Investigation of DNA methylation heterogeneity in cancer

Dissertation

zur Erlangung des Grades eines Doktors
der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

vorgelegt von

Sara Hetzel

Berlin, 2023

Erstgutachter:	Prof. Dr. Knut Reinert
Zweitgutachter:	Prof. Dr. Alexander Meissner
Tag der Disputation:	16.10.2023

Abstract

Within the body, every cell contains the same genetic blueprint, the DNA, which is wrapped around histones and densely packed in the nucleus. Given the same genome, the identity of each cell is in part defined by modifications to the histones but also the genomic sequence itself, such as DNA methylation, that define active and inactive parts of the DNA. In somatic cells, DNA methylation levels are largely bimodal, with a high genome-wide methylation average that predominantly excludes CpG islands (CGIs), features often found near gene promoters that remain free of methylation. These patterns change across the majority of human cancer types, which exhibit global loss of methylation accompanied by a gain of methylation at select CGIs. To date, bisulfite sequencing represents the gold-standard method to profile DNA methylation at single-base resolution and has been widely used to characterize and understand DNA methylation landscapes in healthy and tumor cells. This thesis presents advancements in the computational analysis of bisulfite sequencing data sets, as well as applications to large-scale studies of DNA methylation in cancer. It showcases the adaptation of a local alignment tool to enable homology search for bisulfite-converted sequences, which outperforms established semi-global alignment tools when applied to the search of metagenomic data sets. Additionally, this thesis describes the development of a new application that provides fast and simplified extraction of DNA methylation heterogeneity metrics from single reads of bisulfite sequencing data. The importance of such metrics is demonstrated in the context of two studies that focus on DNA methylation changes within primary tumors and cancer cell lines. Single-read metrics and single-cell methylome profiling show that primary tumors are mainly characterized by heterogeneous, intermediate global and CGI DNA methylation that is intrinsic to the underlying single tumor cells. In contrast, cancer cell lines mostly assume one of two different states, where global DNA methylation levels are either drastically decreased or comparable to healthy tissues, while CGIs become almost fully methylated in both scenarios. Although rarely seen in solid tumors, extremely high genome-wide methylation levels can also be observed in an exceptional primary tumor type, acute lymphoblastic leukemia, where this landscape is influenced by specific epigenetic regulators. Together, the findings of this thesis advance our ability to analyze bisulfite sequencing data sets as well as to apply these more nuanced measurements to understand DNA methylation changes during tumorigenesis and in culture.

Zusammenfassung

Im Körper enthält jede Zelle denselben genetischen Bauplan, die DNA, die um Histone gewickelt und dicht gepackt im Zellkern liegt. Aufgrund des gleichen Genoms wird die Identität jeder Zelle zum Teil durch Veränderungen an den Histonen, aber auch an der Genomsequenz selbst, wie zum Beispiel durch DNA-Methylierung, bestimmt. Diese Modifikationen legen aktive und inaktive Teile der DNA fest. In somatischen Zellen ist die DNA-Methylierung weitgehend bimodal verteilt, mit einem hohen genomweiten Methylierungsdurchschnitt und der Ausnahme von CpG-Inseln (CGI), die häufig in der Nähe von Genpromotoren zu finden sind und frei von Methylierung bleiben. Diese Muster ändern sich bei den meisten menschlichen Krebstypen, die einen globalen Methylierungsverlust bei gleichzeitiger Zunahme der Methylierung an ausgewählten CGIs aufweisen. Bis heute ist die Bisulfit-Sequenzierung der Goldstandard für die Erstellung von DNA-Methylierungsprofilen und wird häufig zur Charakterisierung und zum Verständnis von DNA-Methylierungslandschaften in gesunden und Tumorzellen eingesetzt. In dieser Arbeit werden Fortschritte bei der rechnergestützten Analyse von Bisulfit-Sequenzierungsdatensätzen sowie deren Anwendung in groß angelegten Studien zur DNA-Methylierung bei Krebs vorgestellt. Sie zeigt die Anpassung eines lokalen Alignment-Tools, um eine Homologiesuche für Bisulfit-konvertierte Sequenzen zu ermöglichen, die etablierte semi-globale Alignment-Tools übertrifft, wenn sie bei der Suche von metagenomischen Datensätzen angewendet wird. Darüber hinaus wird in dieser Arbeit die Entwicklung einer neuen Anwendung beschrieben, die eine schnelle und vereinfachte Extraktion von Metriken der Heterogenität von DNA-Methylierung aus einzelnen Reads von Bisulfit-Sequenzierungsdaten ermöglicht. Die Bedeutung solcher Metriken wird im Rahmen von zwei Studien demonstriert, die sich auf DNA-Methylierungsveränderungen in Tumoren und Krebszelllinien konzentrieren. Einzel-Read-Metriken und die Erstellung von Einzelzell-Methylom-Profilen zeigen, dass primäre Tumore hauptsächlich durch heterogene, intermediäre globale und CGI DNA-Methylierung gekennzeichnet sind. Diese betrifft nicht nur den Durchschnitt sondern auch die zugrunde liegenden einzelnen Tumorzellen. Im Gegensatz dazu nehmen Krebszelllinien meist einen von zwei verschiedenen Zuständen an, bei denen das globale DNA-Methylierungsniveau entweder drastisch verringert oder mit der von gesundem Gewebe vergleichbar ist, während die CGIs in beiden Szenarien fast vollständig methyliert sind. Obwohl extrem hohe genomweite Methylierungsniveaus in soliden Tumoren selten zu finden sind, können diese in einem außergewöhnlichen primären Tumortyp, der akuten lymphoblastischen Leukämie, beobachtet werden. Hier wird diese Landschaft durch spezifische epigenetische Regulatoren beeinflusst. Insgesamt verbessern die Ergebnisse dieser Arbeit unsere Fähigkeit, Bisulfit-Sequenzierungsdatensätze zu analysieren und diese differenzierteren Messungen anzuwenden, um DNA-Methylierungsveränderungen während der Tumorentstehung und in Zellkulturen zu verstehen.

Preface

Publications and contributions

This thesis is based on multiple studies, none of which would have been possible without the work and support of many collaborators. The respective contributions are summarized here and highlighted in detail at the beginning of every chapter.

Chapter 3: H. Hauswedell*, S. Hetzel*, S. G. Gottlieb, H. Kretzmer, A. Meissner, and K. Reinert, "Lambda3: homology search for protein, nucleotide, and bisulfite-converted sequences," *Bioinformatics*, Mar 2024 (*equal contribution). DOI: https://doi.org/10.1093/bioinformatics/btae097. Licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

Chapter 4: S. Hetzel, P. Giesselmann, K. Reinert, A. Meissner, and H. Kretzmer, "RLM: Fast and simplified extraction of Read-Level Methylation metrics from bisulfite sequencing data," *Bioinformatics*, Oct 2021. DOI: https://doi.org/10.1093/bioinformatics/btab663. Licensed under CC BY-NC 4.0 (https://creativecommons.org/licenses/by-nc/4.0/).

Chapter 5: S. Hetzel, A. L. Mattei, H. Kretzmer, C. Qu, X. Chen, Y. Fan, G. Wu, K. G. Roberts, S. Luger, M. Litzow, J. Rowe, E. Paietta, W. Stock, E. R. Mardis, R. K. Wilson, J. R. Downing, C. G. Mullighan, and A. Meissner, "Acute lymphoblastic leukemia displays a distinct highly methylated genome," *Nature Cancer*, May 2022. DOI: https://doi.org/10.1038/s43018-022-00370-5. Licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

Chapter 6: S. Hetzel, R. Weigert, N. Bailly, K. Steinmann, A. Gnirke, H. Kretzmer, A. Meissner, and Z. D. Smith, "Redefining DNA methylation landscapes across tumors and cell lines," in preparation.

In addition to the studies listed above, I investigated DNA methylation heterogeneity in the context of extraembryonic development using trophoblast stem cells as a model system during my doctoral studies.

Not included in this thesis: R. Weigert*, S. Hetzel*, N. Bailly, C. Haggerty, I. A. Ilik, P. Y. K. Yung, C. Navarro, A. Bolondi, A. S. Kumar, C. Anania, B. Brändl, D. Meierhofer, D. G. Lupiáñez, F. J. Müller, T. Aktas, S. J. Elsässer, H. Kretzmer, Z. D. Smith, and A. Meissner, "Dynamic antagonism between key repressive pathways maintains the placental epigenome," *Nature Cell Biology*, Apr 2023 (*equal contribution). DOI: https://doi.org/10.1038/s41556-023-01114-y.

Acknowledgments

First and foremost, I would like to thank my supervisor Alexander Meissner for his scientific guidance, support, and mentoring during my years in his lab. I am thankful for the many opportunities I got to work on exciting research topics and his continuous efforts to help me improve my scientific skills and critical thinking.

I would also like to thank Knut Reinert for his efforts to help me improve my understanding of efficient programming and complex data structures. I am grateful for the chances I got to join his lab retreats that opened new opportunities for collaboration and tremendously improved my programming skills.

I am beyond grateful to Helene Kretzmer, who supported me from the first day of my doctoral studies and continues to be a great mentor, colleague, and friend. Spending hours working, thinking, and developing many projects together has taught me what great collaborations look like.

Moreover, I want to thank Zachary Smith for his support and mentoring during some of the challenging projects of my doctorate. His ideas and contributions were essential to shape our work and progress.

I am grateful to Raha Weigert for being a great project partner. Her tireless efforts made our shared collaboration possible, and her spirit motivated everyone involved in our work.

Furthermore, I want to thank Hannes Hauswedell for allowing me to contribute to his application Lambda and teaching me a lot about modern C++ and best practices in programming.

I am grateful to Alexandra Mattei for her support during my first project in the lab, for sharing her immense scientific knowledge, and for her generosity in making me a part of her own work.

Working on a variety of projects is only possible with the help of many colleagues and collaborators. In particular, I would like to thank Nina Bailly and Pay Giesselmann for their support, ideas, and contributions. Special thanks go to the SeqCore team, in particular Sven Klages, Sonia Paturej, Norbert Mages, and Stefan Börno, for their endless support and encouragement over many years. I am grateful to Lisa Barros de Andrade e Sousa for our helpful discussions and her advice. I would also like to thank Svenja Mehringer and Lennard Epping for being the best study group one can imagine during my Bachelor's and Master's degrees that prepared me well for the following years of my doctorate.

Many thanks go to all Meissner and Kretzmer lab members for their continued help, discussions, and scientific input. I also want to thank Lennard Epping, Hannes Hauswedell, Helene Kretzmer, and Zachary Smith for reading and commenting on this thesis.

Finally, I would like to thank my parents, family, and friends for always believing in me and supporting me throughout the past years. In particular, I want to thank my sister Mika, who taught me what it means to never give up, even in the most challenging times.

Contents

List of Figures

List of Tables

Glossary

5caC 5-carboxycytosine

5fC 5-formylcytosine

5hmC 5-hydroxymethylcytosine

5mC 5-methylcytosine

A Adenine

ACC Adrenocortical carcinoma

ALL Acute lymphoblastic leukemia

AML Acute myeloid leukemia

B-ALL B cell acute lymphoblastic leukemia

BLCA Bladder urothelial carcinoma

bp Base pair(s)

BRCA Breast invasive carcinoma

C Cytosine

CESC Cervical squamous cell carcinoma and endocervical adenocarcinoma

CGI CpG island

ChIP Chromatin immunoprecipitation

CIMP CpG island methylator phenotype

CLL Chronic lymphocytic leukemia

COAD Colon adenocarcinoma

CUT&RUN Cleavage under targets and release using nuclease

CUT&Tag Cleavage under targets and tagmentation

DMR Differentially methylated region

DMV DNA methylation valley

DNMT DNA methyltransferase

ECDF Empirical cumulative distribution function

ETP-ALL Early T cell precursor acute lymphoblastic leukemia

FDRP Fraction of discordant read pairs

FISH Fluorescence *in situ* hybridization

G Guanine

GAM Genome architecture mapping

GB Gigabyte(s)

GBM Glioblastoma multiforme

HELP HpaII tiny fragment enrichment by ligation-mediated PCR

hESC Human embryonic stem cell

HMD Highly methylated domain

HNSC Head and neck squamous cell carcinoma

HPC Hematopoietic multipotent progenitor cell

HSC Hematopoietic stem cell

iDMR Imprinted differentially methylated region

k-NN *k* nearest neighbor

kb Kilobase(s)

KIRC Kidney renal cell clear cell carcinoma

KO Knockout

LAML Acute myeloid leukemia

LCA Lowest common ancestor

LGG Brain lower grade glioma

LIHC Hepatocellular carcinoma

LUAD Lung adenocarcinoma

LUSC Lung squamous cell carcinoma

m6A N6-methyladenosine

MB Megabyte(s)

MBD-Seq Methyl-binding domain sequencing

MCL Mantle cell lymphoma

MeDIP-Seq Methylated DNA immunoprecipitation followed by sequencing

MHL Methylation haplotype load

PAAD Pancreatic adenocarcinoma

PBAT Post-bisulfite adaptor tagging

PCA Principle component analysis

PCR Polymerase chain reaction

PDR Proportion of discordant reads

Ph Philadelphia chromosome

PMD Partially methylated domain

PRAD Prostate adenocarcinoma

PRC2 Polycomb repressive complex 2

qFDRP Quantitative fraction of discordant read pairs

READ Rectum adenocarcinoma

RRBS Reduced representation bisulfite sequencing

RSS Resident set size

RTS Read transition score

SKCM Skin cutaneous melanoma

SPRITE Split-pool recognition of interactions by tag extension

STAD Stomach adenocarcinoma

T Thymine

T-ALL T cell acute lymphoblastic leukemia

TAD Topologically associated domain

TCGA The Cancer Genome Atlas

TET Ten-eleven translocation

THCA Thyroid carcinoma

TPLL T cell prolymphocytic leukemia

TPM Transcript per million

TSG Tumor suppressor gene

TSS Transcription start site

UCEC Uterine corpus endometrial carcinoma

UMAP Uniform manifold approximation and projection

UMI Unique molecular identifier

WES Whole-exome sequencing

WGBS Whole-genome bisulfite sequencing

WGS Whole-genome sequencing

WT Wild type

Chapter 1

Introduction

Cancer has historically been viewed to arise from the accumulation of genetic mutations that enable the acquisition of specific properties required for transformation and malignancy. These properties were summarized under the term "hallmarks of cancer" by Douglas Hanahan and Robert A. Weinberg for the first time in 2000 and expanded further in 2011 [1, 2]. The original properties included the evasion of apoptosis, self-sufficiency in growth signals, sustained angiogenesis, insensitivity to anti-growth signals, limitless replicative potential, and tissue invasion [1]. Over the following years, new core properties emerged based on the growing number of cancer studies that highlighted the importance of additional layers to account for the complexity of the diseases. In 2022, new dimensions to the existing and the introduction of emerging hallmarks were presented by Hanahan (see Figure 1 from [3]). Here, one emerging characteristic of tumorigenesis was termed "non-mutational epigenetic reprogramming" and refers to changes in the epigenetic landscape independent of genetic aberrations that can influence gene expression. These can include changes to various layers of epigenetic regulation, such as chromatin accessibility, histone modifications, and DNA methylation [3].

Over the last decades, alterations of the epigenetic landscape during tumorigenesis have been frequently observed, which expanded the traditional view of cancer as purely genetic diseases. This includes a genome-wide decrease in DNA methylation, selective hypermethylation of CpG-dense promoters as well as changes in the distribution of histone modifications [4, 5]. However, understanding and disentangling the interplay between and the effect of epigenetic and genetic aberrations remains a key challenge. Epigenetic alterations can be non-mutational (emerging hallmark) or caused by a genetic alteration in epigenetic regulators or other proteins that can affect the epigenome. On the other hand, epigenetic reconfiguration might sensitize the cancer genome to additional genetic aberrations [6]. For both non-mutational and genetically induced epigenetic alterations, it can be challenging to identify changes that have a regulatory function, such as a direct effect on gene expression [3]. However, besides direct links to gene regulation, the overall characteristics of the cancer epigenome also present interesting avenues to explore. Alterations of the DNA methylation landscape that are frequently observed during tumorigenesis are reminiscent of changes that also occur during aging and extraembryonic development [7–9]. Thus, studying the cancer epigenome could not only reveal tumor-specific gene expression regulation but also shed light on the emergence of similar epigenetic landscapes during different physiological processes.

The high-throughput read-out of epigenetic modifications was facilitated by and is continuously evolving based on next-generation sequencing technologies. For example, chemical modification of the DNA enables the read-out of the methylation status of cytosines across the whole genome during bisulfite sequencing [10, 11]. The resulting genome-wide DNA methylation profiles can be compared across healthy and tumor samples to detect differences of interest that can be integrated with gene expression data or other types of information. With the advent of single-cell sequencing technologies, it is now possible to further deepen our understanding of epigenetic regulation by investigating the cellular heterogeneity within a sample of interest [12]. Thus, the current tools at hand enable the profiling and analysis of the distribution of epigenetic modifications at a continuously larger scale and to a greater detail than previously possible.

This thesis mainly focuses on a specific epigenetic modification, namely DNA methylation, in the context of human cancer. One goal was to advance applications for the analysis of bisulfite sequencing data sets to offer new perspectives for studies that make use of them. Additionally, the use and impact of these applications should be demonstrated in the context of deepening our understanding of DNA methylation changes that occur during tumorigenesis. Following this introduction, chapter 2 therefore provides the necessary biological background. This includes epigenetic modifications, specifically DNA methylation, in mammalian genomes, the known changes of the DNA methylation landscape that occur during tumorigenesis in humans, as well as the basis of sequencing-based read-out and the computational processing of the resulting data sets. Chapter 3 presents the adaptation of a local alignment tool for homology search to accommodate bisulfite-converted query sequences. In chapter 4, the development of an application to extract and analyze read-level methylation metrics from bisulfite sequencing data sets is described. Chapter 5 and 6 present two large studies that focus on DNA methylation dynamics in cancer. First, a single tumor type and its respective DNA methylation landscape are described compared to a pan-cancer context. Second, an extensive study integrating hundreds of newly generated and publicly available data sets investigates the intrinsic properties of DNA methylation landscapes across primary tumors and cancer cell lines. Lastly, chapter 7 provides concluding remarks and discusses open questions following the studies presented in this thesis.

Part I

Background

The first part of this dissertation introduces different layers of epigenetic regulation in mammals, specifically DNA methylation, its role in somatic cells, and changes that occur during tumorigenesis. Additionally, the read-out of DNA methylation using microarray and sequencing technologies is described, and established methods for computational processing from sequencing reads to methylation rates are presented. Due to the main focus of this thesis on DNA methylation in human cancer and respective data sets, especially in chapter 5 and 6, the nomenclature of genes and proteins is used accordingly even though many biological processes will generalize across mammalian species.

Chapter 2

Layers of epigenetic regulation

2.1 Chromosomal architecture

In eukaryotic cells, the DNA in the form of chromosomes is densely packed in the nucleus of each cell, and its organization plays an essential role in different biological processes, including gene expression and genome integrity [13–15]. During interphase, chromosomes occupy different, non-random territories in the nucleus [16]. Additionally, several nuclear compartments exist, including the nuclear lamina, nuclear pore complexes, and different types of nuclear bodies, such as nucleoli. These nuclear compartments can influence and shape the organization of (parts of) chromosomes within the nucleus [17]. For example, gene-dense, frequently transcribed euchromatic regions preferentially reside in the interior part of the nucleus, while gene-poor, inactive, highly condensed heterochromatic areas often associate with the nuclear lamina [18,19].

Due to the three-dimensional organization of DNA in the nucleus, parts of the DNA can physically interact with other regions on the same or different chromosome(s) to form neighborhoods of coordinated gene regulation. Active (euchromatic) and inactive (heterochromatic) chromatin compartments have been shown to preferentially interact with compartments of the same type [18]. Within one chromosome, so-called topologically associated domains (TADs) represent long-range, self-interacting domains with boundaries frequently marked by CTCF and cohesin, two proteins that function in chromatin architecture and insulation [20–22]. These boundaries are highly conserved across cell types, and even species, whereas only a subset of TAD boundaries has been reported to be cell type-specific [20,23,24]. Multiple studies have proposed that the primary function of TADs lies in transcriptional regulation by limiting promoter-enhancer interactions [25,26]. These interactions are considered to be formed by local chromatin loops and are mostly found within the same TAD due to its insulating borders [27]. The disruption of specific TAD boundaries has been subsequently implicated in diseases such as limb malformation, adult-onset demyelinating leukodystrophy, and cancer [28–30].

2.2 Histone modifications

The basic unit of chromatin is the nucleosome, where the DNA is wrapped around (approximately 147 base pairs (bp) of DNA per nucleosome). Each nucleosome consists of two tetramers comprised of the four core histones H2A, H2B, H3, and H4 [31]. These histones can be post-translationally modified to encode regulatory information, primarily within their N-terminal tails, including methylation, acetylation, ubiquitination, and phosphorylation of specific amino acids [32]. Modifications of histone tails act as signals and can be read by other enzymes or complexes, which in turn execute regulatory functions [33]. As a result, chromatin compaction or opening can be facilitated to enable binding by specific transcription factors. Histone modifications thereby play a crucial role in many essential processes in the cell, such as DNA replication, gene expression, DNA repair, and control of the cell cycle [34, 35].

Additionally, aberrant changes in histone modifications have been implicated in developmental defects, and cancer [36, 37]. In the following, the key histone modifications referred to in this thesis are briefly introduced, and their molecular functions are described. The nomenclature of histone modifications includes the histone affected (e.g., H3), the modified amino acid (e.g., lysine at the fourth position or K4), as well as the modification itself (e.g., methylation (me)). In the case of methylation, the number of residues added to the amino acid is indicated (1, 2, or 3 reflecting mono-, di- and trimethylation, respectively) [38].

2.2.1 Active histone marks

H3K4me3 marks active transcription and is therefore commonly found at promoters of actively transcribed genes [39–41]. In contrast to tri-methylation, mono-methylation of the same amino acid (H3K4me1) can be found at active enhancers and enhancers primed for activation [42, 43]. Methylation of H3K4 in mammals is carried out by a variety of histone methyltransferases, including MLL1 to MLL5, SET1A, and SET1B that primarily act as part of larger protein complexes that facilitate H3K4me3 deposition [41, 44]. Additionally, H3K27ac deposited by the histone acetyltransferase p300 marks active enhancers as well as active transcription start sites [45, 46]. The body of transcribed genes is marked by H3K36me3, which is catalyzed by the histone methyltransferase SETD2 in mammals. H3K36me3 safeguards transcription by preventing aberrant transcription initiation and additionally plays a role in DNA damage response [47, 48].

2.2.2 Repressive histone marks

H3K27me3 is a mark reflective of transcriptional repression. It is deposited by the Polycomb repressive complex 2 (PRC2) and catalyzed via its subunit EZH2 [49, 50]. H3K27me3 plays a crucial role during differentiation and development as it is involved in the silencing of tissue-specific genes [31, 51]. Additionally, it is involved in silencing one of the two X chromosome copies in females [52]. H3K9me3 is mainly associated with heterochromatin and long-term silencing. It is commonly found at pericentromeric and other repeat-rich regions in the genome and catalyzed by the proteins SETDB1, SUV39H1, and SUV39H in mammals [50, 53]. H3K9me3-mediated repression is generally associated with the lack of chromatin accessibility, while regions

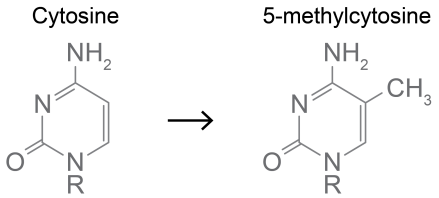

Figure 2.3.1: The addition of a methyl group at the fifth position of cytosine is commonly found in mammals, preferentially in a CpG context.

repressed by H3K27me3 frequently remain accessible for transcription factor binding and paused RNA polymerase [53]. In embryonic stem cells, inactive promoters are also often marked by the active histone modification H3K4me3 in addition to H3K27me3, reflecting a so-called bivalent state. Bivalency has been reported to leave promoters in a poised state, ready for activation, that is essential to timely induce expression upon differentiation into a specific lineage or cell type [54, 55].

2.3 DNA methylation

2.3.1 Function of DNA methylation

In addition to modifications of the histones, the DNA itself can be chemically modified and regulated. Methylation of cytosine (C) at the fifth position (5-methylcytosine or 5mC) is an epigenetic modification found in many organisms and conserved across most animals, plants, and fungi (Figure 2.3.1). In mammals, DNA methylation is primarily found in a CpG context and plays a crucial role in genome stability and transcriptional regulation. Most of the genome is highly methylated to ensure genome stability and silence transposable elements [56, 57]. In contrast, CpG-rich regions, termed CpG islands (CGIs), that are frequently found at promoters and often associated with housekeeping genes, remain mostly free of methylation [56]. Gain of methylation at promoters is often associated with the silencing of the respective genes, although DNA methylation rarely acts as a primary silencing mechanism [58]. Methylation levels of gene bodies are often positively correlated with increased and stable expression. Although initially counterintuitive, DNA methylation is thought to support transcription by nucleosome stabilization for enhanced transcription and regulation of alternative promoters [57, 59, 60].

Furthermore, DNA methylation plays a role in the silencing and inactivation of one of the X chromosome copies in cells with female karyotype as well as genomic imprinting (the inherited expression of a gene exclusively from the maternal or paternal allele) [56, 57]. Here, imprinted differentially methylated regions (iDMRs) control gene expression of the associated imprinted gene(s). These iDMRs are fully methylated on the allele that silences the gene and free of methylation on the allele from which the gene is expressed [61]. Due to its role in genome stability, transcriptional regulation, and imprinting, DNA methylation is essential for embryonic development. It can enable or restrict the differentiation of cells, and aberrant regulation of DNA methylation has been implicated in diseases such as cancer, autoimmune diseases, and metabolic disorders [56, 62, 63].

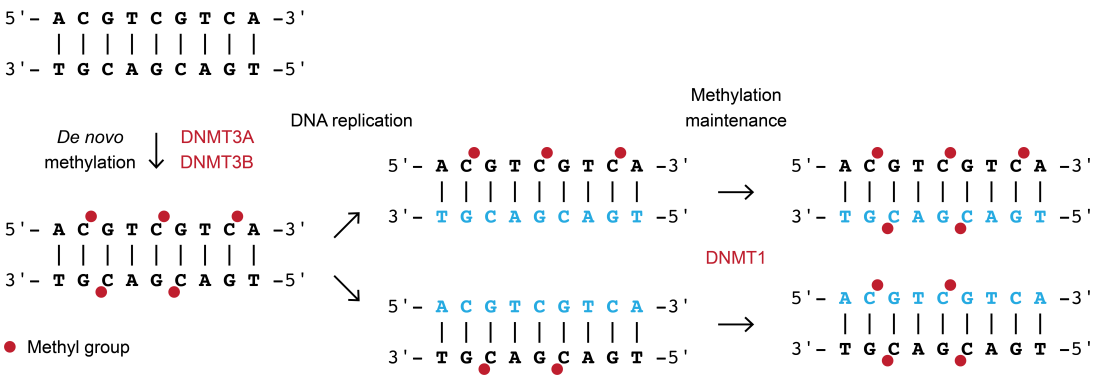

Figure 2.3.2: DNA methylation in mammals can be *de novo* placed by the enzymes DNMT3A and DNMT3B. After replication, the maintenance enzyme DNMT1 recognizes methyl groups in a CpG context on the original strand and places a methyl group at the respective position on the newly generated strand. This figure was adapted from Kretzmer [70].

2.3.2 Regulation of DNA methylation

In mammals, DNA methylation can be deposited by three enzymes: the DNA methyltransferases (DNMTs) 1, 3A, and 3B. While DNMT3A and DNMT3B can place methyl groups *de novo* at cytosines where previously no methylation was present, DNMT1 mainly functions as the DNA methylation maintenance enzyme: After replication, the newly synthesized strand is completely free of methylation (Figure 2.3.2). UHRF1, a co-factor of DNMT1, recruits the enzyme to the replication forks [56,64,65]. As CpGs are symmetric considering the two strands of DNA, DNMT1 can recognize methyl groups at cytosines in a CpG-context of the original strand and add a respective methyl group to the cytosine on the newly synthesized strand (Figure 2.3.2) [56,57]. In turn, DNA methylation can be passively lost over cell divisions if DNMT1 activity is blocked or incomplete [66,67]. Additionally, active removal of methylation can be induced by the ten-eleven translocation (TET) enzymes TET1, TET2, and TET3 (Figure 2.3.3). These enzymes catalyze the hydroxylation of 5mC to 5-hydroxymethylcytosine (5hmC) and subsequent oxidation steps that lead to either 5-formylcytosine (5fC) or further to 5-carboxycytosine (5caC). Both 5fc and 5caC can be replaced by a regular unmethylated cytosine by the base excision repair machinery, or during replication (Figure 2.3.3) [67]. Besides the main enzymes that deposit and remove DNA methylation, other co-factors exist that guide proteins to their respective targets or enhance their activity. An example of such a co-factor is the non-enzymatic protein DNMT3L, which can interact with DNMT3A and DNMT3B and has an essential role during gametogenesis [68,69].

2.3.3 Characteristics of the DNA methylation landscape in somatic cells

Highly and partially methylated domains

The mammalian genome is overall highly methylated, as introduced in the previous sections. However, already during embryogenesis, two types of megabase-scale methylation domains emerge across the genome, which are termed highly and partially methylated domains (HMDs and PMDs)

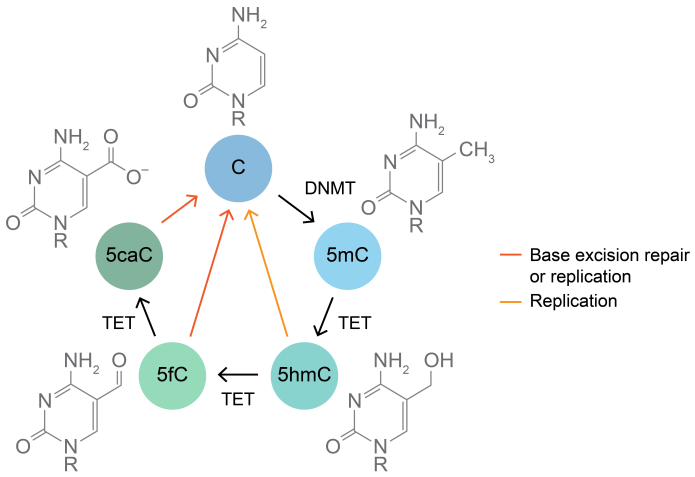

Figure 2.3.3: TET enzymes can induce active de-methylation: Hydroxylation of 5mC leads to 5hmC while further oxidization steps produce 5fC and 5caC. DNMT1 does not recognize the intermediate products, and therefore during replication, no methyl group is added to the respective cytosine on the newly synthesized strand. 5fC and 5caC can also be replaced by an unmethylated cytosine via the base excision repair pathway. This figure was adapted from Wu et al. [71].

