., Doki, Y., and Mori, M. (2010). "Defined factors induce reprogramming of gastrointestinal cancer cells." *Proceedings of the National Academy of Sciences* 107.1, pp. 40–45. DOI: 10.1073/pnas.0912407107.

Molaro, A., Hodges, E., Fang, F., Song, Q., McCombie, W. R., Hannon, G. J., and Smith, A. D. (2011). "Sperm Methylation Profiles Reveal Features of Epigenetic Inheritance and Evolution in Primates." *Cell* 146.6, pp. 1029–1041. DOI: 10.1016/j.cell.2011.08.016.

Morris, K. V. and Mattick, J. S. (2014). "The rise of regulatory RNA." *Nature Reviews Genetics* 15.6, pp. 423–437. DOI: 10.1038/nrg3722.

Moran, S., Arribas, C., and Esteller, M. (2015). "Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences." *Epigenomics*. DOI: 10.2217/epi.15.114.

Muller-Sieburg, C. E., Sieburg, H. B., Bernitz, J. M., and Cattarossi, G. (2012). "Stem cell heterogeneity: implications for aging and regenerative medicine." *Blood* 119.17, pp. 3900–3907. DOI: 10.1182/blood-2011-12-376749.

National Human Genome Research Institute (2016). *DNA Sequencing Costs: Data*. https://www.genome.gov/27541954/dna-sequencing-costs-data/. Accessed: May 2016.

Notta, F., Zandi, S., Takayama, N., Dobson, S., Gan, O. I., Wilson, G., Kaufmann, K. B., Mcleod, J., Laurenti, E., Dunant, C. F., McPherson, J. D., Stein, L. D., Dror, Y., and Dick, J. E. (2016). "Distinct routes of lineage development reshape the human blood hierarchy across ontogeny." *Science (New York, NY)* 351.6269, aab2116. DOI: 10.1126/science.aab2116.

Novershtern, N., Subramanian, A., Lawton, L. N., Mak, R. H., Haining, W. N., McConkey, M. E., Habib, N., Yosef, N., Chang, C. Y., Shay, T., Frampton, G. M., Drake, A. C. B., Leskov, I., Nilsson, B., Preffer, F., Dombkowski, D., Evans, J. W., Liefeld, T., Smutko, J. S., Chen, J., Friedman, N., Young, R. A., Golub, T. R., Regev, A., and Ebert, B. L. (2011). "Densely Interconnected Transcriptional Circuits Control Cell States in Human Hematopoiesis." *Cell* 144.2, pp. 296–309. DOI: 10.1016/j.cell.2011.01.004.

Ohm, J. E., Mcgarvey, K. M., Yu, X., Cheng, L., Schuebel, K. E., Cope, L., Mohammad, H. P., Chen, W., Daniel, V. C., Yu, W., Berman, D. M., Jenuwein, T., Pruitt, K., Sharkis, S. J., Watkins, D. N., Herman, J. G., and Baylin, S. B. (2007). "A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing." *Nature Genetics* 39.2, pp. 237–242. DOI: 10.1038/ng1972.

Okano, M., Bell, D. W., Haber, D. A., and Li, E. (1999). "DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development." *Cell* 99.3, pp. 247–257.

Ong, C.-T. and Corces, V. G. (2011). "Enhancer function: new insights into the regulation of tissue-specific gene expression." *Nature Reviews Genetics* 12.4, pp. 283–293. DOI: 10.1038/nrg2957.

Orkin, S. H. and Zon, L. I. (2008). "Hematopoiesis: An Evolving Paradigm for Stem Cell Biology." *Cell* 132.4, pp. 631–644. DOI: 10.1016/j.cell.2008.01.025.

Otto, C., Stadler, P. F., and Hoffmann, S. (2012). "Fast and sensitive mapping of bisulfite-treated sequencing data." *Bioinformatics (Oxford, England)* 28.13, pp. 1698–1704. DOI: 10.1093/bioinformatics/bts254.