(Figure 2.3.4). Although both domain types are highly methylated, PMDs exhibit slightly less methylation than HMDs, span more than half of the genome in total, and have been shown to be largely conserved across different tissues. In comparison to HMDs, PMDs are overall characterized by low gene and CG density and frequently overlap with late replication timing and lamina-associated domains (see section 2.1) [72,73]. Methylation in PMDs has been shown to decrease during aging, cell culture, and tumorigenesis (see section 2.5.2), and this loss of methylation positively correlates with the number of mitotic divisions. Specifically, isolated CpGs not surrounded by other CpGs nearby and directly flanked by either adenine (A) or thymine (T) termed solo-WCGW CpGs have been shown to be prone to gradual DNA methylation loss in comparison to other CpG contexts [72]. Highly transcribed genes within PMDs pose an exception to this phenomenon where CpGs independent of the sequence context are highly methylated [72,74].

The discovery of these isolated CpGs facilitated more accurate detection of PMDs: Due to the overall still highly methylated nature of both HMDs and PMDs in healthy tissues, segmentation algorithms based on all CpGs within one sample not always successfully identified PMDs. Solo-WCGW CpGs, however, are prone to loss of methylation even in healthy tissue and have been found to exhibit increasingly variable methylation levels when comparing different samples. This variability across samples allowed the detection of common PMDs in healthy and tumor cohorts, which facilitated the characterization of changes in DNA methylation during tumorigenesis (see section 2.5) [72,73].

CpG islands

The human genome comprises around 30,000 CGIs, which are classically defined using the CpG density and GC content of genomic regions [75,76]. A commonly used CGI definition introduced

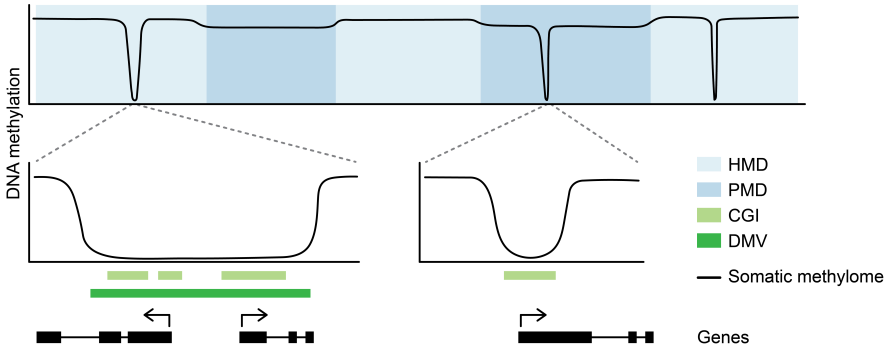

Figure 2.3.4: The somatic methylation landscape is characterized by high genome-wide DNA methylation as well as CGIs and DMVs that usually remain free of methylation. The genome can be separated into HMDs and PMDs, where PMDs exhibit slightly reduced methylation levels compared to HMDs, which is seemingly linked to the number of mitotic cell divisions.

by Gardiner-Garden and Frommer is based on the segmentation of the genome where CGIs are defined as segments with a GC content > 50 %, a length > 200 bp and a ratio of observed over expected CpGs > 0.6 [76]. According to this definition, 7.5% of the approximately 28 million CpGs in the human genome are located in CGIs. Alternative definitions and algorithms have been proposed considering different cut-offs, the identification of CpG clusters based on distances to neighboring CpG sites or hidden Markov models to detect CGIs in the genome [77–79]. CGIs are frequently located in gene promoters (around 70% of human gene promoters are associated with a CGI) but can also be found in intergenic regions where these so-called orphan CGIs have been reported to frequently act as enhancers [75]. CGIs located within gene bodies of actively transcribed genes are often methylated in order to silence alternative transcription start sites and ensure stable gene expression [80].

Methylation of promoter CGIs is rare, restricted to subsets with lower CpG density [56], and is linked to silencing of tissue-specific or germline genes [81–83]. However, in many cases, including X chromosome inactivation, DNA methylation has been reported to act as a secondary silencing mechanism where genes are silenced by chromatin modifications such as H3K27me3 first before methylation marks are deposited for long-term silencing [84, 85]. In contrast, promoter CGIs with high CpG density remain unmethylated even if the associated genes are not expressed. Instead, they are commonly repressed by PRC2 and the deposited H3K27me3 mark [56, 86]. In order to maintain an unmethylated state, *de novo* DNA methylation needs to be constantly repelled from these sites, which is mediated by transcription factors such as SP1 and histone modifications such as H3K4me3 that inhibits *de novo* methyltransferases [87–90]. Additionally, the de-methylation enzymes TET1 and TET3 are recruited to preferentially unmethylated DNA via their CXXC domain, and TET1 exhibits specific enrichment at CGIs with intermediate to high CpG density [91].

The regions flanking CGIs (usually defined as the neighboring two kilobases (kb)) are termed CGI shores, while the regions flanking CGI shores (usually two kb on the outer sides) are termed CGI shelves [92]. The function of methylation levels in CGI shores and shelves has yet to be fully understood. However, studies have shown that methylation in CGI shores is associated

with changes in gene expression. Specifically, tissue or cell type-specific methylation changes that negatively correlate with the expression of neighboring genes can also be found in shores of CGIs instead of or in addition to the island itself [93–95].

DNA methylation valleys

Developmental genes have been reported to frequently reside in larger genomic segments that are depleted of methylation but drastically extend beyond the boundaries of a CGI, which are termed DNA methylation canyons or valleys (DMVs). Promoters of genes located within DMVs are frequently associated with CGIs that are integrated into the larger structure of the respective DMV (Figure 2.3.4) [96–98]. DMVs span multiple kb (between five and 68 kb detected in human embryonic stem cells (hESCs)), mostly include at least one or multiple CGIs, are enriched for transcription factor binding sites, and depleted for repetitive elements. They are frequently shared across different cell types and mostly remain unmethylated across development and even in adult tissues, although a subset of DMVs partially gains methylation in a tissue-specific manner [96, 98]. Additionally, DMVs have been reported to be highly conserved across species such as human, mouse, and zebrafish [97].

Between 700 and 1500 DMVs have been identified when analyzing the methylome of somatic cell types [96–98]. Identifying these regions is commonly based on simple sliding window analyses where consecutive, unmethylated windows of a specific size are merged, and the resulting regions are termed DMVs [97]. Other studies have applied hidden Markov models to detect unmethylated stretches of the genome followed by a size cut-off to enrich for larger regions excluding smaller entities such as single CGIs or transcription factor binding sites [98].

Large DMVs are frequently marked by H3K27me3, extending until the respective regions' borders. Additionally, gene promoters residing in these DMVs are often marked by the activating histone modification H3K4me3, leaving genes in a bivalent state. These associated genes are, therefore, overall not or only very lowly expressed. In contrast, shorter DMVs are often marked exclusively by strong H3K4me3 signal, and genes associated with them are highly and constitutively expressed overall. The distribution of these two marks also underlies tissue- or cell type-specific effects depending on the genes that need to be active in a specific context [96, 98].

Both DNMT3A and PRC2 have been reported to be required to maintain the unmethylated state of DMVs. Loss of DNMT3A in mouse hematopoietic stem cells leads to corrosion of DMV borders with some DMVs extending and others shrinking via aberrant hypo- and hypermethylation respectively [98]. Knockout of PRC2 components in mouse embryonic stem cells and embryos results in aberrant hypermethylation of a subset of DMVs. However, this hypermethylation mostly does not affect CGIs embedded in the respective regions [97, 99].

2.4 Interplay between epigenetic modifications

Although epigenetic modifications can independently influence regulatory processes, as introduced in the previous sections, they can also facilitate the deposition of other epigenetic modifications to jointly act in genome regulation. The maintenance methyltransferase DNMT1 is re-

cruited to and directly interacts with H3K9me3 at heterochromatic regions to ensure stable silencing of associated repetitive elements and maintain genome integrity [100, 101]. The H3K36me3 histone modification that commonly marks active gene bodies interacts with the *de novo* methyltransferase DNMT3B, which is specifically recruited to these regions via its PWWP domain to ensure stable transcription [72, 102, 103]. This dynamic also explains why isolated CpGs within highly transcribed gene bodies in PMDs remain more highly methylated compared to other parts of PMDs: The erosion of methylation associated with accumulating cell divisions is counteracted by active targeting for *de novo* methylation [72]. Different types of histone modifications can also be involved in the recruitment and deposition of each other. Besides PRC2, which deposits the H3K27me3 mark, another type of Polycomb repressive complex, PRC1, deposits the H2AK119ub1 mark. A specific variant of PRC1 can deposit H2AK119ub1 at unmethylated developmental gene promoters [104, 105]. H2AK119ub1, in turn, is recognized by PRC2, which subsequently deposits H3K27me3. The canonical form of PRC1 can then recognize this mark. The two complexes thus act in synergy to establish and maintain a repressed state at their target genes [106, 107].

2.5 Epigenetic changes during tumorigenesis

As described in the previous chapter, Hanahan and Weinberg introduced the concept of cancer hallmarks in 2000 - universal properties that define the transition of healthy to malignant cells, including resistance to cell death, sustained proliferation, and evasion of anti-growth signaling. Molecular changes that lead to the establishment of these properties and cancer formation include activation of oncogenes, the long-term silencing of tumor suppressor genes, and chromosomal aberrations [1]. These properties are frequently established by genetic mutations of the respective genes. However, they can also be induced by epigenetic aberrations. Epigenetic regulators such as histone modifying enzymes, chromatin remodelers, and enzymes involved in the regulation of DNA methylation can be mutated in cancer leading to specific changes in the epigenome [4]. Additionally, non-mutational types of epigenetic reprogramming exist that are induced, for example, by changes in metabolism [3]. However, the cause and consequence of many epigenetic alterations in cancer have not been identified yet. In the following sections, epigenetic changes in tumors focusing on aberrant DNA methylation are introduced that lay the basis of the subsequent chapters in this thesis, specifically chapter 5 and 6.

2.5.1 Changes in chromatin organization and structure

Changes in chromatin conformation and organization are frequently found in cancer cells and include various (often tumor type-specific) mechanisms: Aberrant activity of chromatin remodelers can open chromatin, enabling the accessibility of transcription factors and other epigenetic regulators to previously inaccessible regions [108, 109]. Mutations or aberrant DNA methylation of TAD boundaries can lead to TAD structure disruption, enabling the activation of oncogenes by distal enhancers [30]. Specific mutations in histone (de-)acetyl- or (de-)methyltransferases can change the affinity, localization, or fidelity of these enzymes with tumor-promoting effects [110, 111]. For example, rearrangements of H3K4 methyltransferases of the MLL family lead to

the establishment of aberrant gene activation networks in leukemias resulting in oncogene induction [112]. Additionally, so-called oncohistones can be found in cancer defined by mutations in the respective histone tail, which inhibit the deposition of histone modifications at the mutation site [113]. The H3K27M oncohistone can be found in brain tumors and leads to a decrease in H3K27me3 due to the inability of mutated histone tails to be modified by PRC2 [114]. A different type of epigenetic alteration that causes long-term silencing of already unexpressed genes has been described, a phenomenon termed epigenetic switching. This model is based on observations that genes repressed by H3K27me3 can switch to a more stable silencing mechanism via H3K9 or DNA methylation in cancer, which could potentially ensure more stable repression of tumor suppressor genes [4, 5].

2.5.2 Aberrant DNA methylation

Global hypomethylation

Since the 1980s, researchers have observed characteristic DNA methylation changes in cancer cells across different tumor types. Specifically, in contrast to the highly methylated somatic genome, tumors were reported to exhibit global hypomethylation (Figure 2.5.1) [115–118]. Later studies defined that this loss of methylation preferentially and to a greater extent occurs in PMDs compared to HMDs [119–121]. A study using pan-cancer methylation data sets linked the loss of methylation in PMDs to the accumulation of mitotic cell divisions. Specifically, PMDs were shown to be largely conserved across healthy and tumor tissues and lose methylation already in healthy tissues during aging, which correlated with the corresponding mitotic history. Additionally, the level of PMD hypomethylation was linked to the accumulation of cell divisions across different cancer types.

These observations led to the hypothesis that late replication timing of PMDs reduces the time available for re-methylation by the maintenance DNA methyltransferase DNMT1 [72]. Progressive hypomethylation in PMDs might therefore reflect the inability of DNMT1 to fully remethylate these regions after replication, an effect that gets more pronounced in relation to the number of cell divisions. Highly proliferative tumor cells that evade cell death and undergo more cell divisions than healthy cells, therefore, would show a strong decrease in PMDs compared to healthy tissue [72]. However, it remains unknown whether the characteristic global loss of methylation in tumors represents a tumor-promoting effect or a byproduct of extensive cell division without functional consequences. A drastically different hypothesis was raised by Johnstone et al., who associated global loss of methylation with chromatin reorganization in colon cancer. These topological changes were associated with the repression of oncogenic genes involved in stemness, metastasis, and invasion, reminiscent of a tumor-suppressive state. The study proposed that these epigenetic reconfigurations might reflect a cellular defense mechanism and that the respective observations in cancer might not be linked to tumorigenesis itself [122].

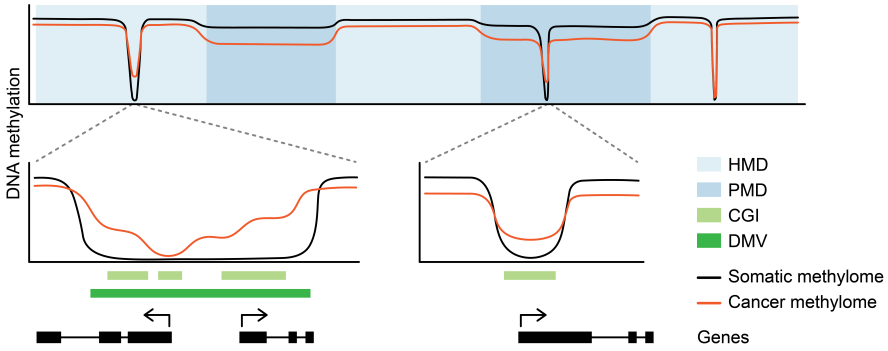

Figure 2.5.1: The methylome of tumor cells is commonly characterized by loss of methylation preferentially in PMDs as well as select CGI hypermethylation.

CGI hypermethylation

In addition to the global decrease in methylation, numerous studies have shown that select previously unmethylated CGIs gain methylation in tumors (Figure 2.5.1) [4, 123, 124]. Although the number and identities of methylated CGIs differ across tumor (sub-)types, CGIs that gain methylation in tumors are frequently targeted by PRC2 in healthy tissues and marked by the corresponding H3K27me3 histone modification. These CGIs and corresponding genes (if the CGI is located in a promoter) are, therefore, already repressed by a different mechanism in healthy cells and remain continuously silenced in tumors [125–127]. Some of these genes are known tumor suppressor genes, which are hypothesized to be more permanently silenced by DNA methylation as part of the previously introduced epigenetic switch model [4, 126]. However, many promoter CGIs targeted for hypermethylation in tumors are linked to developmental genes that are not known to promote tumor-suppressing effects and are not expressed within the cell of origin [128].

Although some tumor types are known to exhibit mutations in epigenetic regulators such as DNA methyltransferases and TET enzymes, these aberrations are not universally observed across cancer types despite the characteristic CGI hypermethylation phenomenon [129]. Additionally, the gain of methylation in some tumors has been reported to exhibit intra-sample heterogeneity where instead of fully methylated alleles, stochastic methylation patterns could be observed across tumor cells [130–132]. Therefore, different models have been proposed to explain the almost universal occurrence of CGI hypermethylation in cancer and its intrinsic features. One model suggests that CGIs previously protected in healthy tissues get exposed to stochastic *de novo* methylation as a consequence of changes in chromatin configurations. Differences in the level of hypermethylation could be explained by tissue-specific chromatin states that result in differential resistance to *de novo* methylation [130]. Chromatin marks implicated in such a transition of epigenetic states are stochastic loss of the PRC2-deposited mark H3K27me3 at these CGIs and the emergence of facultative heterochromatin marked by H3K9me3 together with heterogeneously deposited DNA methylation [4, 5, 126, 128].

A slightly different model assumes that CGI methylation levels result from constant DNA methylation turnover caused by *de novo* methyltransferases and TET enzymes, an equilibrium that can be biased in different directions [130]. Regions with low methylation would therefore be subject

to more efficient DNA methylation removal, while a gain in methylation as observed in tumors would be the result of a shift in the balance between TETs and DNMTs in favor of *de novo* methylation. Finally, it has been hypothesized that both the emergence of PMD hypomethylation as well as PRC2 target CGI hypermethylation could be explained by clonally propagated methylation patterns combined with context-specific rates of DNA methylation turnover [133].

CpG island methylator phenotype

In some tumor types, a so-called CpG island methylator phenotype (CIMP) has been characterized, which affects subsets of patients (CIMP-positive) exhibiting increased CGI methylation levels compared to CIMP-negative patients. CIMP was first described for colorectal cancer and linked to microsatellite instability (mismatch repair deficiency) and mutations in the *BRAF* gene [134, 135]. Additionally, CIMP was associated with different demographic and clinical features such as sex, age, response to treatment, or tumor location in the colon [135]. However, how exactly CIMP is established in colorectal cancer patients remains unknown.

Following the description and early investigations in colorectal cancer, CIMP subtypes have been defined for multiple other tumor types, including glioma, different types of leukemias, melanoma, endometrial and breast cancer [136–141]. However, the definition of CIMP varies widely across cancer types and is commonly based on the methylation level of study- and tumor-specific, variable CpGs located in a subset of CGIs. Therefore, although CIMP has been linked to molecular characteristics and outcomes in each of these tumor types, it remains unclear whether this reflects a pan-cancer phenomenon subjected to similar underlying regulatory principles, also given that the associated covariates do not always align across tumor types. In particular, CIMP was reported to be associated with better (e.g., breast cancer, colon cancer in females, T cell acute lymphoblastic leukemia) as well as poorer prognosis (e.g., endometrial cancer, advanced stage melanomas, renal cell carcinoma) [142]. So far, only one direct cause of CIMP has been identified in gliomas as well as acute myeloid leukemias: Mutations in *IDH1* and *IDH2* cause a decrease in the efficiency of TET enzymes and thus elevated CGI methylation levels [142, 143]. However, other tumor types do not exhibit mutations in these genes, and other respective causal molecular features have not yet been identified [142].

DMV hypermethylation

Similar to the CGIs located within them, DMVs can aberrantly gain methylation in cancer. Pan-cancer investigations have shown that DMVs are prone to hypermethylation across different tumor types and that the genes located within these regions are enriched in oncogenes and genes containing homeobox elements, which are transcription factors involved in cell growth and differentiation [144]. Hypermethylation of DMVs - specifically in the gene body of genes that are usually not or only lowly expressed in healthy tissues - has been shown to be positively associated with gene expression. Aberrant methylation gain within DMVs can therefore induce abnormal expression of genes that potentially provide tumor-promoting effects [144]. Additionally, changes in DMV borders in mouse hematopoietic stem cells upon loss of DNMT3A are associated with expression changes of genes that are also implicated in human leukemias [98].

DNA methylation changes in cultured cancer cells

The establishment of immortalized cancer cell lines has offered researchers a critical tool to study molecular features of cancer, and their response to treatments outside patients [145]. Cancer cell lines and immortalized fibroblasts have been the basis for many epigenetic models of tumorigenesis. Specifically, the epigenetic switch model as well as models of stochastic methylation gain at CGIs have been established using experiments conducted in these *in vitro* model systems [126, 130, 133].

However, studies from the early 2000s reported deviating DNA methylation landscapes between primary tumors and respective cancer cell lines. By inspecting 1,184 CGIs in 114 primary tumors and 24 cancer cell lines of different types, Smiraglia et al. showed that although CGIs targeted for hypermethylation in tumors are also hypermethylated in cell lines, methylation levels rise higher in cell lines compared to tumors. They also found that the CGI methylation levels in cell lines were not uniform but reflected the tumor of origin to a certain degree: If a tumor type exhibited higher methylation levels than another, this relationship was also visible in the corresponding cell lines. Additionally, more CGIs were hypermethylated in cell lines compared to tumors, which the authors attributed to a culture-induced effect independent of cancer itself [146].

Another study compared 70 cancer cell lines and 233 primary tumors from 12 different cancer types at 15 CGIs and also found elevated hypermethylation of cell lines compared to primary tumors. However, a signature of origin could be maintained as cell lines from the same tumor type clustered together based on their CGI methylation profile. Additionally, cell lines profiled in this study were more globally hypomethylated than corresponding primary tumors when measuring the overall methylation content [147]. Later studies confirmed this for both cancer and non-cancer immortalized cell lines [72]. In summary, although frequently used as a model system for epigenetic changes in tumors, cancer cell lines have been early on reported to deviate from corresponding primary tumors specifically based on their DNA methylation profile. A thorough investigation of the universality of these observations across cancer cell lines and potential implications is presented in chapter 6.

2.6 Methods for quantifying epigenetic features

2.6.1 Chromosomal architecture

The read-out of chromosomal interactions can be divided into imaging- and sequencing-based methods. The latter can be further grouped into ligation-based chromosome conformation capture and ligation-free techniques [148]. Imaging-based methods include fluorescence *in situ* hybridization of DNA (DNA-FISH) based on fluorescence-labeled probes designed to hybridize to genomic loci of interest. These can then be visualized in cells using microscopy, allowing the calculation of physical distances between two or multiple loci. However, DNA-FISH is limited by the number of loci that can be visualized at the same time and the inability to accurately measure short-range distances [148]. Ligation-based methods such as 3C, 4C, 5C, and Hi-C measure the chromatin interaction frequencies between genomic loci using sequencing. For this purpose, the chromatin is cross-linked, followed by the digestion of the DNA using restriction enzymes.

The digested fragments are then ligated, and the cross-link is reversed. The different ligation-based assays, which employ distinct additional polymerase chain reaction (PCR) or labeling and fragmentation steps before the actual sequencing, provide the read-out of different types of interactions: Between two known genomic loci (3C), between one known locus and the remaining genome (4C), between a larger genomic region and the remaining genome (5C) and between all pairs of genomic loci (Hi-C) [148]. The pair-wise nature of ligation does, however, limit these techniques, making it difficult to infer if multiple loci are interacting as an ensemble. Newer ligation-free methods such as genome architecture mapping (GAM) or split-pool recognition of interactions by tag extension (SPRITE) have been developed based on cryosectioning and tagging cross-linked chromatin with barcodes, respectively [15, 149].

2.6.2 Histone modifications

The localization of histone modifications in the genome is mainly determined using antibodies specific to the protein of interest. The most commonly used technique to quantify this localization is chromatin immunoprecipitation (ChIP) followed by sequencing (ChIP-seq). Here, DNA is fragmented and selectively enriched based on the proteins attached to it, which are recognized by specific antibodies tailored to histone modifications or other proteins, such as transcription factors [150]. The DNA fragments associated with the protein of interest are then purified, followed by library preparation, and sequencing [150]. The resulting sequencing reads enrich at the genomic locations where the histone modification or transcription factor of interest was localized in the profiled cell population. Other more recently developed sequencing-based methods include cleavage under targets and tagmentation (CUT&Tag) or cleavage under targets and release using nuclease (CUT&RUN), which are based on antibody-guided cleavage of DNA using transposases or nucleases [151, 152]. In contrast to ChIP-Seq, these techniques require less input material and have a higher signal-to-noise ratio, thus requiring fewer sequencing reads and lowering the overall cost [151, 152].

Besides sequencing-based approaches, the overall abundance of histone modifications can be determined experimentally using western blots with specific antibodies recognizing the histone modification or mass spectrometry, which allows to quantify more than 60 modifications simultaneously in the same sample [153].

2.6.3 DNA methylation

Measuring DNA methylation is commonly based on one of three principles: Affinity enrichment, restriction enzymes, or bisulfite conversion. Affinity enrichment methods include methylated DNA immunoprecipitation followed by sequencing (MeDIP-Seq) or methyl-binding domain sequencing (MBD-Seq), which work similarly to ChIP-Seq (see previous section). MeDIP uses a specific antibody against 5mC while MBD-Seq makes use of proteins that recognize methylated DNA and corresponding antibodies [154]. However, the resolution of these methods cannot capture per-base methylation information and is often biased by the CpG density [154]. Restriction enzyme-based methods such as the HpaII tiny fragment enrichment by ligation-mediated PCR (HELP) assay make use of enzymes that cut specific sites (in this case 5'-CCGG-3') either

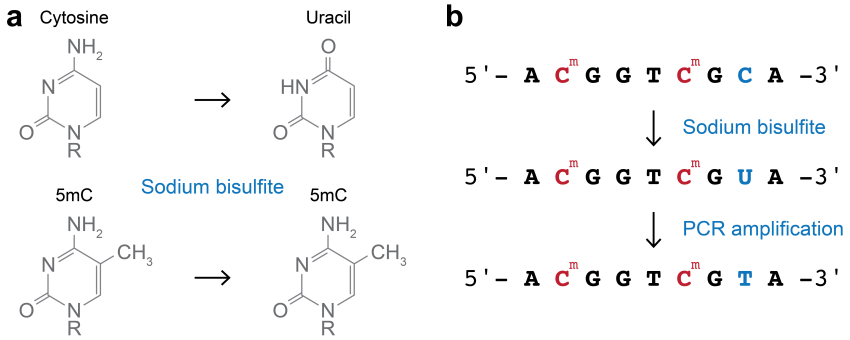

Figure 2.6.1: a) Treatment with sodium bisulfite leads to the conversion of unmethylated cytosine to uracil, while methylated cytosines remain unchanged. b) After bisulfite treatment, the DNA gets amplified (usually via PCR for sequencing-based methods), producing thymines at the position of previously unmethylated cytosines.

methylation-independent (MspI enzyme) or only if the included CpG is unmethylated (HpaII enzyme) [155]. Comparing the fractions of DNA cut by both enzymes then gives an estimate of the methylation rates at the cut sites, which are, however, limited across the genome [155, 156].

The treatment of single-stranded DNA with sodium bisulfite leads to the deamination of unmethylated cytosines to uracils while methylated cytosines remain unchanged. Subsequent amplification of the DNA leads to the synthesis of thymines instead of uracils (Figure 2.6.1) [157]. Therefore, the methylation status of cytosine at a specific position can be determined based on the conversion frequency caused by bisulfite treatment, and this information can be captured using microarrays or sequencing [154, 156]. Third-generation sequencing technologies such as Nanopore sequencing also allow the direct read-out of 5mC. Here, base identities are determined using an electronic current, which is sensitive to modifications of the DNA and allows to distinguish unmethylated from methylated cytosine [158]. Additionally, these so-called long-read sequencing technologies provide the advantage of generating reads in the range of multiple kilo- to megabases, which allow the evaluation of highly repetitive regions that would be difficult to assess with short reads [159].

In this thesis, data sets obtained from different technologies based on bisulfite conversion were used. Therefore, in the following sections, the read-outs that form the basis of the subsequent chapters are introduced.

Methylation microarray

Methylation microarrays are based on slides that contain hybridization probes that allow specific DNA molecules to bind. The most commonly used and established methylation arrays are distributed by Illumina - the Infinium HumanMethylation450 BeadChip (also termed 450k array in this thesis) and the next generation Infinium MethylationEPIC BeadChip [92, 160]. The probes hybridize to bisulfite-converted DNA, and each probe detects the methylation rate of a single cytosine (mostly in a CpG context). The two arrays differ mainly in the number of CpGs that they can measure (more than 480,000 and 850,000 CpGs, respectively, Table 2.6.1 [92, 160].

Two different probe designs exist on both arrays in order to measure the methylation of a single cytosine:

1. Infinium I: Each probe consists of two bead types that span 50 bases where the 3' end corresponds to the CpG of interest. One of the probes is designed to match the methylated cytosine (reflected by a C after bisulfite conversion), while the other probe is designed to match the unmethylated cytosine (reflected by a T after bisulfite conversion and subsequent amplification using multiple displacement amplification). Other CpGs spanned by the sequences of the beads are assumed to have the same methylation state (matching methylated or unmethylated) as previous studies have shown that the methylation status of neighboring CpGs within a close neighborhood is highly correlated [161]. The beads get extended using a fluorescence-labeled nucleotide that matches the base next to the CpG of interest in the DNA fragment that binds to the bead [92].

2. Infinium II: Each probe contains one bead type (50 bases) that allows both unmethylated and methylated sequences to bind using degenerate bases at the cytosine position of CpGs, which bind both C and T. Considering the entire 50 bp sequence, up to three CpGs can be included using this technique. The methylation status of the CpG of interest is determined by the extension of the bead by a single base at the position of the hybridized methylated or unmethylated cytosine: In the case of a methylated cytosine, a fluorescence-labeled G is introduced while an A labeled with a different fluorophore is introduced in case of an unmethylated cytosine (reflected as a T). This probe type has the advantage that it does not rely on the assumption that neighboring CpGs exhibit the same methylation state, which might not always be true [92].

The single-stranded bisulfite-converted DNA is hybridized to the probes, and fluorescence intensity is measured using staining. In the case of the Infinium I probes, the fluorescence intensity of both beads is compared, and depending on the intensity ratio, a so-called beta value that reflects the methylation rate between 0 (unmethylated) and 1 (methylated) can be obtained. For Infinium II probes, the intensity of the different fluorophores that mark methylated or unmethylated CpGs is compared to achieve the same outcome [92].

The probes for both arrays were designed in order to cover regulatory regions of interest, such as CGIs, promoters, gene bodies, and enhancers. Although only a small fraction of the genome can be covered with such an assay, the covered CpGs are consistent across samples. Additionally, methylation arrays are comparably cheap, which makes them a useful tool when studying large cohorts [92].

Whole-genome bisulfite sequencing

Whole-genome bisulfite sequencing (WGBS) is based on sodium bisulfite treatment of the DNA, followed by whole-genome sequencing. Two main protocol types exist: directional and non-directional. Directional refers to only the original forward and original reverse strand being sequenced, while non-directional protocols generate reads from both the original forward and reverse strand but also their respective reverse complements (Figure 2.6.2). MethylC-Seq introduced by Lister et al. is a commonly used directional sequencing protocol [10]. Here, the DNA

Number of probes	HumanMethylation450	MethylationEPIC
Total	485,577	866,895
CpG	482,421	863,904
Non-CpG	3,091	2,932
Random single-nucleotide polymorphism	65	59

Table 2.6.1: Number of probes provided by the Infinium HumanMethylation450 and the Infinium MethylationEPIC BeadChip [92, 160].

is first fragmented, and universal adapters are ligated, followed by bisulfite conversion. The adapters, therefore, have to be fully methylated such that specific primers, which are complementary to the universal adapters, can recognize them in the subsequent amplification. After amplification, the library is subjected to sequencing [10]. This yields two types of reads, one originating from the forward strand (+OS) and one originating from the reverse strand (-OS), where unmethylated cytosines are replaced by thymines (Figure 2.6.2).