Painter, R. C., Osmond, C., Gluckman, P., Hanson, M., Phillips, D. I. W., and Roseboom, T. J. (2008). "Transgenerational effects of prenatal exposure to the Dutch famine on neonatal adiposity and health in later life." *BJOG: An International Journal of Obstetrics & Gynaecology* 115.10, pp. 1243–1249. DOI: 10.1111/j.1471-0528.2008.01822.x.

Park, P. J. (2009). "ChIP-seq: advantages and challenges of a maturing technology." *Nature Reviews Genetics* 10, pp. 669–680. DOI: 10.1038/nrg2641.

Paul, F., Arkin, Y., Giladi, A., Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Winter, D., Lara-Astiaso, D., Gury, M., Weiner, A., David, E., Cohen, N., Lauridsen, F. K. B., Haas, S., Schlitzer, A., Mildner, A., Ginhoux, F., Jung, S., Trumpp, A., Porse, B. T., Tanay, A., and Amit, I. (2015). "Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors." *Cell* 163.7, pp. 1663–1677. DOI: 10.1016/j.cell.2015.11.013.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Second Edition. Cambridge University Press. ISBN: 9781139643986.

Pedersen, B., Hsieh, T.-F., Ibarra, C., and Fischer, R. L. (2011). "MethylCoder: software pipeline for bisulfite-treated sequences." *Bioinformatics (Oxford, England)* 27.17, pp. 2435–2436. DOI: 10.1093/bioinformatics/btr394.

Pembrey, M. E., Bygren, L. O., Kaati, G., Edvinsson, S., Northstone, K., Sjöström, M., Golding, J., and ALSPAC Study Team (2006). "Sex-specific, male-line transgenerational responses in humans." *European Journal of Human Genetics* 14.2, pp. 159–166. DOI: 10.1038/sj.ejhg.5201538.

Pepper, M. and Jenkins, M. K. (2011). "Origins of CD4+ effector and central memory T cells." *Nature Immunology* 12.6, pp. 467–471. DOI: 10.1038/ni.2038.

Perner, J., Lasserre, J., Kinkley, S., Vingron, M., and Chung, H.-R. (2014). "Inference of interactions between chromatin modifiers and histone modifications: from ChIP-Seq data to chromatin-signaling." *Nucleic Acids Research* 42.22, pp. 13689–13695. DOI: 10.1093/nar/gku1234.

Perié, L., Duffy, K. R., Kok, L., Boer, R. J. de, and Schumacher, T. N. (2015). "The Branching Point in Erythro-Myeloid Differentiation." *Cell* 163.7, pp. 1655–1662. DOI: 10.1016/j.cell.2015.11.059.

Phipson, B., Lee, S., Majewski, I. J., Alexander, W. S., and Smyth, G. K. (2016). "Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression." *Ann Appl Stat* 10.2, pp. 946–963. DOI: 10.1214/16-AOAS920.

Pidsley, R., Y Wong, C. C., Volta, M., Lunnon, K., Mill, J., and Schalkwyk, L. C. (2013). "A data-driven approach to preprocessing Illumina 450K methylation array data." *BMC genomics* 14.1, p. 293. DOI: 10.1186/1471-2164-14-293.

Planello, A. C., Ji, J., Sharma, V., Singhania, R., Mbabaali, F., Müller, F., Alfaro, J. A., Bock, C., De Carvalho, D. D., and Batada, N. N. (2014). "Aberrant DNA methylation reprogramming during induced pluripotent stem cell generation is dependent on the choice of reprogramming factors." *Cell Regeneration* 3.4. DOI: 10.1186/2045-9769-3-4.

Plongthongkum, N., Diep, D. H., and Zhang, K. (2014). "Advances in the profiling of DNA modifications: cytosine methylation and beyond." *Nature Reviews Genetics* 15.10, pp. 647–661. DOI: 10.1038/nrg3772.

Qian, L., Huang, Y., Spencer, C. I., Foley, A., Vedantham, V., Liu, L., Conway, S. J., Fu, J.-D., and Srivastava, D. (2012). "In vivo reprogramming of murine cardiac fibroblasts into induced cardiomyocytes." *Nature* 485.7400, pp. 593–598. DOI: 10.1038/nature11044.

Rakyan, V. K., Down, T. A., Balding, D. J., and Beck, S. (2011). "Epigenome-wide association studies for common human diseases." *Nature Reviews Genetics* 12.8, pp. 529–541. DOI: 10.1038/nrg3000.

Reinius, L. E., Acevedo, N., Joerink, M., Pershagen, G., Dahlén, S.-E., Greco, D., Söderhäll, C., Scheynius, A., and Kere, J. (2012). "Differential DNA Methylation in Purified Human Blood Cells: Implications for Cell Lineage and Studies on Disease Susceptibility." *PLoS ONE* 7.7, e41361. DOI: 10.1371/journal.pone.0041361.

Rendeiro, A. F., Schmidl, C., Strefford, J. C., Walewska, R., Davis, Z., Farlik, M., Oscier, D., and Bock, C. (2016). "Chromatin accessibility maps of chronic lymphocytic leukaemia identify subtype-specific epigenome signatures and transcription regulatory networks." *Nature Communications* 7, p. 11938. DOI: 10.1038/ncomms11938.

Restifo, N. P. and Gattinoni, L. (2013). "Lineage relationship of effector and memory T cells." *Current opinion in immunology* 25.5, pp. 556–563. DOI: 10.1016/j.coi.2013.09.003.

Riggs, A. D. (1975). "X inactivation, differentiation, and DNA methylation." *Cytogenetics and cell genetics* 14.1, pp. 9–25.

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

Rosenbauer, F. and Tenen, D. G. (2007). "Transcription factors in myeloid development: balancing differentiation with transformation." *Nature reviews Immunology* 7.2, pp. 105–117. DOI: 10.1038/nri2024.

Rosenfeld, J. A., Wang, Z., Schones, D. E., Zhao, K., Desalle, R., and Zhang, M. Q. (2009). "Determination of enriched histone modifications in non-genic portions of the human genome." *BMC Genomics* 10.1, p. 143. DOI: 10.1186/1471-2164-10-143.

Russ, B. E., Olshanksy, M., Smallwood, H. S., Li, J., Denton, A. E., Prier, J. E., Stock, A. T., Croom, H. A., Cullen, J. G., Nguyen, M. L. T., Rowe, S., Olson, M. R., Finkelstein, D. B., Kelso, A., Thomas, P. G., Speed, T. P., Rao, S., and Turner, S. J. (2014). "Distinct Epigenetic Signatures Delineate Transcriptional Programs during Virus-Specific CD8+ T Cell Differentiation." *Immunity* 41.5, pp. 853–865. DOI: 10.1016/j.immuni.2014.11.001.

Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P. P. (2011). "A ceRNA Hypothesis: the Rosetta Stone of a Hidden RNA Language?" *Cell* 146.3, pp. 353–358. DOI: 10.1016/j.cell.2011.07.014.

Sandoval, J., Heyn, H., Moran, S., Serra-Musach, J., Pujana, M. A., Bibikova, M., and Esteller, M. (2011). "Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome." *Epigenetics : official journal of the DNA Methylation Society* 6.6, pp. 692–702.

Sandoval, J. *et al.* (2013). "A prognostic DNA methylation signature for stage I non-small-cell lung cancer." *Journal of Clinical Oncology* 31.32, pp. 4140–4147. DOI: 10.1200/JCO.2012.48.5516.

Sander, J. D. and Joung, J. K. (2014). "CRISPR-Cas systems for editing, regulating and targeting genomes." *Nature Biotechnology* 32.4, pp. 347–355. DOI: 10.1038/nbt.2842.