In contrast, the non-directional BS-Seq protocol developed by Cokus et al. includes first the ligation of double-stranded unmethylated adapters to the DNA that contain restriction sites recognized by the enzyme DpnI [11]. The DNA is then treated with sodium bisulfite, which also affects the previously unmethylated adapters, followed by PCR with primers that recognize the converted adapter sequences. The resulting double-stranded DNA is then digested with DpnI to remove the first set of adapters, and the actual sequencing adapters are ligated [11]. This two-step approach was designed to remove sequences that might be affected by incomplete bisulfite conversion, which would also affect the sequence of the first adapters. However, it also leads to the generation of four different types of reads that are obtained after sequencing: the original forward and reverse strands as well as their reverse complements, which all differ from each other due to the bisulfite conversion (Figure 2.6.2) [11]. Notably, directional bisulfite sequencing using paired-end reads also leads to the generation of all four read types as the second mate of the pair is obtained from the reverse complement of the fragment (Figure 2.6.2). Nevertheless, both mates carry methylation information of the same fragment they originate from, which is important for subsequent methylation calling (see section 2.7).

Following MethylC-Seq and BS-Seq, other protocols have been developed that aim to optimize different parts of the library preparation process for bisulfite sequencing. Post-bisulfite adaptor tagging (PBAT) is a method to profile whole-genome DNA methylation that can work without PCR or be adapted for experiments with low quantities of DNA as input material, including single-cell sequencing [162]. PBAT circumvents the need for PCR amplification based on the fact that sodium bisulfite treatment is known to be aggressive and can cause the DNA to break. In regular protocols, this means that many fragments with already ligated adapters (sequencing templates) are lost if damaged. In order to increase the input for sequencing, the remaining intact sequencing templates are then amplified by PCR. In contrast, during the PBAT protocol, the adapter ligation follows the bisulfite treatment, thereby omitting this effect [162]. This technique, therefore, has the advantage that only a low amount of input DNA is required as sequencing templates remain intact and that the later sequencing of PCR artifacts that need to be removed during computational processing can be avoided (see section 2.7). However, it has been shown that PBAT

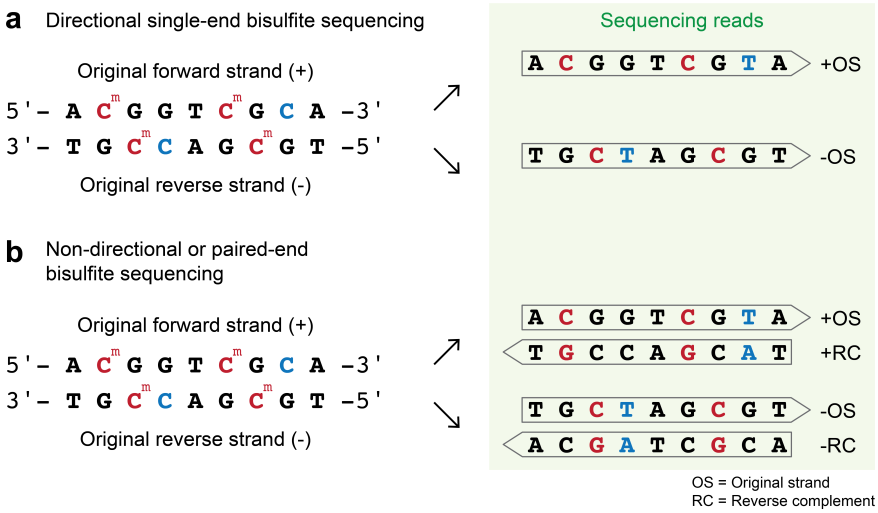

Figure 2.6.2: a) Directional single-end bisulfite sequencing leads to the generation of reads only from the original forward and reverse strand where unmethylated cytosines are reflected by C (reference) to T (read) mismatches. b) For non-directional and paired-end sequencing, four read types resulting from the original two strands and their reverse complements are generated. Unmethylated cytosines are reflected by C to T mismatches for reads resembling the original forward and reverse strands and by G to A mismatches for reads originating from their reverse complements. This figure was adapted from Kretzmer [70].

can lead to the generation of chimeric read pairs during paired-end sequencing. These chimeric read pairs then need to be filtered out during the computational processing of PBAT libraries, which reduces the number of usable reads for later analyses [163].

All bisulfite sequencing libraries suffer from low base composition complexity due to the reduced variability of bases in the sequences (most cytosines are unmethylated in mammalian genomes and therefore converted). This can have an effect on data quality and yield from the sequencing instrument [164]. Thus, higher complexity sequences need to be spiked in for the sequencing of bisulfite-converted libraries, which can either be other library types such as whole-genome sequencing (WGS) or PhiX, a phage that is used to create control libraries with high complexity [165]. Including PhiX as spike-in, however, leads to higher sequencing costs as essentially a certain fraction of reads stems from the spike-in and not the actual samples, and more fragments need to be sequenced to reach a desired coverage [165].

Reduced representation bisulfite sequencing

Reduced representation bisulfite sequencing (RRBS) was introduced by Meissner et al. and offers the possibility to sequence DNA methylation at single-base resolution, but limited to roughly 10% of CpGs in the genome [86, 166]. For this protocol, the DNA is digested with the restriction enzyme MspI that cuts both methylated and unmethylated 5'-CCGG-3' sites upstream of the CpG (Figure 2.6.3). Often a size selection is subsequently applied to enrich for fragments in the range of 40-220 bp. As the fragment size correlates with CpG density, the resulting library is compara-

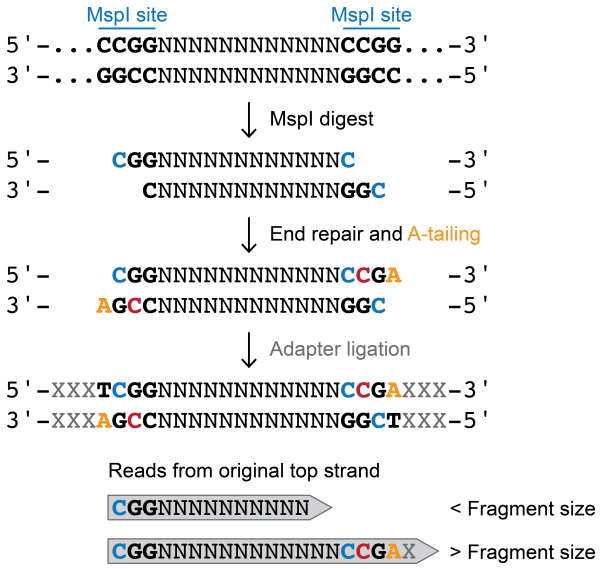

Figure 2.6.3: The different steps of the RRBS protocol. This figure was adapted from the Babraham RRBS guide [167].

tively enriched for features such as CGIs and specific retrotransposon families [86]. The resulting DNA fragments are end-repaired, which introduces an artificial (generally unmethylated) cytosine at the 3' end of both strands. Subsequently, a single adenine is added to the 3' end of both strands in order to enable the ligation of methylated sequencing adapters [86, 167]. Specific kits such as the Ovation RRBS Methyl-Seq System offer so-called diversity adapters that are ligated before the actual sequencing adapters and introduce artificial sequence variability to reduce the fraction of PhiX that needs to be spiked-in [168]. Afterwards, the DNA is bisulfite-converted, amplified, and sequenced (similar to MethylC-Seq) [86, 169].

Due to the small fragment sizes enriched in the RRBS library preparation, sequencing with paired-end is not recommendable as the mates will frequently overlap, leading to a duplicated measurement of the same fragment that could bias later analysis steps. Even a single read might already extend over the length of the entire fragment leading to the sequencing of the adapter sequence and the preceding artificial cytosine that is introduced during the end-repair process (see section 2.7) [167].

Single-cell bisulfite sequencing

Both WGBS and RRBS protocols have been adapted to profile the methylomes of single cells. Here, the greatest limitation is the availability of input DNA material within a single cell compared to bulk sequencing experiments. Therefore, PBAT is frequently used or adapted for whole-genome methylation profiling of single cells due to low input material requirements and the circumvention of sequencing template loss introduced by bisulfite treatment after adapter tagging in standard protocols [170]. Additionally, the RRBS protocol has been optimized in various studies to minimize loss during specific steps such as DNA purification. This allows the use of a more cost-efficient sequencing method adapted for single cells [170].

2.7 From bisulfite sequencing to analysis

As the focus of this thesis lies on DNA methylation-based read-outs, specifically using bisulfite sequencing, in the following sections, common computational processing steps are described that are essential to extract methylation rates from bisulfite sequencing reads. Reads from RRBS or WGBS experiments are subjected to similar processing steps as other next-generation sequencing experiment types, such as WGS or RNA-Seq, before methylation rates can be obtained. However, the effects of the bisulfite conversion have to be taken into account, and processing steps such as read alignment have to be adapted in comparison to other sequencing types. In the following, common processing steps are described, and available tools for bisulfite sequencing data processing are introduced. The specific tools and parameters used within the studies described later in this thesis and processing steps for other types of sequencing data sets (e.g., RNA-Seq) are described in the respective chapters.

2.7.1 Pre-processing

As a first step, the quality of newly generated data sets is assessed. FastQC is a commonly used tool for this purpose, which allows scanning read files in FASTQ format and provides different metrics such as the per-base sequence quality and content, the sequence length distribution, and potential adapter content [171]. These metrics are helpful to inspect to judge whether the sequencing experiment worked as expected and whether additional data cleaning steps should be included in downstream processing, such as the trimming of reads due to low quality or adapter content. Notably, the effect of bisulfite conversion can be observed at this stage already: the majority of cytosines in mammalian genomes are unmethylated since they do not occur in a CpG context, which leads to an almost complete depletion of cytosines or guanines (Gs, depending on the read type and protocol, see section 2.6) in the respective sequences.

Following quality control, trimming low-quality bases or parts belonging to the sequencing adapters can be executed using tools such as cutadapt or trimmomatic [172,173]. In addition, the last two bases at the 3' end of a read prior to the adapter sequence can be clipped for data sets generated using the RRBS protocol to ensure that artificially integrated cytosines are removed from the read. These could later bias the quantification in the case that downstream tools cannot account for this (see following sections).

2.7.2 Alignment

Reads need to be compared to the sequence of a reference genome to determine the origin of every read, a process that is called alignment. This is the central step of every next-generation sequencing processing pipeline. Many tools exist that map short or long reads against a genomic reference genome using data structures such as FM indices and hash tables [174–176]. The search for the location of origin is frequently based on "seed and extend" approaches: First, a substring of a read is searched with no or few errors, and only resulting promising hits are subjected to an extended alignment of the whole sequence using (potentially user-defined) scoring schemes.

Alignment tool	Data structure	Seeding	Extension
Bismark [180]	FM index (Bowtie2 [174] or HISAT2 [181])	3 letter	3 letter
BS-Seeker2 [182]	FM index (Bowtie2)	3 letter	3 letter
BSMAP [183]	Hash table (SOAP [176])	Wild card	Wild card
BWA-meth [184]	FM index (BWA [175])	3 letter	3 letter
GEM3 [185, 186]	FM index	3 letter	3 letter
segemehl [179]	Enhanced suffix array	3 letter	Wild card

Table 2.7.1: List of common bisulfite alignment tools together with the underlying data structures and alignment strategies used.

Reads obtained from bisulfite sequencing require specific alignment tools due to the conversion of unmethylated cytosines to uracils (which are amplified as thymines). In contrast to standard alignment tools, mismatches from C (reference) to T (read) must be considered matches for reads originating from the original forward or reverse strand, while T to C mismatches should still be penalized. Likewise, if reads originate from the reverse complements of the original forward and reverse strands, mismatches from G (reference) to A (read) should be handled as matches. G to A conversions must only be considered for non-directional protocols or directional paired-end sequencing. For WGBS, however, paired-end sequencing is commonly used, which is why in the following, both types of conversion effects will be considered. Two main strategies exist in order to tackle the alignment problem:

1. Three-letter alphabet: Both the reads and the reference genome are converted to a three-letter alphabet, and a regular alignment is executed. To account for reads aligning to the forward and the reverse strand, this needs to be implemented twice, once comprising the letters A, G, and T where all C's are converted to T's and once comprising the letters A, C, and T where all G's are converted to A's [177, 178]. This has the advantage that it can be employed using already existing regular alignment tools such as Bowtie2 or BWA [174, 175]. However, the reduced complexity of a three-letter alphabet can lead to many false positive hits.

2. Wild card: Two variants of this strategy exist. Either both C's and T's are allowed to align to a C in the reference genome (and vice versa for G's and A's for reads originating from the reverse complements of the original forward and reverse strands), or all possible combinations of C's and T's instead of the sequenced T's or G's and A's instead of the sequenced A's are considered. The latter results in a "combinatorial explosion" and might lead to extended runtime to align all possible resulting reads [177, 178].

The read alignment tool segemehl combines both strategies to produce accurate and feasible alignments: Seeds are searched with a three-letter alphabet while valid hits are extended afterwards using a wild card approach where C to T mismatches (or G to A mismatches) are not penalized, which increases the sensitivity of the alignment process (Table 2.7.1) [179].

Some applications, such as BSMAP, also offer a special RRBS mode in which only genomic positions are considered for the alignment that start with the restriction site of the enzyme used in

the digest (typically MspI, see section 2.6) [183]. This reduces the runtime substantially, as only a small fraction of the genome needs to be scanned for potential hits.

Following the alignment, a deduplication step may be applied to remove or flag reads that likely originated from the same fragments as other reads and, therefore, might represent PCR artifacts. GATK offers a commonly used deduplication utility, which is based on the comparison of identical read alignments at the same position [187]. This step is useful to remove reads from the subsequent methylation rate quantification that do not represent biologically different fragments compared to other reads at this position. Considering duplicates for the estimation of methylation rates could artificially bias the quantification as fragments of the same allele appear multiple times. This deduplication strategy is not possible for reads originating from experiments profiled using RRBS due to the restriction enzyme digest: The resulting reads always start at the same genomic positions marked by the restriction site of the enzyme used during library preparation. Therefore, specific RRBS library kits offer the option to use unique molecular identifiers (UMIs) to label single fragments before the PCR step. The labeling can then be used to identify duplicates as PCR products of a specific fragment are marked by the same UMI [168]. Otherwise, it is not possible to properly deduplicate reads sequenced using RRBS.

2.7.3 Calling methylation rates

After alignment and potential deduplication, the methylation rates of single cytosines need to be determined. Although a single cell with two alleles can only exhibit three different methylation values or fractions of methylated reads (Figure 2.7.1, 0 = both alleles are unmethylated at a given cytosine, 1 = both alleles are methylated, 0.5 = one allele is methylated, and one allele is unmethylated), a bulk bisulfite sequencing sample contains reads originating from fragments sampled from a large cell population (Figure 2.7.1). Depending on the underlying cells, the methylation across reads (= alleles) can vary between 0 and 1 (e.g., only 30% of reads could be methylated at a given cytosine due to a cell type mixture or sample heterogeneity leading to a methylation rate of 0.3). Methylation rates obtained from bulk bisulfite sequencing, therefore, always reflect an average measurement across the population. However, since every CpG has its own population-wide estimate, it is not clear what the underlying methylation patterns across multiple CpGs are that lead to a specific methylation rate: A methylation rate of 0.5 at a given CpG could stem from two distinctly methylated cell populations or stochastic methylation distributed across reads at a given locus (Figure 2.7.1).

All reads that cover the position of a specific cytosine are considered, and the read sequence is compared to the reference genome to determine the cytosine's methylation status for every read. Reads that originate from the original forward strand or its reverse complement independent of the strand they align to (+OS and +RC, Figure 2.6.2) contain the information about cytosines on the forward strand, while reads that originate from the original reverse strand or its reverse complement (-OS and -RC) contain the information of cytosines from the reverse strand. For +OS reads, the read sequence needs to be compared to the forward strand where C to T mismatches indicate an unmethylated status of cytosines, while +RC reads need to be compared to the reverse strand where G to A mismatches indicate no methylation. This works analogously for -OS and -RC, where -OS is compared to the reverse strand considering C to T mismatches, while -RC

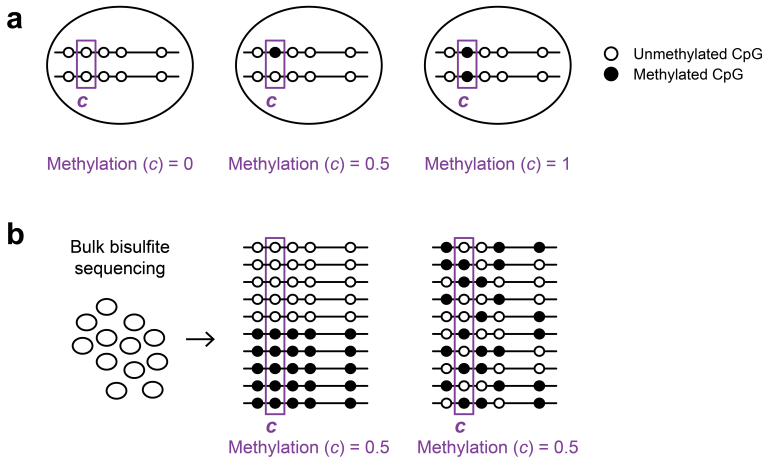

Figure 2.7.1: a) The methylation status of a CpG in a single diploid cell can be either equal to zero (both alleles are unmethylated at this position), one (both alleles are methylated at this position), or 0.5 (one allele is methylated, the other is unmethylated at this position). b) Considering bisulfite sequencing of a bulk of cells, a number of fragments sampled from the underlying cell population get sequenced at a specific CpG or locus. The methylation rate, therefore, presents an average across this population. However, it is not clear what the underlying methylation patterns across multiple CpGs are that lead to a specific methylation rate. For example, a methylation rate of 0.5 can be achieved by two different cell populations that are consistently unmethylated or methylated (left) or a stochastic methylation pattern across reads (right).

reads are compared to the forward strand considering G to A mismatches. This way, the number of reads with unmethylated and methylated status (N_u and N_m) of a specific cytosine c can be determined, and the methylation rate can be calculated:

$$\text{Methylation}\ (c) = \frac{N_m}{N_u + N_m} \tag{2.1}$$

As methylation of cytosines preferentially occurs in a CpG context in mammals, methylation calling is commonly restricted to these positions. However, it can also be applied to non-CpG cytosines. In contrast to non-CpGs, CpGs are symmetric on both strands, which allows DNMT1 to copy the methylation information from one strand to the other after replication (Figure 2.6.2). Therefore, for CpGs, the methylation rates of the forward and reverse strands are frequently combined as they likely reflect the same methylation state within one cell (except just after replication). This strategy allows to increase the coverage when reads originating from both strands can be considered. Standalone tools that are able to calculate the methylation rates from aligned reads are, for example, mcall (compatible with BSMAP alignments) or MethylDackel (compatible with segemehl output) [188, 189]. When using mcall, alignment with BSMAP in RRBS mode is automatically detected, and potential artificial CpGs at the 3' end of reads are omitted for the quantification, which means that trimming these bases before the alignment would not be necessary for this specific tool. Other alignment tools, such as Bismark, already offer built-in options to calculate methylation rates immediately following the alignment.

Part II

Tools for bisulfite sequencing data

In the second part of this dissertation, new tools to process bisulfite sequencing data sets are presented that aim to fill gaps in existing applications and concepts. First, the adaptation of an existing local alignment search tool to incorporate bisulfite-converted sequences is introduced, allowing contamination screening and enabling future metagenomic studies based on DNA methylation read-out. Second, a tool for fast extraction of DNA methylation heterogeneity scores from bisulfite sequencing reads is presented, providing a scalable solution for analyzing methylation heterogeneity on the read level.

Chapter 3

Local alignments for bisulfite-converted sequences

Lambda3 is a local alignment tool for protein as well as nucleotide searches in large sequencing databases developed by Dr. Hannes Hauswedell and implemented in C++ using different libraries for sequence analysis [190, 191]. In this chapter, the adaptation of Lambda3 to include the search of bisulfite-converted sequences is described, and performance compared to classical bisulfite-aware read aligners is assessed.

3.1 Introduction

Read alignment, as described in section 2.7, is based on the search of sequencing reads in databases (usually reference genomes) to find (near-)exact matches of the complete reads in the reference. Here, the objective is to determine the genomic origin of each read [192]. Therefore, semi-global alignments are used that allow a limited amount of differing bases between the read and the reference genome in order to account for potential technical errors or genetic variation within a species (Figure 3.1.1). After matches of each read in the reference genome are located, the best-scoring hit is usually reported based on the assumption that sequencing reads originate from a single genomic location. With the exception of repetitive elements or genomic regions that have not been assembled yet, this location should be uniquely identified in most cases.

In contrast to read alignment, homology search represents a different application of query search in even larger databases. Here, the goal is not to find the exact genomic location of a query (read) but instead to identify sequences of common evolutionary descent often across different species. Therefore, the databases used for the search usually contain not only one reference genome but multiple or even extend to entire collections of sequence databases such as GenBank or RefSeq [193]. In this context, it is less relevant to align a complete sequencing read but instead locate less exact matches or subparts of the read that might be evolutionary conserved. Homology search is therefore often applied in metagenomic studies where samples are not associated with a single organism but instead comprise multiple known or unknown species due to the sampling method or sources of contamination [193]. Instead of semi-global alignments that aim to map

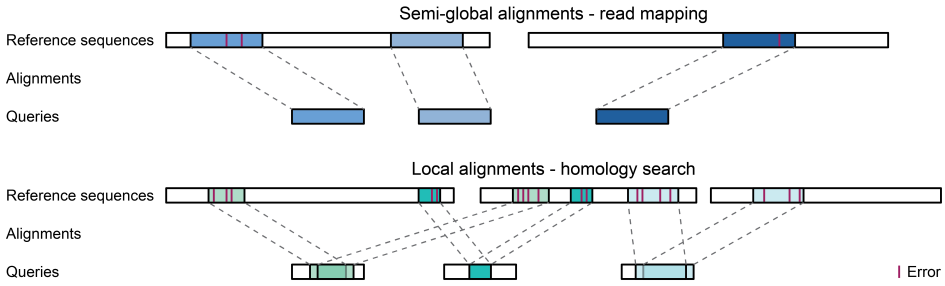

Figure 3.1.1: Semi-global alignments aim to align the complete query (read) sequence to a reference genome with minimal error rates (top). Local alignments require only parts of the query to map to one or multiple reference sequences while generally allowing for more errors, which is desirable for applications using homology search (bottom). This figure has been adapted from Hauswedell [191].

the complete query to the reference genome, homology search uses local alignments. This way, less exact matches of the entire or parts of the query sequences can be identified, and potentially evolutionary conserved domains can be detected. Therefore, not only the best-scoring match is of interest, but identifying many matches per query across the entire database enables the assessment of conservation across species and facilitates downstream taxonomic analyses [193].

The identification of many local matches for a large number of query and reference sequences represents a computational challenge that is typically addressed using heuristic algorithms [194]. This means that it is not guaranteed to find all matches according to the search and alignment parameters but many, while at the same time ensuring the feasibility of such searches. To further reduce the search space, homology search can make use of protein alignments instead of using nucleotides. For this purpose sequencing reads can be translated to amino acids and then searched in a protein database (or translated nucleotide database). The amount of known protein sequences is much smaller than the number of nucleotide sequences due to the fact that only a small subset of the genome is protein-coding (e.g., 1% of the human genome [195]). Additionally, searches in the protein space might reveal different types of conservation that might not be apparent from the DNA itself. The tool BLAST is considered the gold standard for local nucleotide and protein alignments and provided the statistical basis for many tools developed afterwards [194]. Other tools that most importantly improve performance include Lambda [190] and MALT [196] for protein and nucleotide alignments as well as DIAMOND which implements extremely fast and sensitive protein alignments [197, 198].

3.1.1 Aims and scope of the study

The nucleotide search implemented in previous local alignment tools is unsuitable to search reads from bisulfite sequencing experiments due to the introduced nucleotide conversion described in sections 2.6 and 2.7. However, using homology search for this type of data sets can be desirable. In practice, sequencing experiments can be contaminated, which results in lower alignment rates to the reference genome of the original organism. Local alignments in larger databases could therefore help identify the contamination sources and improve experimental set-ups. Additionally, cell-free DNA is frequently profiled using bisulfite sequencing to identify

disease states such as tumors based on the methylation patterns [199, 200]. However, cell-free DNA can also contain remnants of pathogens or other microbes to a low degree, which can also be detected from bisulfite sequencing experiments [201–203]. For this purpose, previous studies have implemented an approach where microbial reference databases have been *in silico* converted by replacing all cytosines with thymines with the subsequent application of BLAST [203]. This way, bisulfite-converted sequences can be detected using a standard local alignment tool. However, this strategy requires pre-processing of the database. Additionally, it does not account for the fact that only C (database) to T (query) mismatches are introduced by bisulfite conversion (not T to C mismatches), which increases the false positive rate. To provide a universal framework to search protein as well as regular and bisulfite-converted sequences, the tool Lambda was adapted to accommodate bisulfite conversion-aware local nucleotide alignments. For this purpose, the third and newest version of the tool Lambda - termed Lambda3 hereafter - was used, which was previously implemented and described by Dr. Hannes Hauswedell [191]. In the following sections, the workflow of Lambda3 is described, followed by the adaptations made to implement a bisulfite mode. Lastly, the performance of Lambda3 compared with other local nucleotide and semi-global bisulfite alignment applications is assessed.

3.2 Lambda3 workflow

Lambda3 is structured into two main parts: the index creation and the search of queries in the index (Figure 3.2.1). Databases for which an index should be built can be supplied as a (potentially compressed) FASTA file and contain nucleotide or amino acid sequences. If nucleotide sequences are provided, but a protein index is requested, the nucleotide sequences are translated into amino acids generating all six possible reading frames (Table 3.2.1). Subsequently, Lambda3 offers the option to reduce the sequence alphabet for index creation and search to decrease the index size and speed up the search process. Amino acids can be reduced to one of two different 10-letter alphabets (defined by Li et al. [204] or Murphy et al. [205]) that group amino acids based on chemical and physical properties. Nucleotides can be reduced from a five-letter alphabet (A, C, G, T, and N) to a four-letter alphabet where the wild card letter N is randomly replaced by one of the other four bases. These reduction techniques do not compromise the quality of the output, as search results are verified during a subsequent alignment step based on the translated and not the reduced alphabet.

Sequences are read and potentially translated or reduced using the BioC++ library for sequence analysis [206]. Optionally, taxonomic information can be provided for the database (or reference) sequences, which are stored with the index and later used to annotate the output of the search. Additionally, Lambda3 offers the option to build and store a taxonomic tree that enables the calculation of the lowest common ancestor (LCA) for all resulting hits of a single query. A uni- or bidirectional FM index is then built using the FMIndex-Collection library [207]. The index, the original reference sequences, potential taxonomic annotations as well as the parameters used for translation or reduction are serialized using the cereal library [208] and stored in a single file on disk.

For the search, the index is deserialized, and information on translation and reduction steps is obtained from the stored parameters. The query sequences can be provided as a (potentially

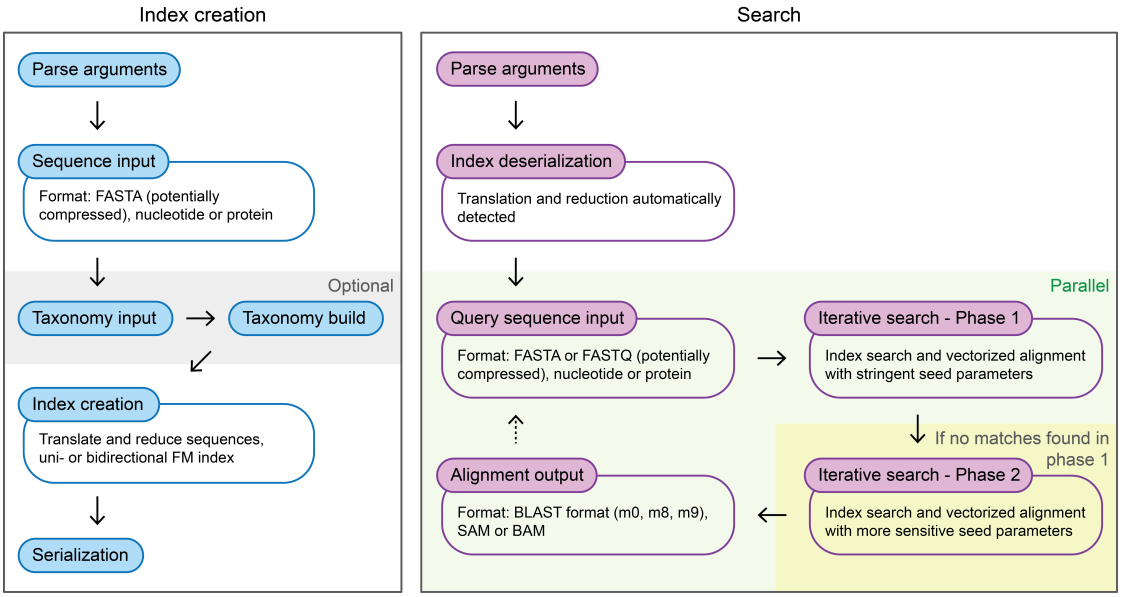

Figure 3.2.1: Overview of the two main parts of Lambda3, the index creation and the search.

compressed) FASTA or FASTQ file. After loading the queries, sequences are first translated and reduced to match the reference sequences (Table 3.2.1). Nucleotide queries are subsequently reverse-complemented before the reduction to allow possible alignments to the reverse strand of the reference database. Each query is then split into seeds, which are searched in the index (Figure 3.2.2). The sensitivity and speed of the search process depend on the length of the seed, the overlap of seeds (defined by the offset of seed start positions), and the number of mismatches allowed during the search (seed delta, Figure 3.2.2). Lambda3 offers different profiles (default, fast, and sensitive) that pre-define these parameters to achieve different trade-offs between speed and sensitivity. For all modes, mismatches are only considered in the second half of the seed in order to accelerate the search. Due to the overlap of seeds, mismatches are still considered along almost the entire query sequence (Figure 3.2.2). If too many hits for a certain seed are found during the search, the seed is automatically elongated to increase the efficiency of the downstream alignment step and avoid the detection of generic hits as they are considered less informative due to their universal occurrence (Figure 3.2.2).

The detected hits are verified using a fast, ungapped alignment around the seed region. Hits passing a specific threshold (defined by the selected mode or user) are considered for downstream analyses, while the remaining hits are discarded. If no hits pass the threshold, the search process is repeated using more sensitive seeding parameters (defined by the different profiles). The hits that finally survive the pre-filtering are sorted by similar size and distributed into batches that are used as input for a fast, vectorized local alignment implemented with the SeqAn2 library [209, 210]. As described above, the alignment steps are performed on the translated reference and query sequences in order to detect and penalize false positive hits that could arise due to the reduced alphabets in the search process.