Sánchez-Castillo, M., Ruau, D., Wilkinson, A. C., Ng, F. S. L., Hannah, R., Diamanti, E., Lombard, P., Wilson, N. K., and Gottgens, B. (2015). "CODEX: a next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities." *Nucleic Acids Research* 43.Database issue, pp. D1117–23. DOI: 10.1093/nar/gku895.

Schlesinger, Y., Straussman, R., Keshet, I., Farkash, S., Hecht, M., Zimmerman, J., Eden, E., Yakhini, Z., Ben-Shushan, E., Reubinoff, B. E., Bergman, Y., Simon, I., and Cedar, H. (2007). "Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer." *Nature Genetics* 39.2, pp. 232–236. DOI: 10.1038/ng1950.

Schmidl, C., Rendeiro, A. F., Sheffield, N. C., and Bock, C. (2015). "ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors." *Nature Methods*. DOI: 10.1038/nmeth.3542.

Schmidt, F., Gasparoni, N., Gasparoni, G., Gianmoena, K., Cadenas, C., Polansky, J. K., Ebert, P., Nordström, K., Barann, M., Sinha, A., Fröhler, S., Xiong, J., Dehghani Amirabad, A., Behjati Ardakani, F., Hutter, B., Zipprich, G., Felder, B., Eils, J., Brors, B., Chen, W., Hengstler, J. G., Hamann, A., Lengauer, T., Rosenstiel, P., Walter, J., and Schulz, M. H. (2016). "Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction." *Nucleic Acids Research*. DOI: 10.1093/nar/gkw1061.

Schübeler, D. (2015). "Function and information content of DNA methylation." *Nature* 517.7534, pp. 321–326. DOI: 10.1038/nature14192.

Scherer, M. (2016). "Dissecting DNA Methylation in Human Aging." MA thesis. Saarland University.

Sender, R., Fuchs, S., and Milo, R. (2016). "Revised Estimates for the Number of Human and Bacteria Cells in the Body." *PLoS Biology* 14.8, e1002533. DOI: 10.1371/journal.pbio.1002533.

Sexton, T. and Cavalli, G. (2015). "The Role of Chromosome Domains in Shaping the Functional Genome." *Cell* 160.6, pp. 1049–1059. DOI: 10.1016/j.cell.2015.02.040.

Sharma, S., Kelly, T. K., and Jones, P. A. (2010). "Epigenetics in cancer." *Carcinogenesis* 31.1, pp. 27–36. DOI: 10.1093/carcin/bgp220.

Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). "Single-cell sequencing-based technologies will revolutionize whole-organism science." *Nature Reviews Genetics* 14.9, pp. 618–630. DOI: 10.1038/nrg3542.

Sheffield, N. C. and Bock, C. (2016). "LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor." *Bioinformatics (Oxford, England)* 32.4, pp. 587–589. DOI: 10.1093/bioinformatics/btv612.

Shih, A. H., Abdel-Wahab, O., Patel, J. P., and Levine, R. L. (2012). "The role of mutations in epigenetic regulators in myeloid malignancies." *Nature Reviews Cancer* 12.9, pp. 599–612. DOI: 10.1038/nrc3343.

Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S. R., Stegle, O., Reik, W., and Kelsey, G. (2014). "Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity." *Nature Methods* 11.8, pp. 817–820. DOI: 10.1038/nmeth.3035.

Smith, A. D., Chung, W.-Y., Hodges, E., Kendall, J., Hannon, G., Hicks, J., Xuan, Z., and Zhang, M. Q. (2009). "Updates to the RMAP short-read mapping software." *Bioinformatics (Oxford, England)* 25.21, pp. 2841–2842. DOI: 10.1093/bioinformatics/btp533.

Smith, Z. D., Gu, H., Bock, C., Gnirke, A., and Meissner, A. (2009). "High-throughput bisulfite sequencing in mammalian genomes." *Methods (San Diego, Calif)* 48.3, pp. 226–232. DOI: 10.1016/j.ymeth.2009.05.003.

Smith, Z. D. and Meissner, A. (2013). "DNA methylation: roles in mammalian development." *Nature Reviews Genetics* 14.3, pp. 204–220. DOI: 10.1038/nrg3354.

Smith, Z. D., Chan, M. M., Humm, K. C., Karnik, R., Mekhoubad, S., Regev, A., Eggan, K., and Meissner, A. (2014). "DNA methylation dynamics of the human preimplantation embryo." *Nature* 511.7511, pp. 611–615. DOI: 10.1038/nature13581.

Smyth, G. K. (2004). "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments." *Statistical Applications in Genetics and Molecular Biology* 3.1, Article3–25. DOI: 10.2202/1544-6115.1027.

Song, K., Nam, Y.-J., Luo, X., Qi, X., Tan, W., Huang, G. N., Acharya, A., Smith, C. L., Tallquist, M. D., Neilson, E. G., Hill, J. A., Bassel-Duby, R., and Olson, E. N. (2012). "Heart repair by reprogramming non-myocytes with cardiac transcription factors." *Nature* 485.7400, pp. 599–604. DOI: 10.1038/nature11139.

Spitz, F. and Furlong, E. E. M. (2012). "Transcription factors: from enhancer binding to developmental control." *Nature Reviews Genetics* 13.9, pp. 613–626. DOI: 10.1038/nrg3207.

Stadler, M. B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., Nimwegen, E. van, Wirbelauer, C., Oakeley, E. J., Gaidatzis, D., Tiwari, V. K., and Schübeler, D. (2011). "DNA-binding factors shape the mouse methylome at distal regulatory regions." *Nature* 480.7378, pp. 490–495. DOI: 10.1038/nature10716.

Stevens, M., Cheng, J. B., Li, D., Xie, M., Hong, C., Maire, C. L., Ligon, K. L., Hirst, M., Marra, M. A., Costello, J. F., and Wang, T. (2013). "Estimating absolute methylation levels at single-CpG resolution from methylation enrichment and restriction enzyme sequencing methods." *Genome Research* 23.9, pp. 1541–1553. DOI: 10.1101/gr.152231.112.

Stegle, O., Teichmann, S. A., and Marioni, J. C. (2015). "Computational and analytical challenges in single-cell transcriptomics." *Nature Reviews Genetics* 16.3, pp. 133–145. DOI: 10.1038/nrg3833.

Stöckel, D., Kehl, T., Trampert, P., Schneider, L., Backes, C., Ludwig, N., Gerasch, A., Kaufmann, M., Gessler, M., Graf, N., Meese, E., Keller, A., and Lenhof, H.-P. (2016). "Multi-omics enrichment analysis using the GeneTrail2 web service." *Bioinformatics (Oxford, England)* 32.10, pp. 1502–1508. DOI: 10.1093/bioinformatics/btv770.

Stricker, S. H., Feber, A., Engström, P. G., Carén, H., Kurian, K. M., Takashima, Y., Watts, C., Way, M., Dirks, P., Bertone, P., Smith, A., Beck, S., and Pollard, S. M. (2013). "Widespread resetting of DNA methylation in glioblastoma-initiating cells suppresses malignant cellular behavior in a lineage-dependent manner." *Genes & Development* 27.6, pp. 654–669. DOI: 10.1101/gad.212662.112.

Stunnenberg, H. G., International Human Epigenome Consortium, and Hirst, M. (2016). "The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery." *Cell* 167.5, pp. 1145–1149. DOI: 10.1016/j.cell.2016.11.007.

Takahashi, K. and Yamanaka, S. (2006). "Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors." *Cell* 126.4, pp. 663–676. DOI: 10.1016/j.cell.2006.07.024.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). "Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors." *Cell* 131.5, pp. 861–872. DOI: 10.1016/j.cell.2007.11.019.

Teschendorff, A. E., Menon, U., Gentry-Maharaj, A., Ramus, S. J., Weisenberger, D. J., Shen, H., Campan, M., Noushmehr, H., Bell, C. G., Maxwell, A. P., Savage, D. A., Mueller-Holzner, E., Marth, C., Kocjan, G., Gayther, S. A., Jones, A., Beck, S., Wagner, W., Laird, P. W., Jacobs, I. J., and Widschwendter, M. (2010). "Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer." *Genome Research* 20.4, pp. 440–446. DOI: 10.1101/gr.103606.109.