For each resulting match, the bit score and *E*-value are calculated based on the local alignment scores, two measurements that are used by BLAST and other local alignment tools to assess the

alignment quality [194]. Given a local alignment score S (the sum of all match, mismatch, and gap scores within the alignment), the bit score S' is calculated as

$$S' = \frac{\lambda * S - ln(K)}{ln(2)} \tag{3.1}$$

Here, higher values of S' are associated with higher quality. The variables λ and K represent scaling factors that ensure comparability of results considering different scoring schemes and different sizes of the search space, respectively [194, 211]. The values of these two variables for gapped alignments and a selection of scoring schemes were determined experimentally by Altschul et al. for amino acid and nucleotide alignments using alignments of random sequences and observing their properties [194]. Subsequently, the reported values of λ and K for different scoring schemes have been used as such in all following tools that implemented bit score and E-value calculations [194]. The E-value can be calculated as the bit score normalized to the query length m and the size n of the database used for the search:

$$E = m * n * 2^{-S'} \tag{3.2}$$

It represents the number of hits that could be expected by chance given a random database of the same size (e.g., $E = 1$ means that one hit with a similar score can be expected by chance in a same-sized database). Therefore, lower E-values are associated with higher alignment quality. In Lambda3, similar to other tools, only hits that pass the pre-defined bit score threshold or E-value cut-off are considered valid. Filtering results based on the E-value is desirable if the overall search space should be taken into account to filter out false positives. However, the E-values of different local alignment tools have been inexplicably reported to frequently deviate from the E-values computed by BLAST [190]. Therefore, for comparability across applications with the same database and query sequences, a bit score threshold should be considered and will be in the scope of this chapter. For this purpose, the bit-score equivalent of an E-value cut-off of interest can be computed as:

$$S' = log2\left(\frac{m * n}{E}\right) \tag{3.3}$$

The matches that survive either the bit score threshold or E-value cut-off for a specific query are sorted, deduplicated, and written to the output file. The LCA is computed for all hits of a specific query if taxonomic annotations were provided during the index creation. Output files can be in BLAST or SAM/BAM format implemented using the SeqAn2 library.

3.3 Bisulfite mode

In the following sections, the specific adaptations made to different parts of Lambda3 to accommodate bisulfite-converted sequences are presented. Similar to the alignment tool segemehl, we chose to search seeds in a three-letter alphabet where either C and T or G and A are treated as the same letter but verify the alignments using an imbalanced scoring scheme [179]. This ensures

BLAST mode	Subject (S) alphabet	Query (Q) alphabet	Translation	Reduction (optional)
BlastN	Nucleotide	Nucleotide	Add reverse complement (Q)	Nucleotide 4-letter
				Bisulfite conversion
BlastP	Amino acid	Amino acid	No translation	Li alphabet
				Murphy alphabet
BlastX	Amino acid	Nucleotide	Six-frame translation (Q)	Li alphabet
				Murphy alphabet
TBlastN	Nucleotide	Amino acid	Six-frame translation (S)	Li alphabet
				Murphy alphabet
TBlastX	Nucleotide	Nucleotide	Six-frame translation (S+Q)	Li alphabet
				Murphy alphabet

Table 3.2.1: Search modes implemented in Lambda3 and the respective input alphabets, translation, and optional reduction steps. The first column indicates the respective BLAST mode.

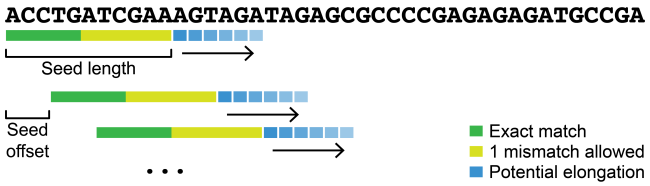

Figure 3.2.2: Seeding strategy in Lambda3. The query is split into seeds of a pre-defined length. The distance between seed start positions is defined by the seed offset. By default, mismatches are only considered in the second half of the seed, and the seed will be automatically elongated if too many hits are found. Here, an example using seed length = 11, seed offset = 3, and seed mismatch = 1 is shown.

that only C (reference) to T (query) mismatches are considered as matches, while T (reference) to C (query) mismatches are still penalized (analogously for G and A mismatches, see section 2.7.2). The bisulfite mode is available within Lambda3 under two different sub-commands (`mkindexbs` and `searchbs`) that build the index and perform the search, respectively.

3.3.1 Alphabet reduction

As described in the previous section, Lambda3 offers the possibility to reduce the alphabet of the input sequences and therefore enhance the performance of the index-based search. In order to account for the effects of bisulfite conversions in nucleotide sequences during the index search, this option was adapted to implement a new bisulfite conversion-aware reduction mode. For this purpose, a new alphabet was implemented that incorporates the two three-letter alphabets that are required to search bisulfite-converted sequences (Figure 3.3.1): one to search queries originating from the original forward and reverse strands (accounting for C to T mismatches) and one

to search queries originating from the reverse complements of the original strands (accounting for G to A mismatches). The first three-letter alphabet contains three artificial characters, two that equal the bases A and G and one that equals both C and T bases. The second alphabet includes two characters that equal the bases C and T, respectively and one that equals both G and A. These are then all combined in a single, new alphabet that the reference and query sequences are converted (reduced) to before the search (Figure 3.3.1): Before the index is built, reference sequences are duplicated and for each sequence, one copy is reduced reflecting the effect of bisulfite conversion on the original strands (first three characters in the new six-letter alphabet), while the second copy is reduced reflecting the effect on the reverse complements of the original strands (last three characters in the new six-letter alphabet, Figure 3.3.1). The query sequences are treated accordingly after a reverse complement for every query sequence is added, similar to the actual nucleotide mode (Figure 3.3.1). This accounts for every possible origin and alignment of the query sequences to the database and results in a total of four sequences per original query sequence.

3.3.2 Index construction

The index is constructed based on the reduced reference sequences analogously to the regular nucleotide mode. The new six-letter alphabet that represents the two bisulfite-specific three-letter alphabets, therefore, allows the creation of a single index in contrast to some bisulfite alignment tools that build two indices, one for each three-letter alphabet [179, 180]. In addition to the reduced disk space, this required minimal changes to the existing code base as the bisulfite-aware index was implemented based on the already existing reduction concept in Lambda3.

3.3.3 Search

Seeds are searched using the reduced six-letter alphabet. Thus as in the normal nucleotide mode, equal letters between query and reference are considered matches, while differing letters are considered mismatches. Due to the reduced bisulfite-specific alphabet, C and T or G and A are considered the same letter and are therefore treated as matches, which increases the number of detected seeds in comparison to the normal nucleotide search. Here, the adaptive seeding strategy previously implemented in Lambda3 becomes even more important as seeds will be automatically elongated if too many hits are located to make the resulting hits more meaningful. During the implementation, we realized that if the seed is located at the end of the query, the number of located hits can drastically increase as the seed can no longer be extended. This, again, is a much larger problem for hits located in a reduced bisulfite conversion-aware alphabet than for a regular four-letter DNA alphabet due to the decreased complexity of the seeds (the detected number of seeds was frequently one order of magnitude larger than for the respective nucleotide search). Therefore, we implemented a simple filter that discards any seeds for which too many hits are found after the elongation step.

3.3.4 Alignment

During the alignment of query and reference sequences, a scoring scheme is used to weight matches, mismatches, and gaps (different costs for gap opening and extension). The sum of the individual scores represents the alignment score that is used to calculate the bit score, defining the quality of the located hit. In contrast to the search, the subsequent local alignments are calculated using the translated and not the reduced reference and query sequences to eliminate potential false positives. As described in section 2.7.2, here, the asymmetric effect of the bisulfite conversion needs to be taken into account. Therefore, an imbalanced scoring scheme was implemented for the alignment that considered mismatches from C (reference) to T (query) as matches for reads reflecting the original forward and reverse strands, while mismatches from G to A were considered as matches for reads from the reverse complements of the original strands. For this purpose, the vectorized alignment is carried out in two steps, separating the hits that were located in reference sequences reduced to the first and the last three letters of the six-letter bisulfite conversion-aware alphabet (Figure 3.3.1). For each step, the respective imbalanced scoring scheme is used.

3.3.5 Output

The matches detected by the two local alignment steps are merged per query before the final sorting and deduplication step. Based on the BLAST output format, it is not possible to infer which of the four read types that arise during bisulfite sequencing aligned. Instead, it is only possible to infer whether the read or its reverse complement aligned (start and end positions are swapped in the second case). However, the information about the actual read type can be obtained if SAM or BAM output is requested. Although the mandatory positions in a SAM file are defined and need to follow the sequence alignment format specification [212], alignment tools can define their own additional tags to provide custom information. Such a tag is commonly defined by bisulfite alignment tools to indicate which of the four bisulfite read types a specific read belongs to. In line with this, Lambda3 outputs a custom tag that was initially designed to report the reading frame for translated protein alignments but now also reports the read type for bisulfite searches.

3.4 Benchmarks

3.4.1 Data sets

To assess the performance of Lambda3 and semi-global bisulfite alignment tools, we made use of different simulated and real sequencing data sets (Table 3.4.1). First, we selected three DNA data sets (q1-q3) that were used to test and evaluate the bisulfite mode in comparison to the nucleotide mode of Lambda3, BLAST (the gold standard for local alignments), and MALT (another local alignment tool that allows nucleotide searches). The data sets q1 and q2 (simulated) were obtained from the CAMI challenge II, an initiative that aims to provide gold-standard data sets to benchmark tools used in metagenomic studies [213]. The data set q3 was obtained from a

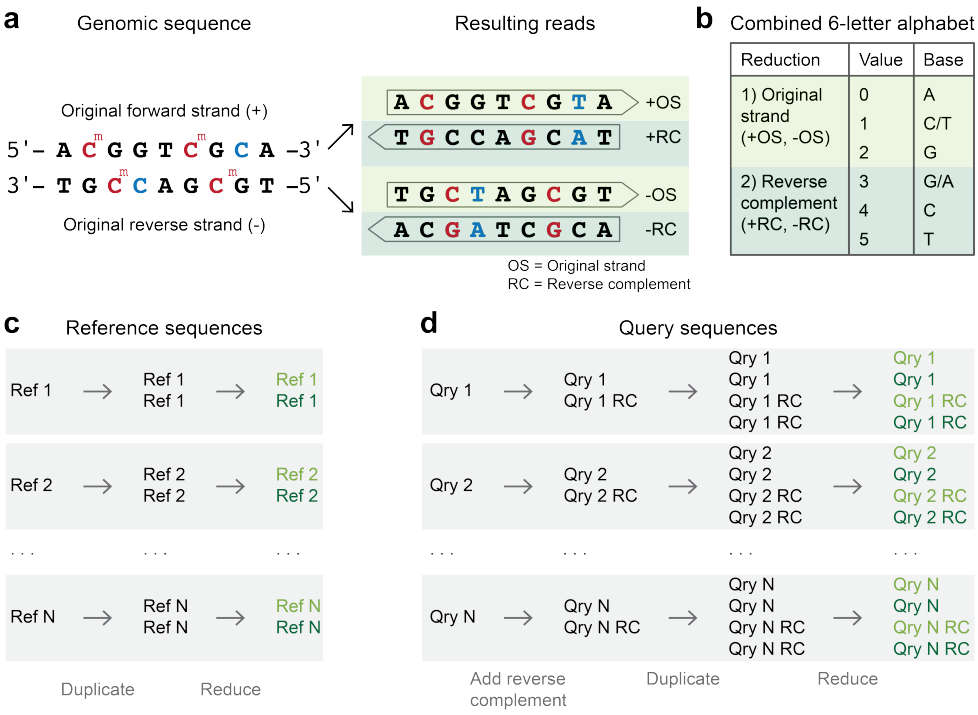

Figure 3.3.1: a) Sodium bisulfite treatment leads to the conversion of unmethylated cytosines to uracils. During subsequent amplification steps, uracils get replaced with thymines. As a result, four different read types can be distinguished that stem from the original forward and reverse strand or their reverse complements. b) To build a single index for bisulfite-aware alignments, a new six-letter alphabet was implemented in Lambda3, where the first three characters represent the reduced alphabet associated with reads from the original forward and reverse strand (C and T are considered identical). The last three characters represent the reduced alphabet for reads originating from the reverse complements of the original forward and reverse strands (G and A are considered identical). c) Before the index is built, reference sequences are duplicated. The first copy is subsequently reduced using the characters of the combined alphabet associated with the original strands, while the second copy is reduced using the characters associated with the reverse complements. This way, the index is built across the same alphabet that accounts for both reductions without the necessity to build two indexes. d) Query sequences are reduced analogously to the reference sequences after the sequences have been reverse complemented to account for every possible alignment according to the four read types.

metagenomic study by Bahram et al. [214]. To test Lambda3's bisulfite mode in comparison to regular nucleotide searches, we *in silico*-converted these query data sets (q1-q3) into sequences mimicking a bisulfite conversion experiment where cytosines were converted to thymines at a conversion rate of 99% of all cytosines (similar to high-quality mammalian bisulfite sequencing data sets, scripts obtained from [215]). For simplicity, we converted all query sequences accordingly, which reflects the effect of bisulfite conversion of reads from the original strands (Figure 3.3.1). To ensure that Lambda3 can also detect reads both from reverse complements of the original strands, we additionally created a version of the two data sets where guanines were

ID	Query set	Length	Molecule	Source
q1	Strain diversity	125 bp	DNA (simulated, *in silico* bisulfite-converted)	CAMI II challenge (2022)
q2	Plant-associated	125 bp	DNA (simulated, *in silico* bisulfite-converted)	CAMI II challenge (2022)
q3	Topsoil	251 bp	DNA (*in silico* bisulfite-converted)	Bahmram et al. (2018)
q4	Xenograft breast tumor	125 bp	bisulfite-converted cell-free DNA	Liu et al. (2021)
q5	Fungi	76 bp	bisulfite-converted DNA	Bewick et al. (2019)

Table 3.4.1: Query data sets used to evaluate the performance of Lambda3 and semi-global bisulfite read alignment applications. Data sets q1-q3 were *in silico* bisulfite-converted to compare Lambda3's bisulfite mode with the original nucleotide mode and other local alignment tools.

converted to alanines with the same conversion rate, which led to comparable results. The *in silico* conversion provides the advantage that the original sequences can be searched with regular local nucleotide alignment tools, while the matching bisulfite-converted sequences can be searched with bisulfite alignment tools, allowing a direct comparison.

In addition to these *in silico*-converted data sets, we made use of real-world bisulfite sequencing samples. We selected data from a breast tumor xenograft model, where the cell-free DNA of the xenograft model is expected to contain DNA fragments of both organisms, the host mouse model and the engrafted human cells, but also potential remnants of microbes [216]. Additionally, we sampled a pan-fungi data set from a study that profiled different fungi species to mimic a cross-species sequencing experiment [217]. Of each data set, 200 megabytes (MB) were extracted and used for performance assessment (see appendix A for more information on the query set selection). All data sets are single-end, as local alignment tools do not process reads in pairs.

The *in silico*-converted data sets (q1-q3) were compared against the collection of microbial genomes assembled within the Human Microbiome Project (5.8 gigabytes (GB), download June 7, 2022, [218]). Reads from the xenograft model (q4) were compared against a database consisting of the human (hg19) and mouse (mm10) genomes as well as the Human Microbiome Project (12 GB). For the fungi data set (q5), all fully assembled fungi reference genomes (download June 7, 2022) were used as the database (4.4 GB).

3.4.2 Parameter selection and comparison with nucleotide search

After the implementation of the bisulfite mode, default values for the different parameters that constrain and guide the search process needed to be selected, including distinct sets for a fast and a sensitive mode (also available for nucleotide and protein searches). The most important parameters are listed in the following:

- **seed length** - The initial length of the seed.

- **seed offset** - The distance between seed start positions in the same query.
- **seed delta** - The number of mismatches allowed in the (second half of the) seed.
- **pre-scoring threshold** - The minimum average score per position in the region that is used to score each seed after the search but before the alignment.
- **bit score threshold** - The minimum bit score with which a hit is accepted and reported.

For the regular nucleotide mode, the parameters were selected based on the comparison to other existing tools such as BLAST, DIAMOND (protein), and MALT (nucleotide), comparing runtime and the number of detected queries, which is a common measure of sensitivity for local alignment tools [219, 220]. The effect of the bisulfite conversion has implications that require these parameters to be adjusted for the bisulfite mode. Given the regular nucleotide mode parameters, the number of hits found during the search will likely increase due to the reduced three-letter alphabets (up to 43% measured for q1-q3). Of these hits, many more will pass the pre-scoring step, namely the ungapped alignment performed right after the search in a region around the seed location (up to 2,350% increase measured for q1-q3 using the same seeding parameters for nucleotide and bisulfite mode). This is expected because C to T (or G to A) mismatches arising from the bisulfite conversion are indistinguishable from actual mismatches of the same kind introduced, for example, via single nucleotide polymorphisms or larger differences in the genomic sequences of different species. Therefore, the bisulfite mode will also generally lead to higher alignment scores.

As introduced in section 3.2, the bit score (and therefore also the *E*-value) is based on the alignment score but additionally relies on the variables λ and K that ensure the comparability of results across different scoring schemes and search space sizes. The values of these variables were experimentally determined for amino acid and nucleotide searches using various scoring schemes. However, these experimental conditions are not likely to reflect the search with bisulfite-converted sequences. Thus, the increased alignment scores of the bisulfite mode will lead to overall higher bit scores that will not equal alignments with higher quality but instead reflect a shift in the bit score distribution due to λ and K not being measured for bisulfite-converted sequences. Therefore, we decided to adjust the bit score threshold for the bisulfite mode to account for this effect. Finally, even after the adjustment of different parameters, the overall false positive rate is expected to be increased with respect to regular nucleotide searches due to actual mismatches that are shadowed by the bisulfite conversion effect, which likely have a higher impact for local in comparison to semi-global alignments.

We chose to select parameters for the bisulfite mode in comparison to the regular nucleotide mode of Lambda3 (version 3.0.0) and BLAST (version 2.13.0). For this purpose, we made use of the query data sets q1-q3, of which we established regular nucleotide and *in silico* bisulfite-converted versions. Since the bisulfite mode also performs a type of nucleotide search, we aimed to generate results similar to that of the regular nucleotide mode. The larger number of hits that arise due to the reduced alphabet complexity required the selection of stricter seeding parameters than for the nucleotide mode to enable sufficient performance and reduce the false positive rate. However, different search parameters and filtering criteria might not only lead to a decrease in false positive hits but could also result in missing matches located by the nucleotide mode or finding new hits that the nucleotide mode does not detect. To evaluate whether the overall results detected by the

bisulfite mode with different parameters are valid, we used the results of BLAST as an additional comparison. BLAST still represents the most sensitive tool for local alignments [198]. Therefore, we reasoned that it could be used as a gold standard for our comparisons. This approach is an approximation as it is not guaranteed that a query that BLAST cannot detect represents an actual negative (i.e., a query that has truly no meaningful hits in the database). Therefore, we used BLAST with the same scoring scheme as Lambda3 and a relaxed cut-off of $E < 1$ to define an inclusive set of true positives with the assumption that any query that cannot be detected with such a cut-off would be likely a false positive if detected by other tools.

For the parameter search, we tested different combinations of seed length, seed offset, seed delta, pre-scoring threshold, and bit score threshold using the query data sets q1-q3 with their respective database defined in the previous section. First, the results were compared against the gold standard defined by BLAST with the objective of increasing the true positive rate TPR and minimizing the false positive rate FPR. Here, we compared the queries that were detected with any hit between BLAST and Lambda3's bisulfite mode. The true positive rate TPR was defined as

$$TPR = \frac{TP}{P} \tag{3.4}$$

where TP denotes the number of queries that were detected by both BLAST and Lambda3, while P denotes the number of queries that were overall detected by BLAST. Accordingly, the false positive rate FPR was defined as

$$FPR = \frac{FP}{N} \tag{3.5}$$

where FP denotes the number of queries that were detected by Lambda3 but not BLAST, while N denotes the number of queries that were not detected by BLAST. In addition to TPR and FPR, we measured the false discovery rate FDR to assess the fraction of false positives within the overall results of Lambda3:

$$FDR = \frac{FP}{FP + TP} \tag{3.6}$$

Additionally, the results were compared to the default, fast, and sensitive profile of Lambda3's nucleotide mode, where we considered the same measurements. As an additional constraint, we aimed to select parameters that would result in increasing numbers of queries found from fast to default to sensitive mode, with runtime increasing accordingly. We then selected a pre-scoring threshold for which adequate seeding parameters could be detected for all three profiles (Table A.2.1).

In order to determine the bit score threshold for each query data set, we employed the following strategy: For Lambda3's nucleotide mode, we considered an E-value cut-off of 0.01, which implies that less than 1% of the reported hits are expected to have occurred by chance. Given the query sequence lengths and database size, this led to the bit score thresholds 46, 46, and 47 for q1-q3 (due to the longer queries in q3) according to equation 3.3. We then selected a bisulfite-

specific bit score threshold for q1 and q2 (that have the same database size and query length) for which the TPR was comparable to that of Lambda3's nucleotide mode while minimizing the FPR (resulting bit score 68). Based on the associated bisulfite-specific *E*-value that resulted from this bit score according to equation 3.2 and the respective database sizes and query sequence lengths, we computed the bit score thresholds for q3-q5 (68, 68, 66, equation 3.3).

The evaluation of the resulting modes is presented in the following section in combination with the comparison to semi-global alignment tools.

3.4.3 Comparison with bisulfite alignment applications

We then selected commonly used bisulfite alignment tools (BSMAP (version 2.90), Bismark (version 0.24.0), and GEM3 (version 3.6.1)) in order to compare the performance of Lambda3's bisulfite mode with classic semi-global alignments. Of these tools, Bismark is the only aligner offering a local alignment mode. Initially, we also selected segemehl for the comparison. However, it was not possible to build an index for our databases due to the large number of reference sequences included that were not supported by segemehl. Performance was measured on an Intel(R) Xeon(R) Gold 6248 CPU @ 2.50GHz. Every tool was executed with 40 threads except Bismark, which does not offer the option to limit the number of threads reliably. The user can only specify the number of instances of Bismark that will be started in parallel that, according to the manual, start between two and six threads each. Therefore, the number of parallel instances was set to eight to approximate 40 threads. Bisulfite alignment tools usually, by default, assume directional bisulfite sequencing libraries, which means that in the single-end mode, only conversions from C to T are considered (not G to A, see section 2.6). As Lambda3 only searches bisulfite-converted sequences in a non-directional fashion, this was also enabled for semi-global applications. Otherwise, alignment tools were executed with default parameters. As an additional comparison, performance for the query data sets q1-q3 was also measured with Lambda3's nucleotide mode and the local nucleotide aligner MALT (version 0.6.1) using the unconverted query sequences. For this purpose, MALT was executed using the same scoring scheme and bit score threshold as Lambda3's nucleotide mode to ensure comparability.

In order to assess the performance of different tools, we first compared the results of q1-q3 against BLAST (relaxed search parameters as described in the previous section). Overall, Lambda3's nucleotide mode exhibits higher true positive rates than MALT (default and sensitive profile, Figure 3.4.1). Additionally, the local nucleotide alignments performed by Lambda3 and MALT exhibit very low false positive and discovery rates. Similarly, the semi-global alignment applications present very low false positive rates, but true positive rates are often comparatively decreased (sometimes close to zero). In comparison, Lambda3's bisulfite mode shows true positive rates comparable to and sometimes exceeding Lambda3's nucleotide mode and MALT (Figure 3.4.1). As expected, the false positive rates increase for the bisulfite mode but never exceed 0.25 with default parameters. The same observation can be made for the local alignment mode of Bismark, which showcases that, as expected, local bisulfite alignments are generally more prone to include false positives (Figure 3.4.1). Compared to Bismark using local alignments, Lambda3's bisulfite mode detects more true positives ranging from 1.6 to 9.7 fold. The FDR is low for all applications and modes, which is influenced by the overall small amount of queries not detected by BLAST

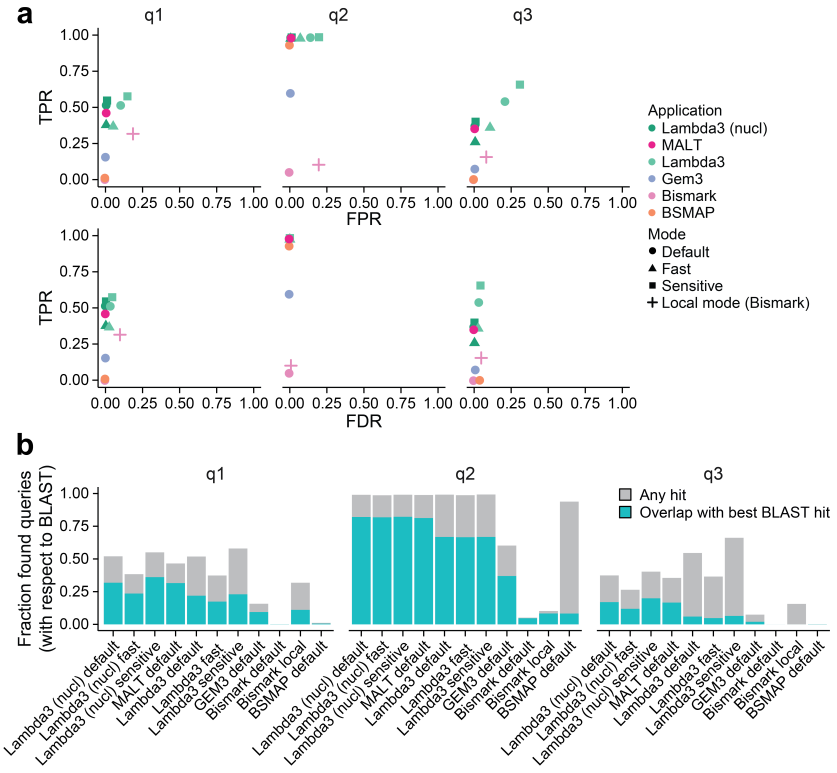

Figure 3.4.1: a) True positive rate compared to the false positive rate (top) and false discovery rate (bottom) for every application with respect to the results of BLAST as the gold standard. b) Bar plots visualizing the fraction of queries found by BLAST that were also found by the other applications. Additionally, the fraction of queries is visualized for which the best hit overlaps with the best hit located by BLAST.

(Figure 3.4.1).

In addition to the queries detected with at least one hit by each application, we compared the fraction of queries for which the best hit detected by each application overlapped the best hit reported by BLAST (Figure 3.4.1). The overall fractions determined are rather low, including for regular nucleotide searches. The best hits detected by Lambda3's bisulfite mode exhibit slightly less overlap with BLAST's best hits than the nucleotide mode. When comparing the results of the bisulfite mode of Lambda3 with the respective nucleotide mode instead of BLAST, we only find moderately increased fractions of hits that overlap (Figure A.2.1). This is not unexpected as the different choice of parameters, together with potentially real mismatches considered as matches in the bisulfite mode, can likely lead to differences in the reported hits as well as the associated scoring (as described in the previous section).

Lastly, we aimed to compare the performance across all bisulfite alignment tools, measuring the number of queries with at least one hit, the runtime, and memory consumption for all five query data sets (*in silico*-converted and real-world bisulfite sequencing data sets). Runtime and memory consumption were measured as the fastest of three runs. Lambda3's bisulfite mode consistently outperforms semi-global alignment tools based on the number of queries detected, with the exception of q5, where Bismark's local mode detects more queries (Figure 3.4.2). How-

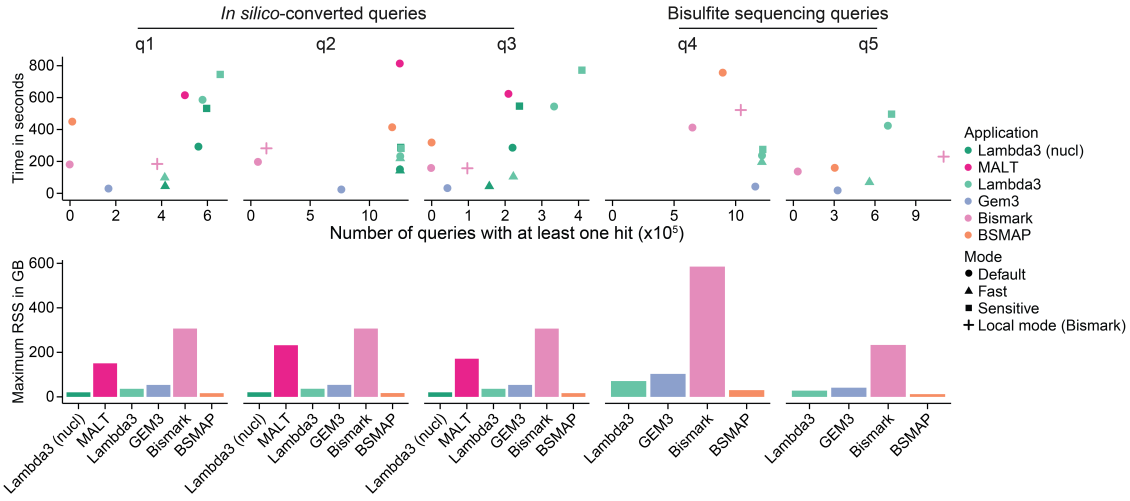

Figure 3.4.2: Top: Comparison of runtime and number of detected queries across local nucleotide alignment as well as local and semi-global bisulfite alignment applications. Bottom: Memory consumption measured for each application (default profile) and query data set shown above. The memory consumption is primarily linked to the size of the database, which is why the highest values were measured for q4 as the corresponding database is the largest (see section 3.4.1).

ever, across all data sets, the number of queries found with at least one hit varies greatly across the other bisulfite alignment tools, while Lambda3 consistently detects hits for large fractions of the query sequences outperforming most (if not all) of the bisulfite-aware applications (Figure 3.4.2). The default and sensitive profiles of Lambda3 are, in some cases, slightly slower than semi-global aligners, which can be expected due to the larger number of seeds and hits that need to be processed within a local alignment tool. However, the fast mode shows comparable or faster runtimes to BSMAP and Bismark while still detecting more queries for most data sets (Figure 3.4.2). GEM3 is the fastest tool (its runtime depends mostly on the time required to load the database). As a comparison, the performance of Lambda3's nucleotide mode and MALT were visualized for q1-q3. BLAST was omitted here due to its extensive runtime (for q1, BLAST is more than 50 times slower than the default nucleotide mode of Lambda3). As expected, the nucleotide search is faster than the bisulfite search.

Of all tools, Bismark consumes the most memory (up to 585 GB for q4 with the largest database, measured as the maximum resident set size (RSS)), while GEM3 and Lambda3 only use up to 103 GB and 71 GB of RAM, respectively. Lambda3's bisulfite mode uses more memory than the corresponding nucleotide mode, which is expected due to the doubling of reference sequences. BSMAP is the most memory-efficient tool (30 GB for the largest database), which could be attributed to the fact that it is the only tool that does not build an FM index but instead is based on hash tables (Figure 3.4.2).

3.5 Discussion

Lambda3 has been adapted to accommodate the search of bisulfite-converted nucleotide sequences and thereby enabled BLAST-like searches for reads from bisulfite sequencing experiments. The bisulfite mode shows comparable sensitivity to other local nucleotide searches (Lambda3's fast nucleotide mode, MALT), while the false positive rate slightly increases as expected for local bisulfite alignments. Our benchmarks showcased that the bisulfite mode of Lambda3 consistently detects more queries compared to standard semi-global alignment tools, which was most pronounced for actual metagenomic data sets (q1-q3). These results show that standard bisulfite alignment tools are not suitable for performing this type of search, even though some of them perform seemingly well in a query-dependent fashion. However, semi-global alignment tools do not provide a measurement of the statistical significance of a located hit such as the bit score or *E*-value. This means that even in the case of q5, where the local mode of Bismark detects more queries than Lambda3's bisulfite mode, these additional hits might be detected by Lambda3 with more relaxed settings. At the same time, the fast bisulfite mode exhibits runtimes comparable to or exceeding most semi-global alignment tools with less memory consumption. In summary, Lambda3 now represents a universal tool to perform local alignment searches with a protein, nucleotide, and bisulfite mode.

The search parameters of Lambda3's bisulfite mode have been selected such that the true and false positive rates were optimized with respect to BLAST but also Lambda3's nucleotide mode. This included the selection of different bit score thresholds to account for the increased alignment scores and generally higher false positive rates expected from local bisulfite alignments. For future versions of Lambda3, it might be desirable to set up an *in silico* experiment in order to determine values for λ and K in a bisulfite-converted sequence context given different scoring schemes. This might improve the quantification of the bit scores (and *E*-values) and thus might lead to more accurate quantification of alignment quality and filtering of false positive hits.

In order to further extend Lambda3 as a universal local alignment tool, Lambda3 could be expanded to include a long-read mode that is able to process long query sequences, such as reads generated by Nanopore or PacBio sequencing. The resulting kilo- to megabase-scale reads are more error-prone than short reads derived from Illumina sequencing and can include frameshift errors [221]. These errors occur during the sequencing process and alter the reading frame within a gene of interest. This is not relevant for nucleotide searches but has an impact on homology searches in the protein space, where the alignment process needs to be able to account for that. Such a long-read mode is already implemented in the protein alignment application DIAMOND and would further improve the usability of Lambda3 [197, 198].

Chapter 4

Measuring DNA methylation heterogeneity from bisulfite sequencing reads

This chapter presents RLM, a tool providing fast and easy extraction of read-level methylation metrics as a measure of DNA methylation heterogeneity. This work was published in *Bioinformatics* in November 2021 [222], and the chapter follows the publication closely, including figures that have been adapted from it. The chapter introduces the concept of DNA methylation heterogeneity measurements, including a description of commonly used metrics. Subsequently, the workflow of RLM is described, and test cases and benchmarks are presented, followed by a discussion of use cases, drawbacks, and biases of read-level methylation metrics.

Dr. Pay Giesselmann and Dr. Helene Kretzmer helped prepare input files for integration tests, and Dr. Pay Giesselmann assisted in benchmarking RLM.

4.1 Introduction

4.1.1 Sources of DNA methylation heterogeneity

CpG methylation, as measured by short-read bisulfite sequencing methods such as WGBS or RRBS, represents an average measurement across the underlying cell population (see section 2.7). However, the cells or alleles in the pool subjected to sequencing do not necessarily need to exhibit the same methylation profile across a specific region. Although most of the genome is consistently methylated or unmethylated across cell types, deviations exist from these patterns. An example of strongly deviating alleles within a cell population is genomic imprinting, where only one allele expresses the imprinted gene, and the other allele is silenced via DNA methylation of the iDMR (see section 2.3). In this case, half of the respective reads should be unmethylated while the other half should be methylated, resulting in average CpG methylation values of 0.5 merging the two allelic conditions (Figure 4.1.1).

Similarly, heterogeneous tissue could include different cell types with specific methylation patterns at promoters or enhancers that would be included in the CpG-wise methylation rate and might be reflected by rather intermediate DNA methylation levels (Figure 4.1.1). This could

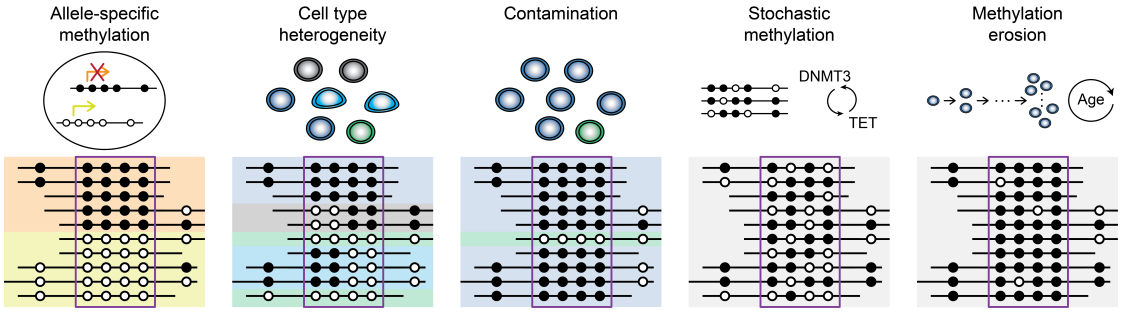

Figure 4.1.1: Schematic of read compositions underlying different sources of DNA methylation heterogeneity. White circles represent unmethylated CpGs, and black circles represent methylated CpGs. The region of interest that is spanned by all displayed reads is marked by a square. This figure has been adapted from Scherer et al. [223].

also apply to contamination of a probe by different cell types, for example, during cell culture where feeders or differentiating cells could accidentally be included in the sequenced cell pool (Figure 4.1.1). A study using 25 primary human tissue and purified cell types also suggested that 2% of CpGs in the human genome consistently exhibit intermediate methylation and are enriched near genes and in enhancers, an evolutionarily conserved state associated with intermediate expression of the associated genes [224]. These intermediate methylation states were found largely independent of allele-specific methylation, which could imply that, instead, cells exhibit an intrinsically heterogeneous methylation profile along these regions [224]. Stochastic, heterogeneous gain of methylation has also been found to accumulate in tumors and cultured fibroblasts at previously unmethylated CGIs [130]. These stochastic methylation patterns raised the hypothesis that cell type-intrinsic heterogeneity might stem from active DNA methylation turnover guided by *de novo* methyltransferases and TET enzymes, a balance that can be biased towards different overall methylation levels [130] (see section 2.5.2). Another possible explanation for heterogeneous DNA methylation levels across cell populations is DNA methylation erosion that could be induced by the infidelity of the maintenance methyltransferase DNMT1 to stably maintain DNA methylation across many cell divisions preferentially in PMDs [72,223] (see section 2.3 and Figure 4.1.1).

Cell types with different methylation profiles present within a tissue can be disentangled using experimental approaches such as sorting and separately sequencing different cell types [223]. However, the cell types present in a tissue might not always be known or lack sufficiently defined marker genes to sort them. Additionally, computational approaches exist that use deconvolution techniques such as non-negative matrix factorization, usually applied to data sets generated with methylation arrays [225, 226]. These algorithms also frequently require previously established datasets from pure cell types to infer cell type compositions, although reference-free algorithms exist [226]. Newer techniques, such as single-cell sequencing approaches, allow for inspecting the methylome of individual cells; however, at currently high costs providing only limited coverage of the genome of each cell. These limitations make single-cell technologies currently not always suitable, especially for larger cohorts [223].

As an alternative solution, many studies have introduced metrics that aim to quantify the extent of DNA methylation heterogeneity from the single reads of bisulfite sequencing experiments

spanning the same regions or loci [130, 131, 223, 227, 228]. This has the advantage that regular WGBS or RRBS can be used to not only profile the average methylation signature of a cell population but also offer another layer of information regarding the underlying within-sample heterogeneity. Studies in cancer have shown that using such scores to quantify the DNA methylation heterogeneity across cells of a population can offer important insights into specific phenotypes as different levels of methylation heterogeneity correlate, for example, with transcription levels or patient survival [130, 131].

4.1.2 Read-level methylation metrics

In the following sections, the underlying properties of single reads are described, and different read-level methylation metrics are introduced that were considered in the scope of this study.

Properties of single reads

Read-level methylation metrics are based on the DNA methylation patterns found on single reads covering a specific CpG or *k*-mer. A single read reflects one allele of one cell present in the pool of cells that was subjected to sequencing. The methylation status of a CpG on a single read can, therefore, only be methylated or unmethylated (Figure 4.1.2). If a read spans multiple CpGs, the pattern of methylation across these CpGs can contain valuable information with respect to the methylation heterogeneity of a single allele. All CpGs on a read can be homogeneously unmethylated or methylated, a state that is termed concordant. The read can also contain both methylated and unmethylated CpGs, a state also referred to as discordant, and might reflect turnover dynamics or methylation erosion at this specific allele fragment (Figure 4.1.2). If a read spans methylated and unmethylated CpGs, high heterogeneity within this read can be reflected by a high fraction of transitions from methylated to unmethylated CpGs and vice versa (Figure 4.1.2).

Proportion of discordant reads

The proportion of discordant reads (PDR) describes the number of reads covering a specific CpG that are discordant. The PDR equals one if all reads covering a CpG are discordant and zero if all reads are concordant (Figure 4.1.3). The metric has been introduced by Landau et al. [131] to study local disordered methylation leading to intra-tumor heterogeneity in the context of chronic lymphocytic leukemia, which they associated with low-level transcription and a poorer prognosis for patients. PDR is a metric that is meant to reflect intra-molecule heterogeneity (an entire read is classified as discordant or concordant, and the population average of this measurement is considered) and, therefore, useful to quantify DNA methylation erosion or stochastic methylation within single molecules at a given locus [131, 223].

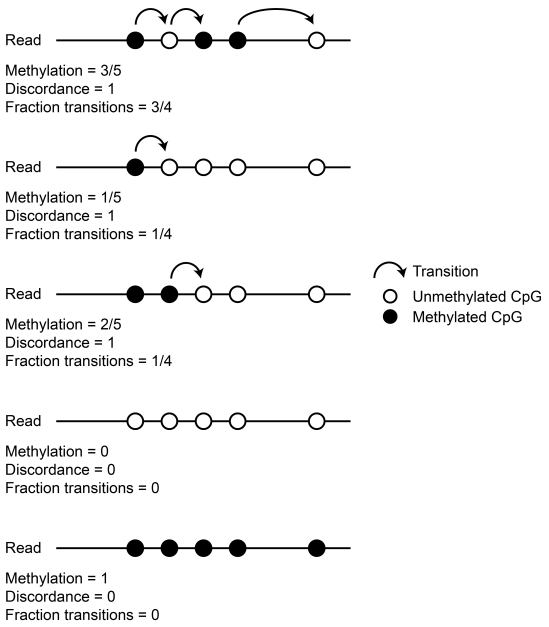

Figure 4.1.2: Properties of single reads from a bisulfite sequencing experiment. Reads span multiple CpGs that can be unmethylated or methylated. This pattern can be concordant (all CpGs on a read are unmethylated or methylated) or discordant (read contains both methylated and unmethylated CpGs). Additionally, the pattern of methylated and unmethylated CpGs can be characterized by the number or fraction of transitions from methylated to unmethylated CpGs.

Read transition score

Charlton et al. introduced the frequency of transitions between the unmethylated and methylated state across CpGs measured on the same read as a metric to define the relationship of neighboring CpGs in phase [227] (Figure 4.1.2). In this study, the authors investigated the re-methylation dynamics after replication in human embryonic stem cells and found that methylation of the nascent DNA strand progresses in two phases: Within the first hour after replication, a strong methylation increase can be observed that the authors attributed to active DNMT1 recruitment [227]. In the second phase, which occurs within the next few hours, methylation levels slowly increase until they reach levels quantified in the bulk cell pool. Charlton et al. hypothesized that this second phase could be independent of the replication timing and instead might reflect a search for targets missed previously due to limitations of DNMT1 close to the replication fork. The authors used the fraction of transitions per read and its decline over time after replication to verify previous reports that DNMT1 acts processively on the same DNA fragments [227, 229]. This fraction of transitions per read can also be aggregated per CpG by calculating the average across all reads spanning its position, which is termed read transition score (RTS) in this thesis (Figure 4.1.3). It reflects similar properties as PDR: It also measures intra-molecule heterogeneity. However, it offers more resolution with respect to the dynamics of neighboring CpGs.

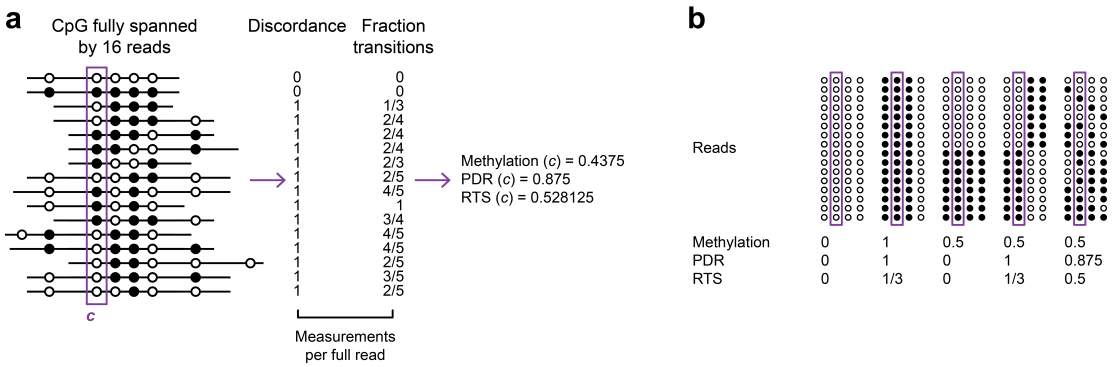

Figure 4.1.3: Schematic of read-level methylation metrics per single CpG. a) Every read that spans a specific CpG can be characterized by either its discordance (0 or 1) or its fraction of transitions from unmethylated to methylated CpGs and vice versa (0 = no transitions, 1 = alternating pattern of methylated and unmethylated CpGs). The average of these single-read metrics reflects the PDR and the RTS, respectively. b) Examples of different read configurations and the resulting PDR and RTS for the CpG marked with a square.

Entropy

The measurement of DNA methylation entropy has been established by Xie et al. to measure the randomness of specific DNA methylation patterns within a population at a given locus [228]. They found that different loci along the genome can exhibit different methylation entropy values but that this is relatively consistent across different samples at the same loci, with the exception of primary tumors compared to their healthy tissue. Methylation entropy is calculated per 4-mer w of consecutive CpGs spanned by N reads and considers the number of occurrences n_k of each possible configuration (termed epiallele) k of methylated and unmethylated CpGs across the 4-mer (16 epialleles possible):

$$\text{Entropy}\ (w) = \frac{1}{4}\sum_{k=1}^{16}(-\frac{n_k}{N} log2 \frac{n_k}{N}) \tag{4.1}$$

Methylation entropy is equal to zero if all reads exhibit the same epiallele at a given 4-mer and one if all 16 epialleles are represented with the same number of occurrences (Figure 4.1.4). In contrast to PDR and RTS, this score is not computed per CpG but per 4-mer and offers the possibility to detect variance in methylation patterns across reads such as generated by tissue heterogeneity, contamination, or stochastic methylation patterns.

Epipolymorphism

Epipolymorphism is a measurement that — similar to methylation entropy — is based on the assessment of epiallele configurations at a given 4-mer. The metric uses the concept of Tallis entropy (in contrast to methylation entropy that is based on Shannon's entropy) for a 4-mer w spanned by N reads given the number of occurrences n_k of each epiallele k:

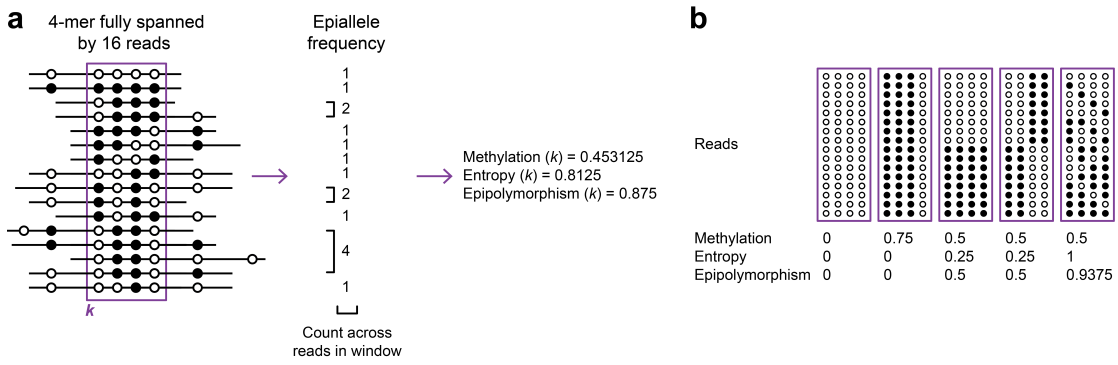

Figure 4.1.4: Schematic of read-level methylation metrics per 4-mer. a) Every read that spans a specific 4-mer exhibits one of 16 configurations of methylated and unmethylated CpGs across the respective four CpGs. These configurations are termed epialleles, and the number of each distinct epiallele present in the population is used to calculate methylation entropy or epipolymorphism. The higher the entropy or epipolymorphism, the higher the within-sample heterogeneity. b) Examples of different read configurations and the resulting entropy and epipolymorphisms for the 4-mer marked with a square.

$$\text{Epipolymorphism } (w) = 1 - \sum_{k=1}^{16} \left(\frac{n_k}{N}\right)^2 \tag{4.2}$$

The maximum value of epipolymorphism associated with the highest heterogeneity is 0.9375 in contrast to methylation entropy that spans the range from zero to one (Figure 4.1.4). The metric was introduced by Landan et al., who used it to analyze the heterogeneity of DNA methylation patterns at regions that become differentially methylated over time in cultured immortalized fibroblasts [130]. The authors then established the hypothesis that hyper- and hypomethylation, as observed in cancer, occur in a stochastic fashion as described in section 2.5.2.

Other scores

Besides the previously described scores, other metrics exist that aim to quantify different aspects of within-sample methylation heterogeneity but were not implemented as part of RLM. Briefly, these include methylation haplotype load (MHL), the fraction of discordant read pairs (FDRP), and quantitative FDRP (qFDRP). MHL is based on the identification of blocks of consecutively methylated CpGs per read, while FDRP and qFDRP aim to offer insights into the within-sample heterogeneity at the single CpG level [223, 230]. For this purpose, FDRP and qFDRP are based on pair-wise comparisons of all reads that span a given CpG [223]. A pair is called discordant if the methylation status of any CpG that is spanned by both reads differs between them, and the overall score is normalized by the number of pairs. qFDRP additionally takes the Hamming distance of a pair into account. The pair-wise comparison strategy makes this score increasingly time-consuming to compute depending on the coverage, which is why the authors propose a down-sampling approach in case of too many reads that span a given CpG [223].

4.1.3 Aims and scope of the study

Previous studies using read-level methylation metrics frequently used custom scripts or offered tools with limited usability that only provide one specific metric, are limited to alignment files of a single read mapping tool, or are only compatible with specific reference genomes (detailed comparisons in section 4.3.3). Therefore in this study, we developed RLM, a new tool performing fast and simplified extraction of read-level methylation metrics from bisulfite sequencing data sets. Our main goal was to provide a generic tool that would allow us to fastly compute multiple metrics and make it accessible to many users and projects. This also included the ability to use RLM with alignment files from any reference genome, support for multiple commonly used bisulfite read mapping tools, as well as compatibility with different sequencing set-ups such as WGBS, RRBS, and enrichment strategies.

4.2 RLM workflow

RLM was implemented as a standalone tool using the C++-based software library SeqAn3 [191, 231]. The application is available via GitHub (https://github.com/sarahet/RLM) and can be easily installed using cmake. The following sections describe the workflow illustrated in Figure 4.2.1.

4.2.1 Input

As input, RLM takes already aligned reads in SAM or BAM format [212]. RLM can process alignments from all common short-read sequencing set-ups such as WGBS, RRBS, and enrichment strategies (e.g., amplicon sequencing). Additionally, both single- and paired-end reads can be handled (set-up defined via runtime parameter). The alignment file does not need to be sorted, although for paired-end data it would be an advantage if the input is sorted according to the genomic position or read name since the first read in a pair that is read will be kept in memory until the second read is found (see section 4.2.3). If the paired-end input alignments are not sorted, this could lead to an unnecessary overhead in memory consumption.

It is recommended to deduplicate the alignment file after mapping in order to remove PCR duplicates. These are reads that originate from the PCR step and do not represent different cells within the population but copies of a specific fragment (see section 2.7.2). Calculating read-level metrics could therefore be biased by such artifacts because these reads will be mistakenly considered as biologically independent.

As described in chapter 3, a SAM tag is commonly defined by bisulfite alignment tools in order to indicate, which of the four bisulfite read types has been detected for every record (see section 2.7.2). Within RLM, this tag is used accordingly to decide whether the position of the cytosine on the forward or on the reverse strand is needed to define the methylation status of a CpG. RLM recognizes the tags and therefore accepts alignment files from the commonly used bisulfite alignment tools BSMAP, Bismark, segemehl and GEM3. The aligner needs to be specified via a runtime parameter in order to ensure that RLM considers the correct tag.

In addition to the SAM or BAM file, RLM requires the reference genome used to align the reads as additional input file in FASTA format. The reference genome is necessary to find the location of CpGs and compare the sequence of reads overlapping them.

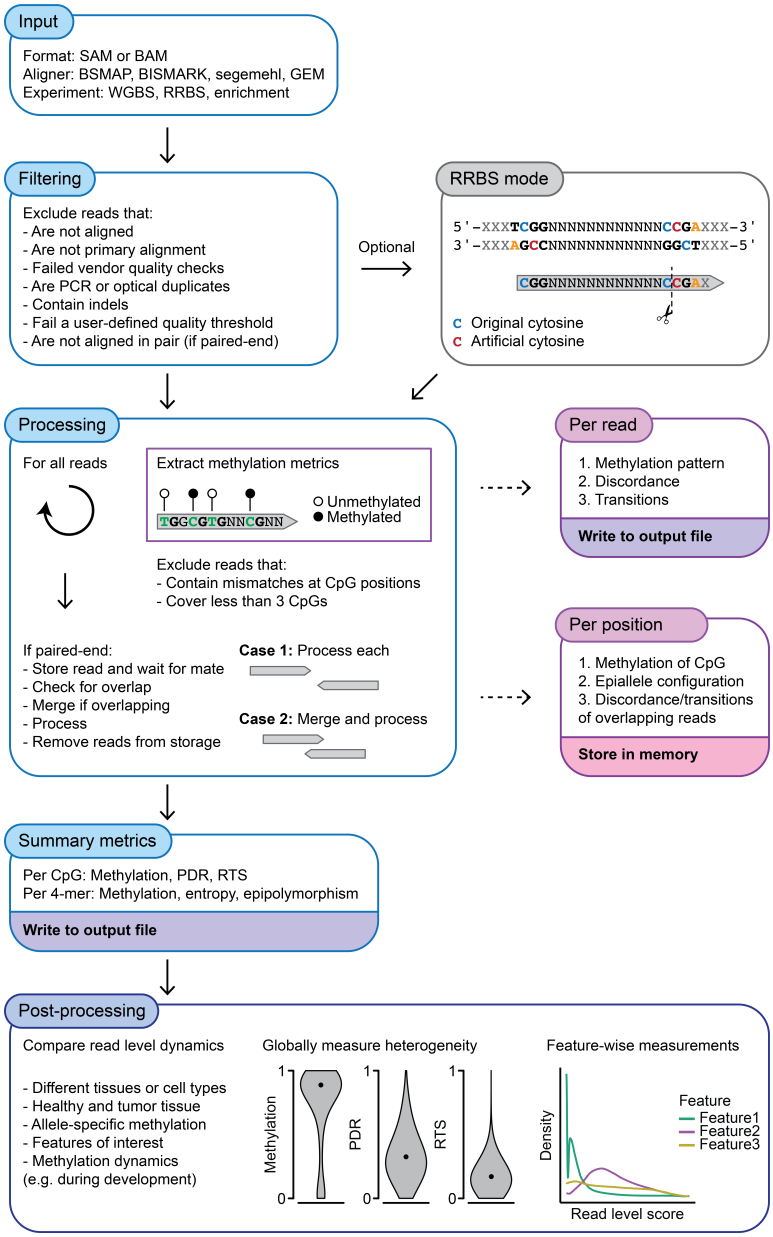

Figure 4.2.1: Workflow of RLM summarizing the input requirements, filtering steps, optional RRBS mode, processing of paired-end reads, possible output formats, and optional post-processing as provided by the tool.

4.2.2 Read filtering

RLM streams across the input lines of the BAM or SAM file and processes reads directly to minimize memory consumption (temporary storage of read information is an exception for paired-end

reads; see next section 4.2.3). Reads that fulfill one of the following criteria (identified via the SAM flag if not stated otherwise) are skipped and not considered for downstream analyses:

- Unaligned
- Not primary alignment
- Supplementary alignment
- Aligned not in pair if paired-end sequencing set-up
- Marked as duplicate
- Failed the vendor quality checks
- Mapping quality below user-defined threshold (runtime parameter, default: 30)
- Contains indels (extracted from CIGAR string)

Unaligned reads cannot be processed because the genomic origin is unknown, while secondary alignments and duplicates artificially bias the methylation-based quantifications. Alignments of only one read of a pair might stem from low quality, where one read did not pass previous QC steps. Reads that do not pass vendor quality checks or mapping quality thresholds are likely to include sequencing errors, span heavily mutated sites, or originate from a different source than the alignment location. Therefore, such reads are excluded in order to calculate metrics only based on high-confidence reads. Additionally, reads that contain indels might delete or introduce a CpG. This cannot be accurately reflected by read-level metrics as such modifications might not affect the whole cell population and, in case of an insertion, lack a genomic reference for comparison. Thus these CpGs cannot confidently be detected from the bisulfite-converted reads.

The CIGAR string is not only scanned for indels but also to determine whether a read was soft-clipped during the alignment. While hard-clipping leads to the actual read sequence that is stored in the BAM file being shortened according to the number of bases clipped, soft-clipping only leads to a respective entry in the CIGAR string, but the actual read sequence remains unchanged. However, the start position of the read with respect to the reference genome always reflects the first base that is not clipped. Therefore, in the case of soft-clipped reads, the read sequence gets shortened by RLM before the actual processing to allow an accurate comparison of the read sequence and the corresponding reference genome sequence.

If the processed reads originate from an experiment profiled using RRBS, another modification to the read sequence may be applied. As described in section 2.6, during the library preparation process, artificial (usually unmethylated) cytosines are introduced at the 3' end of fragments after restriction enzyme digest. If the DNA fragment is shorter than the read length, the artificial cytosine is sequenced and biases any methylation-based quantifications because it does not reflect the actual methylation status of that cytosine. These bases can be either trimmed during pre-processing or need to be excluded from quantifications such as methylation rate calling later (see section 2.7). If trimming of these bases has been performed, the RRBS reads can be treated as any other experiment. If not, RLM offers the option to activate a specific RRBS mode that clips either the first two or last two bases of the read sequence (depending on the read type, see section

2.6). The potential downside of this approach is that a natural CpG cannot be distinguished from an artificial CpG, and information can potentially get lost.

The options specifying the used aligner, single- or paired-end sequencing design, and the RRBS mode are runtime parameters as they depend on the actual input alignments. However, in order to reduce the overall runtime, the respective branches are already established at compile-time using templates.

4.2.3 Paired-end reads

For paired-end sequencing, reads are pre-processed before the read-level information can be extracted. Depending on the fragment size and read length, the two reads of a pair can overlap if the fragment is shorter than the combined read lengths. In this case, it is not favorable to process the two reads separately because CpGs covered by both would be considered twice for the same fragment/allele, which biases the later quantifications. Considering the first read of a pair that is read, there are two options:

1. The read passes all filters. It is then kept temporarily in memory until the second read of the pair is read.
2. The read does not pass a filter. In the case of artifacts like duplicates, mapping quality, or vendor check fails, the second read is usually affected as well, and the respective artifact reflects generally problematic properties. Therefore, the pair is not considered further. In the case of an indel, the second read might not span a similar site and could be used to calculate read-level methylation metrics. The read name is therefore stored in memory to allow the processing of the second read as a single read (both reads of a pair share the same name).

Considering the second read of a pair that is read, the program continues as follows:

1. If the corresponding read of the pair has been read before and has been excluded due to indels, the current read is immediately processed further (see next section). The read name is then removed from storage.
2. If the corresponding read of the pair has been read before and is stored in memory, the overlap between both reads is calculated using the start coordinates and read lengths. If the two reads do not overlap, they are processed independently. In the case of an overlap, the reads are merged and processed as one record. For this purpose, the start coordinate of the read with the smaller genomic coordinate is used, and the read sequence is merged. Afterwards, the previously stored read is removed from storage to not unnecessarily increase memory consumption over time.

4.2.4 Extracting methylation information per read

The individual reads are then compared to the respective reference genome sequence. First, the number of CpGs covered by a read is determined. If the read spans less than three CpGs or includes a sequencing error or mutation at a CpG position, it is discarded, and the next read is

read from the input file. Too few CpGs do not allow meaningful interpretation of metrics such as discordance or fraction of transitions. Although metrics based on 4-mers, such as entropy, cannot be calculated from reads that span only three CpGs, they are admitted to allow CpG-based metrics, such as PDR, to be calculated for less CpG-dense genomic regions. Mismatches at CpG positions, on the other hand, disturb metrics such as entropy because one position of the 4-mer would be missing.

If a read passes these additional filtering steps, the methylation status of every CpG and 4-mer is extracted using the information of the bisulfite tag that indicates the read origin and mapping orientation of the read sequence in comparison to the reference genome. Per CpG, the methylation of the read at this position, the discordance, and the fraction of transitions of the full read spanning the CpG are stored. Per 4-mer, the epiallele represented by the read is stored. Finally, the individual read-specific metrics (methylation pattern and average of the read, discordance, fraction of transitions) are written to an output file in BED format before the next read is read from the input file.

4.2.5 Score computation

After the complete alignment input file is read, per CpG, the mean methylation, PDR, and RTS are computed. Mean methylation, entropy, and epipolymorphism are calculated per 4-mer using the information that was stored from the associated reads. For 4-mers, additionally, the count of all epialleles underlying the entropy and epipolymorphism calculations is reported. The metrics per CpG and 4-mer are reported in a separate BED file, respectively. Only CpGs or 4-mers covered by a user-defined minimum number of reads are reported (runtime parameter, default: 10). The calculation of these read-level metrics and the generation of the corresponding output BED files is optional. By default, all three BED files (single read information, per-CpG, and per-4-mer metrics) are generated. However, the user can also specify to only generate the CpG-specific or 4-mer-specific output in addition to the single read output (the output file with information per single read is always provided).