Teschendorff, A. E., Zhuang, J., and Widschwendter, M. (2011). "Independent Surrogate Variable analysis to deconvolve confounding factors in large-scale microarray profiling studies." *Bioinformatics (Oxford, England)* 27.11, pp. 1496–1505. DOI: 10.1093/bioinformatics/btr171.

Teschendorff, A. E., Marabita, F., Lechner, M., Bartlett, T., Tegnér, J., Gomez-Cabrero, D., and Beck, S. (2013). "A Beta-Mixture Quantile Normalization method for correcting probe design bias in Illumina Infinium 450k DNA methylation data." *Bioinformatics (Oxford, England)* 29.2, pp. 189–196. DOI: 10.1093/bioinformatics/bts680.

Teschendorff, A. E. (2015). *Computational and Statistical Epigenomics*. Ed. by A. E. Teschendorff. Vol. 7. Translational Bioinformatics. Springer. DOI: 10.1007/978-94-017-9927-0.

Thompson, P. J., Macfarlan, T. S., and Lorincz, M. C. (2016). "Long Terminal Repeats: From Parasitic Elements to Building Blocks of the Transcriptional Regulatory Repertoire." *Molecular Cell* 62.5, pp. 766–776. DOI: 10.1016/j.molcel.2016.03.029.

Tobi, E. W., Goeman, J. J., Monajemi, R., Gu, H., Putter, H., Zhang, Y., Slieker, R. C., Stok, A. P., Thijssen, P. E., Müller, F., Zwet, E. W. van, Bock, C., Meissner, A., Lumey, L. H., Eline Slagboom, P., and Heijmans, B. T. (2014). "DNA methylation signatures link prenatal famine exposure to growth and metabolism." *Nature Communications* 5, p. 5592. DOI: 10.1038/ncomms6592.

Triche, T. J., Weisenberger, D. J., Berg, D. van den, Laird, P. W., and Siegmund, K. D. (2013). "Low-level processing of Illumina Infinium DNA Methylation BeadArrays." *Nucleic Acids Research* 41.7, e90–e90. DOI: 10.1093/nar/gkt090.

Varley, K. E., Gertz, J., Bowling, K. M., Parker, S. L., Reddy, T. E., Pauli-Behn, F., Cross, M. K., Williams, B. A., Stamatoyannopoulos, J. A., Crawford, G. E., Absher, D. M., Wold, B. J., and Myers, R. M. (2013). "Dynamic DNA methylation across diverse human cell lines and tissues." *Genome Research* 23.3, pp. 555–567. DOI: 10.1101/gr.147942.112.

Veenendaal, M. V. E., Painter, R. C., Rooij, S. R. de, Bossuyt, P. M. M., Post, J. A. M. van der, Gluckman, P. D., Hanson, M. A., and Roseboom, T. J. (2013). "Transgenerational effects of prenatal exposure to the 1944-45 Dutch famine." *BJOG: An International Journal of Obstetrics & Gynaecology* 120.5, pp. 548–553. DOI: 10.1111/1471-0528.12136.

Venter, J. C. *et al.* (2001). "The Sequence of the Human Genome." *Science (New York, NY)* 291.5507, pp. 1304–1351. DOI: 10.1126/science.1058040.

Vierbuchen, T., Ostermeier, A., Pang, Z. P., Kokubu, Y., Südhof, T. C., and Wernig, M. (2010). "Direct conversion of fibroblasts to functional neurons by defined factors." *Nature* 463.7284, pp. 1035–1041. DOI: 10.1038/nature08797.

Waddington, C. H. (1942). "The Epigenotype." *Endeavour*, pp. 18–20.

Waddington, C. H. (1957). *The Strategy of the Genes: A Discussion of Some Aspects of Theoretical Biology*. London: Allen & Unwin.

Wallner, S., Schröder, C., Leitão, E., Berulava, T., Haak, C., Beisser, D., Rahmann, S., Richter, A. S., Manke, T., Bönisch, U., Arrigoni, L., Fröhler, S., Klironomos, F., Chen, W., Rajewsky, N., Müller, F., Ebert, P., Lengauer, T., Barann, M., Rosenstiel, P., Gasparoni, G., Nordström, K., Walter, J., Brors, B., Zipprich, G., Felder, B., Klein-Hitpass, L., Attenberger, C., Schmitz, G., and Horsthemke, B. (2016). "Epigenetic dynamics of monocyte-to-macrophage differentiation." *Epigenetics & Chromatin* 9.1, p. 33. DOI: 10.1186/s13072-016-0079-z.

Wang, T., Zeng, J., Lowe, C. B., Sellers, R. G., Salama, S. R., Yang, M., Burgess, S. M., Brachmann, R. K., and Haussler, D. (2007). "Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53." *Proceedings of the National Academy of Sciences* 104.47, pp. 18613–18618. DOI: 10.1073/pnas.0703637104.

Wang, J., Huda, A., Lunyak, V. V., and Jordan, I. K. (2010). "A Gibbs sampling strategy applied to the mapping of ambiguous short-sequence tags." *Bioinformatics (Oxford, England)* 26.20, pp. 2501–2508. DOI: 10.1093/bioinformatics/btq460.

Wang, Z. and Kunze, R. (2015). "Transposons in Eukaryotes (Part A): Structures, Mechanisms and Applications." *eLS*. John Wiley & Sons, Ltd. DOI: 10.1002/9780470015902.a0026264.

Wei, Y., Schatten, H., and Sun, Q.-Y. (2015). "Environmental epigenetic inheritance through gametes and implications for human reproduction." *Human reproduction update* 21.2, pp. 194–208. DOI: 10.1093/humupd/dmu061.

Weisenberger, D. J. (2014). "Characterizing DNA methylation alterations from The Cancer Genome Atlas." *The Journal of clinical investigation* 124.1, pp. 17–23. DOI: 10.1172/JCI69740.

Whitaker, J. W., Chen, Z., and Wang, W. (2015). "Predicting the human epigenome from DNA motifs." *Nature Methods* 12.3, pp. 265–272. DOI: 10.1038/nmeth.3065.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York Inc. ISBN: 978-0-387-98140-6.

Widschwendter, M., Fiegl, H., Egle, D., Mueller-Holzner, E., Spizzo, G., Marth, C., Weisenberger, D. J., Campan, M., Young, J., Jacobs, I., and Laird, P. W. (2007). "Epigenetic stem cell signature in cancer." *Nature Genetics* 39.2, pp. 157–158. DOI: 10.1038/ng1941.

Williams, G. T. and Farzaneh, F. (2012). "Are snoRNAs and snoRNA host genes new players in cancer?" *Nature Reviews Cancer* 12.2, pp. 84–88. DOI: 10.1038/nrc3195.

Wilson, G. A., Dhami, P., Feber, A., Cortázar, D., Suzuki, Y., Schulz, R., Schär, P., and Beck, S. (2012). "Resources for methylome analysis suitable for gene knockout studies of potential epigenome modifiers." *GigaScience* 1.1, p. 3. DOI: 10.1186/2047-217X-1-3.

Woolthuis, C. M. and Park, C. Y. (2016). "Hematopoietic stem/progenitor cell commitment to the megakaryocyte lineage." *Blood* 127.10, pp. 1242–1248. DOI: 10.1182/blood-2015-07-607945.

Wu, T. D. and Nacu, S. (2010). "Fast and SNP-tolerant detection of complex variants and splicing in short reads." *Bioinformatics (Oxford, England)* 26.7, pp. 873–881. DOI: 10.1093/bioinformatics/btq057.