4.2.6 Post-processing

Although RLM is a standalone C++-based application, the GitHub repository also includes an R markdown script that offers basic summary statistics and overview figures based on the RLM output files. The script is provided with the intention of allowing users with limited experience to easily visualize their data sets and offer a selection of possible representations to more experienced users to be adapted further. It requires multiple R packages to be installed that are commonly used for visualization by the community, such as ggplot2, vioplot, and RColorBrewer. The script generates a PDF report including summary statistics (minimum, maximum, quantiles, average) of the number (total or methylated) of CpGs per read, the fraction of transitions per read, and the single-read methylation. Additionally, matching histograms are provided, including the coverage of reads per CpG or 4-mer. Finally, different types of density plots visualizing the distribution of read-level methylation metrics separately or in relation to the average methylation across the entire output or within pre-defined regions are provided.

To provide further guidelines, the documentation of RLM includes detailed descriptions of the workflow and the content of the output files. Additionally, code snippets are provided to easily generate data formats that allow the visualization of the resulting read-level methylation metrics in genome browser tools such as IGV or UCSC using commonly used tool suites such as UCSC-tools [232, 233]. Finally, an example report generated using the R markdown post-processing script is shown, and explanations, as well as possible interpretations with respect to the somatic methylation landscape, are offered.

4.3 Benchmarks

4.3.1 Test cases

In order to ensure the correctness of RLM functions and output, a variety of test cases were designed that are part of the GitHub repository and automated using continuous integration with GitHub Actions. For the design, the SeqAn3 application template was used. Unit tests were designed that test the calculation of entropy, epipolymorphism, and methylation per 4-mer as well as the fraction of transitions and discordance per read. For each score, different examples spanning the range of possible configurations were used. Additionally, the functionality of RLM was evaluated using integration tests that assess the correctness of the output given a specific input. The following functionalities of RLM were tested:

1. Argument parsing: In order to ensure that options are correctly parsed, tests were designed that ensure that the help message appears correctly if RLM is executed without arguments and that the correct error message appears if an input file is missing.
2. Handling of different read types: Different input SAM files were designed that include all read configurations that should be covered by the functionality of RLM, ranging from reads that do not pass the filtering steps to read pairs with and without overlap as well as single invalid reads. Additionally, test records, including soft and hard clipping as well as records for the RRBS mode, were designed. The expected matching output files were generated manually and then used to test the correctness of the RLM output.
3. Read mapping tools: A test read file was generated and aligned with all supported read mapping tools (BSMAP, Bismark, segemehl, and GEM3). The resulting BAM files were then used as input for a test case that executes RLM and produces the single-read output file. The corresponding output files for all alignment tools were then pair-wise compared for each read that aligned at the same position (this was necessary as some alignment tools don't support edit distance or use different scoring schemes or clipping strategies).

4.3.2 Performance

In order to measure the performance of RLM (version 1.0.0), a publicly available WGBS data set was downloaded (mouse epiblast, GSM4075619) and aligned to the mouse reference genome mm10 using BSMAP. Different numbers of reads were randomly sampled from the resulting BAM file covering the number of reads from small-scale experiments such as RRBS up to high-coverage

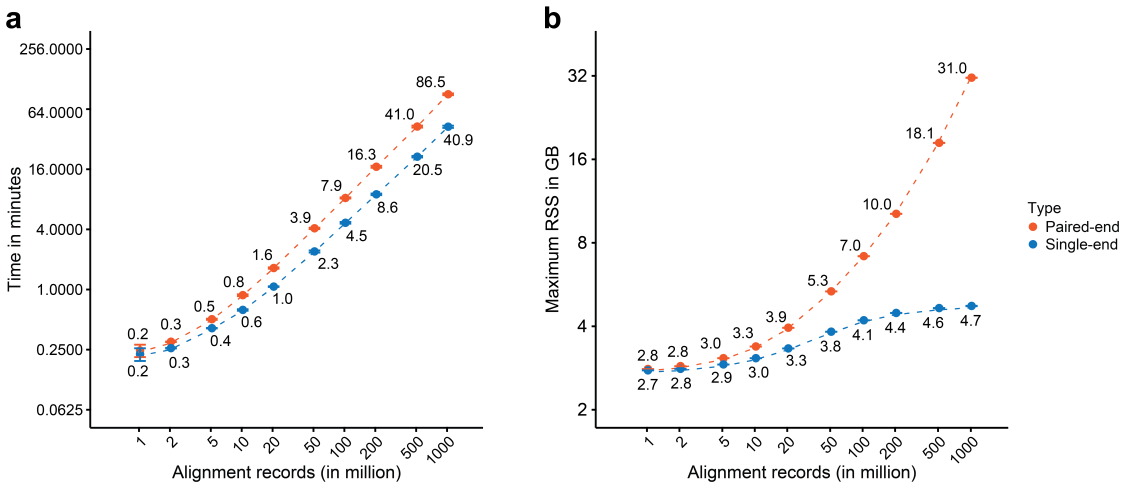

Figure 4.3.1: Runtime a) and memory consumption b) measured for RLM in single-end and paired-end mode for different numbers of randomly sampled BAM records. Dots represent the average runtime across five iterations. Error bars represent the standard deviation.

WGBS data sets. Performance of RLM in single- and paired-end mode was measured on an Intel Xeon Gold 6242 @ 2.80GHz. Runtime and memory consumption were measured as average across five runs, and the corresponding standard deviation was calculated (Figure 4.3.1). The runtime scales linearly with the input read number for both paired-end and single-end modes (e.g., 41 minutes for 500 million records and 87 minutes for one billion records in paired-end mode). However, RLM in paired-end mode is more than two times slower than in single-end mode (41 and 87 minutes for one billion records in single- and paired-end mode, respectively). This gap can be explained by the overhead that is introduced when temporarily storing the first read of paired-end reads and the subsequent resolution of potential overlaps, which is also reflected in memory consumption: While the single-end mode uses less than five GB for the largest input size, the memory consumption in paired-end mode increases stronger the more reads are processed.

4.3.3 Comparison with existing tools

Only a few studies that previously used read-level methylation metrics also provided tools that would allow users to extract these scores from their own data sets. The previously developed tools are frequently limited based on the input features they allow and the number of metrics they provide, which is summarized in the following section.

Features

DMEAS, CluBCpG, and WSH are tools that were published previously and provide frameworks for the computation of read-level methylation from bisulfite sequencing data sets (Table 4.3.1) [223, 234, 235]. DMEAS and CluBCpG are standalone applications (although CluBCpG requires

Tool	Scores	Compatible alignment tools	Reference genome	RRBS mode
RLM	Single read discordance and transitions Entropy Epipolymorphism PDR RTS Matching mean methylation per score	Bismark BSMAP segemehl GEM3	Any	Yes
DMEAS	Entropy	Bismark	Any	No
CluBCpG	Clustering-based read-level analysis	Bismark	Any	No
WSH	Entropy Epipolymorphism PDR MHL FDRP qFDRP	Bismark only for entropy and epipolymorphism	Preferably hg38	No

Table 4.3.1: Features and input requirements of different tools that compute read-level methylation metrics from bisulfite sequencing data sets.

samtools to be installed and available via the PATH variable), whereas WSH is implemented as an R package. As output, DMEAS computes methylation entropy, CluBCpG extracts a clustering-based read-level analysis while WSH provides multiple scores (PDR, MHL, entropy, epipolymorphism, FDRP, and qFDRP). For some scores like PDR, WSH requires additional, pre-computed input in the form of the exact position of all CpGs in the reference genome for which the score should be computed. DMEAS and CluBCpG only support Bismark alignments. WSH, on the other hand, is restricted to Bismark alignment files only for entropy and epipolymorphism calculations. Additionally, according to the documentation, WSH should be preferentially used with reads aligned to the human genome hg38 (although it remains unclear whether that means that no other reference genome is supported) [236]. None of the existing tools provides a specific mode to handle potentially artificial bases as introduced by the RRBS protocol.

In contrast, RLM supports multiple commonly used alignment tools and any reference genome that was used to align the reads. In addition to the four read-level methylation metrics (PDR, RTS, entropy, and epipolymorphism), it also outputs matching methylation rates per CpG and 4-mer as due to the filtering steps only a subset of reads at a given position is considered, which might not reflect the methylation rates as provided by standard methylation calling tools. RLM also outputs methylation metrics (methylation pattern, discordance, fraction of transitions) for every single read that passes the filtering steps. It also offers a specific RRBS mode, as described in the previous section.

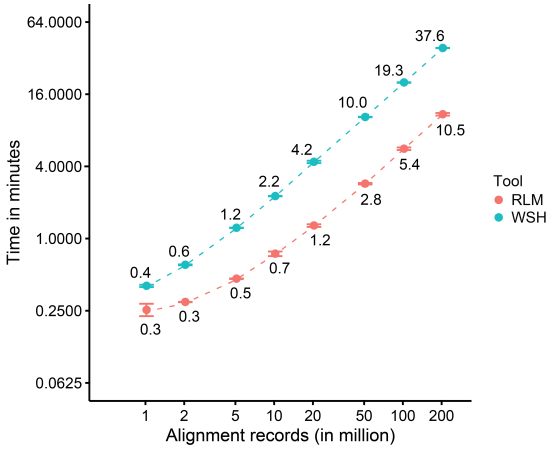

Figure 4.3.2: Runtime measured for RLM and WSH for different numbers of randomly sampled BAM records. Dots represent the average runtime across five iterations. Error bars represent the standard deviation.

Runtime comparison

Out of the existing tools, WSH is the most similar tool to RLM as it also offers multiple scores that RLM provides. As WSH only accepts alignment files produced by Bismark and preferentially aligned to hg38, a publicly available WGBS data set (human embryonic stem cell line HUES8, GSM3618718) was downloaded and aligned to the human reference genome hg38 using Bismark. In order to measure the performance, increasing numbers of BAM records were randomly sampled from the complete alignment file. RLM was executed in single-end mode as WSH does not offer specific treatment for paired-end reads. Both tools were otherwise executed with default parameters (WSH does not provide versioning, the package was installed from GitHub in May 2021). RLM uses two threads, one for the main program and one for the decompression of the input BAM file. For WSH, the documentation does not provide information about potentially underlying parallelizations. For both tools, the computation of methylation entropy was measured as WSH requires the pre-computed exact positions of CpGs for the PDR calculation. Runtime was measured as described in section 4.3.2. While the runtime is relatively comparable for very few alignment records, RLM extracts methylation entropy measurements more than three times faster than WSH for larger read numbers > 10 million (a typical WGBS experiment of the human genome consists of around 300 million fragments Figure 4.3.3).

Score comparison

RLM includes multiple filtering steps before the read-level calculations to ensure that only high-quality reads are considered for the analysis. This leads to a subset of reads being used as a basis for read-level methylation metrics. For example, out of the 50 million reads shown in Figure 4.3.3, only 8.2 million reads contain at least three CpGs, which is the minimum requirement for PDR and RTS calculations. Of these, however, only 82% survive the filtering steps before the read-level calculations. These filters are not implemented in the WSH package, which makes the resulting scores not comparable as they are based on different read sets. This means that

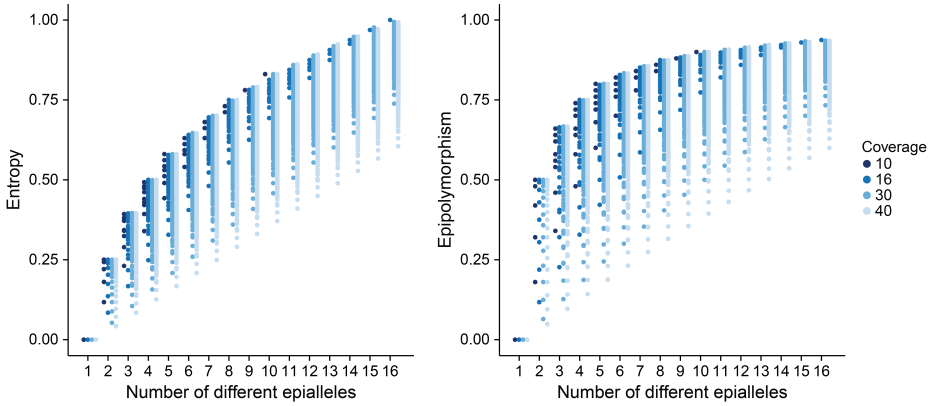

Figure 4.4.1: Entropy (left) and epipolymorphism (right) values that are associated with all possible distributions of one to 16 epialleles in a population, depending on the coverage.

for a given CpG or 4-mer, the read-level score could be drastically different between RLM and WSH due to the different sets of reads used for the calculations. Therefore, the resulting output of RLM and WSH was not compared, and the correctness of the RLM calculations was instead verified by the previously described test strategies.

4.4 Discussion

RLM was developed to provide a standalone, fast, and scalable tool that can extract multiple read-level methylation metrics, is compatible with many experimental set-ups and read mapping tools, and allows easy integration into existing bisulfite sequencing processing pipelines. In contrast to previously published tools that frequently offer single scores and support only a specific read mapping tool or reference genome, RLM is more flexible and thus suitable for a wider user group. Large consortia like the International Human Epigenome Consortium (IHEC) have standardized pipelines that include alignment tools such as GEM3 instead of Bismark, for which previously developed read-level methylation tools are not suitable.

A central question to future studies using metrics scoring the within-sample methylation heterogeneity of bisulfite sequencing data sets is which score to use. Different metrics were developed to study specific aspects of methylation heterogeneity, and they are subject to different biases and vulnerabilities. Scherer et al. provided a detailed comparison of different within-sample heterogeneity metrics with respect to their suitability for potential applications and their drawbacks concerning sources of noise and bias [223]. In the scope of this thesis, only the scores implemented in RLM are mentioned in the following paragraphs.

In contrast to entropy and epipolymorphism, PDR was developed to quantify intra-molecule heterogeneity of single reads (discordant or concordant) followed by the aggregation of this metric across all reads spanning a CpG. Scherer et al. defined the assessment of methylation erosion as the main strength of this score, together with the possibility to compute CpG-wise measurements in contrast to epiallele-based metrics that can only be computed per 4-mer. Based on simulated

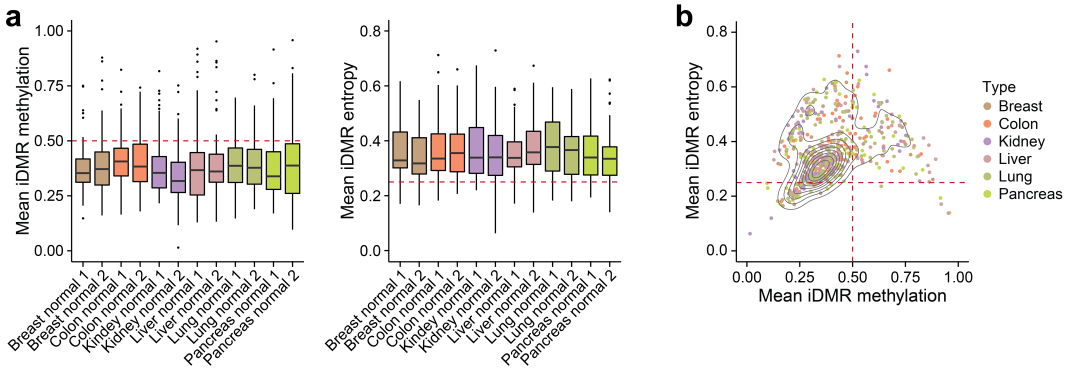

Figure 4.4.2: a) Distribution of the average methylation (left) and entropy (right) of iDMRs across healthy human tissues (samples from chapter 6). b) Scatterplot showing the relationship between average methylation and entropy of each iDMR across healthy human tissues. Lines reflect the density.

data sets, however, PDR was dependent and influenced by CpG density and read length as long read lengths and high CpG density increased the chance for a read to be discordant [223]. RTS was not covered in this comparative study but reflects similar properties as PDR and is, therefore, also likely to be influenced by read length and CpG density. These biases, however, are neglectable when comparing different samples generated with the same sequencing set-up at the same genomic loci.

Methylation entropy and epipolymorphism were both developed to quantify inter-molecule variability and were able to detect and quantify heterogeneity in simulated cell type mixtures and contamination data sets [223]. Scherer et al. showed that PDR was less sensitive in these scenarios. Interestingly, although PDR was developed specifically to detect methylation erosion, also entropy and epipolymorphism were able to detect simulated erosion events. Overall, inspecting multiple simulated scenarios of heterogeneous DNA methylation, Scherer et al. showed a relatively high correlation between epiallele-based metrics and PDR, suggesting that inter-molecule and intra-molecule variability are partially associated across the genome.

The drawback of epiallele-based metrics is that they can only be computed for a limited number of CpG-rich loci due to the requirement of four consecutive CpGs covered by the same read. Depending on the read length, this is mainly possible in CGIs, CpG-dense promoters, or CpG-rich repeat elements such as the 5' UTR of LINE elements [237]. Repetitive sequences, on the other hand, are frequently the subject of multi-mapping reads, which makes read-level methylation quantifications less reliable as the detected origin of a read might be ambiguous. In contrast to PDR, methylation entropy as a measurement itself is independent of the read length, as a fixed CpG-based window size is evaluated. However, it could be affected if reads get too short and are less likely to span four consecutive CpGs. However, methylation entropy and epipolymorphism have been found to increase with higher sequencing coverage [223]. A reason for this bias could be that the chance of noise increases with higher coverage. This is reflected by the increasing number of possible epiallele configurations that can lead to the same entropy or epipolymorphism values depending on the coverage (Figure 4.4.1).

Additionally, some practical biases should be considered when using read-level methylation met-

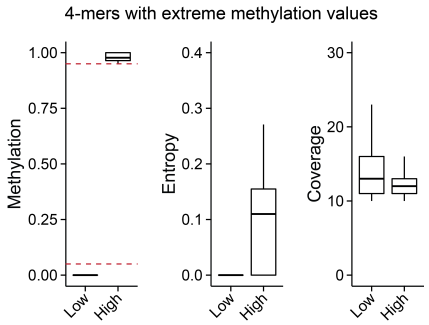

Figure 4.4.3: Distribution of methylation (left), entropy (middle), and coverage (right) for 4-mers present in CGIs of healthy breast tissue (sample from chapter 6), which have extreme methylation values (< 0.05 and > 0.95 termed low and high respectively).

rics. A good example to demonstrate this is genomic imprints. As described in the introduction, iDMRs should be spanned by reads, of which 50% are unmethylated while the other 50% are methylated, reflecting the methylation status of the two alleles within each cell. These regions should also be stably maintained across different healthy cell types. In theory, one would expect a corresponding PDR of zero and a methylation entropy value of 0.25, as shown in Figures 4.1.3 and 4.1.4. In reality, this is frequently not the case, as exemplified in Figure 4.4.2. First, the average methylation at the shown iDMRs is not accurately reflecting 50% but seems biased towards lower methylation levels. It has been reported previously that different steps during the library preparation, such as bisulfite conversion and PCR, can lead to biases that, depending on the protocol used, could either favor unmethylated or methylated CpGs leading to slightly shifted distributions in sequencing reads [163]. Additionally, sources of noise, such as sequencing errors or cells at different cell cycle phases that might be captured in the middle of replication, can increase the measured heterogeneity. For these reasons, the distribution of iDMR methylation entropy enriches close to 0.25; however, it also frequently deviates from the theoretical value (Figure 4.4.2).

The mentioned sources of noise can also lead to different types of artifacts. In theory, read-level methylation metrics are independent of low or high methylation levels meaning that fully methylated reads lead to the same score as fully unmethylated reads. The erosion of these patterns to the same extent should be reflected by the same entropy or PDR. In reality, regions with extremely high methylation are more likely to be affected by noise than extremely lowly methylated regions: During replication, the nascent strand is completely unmethylated, and methylation has to be copied by DNMT1. Cells captured during replication, therefore, might still contain hemi-methylated DNA that does not affect unmethylated regions such as CGIs but regions with intermediate or high methylation levels. At the same time, sequencing errors or mutations that result, for example, in deamination would appear as unmethylated CpGs in a bisulfite sequencing data set as such an event is not distinguishable from a bisulfite conversion of unmethylated cytosines. This means that more mechanisms exist that could lead to the (seeming) removal of a methylated CpG, including incomplete maintenance, deamination, mutations, or active removal by TET enzymes, whereas *de novo* methylation of a CpG can only be carried out by active targeting by DNMT3 enzymes. This is exemplified in Figure 4.4.3 where the most extremely methylated 4-mers (< 0.05 and > 0.95 termed low and high respectively) within CGIs were extracted from a

healthy breast tissue sample. When considering the distributions of associated methylation and entropy, one can observe that the highly methylated 4-mers overall tend to show methylation values < 1 associated with higher entropy. In contrast, the lowly methylated 4-mers are mostly completely unmethylated, resulting in zero entropy.

In summary, read-level methylation metrics have different strengths and weaknesses that might be relevant to consider, especially when interpreting corresponding results. With RLM, a new application has been developed that provides fast and easy access to a variety of metrics for the DNA methylation community based on frequently used experimental setups and processing pipelines. This new availability enables the standard usage of this additional layer of information from bisulfite sequencing experiments, which has been associated with important phenotypic aspects, specifically in cancer.

Part III

DNA methylation in cancer

The third part of this dissertation describes the application of previously introduced concepts and tools for DNA methylation analysis in cancer. First, a methylation study of a single tumor type, acute lymphoblastic leukemia, is presented, combining epigenetic information with transcriptional and mutational data to gain insights into the specific regulation of its DNA methylation landscape. Second, a large-scale pan-cancer study comparing the methylome of primary tumors and cancer cell lines is presented, integrating newly generated high-coverage sequencing data sets as well as thousands of publicly available samples.

Chapter 5

The distinct DNA methylome of acute lymphoblastic leukemia

This chapter presents the largest to-date study of whole-genome bisulfite sequencing data sets from patients with acute lymphoblastic leukemia (ALL). This work was published in *Nature Cancer* in May 2022 [238], and the chapter follows the publication closely, including figures that have been adapted from it. Different subtypes of ALL were analyzed and compared both with each other and to a variety of publicly available hematopoietic and solid tumor types. Methylation data were integrated with transcriptomic and mutational data sets to identify potential regulators of the DNA methylation landscape in ALL. Finally, methylation and transcription of multiple cancer cell lines, including a perturbation experiment, have been profiled to provide further evidence for the role of epigenetic regulators identified in primary ALL cases.

All patient data sets have been acquired and generated as part of the St. Jude Children's Research Hospital - Washington University Pediatric Cancer Genome Project. Most methylation data sets of cancer cell lines have been generated at the St. Jude Children's Research Hospital (Memphis, TN) in the lab of Charles G. Mullighan MBBS (Hons), MSc, MD. Methylation and transcriptional data sets for two cancer cell lines have been generated at the Max Planck Institute for Molecular Genetics by Dr. Alexandra L. Mattei (cell culture and library preparation) and the Sequencing Core Facility. Dr. Alexandra L. Mattei performed perturbation experiments.

5.1 Biological background

5.1.1 Acute lymphoblastic leukemia

Acute lymphoblastic leukemia is the most common pediatric cancer type, which develops from B or T lymphoblasts resulting in the rapid accumulation of immature lymphocytes that can lead to death from bone marrow failure (Figure 5.1.1) [239, 240]. Around 50% of the ALL cases affect adults where the disease has a relatively poor prognosis: While more than 90% of pediatric patients are alive five years after diagnosis, only 25% of adult patients survive more than five years (data based on ALL cases in the USA, 2009) [241].

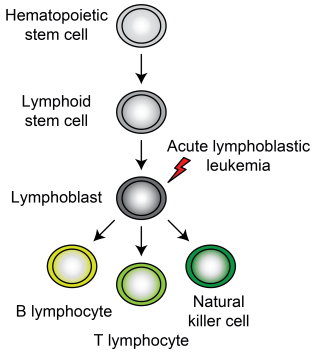

Figure 5.1.1: Development of lymphocytes from hematopoietic stem cells: Acute lymphoblastic leukemia develops from immature lymphocytes called lymphoblasts (adapted from [242]).

ALL consists of different subtypes characterized by specific mutations, chromosomal rearrangements, aneuploidy, or overexpression of oncogenes. In B cell acute lymphoblastic leukemia (B-ALL), amongst others, subtypes have been described based on fusion genes such as *BCR-ABL1*, *DUX4* rearrangements, *PAX5* mutations, hypo- and hyperdiploidy [243]. T cell acute lymphoblastic leukemia (T-ALL) originates from different stages of immature thymocytes and can be classified based on the stage where normal development arrested: Early T cell precursor ALL (ETP-ALL), early and late cortical T-ALL [244]. Different mutation and expression patterns characterize these stages. ETP-ALL frequently carries mutations in *NRAS*, *FLT3*, *ETV6*, and *RUNX*. Early cortical T-ALL often exhibits overexpression of the oncogenes *TLX1* and *TLX3*, mutations of *NOTCH1* and *WT1*, loss of *CDKN2A* as well as *NUP214-ABL1* fusion. Late cortical T-ALL cases are characterized by aberrant expression of *TAL1*, *LMO1*, or *LMO2* frequently induced via fusion with other genes in addition to *NOTCH1* and *CDKN2A* mutations [244].

5.1.2 Previous studies on DNA methylation in ALL

Previous studies in ALL have mainly used the Infinium HumanMethylation450 BeadChip (450k array) or enrichment-based sequencing approaches to study methylation patterns across ALL subtypes [245–250]. These assays have the advantage that they are relatively low in cost, thus enabling the use of larger patient cohorts. However, they are also limited in the information that can be obtained as they prioritize CpG-dense regions such as CGIs and regulatory regions such as promoters, gene bodies, and enhancers and, therefore, cannot provide a representation of the complete genome (see section 2.6). Few studies have used genome-wide sequencing and provided exemplary samples from subtypes such as ETV6-RUNX1, high hyperdiploidy, unclassified B-ALL, and T-ALL. Here, contradicting results about the global methylation landscape of ALL have been reported ranging from mild genome-wide hypermethylation to significant hypomethylation [251–253].

Based on methylation array cohorts, a subset of patients in T-ALL has been reported to exhibit a CpG island hypermethylator phenotype (see section 2.5.2) [254]. The subtyping into CIMP-positive and -negative patients has been established by clustering the patients based on a selection of array probes in CGIs. The CIMP-positive T-ALL cases have been shown to coincide with

increased expression of ANTP homeobox genes, shorter telomere length, and higher mitotic age, which has led to the hypothesis that these patients might follow a different route to tumorigenesis compared to CIMP-negative patients. Additionally, patients of the CIMP-positive group have been associated with better prognosis and overall survival [254]. Another study using mouse models of T-ALL combined with patient-derived data sets also found an association of CIMP with the proliferative history of cells and suggested that the phenotype could arise following a preleukemic phase that leads to increased mitotic age and CGI hypermethylation [255].

In previous studies, no link between CIMP subtypes and recurrent mutations or expression differences in epigenetic regulators could be identified [254]. A study using mouse models and T-ALL cancer cell lines described the oncogene MYC to orchestrate expression changes of the key DNA methylation regulators TET1, TET2, DNMT3B, and DNMT1 in T-ALL where the loss of MYC led to changes in the DNA methylation landscape. However, the observed dynamics were not investigated with respect to a CIMP subtyping in patients [256]. Therefore, further studies are needed to link findings from model systems to reported CGI-based methylation phenotypes in primary T-ALL cases.

5.1.3 Aims and scope of the study

In this study, we aimed to address open questions in the field of DNA methylation in ALL. We generated whole-genome bisulfite data sets of a large cohort comprising different ALL subtypes to study the genome-wide methylation landscape of these tumors and place the ALL methylome with respect to known global changes of DNA methylation in a pan-cancer comparison. We made additional use of the higher CpG coverage and revisited the CGI hypermethylation dynamics previously observed in T-ALL integrated with mutational and transcriptomic data sets. Based on our findings, we selected a different computational approach compared to previous studies based on correlation tests to establish links between the expression of epigenetic regulators and global as well as CGI methylation levels. Finally, we used a perturbation experiment in T-ALL cell lines specifically selected to resemble different patient-derived methylation features to solidify observations from T-ALL patients.

5.2 Materials and methods

5.2.1 Cohort overview

WGBS data sets of leukemic cells from 82 patients were generated as part of the St. Jude Children's Research Hospital - Washington University Pediatric Cancer Genome Project. Patients were diagnosed with one of three B-ALL subtypes (DUX4-rearranged/ERG-deregulated, hypodiploid and Philadelphia chromosome (Ph)-like ALL) or T-ALL (Figure 5.2.1). Since ALL is a common pediatric tumor, most samples came from pediatric or adolescent patients, and only for T-ALL one adult patient was included in the cohort. For a large number of patients, RNA-Seq and whole-genome or -exome sequencing (WES) data were already generated and published previously, which could be re-used in this study (accession numbers: EGAS00001005203, EGAS00001004810, EGAS00001005250, EGAS00001005084, EGAS00001001923, phs000218,

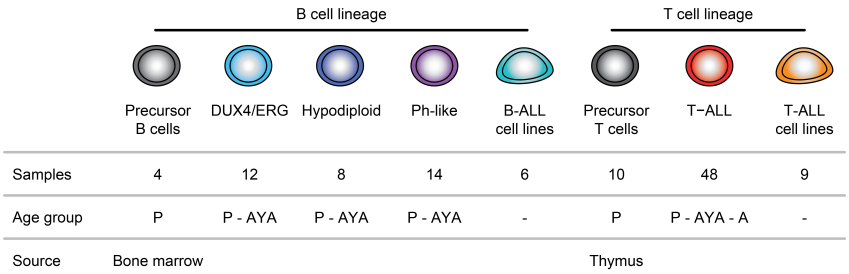

Figure 5.2.1: The number of samples of healthy precursor B and T cells, ALL subtypes, and ALL cancer cell lines profiled with WGBS. Samples are grouped according to their age into pediatric (P, 0-15 years), adolescents and young adults (AYA, 16-39 years), and adults (A, $\geq$ 40).

EGAS00001003266, and EGAS00001000654). As control samples, healthy precursor B and T cells of different developmental stages were sorted from bone marrow and neonatal thymi, respectively, and profiled with WGBS (four precursor B and 10 precursor T cell samples). Additionally, WGBS data for five B-ALL and nine T-ALL cell lines were generated. The cell lines were selected matching the primary patient subtypes: NALM-6 (DUX4/ERG, two replicates), MUTZ5 and MHH-CALL-4 (Ph-like), NALM-16 and MHH-CALL-2 (hypodiploid B-ALL), and the T-ALL cell lines MOLT-16 (*TRA-MYC* fusion), PEER (*NUP214-ABL1* fusion), PER-117 (*STAG2-LMO2* fusion), RPMI-8402 (*STIL-TAL1* fusion), LOUCY (*SET-NUP214* fusion), TALL-1 (*STAG2-PTGER3* fusion), ALL-SIL (*NUP214-ABL1* fusion), Jurkat (TAL1 overexpression via enhancer mutation), and DND41 (*TLX3-BCL11B* fusion). All cell lines were ordered from DSMZ with the exception of PER-117, which was gifted by Prof. Dr. Ursula Kees.