Xi, Y. and Li, W. (2009). "BSMAP: whole genome bisulfite sequence MAPping program." *BMC Bioinformatics* 10.1, p. 232. DOI: 10.1186/1471-2105-10-232.

Xi, Y., Bock, C., Müller, F., Sun, D., Meissner, A., and Li, W. (2012). "RRBSMAP: a fast, accurate and user-friendly alignment tool for reduced representation bisulfite sequencing." *Bioinformatics (Oxford, England)* 28.3, pp. 430–432. DOI: 10.1093/bioinformatics/btr668.

Xie, M., Hong, C., Zhang, B., Lowdon, R. F., Xing, X., Li, D., Zhou, X., Lee, H. J., Maire, C. L., Ligon, K. L., Gascard, P., Sigaroudinia, M., Tlsty, T. D., Kadlecek, T., Weiss, A., O'geen, H., Farnham, P. J., Madden, P. A. F., Mungall, A. J., Tam, A., Kamoh, B., Cho, S., Moore, R., Hirst, M., Marra, M. A., Costello, J. F., and Wang, T. (2013). "DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape." *Nature Genetics* 45.7, pp. 836–841. DOI: 10.1038/ng.2649.

Yamanaka, S. and Blau, H. M. (2010). "Nuclear reprogramming to a pluripotent state by three approaches." *Nature* 465.7299, pp. 704–712. DOI: 10.1038/nature09229.

Yoder, J. A., Walsh, C. P., and Bestor, T. H. (1997). "Cytosine methylation and the ecology of intragenomic parasites." *Trends in genetics : TIG* 13.8, pp. 335–340.

Yu, H., Zhu, S., Zhou, B., Xue, H., and Han, J.-D. J. (2008). "Inferring causal relationships among different histone modifications and gene expression." *Genome Research* 18.8, pp. 1314–1324. DOI: 10.1101/gr.073080.107.

Zentner, G. E. and Henikoff, S. (2014). "High-resolution digital profiling of the epigenome." *Nature Reviews Genetics* 15.12, pp. 814–827. DOI: 10.1038/nrg3798.

Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T., and Flicek, P. R. (2015). "The Ensembl Regulatory Build." *Genome Biology* 16.1, p. 56. DOI: 10.1186/s13059-015-0621-5.

Ziller, M. J., Müller, F., Liao, J., Zhang, Y., Gu, H., Bock, C., Boyle, P., Epstein, C. B., Bernstein, B. E., Lengauer, T., Gnirke, A., and Meissner, A. (2011). "Genomic Distribution and Inter-Sample Variation of Non-CpG Methylation across Human Cell Types." *PLoS Genetics* 7.12, e1002389. DOI: 10.1371/journal.pgen.1002389.

Ziller, M. J., Gu, H., Müller, F., Donaghey, J., Tsai, L. T. Y., Kohlbacher, O., De Jager, P. L., Rosen, E. D., Bennett, D. A., Bernstein, B. E., Gnirke, A., and Meissner, A. (2013). "Charting a dynamic DNA methylation landscape of the human genome." *Nature* 500.7463, pp. 477–481. DOI: 10.1038/nature12433.

Ziller, M. J., Edri, R., Yaffe, Y., Donaghey, J., Pop, R., Mallard, W., Issner, R., Gifford, C. A., Goren, A., Xing, J., Gu, H., Cacchiarelli, D., Tsankov, A. M., Epstein, C., Rinn, J. L., Mikkelsen, T. S., Kohlbacher, O., Gnirke, A., Bernstein, B. E., Elkabetz, Y., and Meissner, A. (2015). "Dissecting neural differentiation regulatory networks through epigenetic footprinting." *Nature* 518.7539, pp. 355–359. DOI: 10.1038/nature13990.

Zou, J., Lippert, C., Heckerman, D., Aryee, M., and Listgarten, J. (2014). "Epigenome-wide association studies without the need for cell-type composition." *Nature Methods* 11.3, pp. 309–311. DOI: 10.1038/nmeth.2815.