External data sets

In order to place the characteristics of the ALL methylome found in this study with respect to previously published results in ALL but also other cancer types, a selection of publicly available WGBS data sets was used:

1. B-ALL: Seven B-ALL samples of unknown subtype published by the Blueprint Epigenome project [257].

2. Other hematopoietic tumor types: 12 acute myeloid leukemia (AML), three T cell prolymphocytic leukemia (TPLL), six chronic lymphocytic leukemia (CLL), and five mantle cell lymphoma (MCL) samples published by the Blueprint Epigenome project [257].

3. Healthy hematopoietic cell types: Different maturation stages of B and T cells from multipotent progenitor cells to terminally differentiated memory B and T cells published by the Blueprint Epigenome project [257]. This included memory B cells as control samples for CLL and MCL, hematopoietic multipotent progenitor cells (HPCs) as a control sample for AML, and CD4 and CD8 single positive alpha beta T cells as control samples for TPLL.

4. Solid tumor types: six bladder urothelial carcinoma (BLCA), five breast invasive carcinoma (BRCA), two colon adenocarcinoma (COAD), five lung adenocarcinoma (LUAD), four lung

squamous cell carcinoma (LUSC), two rectum adenocarcinoma (READ), four stomach adenocarcinoma (STAD), and five uterine corpus endometrial carcinoma (UCEC) samples from The Cancer Genome Atlas (TCGA) [72]. One corresponding normal control was available for each tumor type and included in the pan-cancer comparison.

5.2.2 Initial data processing

The experimental procedures used to generate methylation and expression data sets are described in the appendix B.

Whole-genome bisulfite sequencing

The quality of the sequencing runs was inspected using FastQC (version 0.11.9) [171]. Reads were subjected to trimming using TrimGalore (version 0.4.4) [258]: Bases with a quality score less than 30 were removed, as well as adapter content, which could otherwise impact the alignment of the reads. Additionally, 10 bases were trimmed from each end of the read pairs. The trimmed reads were then aligned to the human reference genome (hg19) using the alignment tool BSMAP with default parameters (version 2.90) [183]. Following the alignment, PCR duplicates were removed using GATK with the 'MarkDuplicates' command (version 4.1.4.1) to avoid technical bias or artifacts in the subsequent analysis steps [187]. Afterwards, methylation rates were called with mcall (MOABS package, version 1.3.2) [188]. The resulting methylation rates were filtered such that only CpGs covered by at least 10 and at most 150 reads on autosomes were considered for downstream analyses resulting on average in around 22 million CpGs per sample. The sex chromosomes were omitted to not bias the analysis by sex-specific differences (missing Y chromosome and one inactivated, fully methylated second X chromosome in females).

RNA sequencing

Both patient and cell line samples were processed the same way. The quality of the sequencing runs was assessed using FastQC (version 0.11.9) [171]. The reads were then trimmed using cutadapt (version 2.4) [172]: Low-quality bases (less than Q20) and adapter content were removed. Additionally, poly-A tails were trimmed as these are added post-transcription to the RNA and, therefore, cannot be aligned to the reference genome. Afterwards, reads were aligned to the human reference (hg19) using STAR (version 2.7.5a) [259] and gene as well as transcript expression was quantified using stringtie (version 2.0.6) [260] with the gene annotation obtained from GENCODE (release 19). Fusion gene calls were provided by the group of Charles G. Mullighan (St. Jude Children's Research Hospital).

Whole-genome and whole exome sequencing

WGS and WES data sets of patients were generated and analyzed at the St. Jude Children's Research Hospital as described previously, and most data sets were already published in earlier

studies (see section 5.2.1). Mutation calls were provided by the group of Charles G. Mullighan (St. Jude Children's Research Hospital).

5.2.3 ALL subtype and pan-cancer DNA methylation analysis

The following section outlines how the methylome of ALL subtypes was characterized and placed with respect to other tumor types in a pan-cancer comparison.

Global quantifications

In order to broadly classify healthy and patient samples in terms of their global and CGI methylation levels, the arithmetic mean across all sufficiently covered CpGs (between 10 and 150 reads) except those in CGIs (global) or exclusively overlapping CGIs was computed. As a first step, this allowed us to compare the overall methylation levels of the two broad features differently affected in cancer methylomes (global hypomethylation and CGI hypermethylation).

Genomic features

Different features were defined and subsequently compared between samples or cancer types (introduced in the following paragraphs). To compare different cancer types, for specific analyses, a tumor type-specific CpG-wise methylation profile was computed by calculating the arithmetic mean per CpG across all samples of the respective ALL subtype or cancer type. Only CpGs covered by at least 80% of the respective samples were considered for the average subtype methylation profile. The average methylation per sample or subtype of each feature was then calculated using the arithmetic mean of all CpGs within the defined feature only if at least three covered CpGs overlapped the region. This ensures that features are not defined based on single CpGs due to low coverage and, by that, could be more susceptible to noise. In the following, the definition of the different features used in this study is presented.

Genomic tiles Genomic tiles of size one kb were generated by segmenting the genome into non-overlapping, consecutive windows using bedtools (makewindows) [261].

CpG islands The annotation of CGIs was downloaded from UCSC for the human reference genome (hg19). CGI shores were defined as the two kb flanking each island on each side, while CGI shelves were defined as the two kb flanking each shore.

Highly and partially methylated domains Zhou et al. defined HMDs and PMDs previously based on the variability of isolated CpGs (solo-WCGW CpGs) across a pan-cancer WGBS cohort (see sections 2.3 and 2.5.2) [72]. They showed that their genome segmentation into HMDs and PMDs per 100 kb tile also generalized to healthy tissues and external tumor cohorts with high overlap, including a public data set of B-ALL cases (Blueprint, unknown subtype). Therefore,

HMDs and PMDs in this study were assigned using the annotation by Zhou et al., and for exemplary heatmap representation, the average of all solo-WCGW CpGs within an HMD or PMD was calculated.

Sliding windows In order to compare genome-wide methylation distributions more thoroughly, sliding windows of the genome were computed using bedtools (makewindows) [261]. Windows were defined to have a size of one kb with a step size of 250 bp (consecutive windows overlapping by 750 bp). Every window was assigned to a PMD or HMD based on the largest overlap with these previously defined regions (see previous paragraph). The average of each sliding window per subtype was calculated excluding CpGs in CGIs as this analysis was used to assess global hypomethylation levels, which could be biased by CGIs that generally deviate from the genomic background behavior (see 2.3). Sliding windows per subtype or tumor type were compared to the respective healthy control (see section 5.2.1), and windows with a delta methylation < -0.1 or > 0.1 were defined as hypo- or hypermethylated respectively.

DNA methylation valleys DMVs were defined separately for precursor B and T cells. For this purpose, the average methylation of the respective healthy samples was calculated in sliding windows of five kb size with a step size of one kb (consecutive windows overlapping by four kb). For this purpose, the average CpG-wise methylation signature for healthy B and T cells was used, and CpGs overlapping CGIs were excluded. Sliding windows were computed using bedtools (makewindows) [261]. The window and step size were chosen based on previous studies on DMVs [96, 97]. The average methylation of CGIs was computed separately. Next, sliding windows and CGIs were filtered to select candidates with an average methylation < 0.15 and ≥ 10 CpGs. These candidates were merged if overlapping, and regions were retained as DMVs that did not consist of single CGIs but included unmethylated flanking parts defined by overlapping/neighboring sliding windows.

Promoters Promoters were defined as the 1.5 kb upstream and 500 bp downstream of the transcription start site (TSS) defined by the GENCODE gene annotation.

DMR calling

Differentially methylated regions (DMRs) between ALL subtypes and healthy precursor lymphocytes (precursor T cells as control for T-ALL, precursor B cells as control for DUX4/ERG, hypodiploid and Ph-like ALL) were computed using the tool metilene (version 0.2-8) [262]. DMRs were required to contain at least 10 CpGs not further than 300 bp apart. The average methylation difference between the two groups (healthy control and ALL subtype) was required to be at least 0.2. The selection of parameters ensures that only strong, local changes consistently affecting multiple CpGs that distinguish tumor from healthy control cells are detected. The output of metilene consists of candidate regions fulfilling these requirements with a *p*-value assigned (calculated using a two-dimensional Kolmogorov–Smirnov test). Only DMRs with an adjusted p-value < 0.05 (corrected for multiple testing using Bonferroni correction) were selected. DMRs

were classified into hyper- and hypomethylated (positive or negative difference with respect to the control cells).

DMRs were then annotated and assigned to specific features if either 20% of the DMR or 20% of the feature overlapped. The overlap was calculated using bedtools (intersectBed) [261]. The following features were used:

1. Methylation-based features: CGIs, CGI shores, CGI shelves, DMVs, and PMDs (see 2.3).
2. Chromatin states: Segmentation of the genome based on histone modifications into 15 different chromatin states by ChromHMM defined in hematopoietic stem cells and obtained from Roadmap [263]. The 15 states were collapsed into the following groups:
    - Active TSS (marked mainly by H3K4me3 and H3K27ac; states: 1_TssA and 2_TssAFlnk)
    - Bivalent TSS (marked mainly by H3K4me3 and H3K27me3; states: 10_TssBiv and 11_BivFlnk)
    - Transcript (marked mainly by H3K36me3; states: 3_TxFlnk, 4_Tx, and 5_TxWk)
    - Enhancer (marked mainly by H3K27ac and H3K4me1; states: 6_EnhG, 7_Enh, and 12_EnhBiv)
    - Heterochromatin (marked mainly by H3K9me3; states: 8_ZNF/Rpts and 9_Het)
    - Repressive (marked mainly by H3K27me3; states: 13_ReprPC and 14_ReprPCWk)
    - Quiescent (no modifications; states: 15_Quies)

Random background DMRs were defined to determine whether the number of DMRs overlapping with a specific feature follows the genome-wide distribution or is enriched or depleted in comparison. For this purpose, regions were selected that fulfill the same requirements as applied during the DMR calling (at least 10 CpGs, not more than 300 bp apart, same sizes as the called DMRs). Out of these candidates, an equal number of background DMRs compared to the actual DMRs was randomly sampled with 1000 repetitions. The resulting set of background DMRs was annotated with respect to genomic features in the same way as the actual DMRs. The enrichment E_F of DMRs in a specific feature class F compared to the genomic background was calculated as

$$E_F = \frac{n_F}{N} / \frac{m_F}{M} \tag{5.1}$$

where n_F and m_F denote the number of DMRs and background DMRs, respectively, overlapping a feature of class F. N and M represent the total number of DMRs and background DMRs, respectively. Values > 1 imply enrichment of DMRs in feature class F, while values < 1 reflect depletion.

5.2.4 CGI cluster analysis

The following section describes how CGIs were clustered based on their methylation profile in T-ALL and healthy precursor T cell samples and annotated subsequently using different types of

genomic features.

Consensus clustering

Defining the optimal number of clusters and assessing their robustness and stability are fundamental challenges of clustering data. Consensus clustering allows assessing the optimal number of clusters and their stability by generating a consensus across multiple clustering runs. Using a clustering algorithm and distance metric of choice (e.g., hierarchical or k-means clustering with Euclidean or correlation distance), clustering is repeated n times (where n is user-defined). In each iteration, a subset of features or samples (or both) is randomly sub-sampled and used as input for the clustering. A consensus of all iterations is then computed for the final assignment of samples to clusters based on the presented features. This procedure is performed for different cluster numbers k for $k = 2, 3, ..., k_{max}$. The optimal number of clusters is selected based on the number that produced the most stable result across all iterations [264].

To cluster CGIs based on the methylation levels present in T-ALL and precursor T cells, only CGIs that were covered by all samples with at least three CpGs were considered (48 T-ALL and 10 precursor T cell samples). Additionally, CGIs that remained consistent across all samples were excluded from the clustering to remove features with low variability. These excluded CGIs were defined as "low" if the mean methylation for all samples was < 0.2 or as "high" if the mean methylation for all samples was > 0.8. The remaining islands were clustered using consensus clustering with partitioning around medoids, Euclidean distance, and 100 repetitions using the R package "ConsensusClusterPlus" [265] (Figure 5.2.2). After the consensus clustering, this package reports the empirical cumulative distribution functions (ECDFs), which show the measured consensus distribution for each value of k that can be used to select the optimal number of clusters with maximum stability [265]. For this purpose, also the relative change in the area under the ECDF curve between k and $k-1$ is reported to allow the selection of k, after which only minor increases can be observed [265]. Based on these visualizations, the optimal number of clusters (in this case, $n = 4$) was determined. The methylation of CGIs across patients per cluster was visualized using the ComplexHeatmap package [266].

Characteristics of CGIs

In order to characterize the CGIs associated with each cluster and the two previously excluded groups of CGIs (low and high), the length of each CGI was computed in bp, the number of CpGs within each island was counted, and the GC content was assessed using

$$GC_{content} = \frac{\#C + \#G}{\#A + \#C + \#G + \#T} \tag{5.2}$$

where $\#A$, $\#C$, $\#G$, and $\#T$ denote the number of the respective base within a CGI.

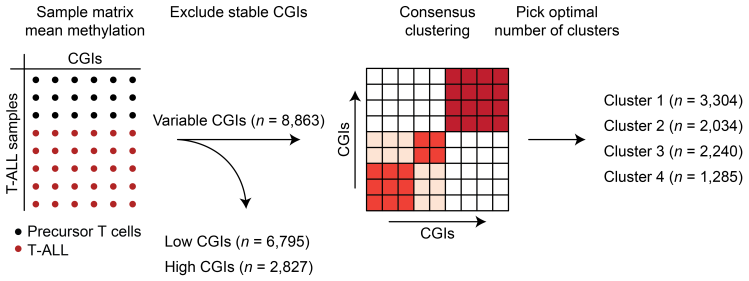

Figure 5.2.2: Schematic of the consensus clustering of CGIs based on precursor T cell and T-ALL samples. Starting with all commonly covered CGIs ($n = 18{,}485$), stably lowly or highly methylated CGIs were excluded before the clustering (< 0.2 and > 0.8 in all samples, respectively). The average methylation of each of the remaining variable CGIs per sample was used as input for the consensus clustering. In 100 iterations, 80% of CGIs were randomly sampled and clustered based on all samples using partitioning around medoids with Euclidean distance. The consensus of all 100 repetitions was reported, and the optimal number of clusters was picked based on the most stable clusters across all repetitions (in this case, four clusters).

Association with genomic features

For downstream analysis, CGIs associated with each cluster or previously excluded group were annotated using genomic features as described for DMRs (see 5.2.3).

5.2.5 T-ALL methylation-based subtyping

The following section describes how T-ALL samples were grouped into methylation-based subtypes using CGI methylation levels to compare samples of relatively low or high methylation levels with respect to covariates.

Clustering of samples

T-ALL samples were grouped into CGI methylation-based subtypes using hierarchical clustering with Euclidean distance based on the variable CGIs defined in 5.2.4. The top three clusters were selected, which reflected groups of samples with rather low, intermediate, and high CGI methylation and termed T-ALL^{LM}, T-ALL^{IM}, and T-ALL^{HM} respectively. These groups were then subsequently used to test the association of CGI methylation levels with different covariates.

Association with genetic and transcriptomic drivers

Based on available mutation and fusion gene calls, as well as demographic data provided by the St. Jude Children's Research Hospital, the following covariates were tested for association with the CGI methylation-based subtypes defined in T-ALL:

1. Demographic data (age group and sex)

2. Commonly mutated genes (NOTCH1, NRAS, WT1, MED12, SUZ12, ETV6, FLT3)
3. Genetic subtypes based on fusion genes or overexpression (HOXA, TLX3)

The association with either all three methylation groups or only T-ALL^{LM} and T-ALL^{HM} as the most extreme groups were tested using Fisher's exact test, which is suitable for small counts in contingency tables.

Methylation entropy analysis

Methylation entropy per 4-mer in CGIs was calculated using RLM [222], which has been introduced in chapter 4. The entropy of all 4-mers within a CGI was subsequently averaged.

5.2.6 Correlation of DNA methylation and gene expression

In this section, the methods used to identify genes associated with different methylation levels within and across ALL subtypes are presented.

Correlation test

In order to detect genes whose expression is associated with different levels of CGI or global methylation in ALL patients in an unbiased manner (without pre-grouping patients into methylation-based subtypes), a correlation test between the log2-transformed transcript per million (TPM) of each gene and the global average methylation (excluding CpGs in CGIs) or the overall CGI methylation across patients was conducted. Only active genes (average TPM across all samples ≥ 0.5) were considered to not unnecessarily inflate the number of tests. The correlation test was conducted using Spearman's correlation coefficient. In contrast to Pearson's correlation coefficient, which evaluates linear relationships, Spearman's correlation coefficient evaluates monotonic relationships, which is essential as DNA methylation and expression levels are not necessarily linearly associated [267]. *P*-values were corrected for multiple testing using FDR. Genes with an adjusted p-value < 0.01 were termed significant. For this analysis, T-ALL, DUX4/ERG, and Ph-like B-ALL patients were considered. Hypodiploid B-ALL samples were excluded from the analysis due to their high aneuploidy, which can affect gene expression (monoallelic instead of biallelic) and, therefore, might bias the correlation analysis. In this section, the methods used to identify genes associated with different methylation levels within and across ALL subtypes are presented.

Promoter methylation analysis of epigenetic regulators

In addition to the correlation test between gene expression and per-patient methylation levels, the promoter methylation of a selection of epigenetic regulators was inspected to identify potential silencing events of these genes by methylation. These events might be rare, only affecting a smaller subset of patients, and thus not be picked up by the correlation analysis. A set of epigenetic regulators was selected based on the following characteristics:

1. Proteins that directly regulate DNA methylation, such as methyltransferases and TET enzymes (DNMT1, DNMT3A, DNMT3B, TET1, TET2, TET3).
2. Proteins that have been reported to be involved in the recruitment or other regulation of the above-named enzymes in a cancer context:
    - MYC: Found in T-ALL mouse models to influence the expression of DNA methylation regulators such as TET1, TET2, DNMT3B, and DNMT1 [256].
    - WT1: Recruits TET2 (shown in AML) [268].
    - IDH1/IDH2: Mutations of these genes cause a CIMP phenotype (shown in gliomas) [143].
3. Polycomb group proteins (EED, EZH2, SUZ12, RING1, RNF2, KDM2B, BAP1).
4. H3K9 histone methyltransferases (SUV39H1, SUV39H2, EHMT2). The cross-talk between H3K9me3 and DNMT1 has been shown to enable stable methylation maintenance [100].
5. Chromatin remodeler (ARID1A, ARID1B, ARID2, PBRM1, SMARCA4, SMARCB1).
6. Other genes:
    - QSER1: Shields a subset of DMVs from *de novo* DNA methylation [269].
    - HELLS: Shown to regulate methylation at repetitive elements while its loss leads to global hypomethylation [270].

For each of these genes, the methylation of the promoter was assessed the following way: If a CGI overlapped the promoter by at least 20% of either the CGI or the promoter, the average methylation of the promoter was defined as the CGI methylation level. If multiple CGIs overlapped the promoter, the average of both CGIs was used. If no CGI overlapped the promoter (only ARID1B had no promoter CGI), the average methylation of the entire promoter region (1500 bp upstream and 500 bp downstream of the TSS) was used.

5.2.7 ALL cell line analysis

The following section describes analyses conducted using cancer cell lines as *in vitro* models for ALL. The cell lines selected for this purpose have been listed in section 5.2.1. Experimental methods are described in the appendix B.

Clustering of samples

Healthy precursor B and T cell, ALL patient and cell line samples were clustered using hierarchical clustering with Euclidean distance based on the most variable CpGs (top 5% CpGs with the highest standard deviation across all considered samples) in order to visualize the distance between samples based on their DNA methylation profile.

Differential expression analysis

Differential gene expression between Jurkat and DND41 cells, as well as between Jurkat cells with and without TET2 knockout, was carried out using the R package DESeq2 [271]. For this purpose, three replicates of each condition were used. Only genes with at least 10 reads across all replicates used in each comparison were considered. Genes were termed differentially expressed with an absolute log2 fold change > 1, a p-value adjusted for multiple testing < 0.05, and an average TPM across all considered replicates ≥ 0.5. The last filter step was applied to remove genes with a high log2 fold change but overall very low expression levels. A change from 0.2 to 0.8 TPM in expression might be determined as a significant rise. However, the overall low expression levels might not be biologically meaningful.

5.3 Results

5.3.1 Genome-wide methylation of ALL subtypes

Global methylation levels

Global DNA methylation has been reported to decrease during tumorigenesis as described in section 2.5.2. Given the few whole-genome bisulfite sequencing data sets published previously for patients with ALL and the sometimes contradicting results reported, we first analyzed and displayed genomic background methylation of ALL patients and respective controls in different ways to assess how ALL positions compared to other tumor types studied previously. For this purpose, we additionally used publicly available data sets of various solid and hematopoietic tumor types (overview of the data sets provided in section 5.2.1). Examining representative samples from two tumor types that have been previously reported to present with global hypomethylation - CLL and COAD -, we observed the characteristic loss of methylation compared to the respective healthy tissue (memory B cells and healthy colon tissue) in genome browser tracks as well as CpG-wise density plots (Figure 5.3.1 and 5.3.2). When we compared this to patients from different ALL types, we instead observed that especially T-ALL remained highly methylated comparable to precursor T cells, while B-ALL samples only mildly lost methylation in contrast to the more drastic decrease in CLL and COAD.

To improve the quantification and additionally show the indication-intrinsic variability across samples from our and external cohorts, we compared ALL and other tumor samples per type using sample-wise average CpG methylation excluding CpGs in CGIs (Figure 5.3.3). CGIs are CpG-rich regions (8% percent of CpGs on human autosomes are located in CGIs) but are usually unmethylated in healthy tissues with the possibility to hypermethylate in tumors. CpGs in CGIs were therefore excluded in this analysis to not skew the genomic background quantification. Hematopoietic tumor types and healthy blood cells generally exhibited higher methylation levels than solid tissues and tumors. However, when comparing tumors with their respective control tissue, ALL subtypes, as well as AML, preserved unusually high genome-wide methylation, which for some T-ALL samples even exceeded the levels of precursor T cells. B-ALL samples exhibited a mild decrease but remained highly methylated (of all B-ALL subtypes Ph-like ALL showed the

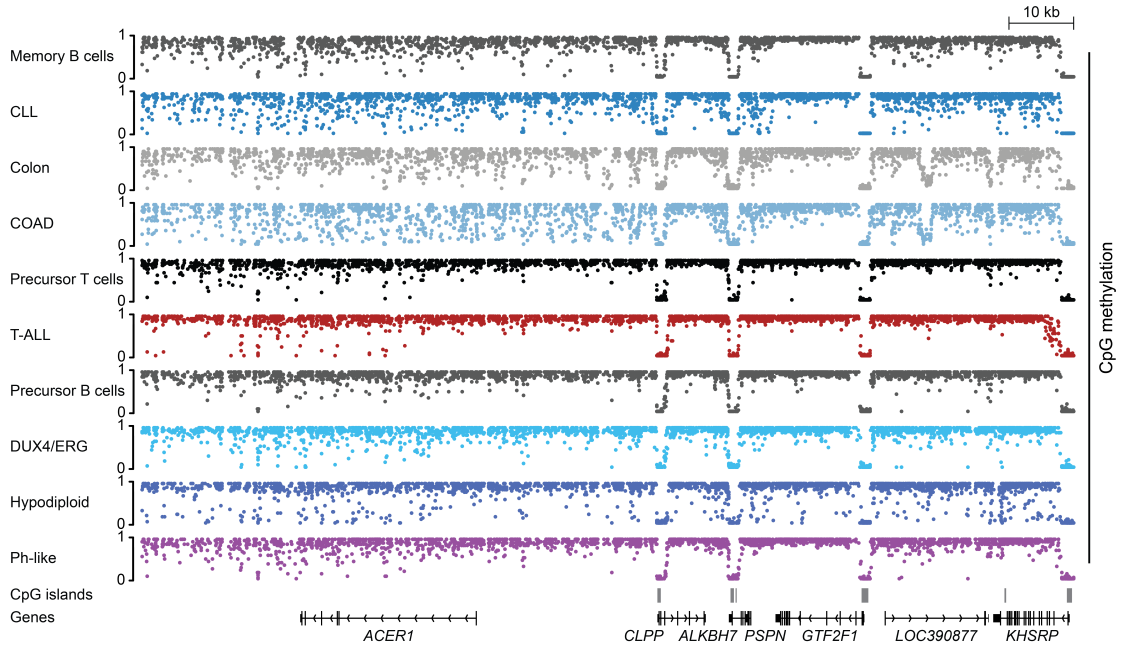

Figure 5.3.1: Genome browser tracks of memory B cells, CLL, healthy colon, COAD, precursor T cells, T-ALL, precursor B cells, and B-ALL subtypes. Both CLL and COAD exhibit loss of methylation in previously highly methylated regions, while the T-ALL sample remains highly methylated comparable to healthy precursor T cells. B-ALL samples exhibit minor loss of methylation compared to precursor B cells.

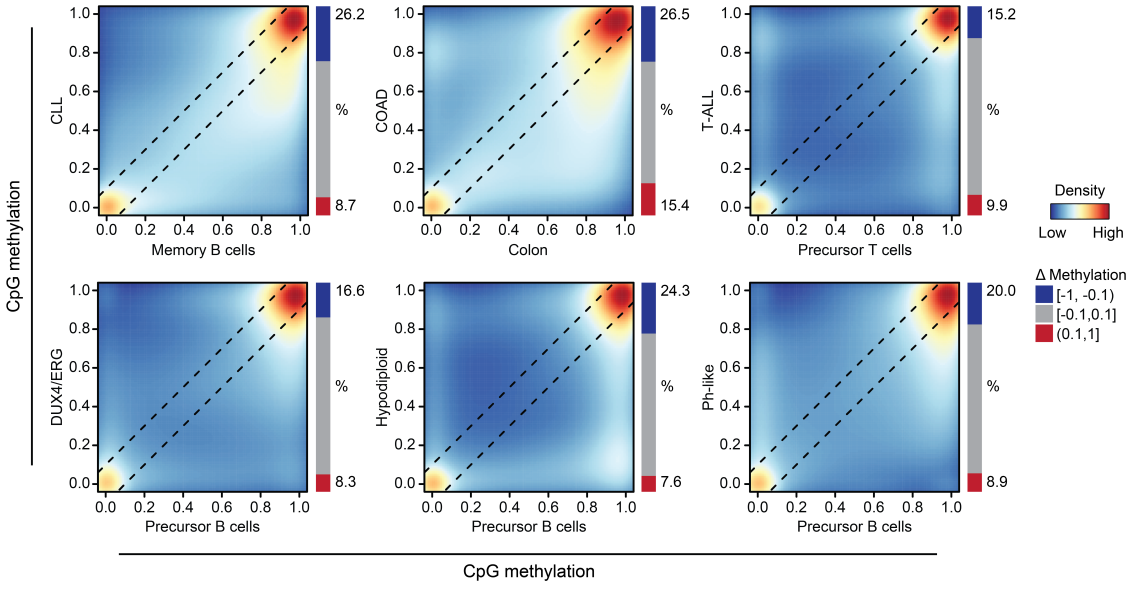

Figure 5.3.2: Density plot showing the CpG-wise comparison between examples of CLL, COAD, ALL subtypes, and their respective healthy tissue. The barplot indicates the percentage of CpGs that are hyper- or hypomethylated in the cancer samples compared to their healthy counterparts. T-ALL, DUX4/ERG, and Ph-like ALL show smaller proportions of hypomethylated CpGs compared to CLL and COAD.

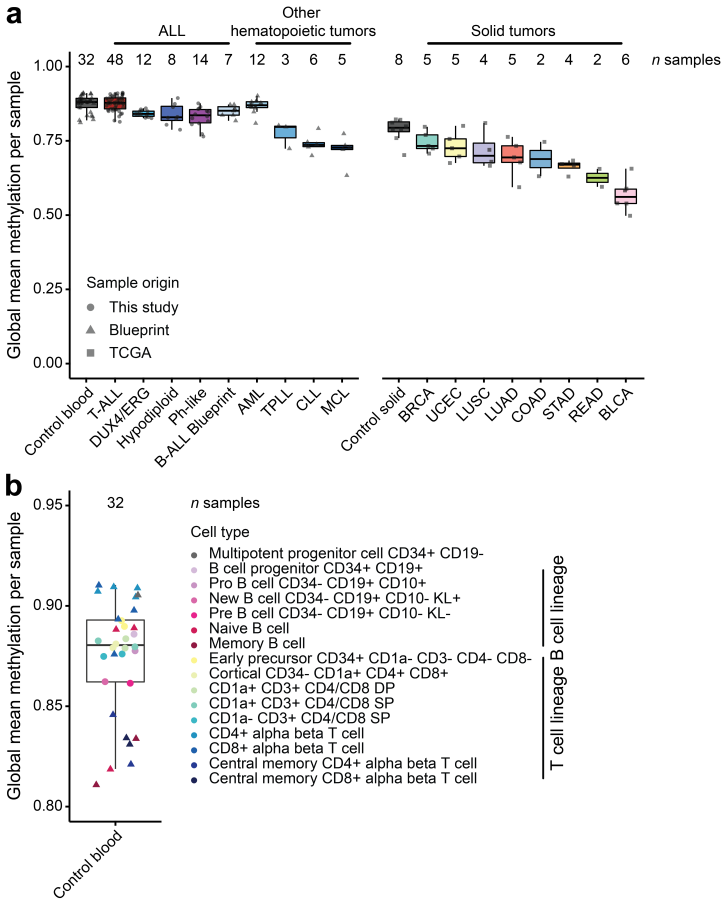

Figure 5.3.3: a) Global average methylation per sample measured excluding CpGs in CGIs. b) Healthy lymphoid cell types are additionally shown with a limited y-axis.

most substantial decrease). In contrast, TPLL, CLL, and eight different solid tumor types exhibited strong hypomethylation that, in the case of BLCA, reached levels as low as 0.5.

The unusually high methylation levels of B-ALL (unknown origin) and AML samples from the Blueprint consortium have been observed previously by Zhou et al. (Blueprint B-ALL cases are shown in Figure 5.3.3 for comparison) [72]. The authors speculated that in the case of B-ALL, the high methylation levels might be associated with the pediatric age of the patients: Global DNA methylation also diminishes as an effect of aging, and Zhou et al. additionally reported a general association of hypomethylation degree and mitotic cell divisions [72]. Therefore, we used our T-ALL cohort, which presented with strikingly high methylation levels and included patients ranging from two to 64 years. Using a two-sided Wilcoxon rank-sum test, we tested the alternative hypothesis that global methylation levels of pediatric T-ALL patients differed from that of the older T-ALL patients. However, no significant difference could be observed, and, additionally, no significant difference based on sex could be determined ($p = 0.2$ and $p = 0.44$ respectively, Figure B.2.1).

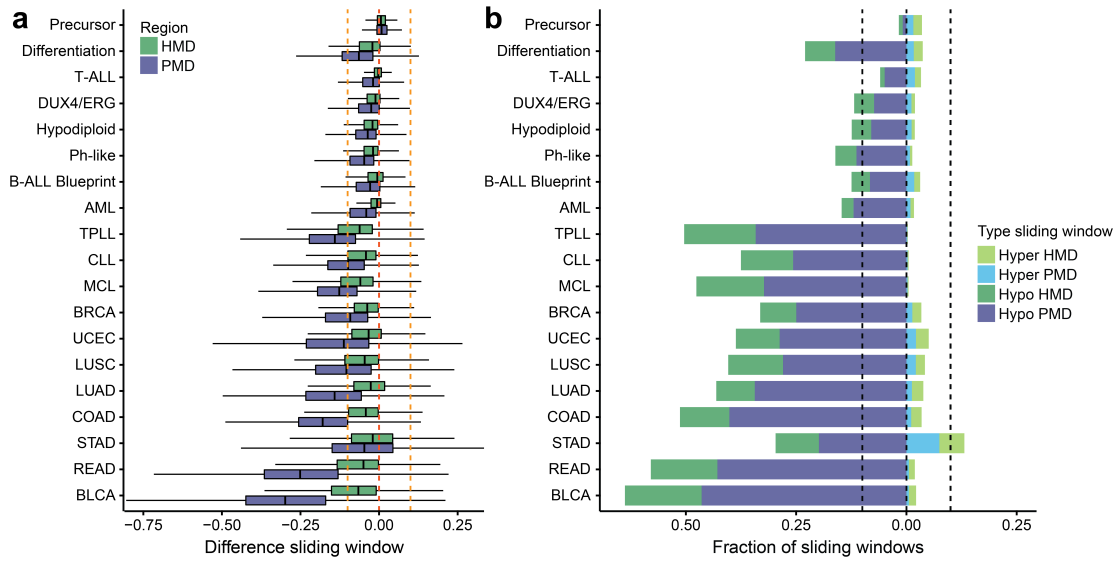

Figure 5.3.4: a) Boxplots showing the delta mean methylation of sliding windows in HMDs and PMDs for different tumor types compared to their respective healthy tissue. Comparisons between precursor T and B cells (precursor control), as well as between memory B and precursor B cells (differentiation control), were added to put the decrease of methylation across tumor types into perspective. b) Fractions of sliding windows that are hyper- or hypomethylated in each tumor type split by location into HMDs and PMDs.

Partially methylated domains

Previous reports have shown that global hypomethylation in tumorigenesis accumulates primarily in PMDs (see section 2.5.2). To further explore the absence of a strong global methylation decrease in ALL subtypes, we segmented the genome into sliding windows assigned to either HMDs or PMDs (see section 5.2.3). We then used the difference in methylation per window between a tumor type and its matching control to visualize the extent and variability of the loss of methylation along the genome (Figure 5.3.4). Again, CpGs in CGIs were excluded. For this analysis, we used two controls: First, the difference between two highly methylated precursor lymphocyte types that measure cell type-specific changes. Second, the difference between memory and precursor B cells as a natural differentiation control (during B cell maturation, global methylation decreases until reaching the memory B cell or plasma cell stage [272]). As in our previous analysis, T-ALL showed similar methylation levels compared to precursor T cells. B-ALL subtypes and AML exhibited a minor methylation decrease. However, this loss was less than the hypomethylation occurring during normal B cell development. Solid tumor types, as well as TPLL and CLL, again showed a more drastic decrease in methylation. This was also reflected in the fraction of windows decreasing by more than 0.1 in methylation (Figure 5.3.4, right). Here, 6-16% of windows were hypomethylated according to this criterion in ALL subtypes and AML, while 30-64% of windows were hypomethylated for other cancer types. For every tumor type, hypomethylation was most pronounced in PMDs compared to HMDs. T-ALL, B-ALL, and AML samples exhibited no or only little loss of methylation in HMDs compared to PMDs, while the other cancer types also lost methylation in HMDs.

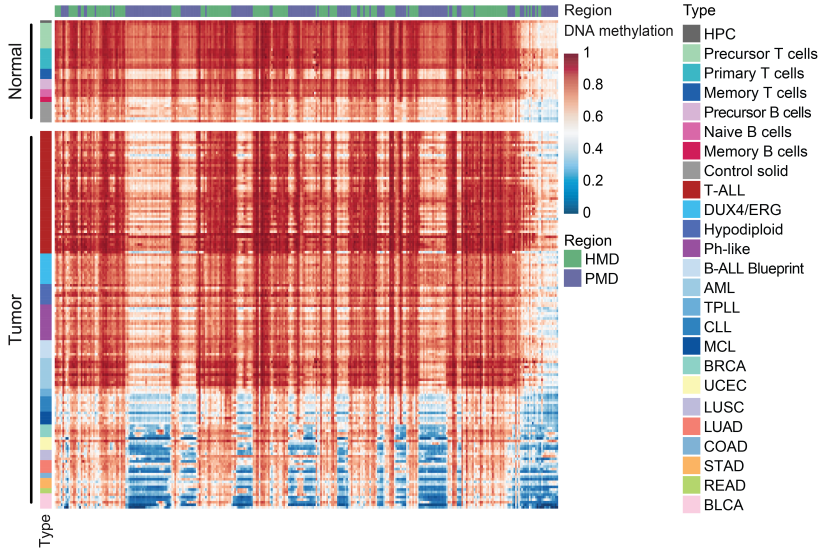

Figure 5.3.5: Heatmap showing the average solo-WCGW CpG methylation of pan-cancer samples in HMDs and PMDs along chromosome 16p.

Zhou et al. showed that the loss of methylation in PMDs is most pronounced in CpGs that are flanked by an adenine or thymine and without another CpG nearby (35 bp on either side), the so-called solo-WCGW CpGs (see section 2.3) [72]. Therefore, in line with the analysis presented by Zhou et al., we analyzed average solo-WCGW CpG methylation in HMDs and PMDs along chromosome 16p as an example (Figure 5.3.5). These CpGs that are most susceptible to hypomethylation also remained highly methylated in T-ALL samples compared to healthy control cells, which was strongly opposed by the effect seen in solid and most other hematopoietic tumor types. Not only pediatric but also adolescents and adult T-ALL patients lacked a strong methylation decrease at solo-WCGW CpGs in PMDs. This contrasts the age-related hypothesis raised by Zhou et al. suggesting that the stable genome-wide methylation levels did not occur due to specifically young ages at tumor formation (two-sided Wilcoxon rank sum test using the average solo-WCGW methylation in PMDs, $p = 0.21$).

DMR analysis of ALL subtypes

Following the analysis of the overall genomic background levels in ALL subtypes, we aimed to identify local but significant changes across the genome compared to the respective healthy tissue. Therefore, we called DMRs between all samples of a subtype and its control cell type (see section 5.2.3). ALL subtypes exhibited between approximately 13,000 and 26,000 DMRs (Table 5.3.1). T-ALL showed a strong bias towards hypermethylation, with around 88% of its DMRs exhibiting a gain of methylation in line with the overall highly methylated genome. B-ALL subtypes showed higher fractions of hypomethylated DMRs, which in the case of the DUX4/ERG subtype reached almost 50% of the overall number of DMRs.

When comparing the DMRs to random control regions with comparable properties, we observed enrichment of DMRs in similar genomic features across ALL subtypes (Figure 5.3.6). Hypermethylated DMRs are enriched in CpG-dense features frequently unmethylated or lowly methylated

Type	# Hyper DMRs	# Hypo DMRs
T-ALL	23,327	3,111
DUX4/ERG	11,693	11,299
Hypodiploid	10,976	2,423
Ph-like	9,526	5,931

Table 5.3.1: Number of hyper- and hypomethylated DMRs called per subtype.

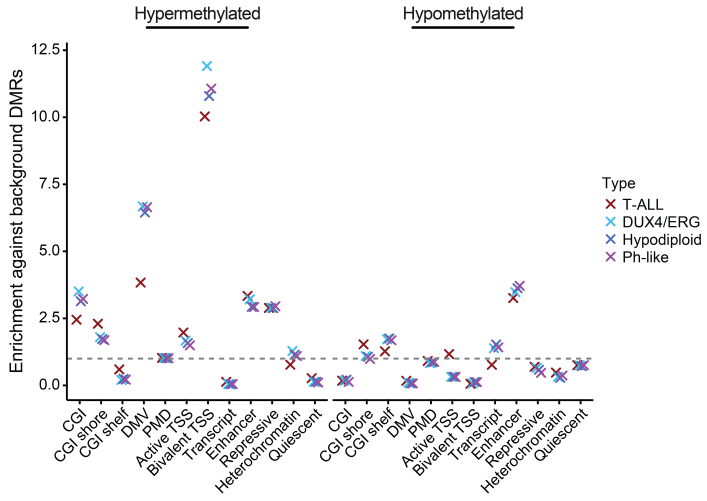

Figure 5.3.6: Enrichment of DMRs in genomic features and chromatin states per ALL subtype.

in healthy tissues such as CGIs, CGI shores, and DMVs. Additionally, regions marked by the repressive H3K27me3 mark in hematopoietic stem cells (bivalent TSS and Polycomb-repressed) are targets of hypermethylation similar to previously reported changes in other tumor types [127]. Enhancers are targets of both hypo - and hypermethylation, which could potentially cause activation and deactivation of specific enhancers regulating tumor-specific genes or representing byproducts of generally misregulated pathways. Together these results show that the absence of global hypomethylation specifically in T-ALL can also be found at the local level. At the same time, these changes affect similar types of regulatory regions as in B-ALL subtypes.

CGI methylation levels

CGI hypermethylation is one of the characteristic DNA methylation changes occurring during tumorigenesis. In line with this, hypermethylated DMRs of all ALL subtypes were enriched in CGIs, as shown in the previous section. Therefore, we wanted to explore this phenomenon further in a pan-cancer context. Zooming into exemplary CGIs comparing patients from ALL subtypes as well as CLL and COAD, all samples showed hypermethylation compared to the healthy tissue but with a distinct architecture across the island (Figure 5.3.7). Similar to the analysis of global CpG methylation, we then compared CGI methylation between ALL subtypes and other hematopoietic

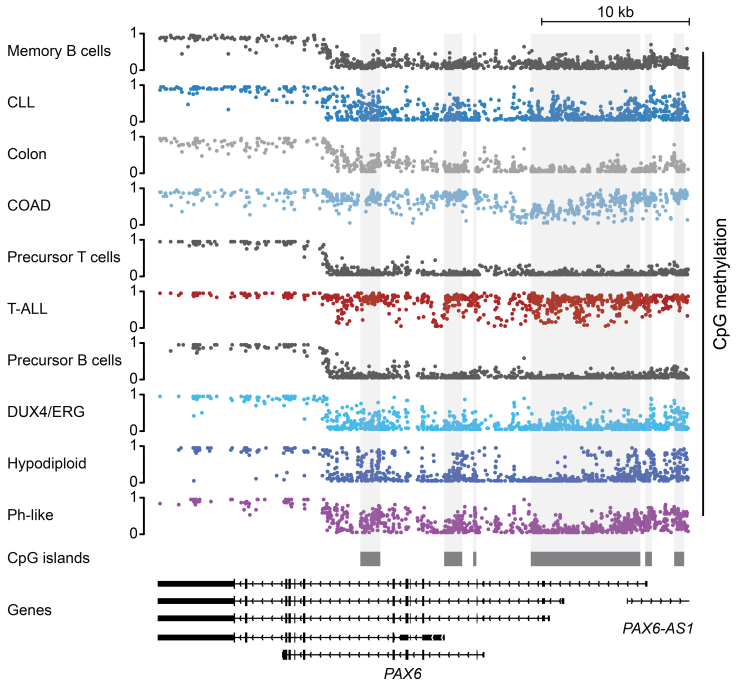

Figure 5.3.7: Genome browser tracks of memory B cells, CLL, healthy colon, COAD, precursor T cells, T-ALL, precursor B cells, and B-ALL subtypes at the *PAX6* locus.

and solid tumor types (Figure 5.3.8). All ALL subtypes exhibited CGI hypermethylation to varying degrees, with T-ALL standing out due to extremely high methylation levels but also a high variability across all samples, with some close to the healthy control cells and others reaching average methylation levels of 0.5 across all CpGs in CGIs. This pattern was unique even compared to most other hematopoietic and solid tumors. Comparing the average CGI methylation with the average methylation of the genomic background excluding CGIs across ALL patients and healthy cells, we observed that while precursor B and T cells exhibited low CGI and high background methylation as expected, the methylation levels of both CGIs and genomic background were positively associated with each other across ALL patients (Figure 5.3.9). Specifically for T-ALL, patients with the highest genome-wide methylation levels also tended to exhibit the highest CGI methylation levels.

The high variability of T-ALL CGI methylation levels was reminiscent of the observation of a CIMP subtyping defined by previous studies. This classification divided T-ALL patients into samples with extreme CGI hypermethylation (CIMP positive) and lower CGI hypermethylation (CIMP negative). These classifications were previously established by clustering patients based on a selection of CpGs in CGIs covered by the 450k array [254]. Instead of directly clustering our samples, we used a principle component analysis (PCA) based on the average methylation of the commonly covered and variably methylated CGIs ($n = 8,863$) of healthy precursor T cell and T-ALL samples (Figure 5.3.10, left). T-ALL samples did not separate into groups defined by low or high CGI methylation but, in contrast, distributed according to the CGI methylation levels (here indicated by the median methylation across all variable CGIs). This can be recapitulated when using the status of methylation (unmethylated/methylated, methylation > 0.2) as input for the PCA, underlining that not only the CGI methylation levels display a continuous range across

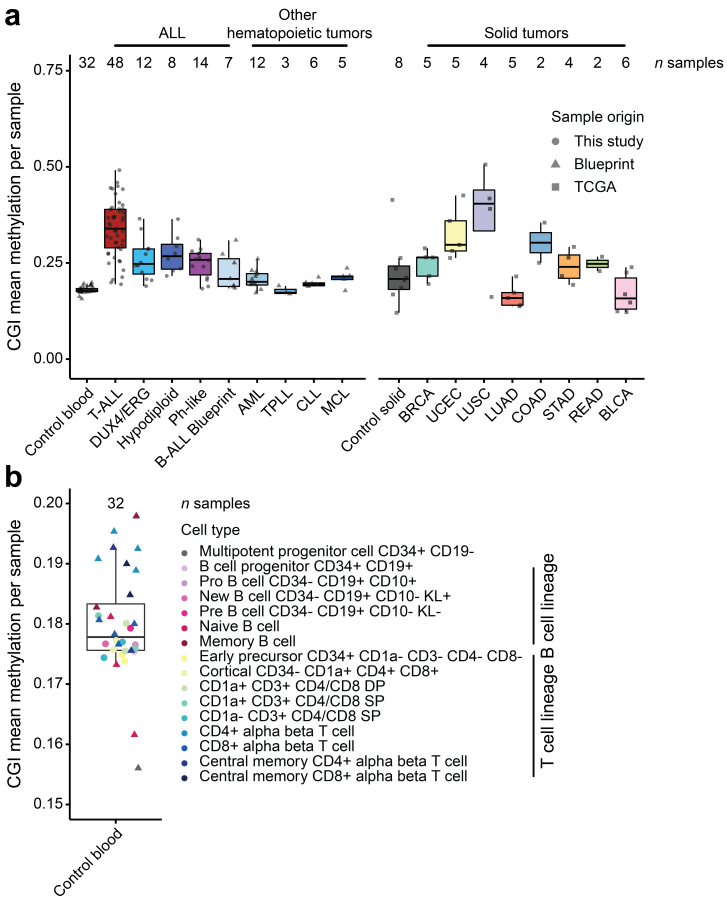

Figure 5.3.8: a) CGI average methylation per sample for ALL subtypes, other tumor types, and their respective controls. b) Healthy lymphoid cell types are additionally shown with a limited y-axis.

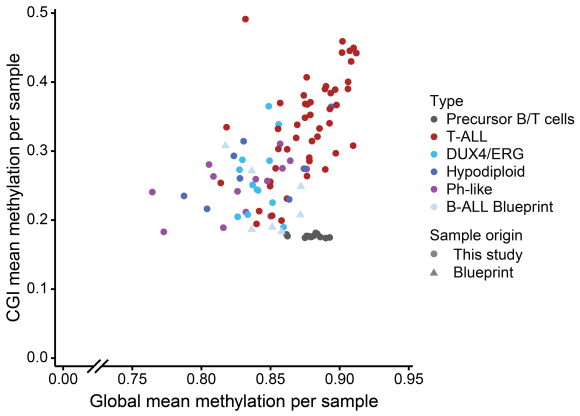

Figure 5.3.9: Correlation of global and CGI methylation across ALL samples.

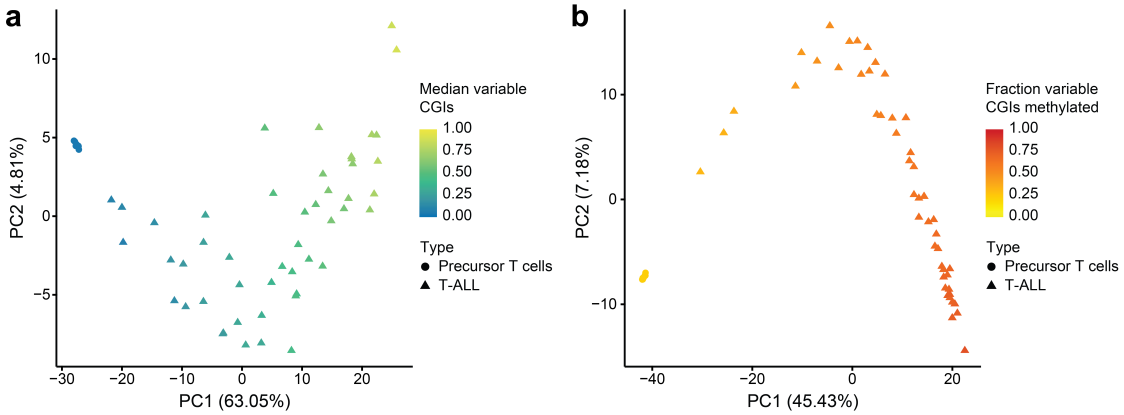

Figure 5.3.10: a) PCA based on the mean methylation of variable CGIs of precursor T cell and T-ALL samples. b) PCA based on the methylation status (methylated/unmethylated) of variable CGIs of precursor T cell and T-ALL samples.

samples but also the number of targets as indicated by the color scale (Figure 5.3.10, right). Both measurements covered a wide range from patients with CGI methylation close to healthy precursor cells and others reaching almost complete methylation of all variably methylated CGIs. Therefore, the previous classification into CIMP groups might not fully resemble the underlying dynamics of CGI hypermethylation in T-ALL.

5.3.2 Clustering of CGIs based on T-ALL patients

The PCA showed that patients with T-ALL distribute according to their CGI methylation levels and the number of targets. To characterize which targets are affected by different methylation levels, we used a consensus clustering approach based on the variably methylated CGIs (see section 5.2.4). This analysis resulted in four clusters of CGIs exhibiting different methylation dynamics across our T cell lineage cohort (Figure 5.3.11). Cluster 1 consists of CGIs that are unmethylated in healthy cells but also, to a large extent, in T-ALL samples. Hypermethylation targets are sporadic and rather sample-specific, while their methylation levels rarely reach 100%. Clusters 2 and 3 also contain CGIs that are unmethylated in precursor T cells and mostly hypermethylated in T-ALL samples. This methylation increases from low to high according to the overall variably methylated CGI levels, with some patients almost resembling healthy cells and others reaching complete methylation across almost all CGIs. For cluster 2, the methylation gain is rather heterogeneous and sample-specific, while cluster 3 shows a more homogeneous trend across islands for all T-ALL patients. Methylation levels are generally higher in cluster 3 compared to cluster 2, and more patients show levels close to 100%. Cluster 4 contains CGIs that are already highly methylated in healthy samples and reach even higher levels in T-ALL patients. Also, the most substantial gain is observed here in patients with overall high CGI methylation.

Analyzing the four clusters together with the previously excluded groups of stably lowly and highly methylated CGIs (group low/high), we observed that group low and cluster 1 overall contain CGIs with high numbers of CpGs, high GC content and large region size (Figure 5.3.12). All three characteristics (CpG number, GC content, length) decrease along the clusters as methy-

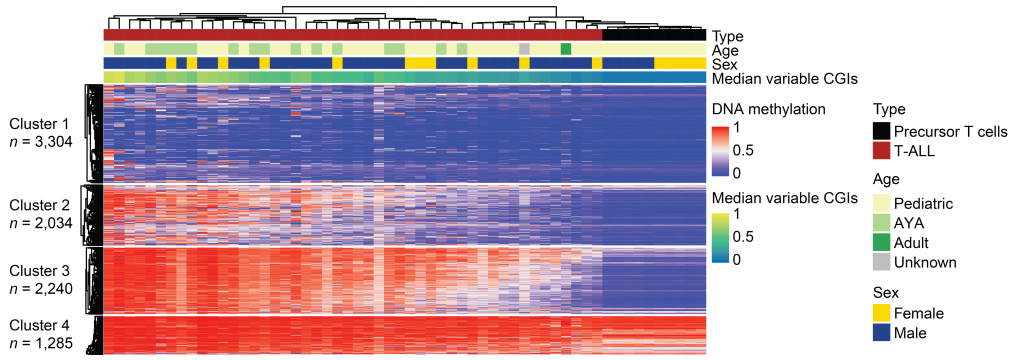

Figure 5.3.11: Methylation of CGIs per sample and cluster identified by consensus clustering.

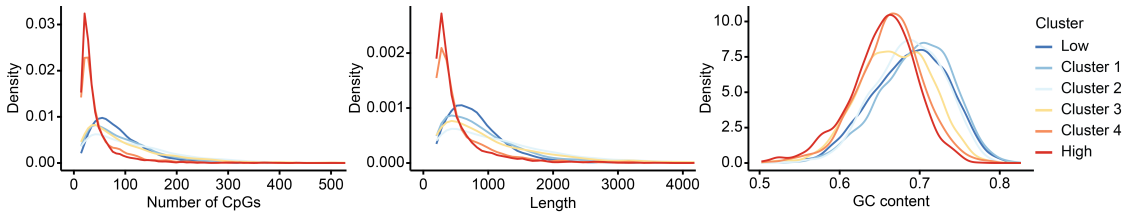

Figure 5.3.12: Distribution of CpG number (left), length (middle), and GC content (right) for the defined CGI groups/clusters.

lation levels increase, with the smallest numbers present for cluster 4 and the group of stably highly methylated CGIs. This is in line with previous findings showing that large, GC-rich islands usually remain free of methylation even if the corresponding genes are turned off [56,86].

We then aimed to characterize the defined CGI clusters further with respect to their associated (overlapping) genomic features (Figure 5.3.13). Lowly methylated CGIs (group low and cluster 1) are frequently located in promoters (also often associated with active genes in precursor T cells) but rarely in PMDs, gene bodies, or intergenic regions. The fraction of CGIs overlapping with promoters decreases from group low to high along the clusters, which is most pronounced for active promoters. At the same time, the fraction of CGIs associated with gene bodies rises until reaching more than 75% in highly methylated CGIs, which is presumably linked to their transcriptional activity in normal and tumor cells. Methylation in gene bodies is known to frequently positively correlate with and stabilize transcription due to its implicated role in proper transcription initiation and nucleosome stability [57,59,60]. Additionally, the fraction of CGIs in PMDs, DMVs, and intergenic regions also rises from low to high. However, the fraction of CGIs in PMDs and DMVs peaks in cluster 3 and decreases afterwards. This is in line with previous studies showing preferential hypermethylation of CGIs in PMDs, whereas cluster 4 and group high contain CGIs already highly methylated in healthy T cells [273]. Similarly, DMVs frequently contain PRC2-marked CGIs overlapping developmental gene promoters [97]. These CGIs remain unmethylated in healthy tissues as part of the DMVs and therefore are not expected to be found in the clusters highly methylated in precursor T cells.

In addition to genomic features, we used publicly available chromatin states for hematopoietic stem cells (HSCs) and the T-ALL cell line DND41 (as a proxy for T-ALL) to assign a chromatin

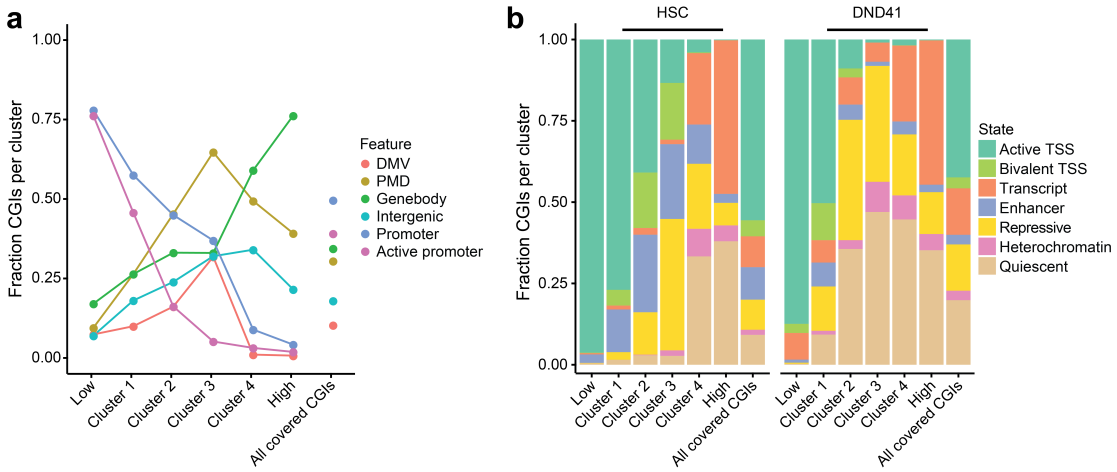

Figure 5.3.13: a) Fraction of CGIs per cluster overlapping different genomic features. Fractions for all considered CGIs are shown for comparison. b) Fraction of CGIs per cluster in chromatin states defined for hematopoietic stem cells (left) and the T-ALL cell line DND41 (right).

state to each CGI and compare the fractions of states across CGI clusters and between healthy and cancerous cells (Figure 5.3.13). The chromatin states resembling active transcription start sites and transcripts in HSCs reflect the findings from the genomic feature analysis: The fraction of CGIs assigned to heterochromatin and quiescent regions is highest in the rather methylated CGI clusters in HSCs. Enhancers are mainly present in the variably methylated clusters 1 to 4, while Polycomb-repressed states are primarily associated with clusters 2 to 4. Expression of genes in precursor T cells supports the observed chromatin state dynamics in HSCs: Genes with a promoter CGI of group low or cluster 1 are frequently expressed, while genes associated with cluster 2 to 4 and group high in repressed chromatin states in healthy HSCs are mainly already unexpressed in the healthy state (Figure 5.3.14). When conducting an overrepresentation analysis of the genes associated with promoter CGIs of each CGI group, only the unmethylated or lowly methylated clusters lead to significant enrichments (Figure 5.3.15): Unmethylated promoter CGIs of group low are associated with genes implicated in cell maintenance such as protein translation, DNA replication and cell cycle regulation, which explains why these promoters stay unmethylated across healthy and tumor samples. Turning off these genes would also be unfavorable for a tumor cell as it would inhibit pathways essential to its survival. Genes with a promoter CGI assigned to cluster 1 are enriched in MAPK and JNK signaling pathways, which are frequently disturbed or deregulated in cancer [274].

When comparing the distribution of chromatin states per cluster between HSCs and DND41 (resembling the healthy and cancerous state), we observed that chromatin states associated with the low and high groups stayed relatively stable. In contrast, for the variable clusters 1 to 4, the fraction of quiescent and heterochromatic regions increased in the cancerous state. At the same time, bivalent promoters disappear almost entirely in DND41 compared to HSCs. These findings resemble previous reports on chromatin changes observed in primary tumors [4]. Using a two-sided Chi-squared test, we assessed that changes in chromatin states between HSCs and DND41 are statistically significant for every CGI cluster (Table B.2.1). However, the effect size

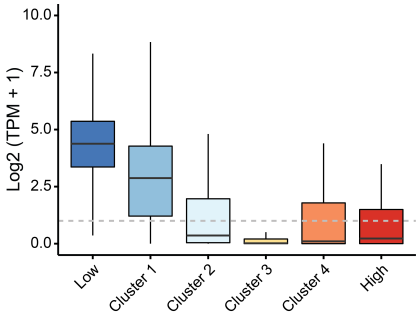

Figure 5.3.14: Expression of genes associated with promoter CGIs per cluster.

was large only for clusters 1 to 3 (Table B.2.1), suggesting that the changes in chromatin states for these three clusters were most pronounced.

Lastly, we used the CGI clusters defined in T-ALL to visualize methylation levels of the respective CGIs in other hematopoietic and solid tumor types (Figure 5.3.16). Although defined based on a different tumor type, we found that methylation levels changed similarly from low to highly methylated CGI clusters across tumor types. Underlining the housekeeping role of genes associated with CGIs of group low, these CGIs remained unmethylated across all tumor types and healthy tissues. Similarly, CGIs of the high group were highly methylated but exhibited slightly less methylation in tumor types that show global hypomethylation, such as TPLL, CLL, and solid tumors, suggesting that a genome-wide effect might influence their methylation. Besides the two extreme groups, methylation rose from lower to higher levels along clusters 1 to 4. However, the exact levels appeared tumor-specific, where some tumor types like TPLL and LUAD generally exhibited lower CGI methylation than others. These results suggest that although CGI methylation has been shown to be tumor-specific, a pan-cancer mechanism might exist that primes certain groups of CGIs for specific hypermethylation levels.

5.3.3 Relation of DNA methylation with other characteristics in T-ALL

We aimed to investigate the association between T-ALL CGI methylation levels and genetic as well as transcriptomic drivers and demographic covariates. For this purpose, we used hierarchical clustering to cluster T-ALL patients based on the methylation levels of variably methylated CGIs ($n = 8,863$). We extracted the three top-level clusters that represented patients with extreme methylation levels (very low or high CGI methylation levels) as well as intermediate CGI methylation and termed them accordingly T-ALL$^{\text{LM}}$ (cluster 1), T-ALL$^{\text{IM}}$ (cluster 2), and T-ALL$^{\text{HM}}$ (cluster 3) (Figure 5.3.17). We then used these clusters to test the association of rather different CGI methylation levels in T-ALL patients with age group, sex, genetic subtypes, and recurrent mutations using Fisher's exact test (Figure 5.3.18, Table B.2.2). None of the covariates were significantly associated with our T-ALL methylation-based subtypes, although HOXA and TLX3 subtypes seemed to co-occur frequently with T-ALL$^{\text{IM}}$ (cluster 2) and T-ALL$^{\text{HM}}$, which has been observed for previously defined CIMP-positive T-ALL cases [254]. However, it should be noted that the cohort was not initially designed to allow a thorough analysis of clinical features, genomic alterations, and methylation states and therefore did not include enough samples for these types of analysis (transcription and mutation data sets were also not available for all patients).

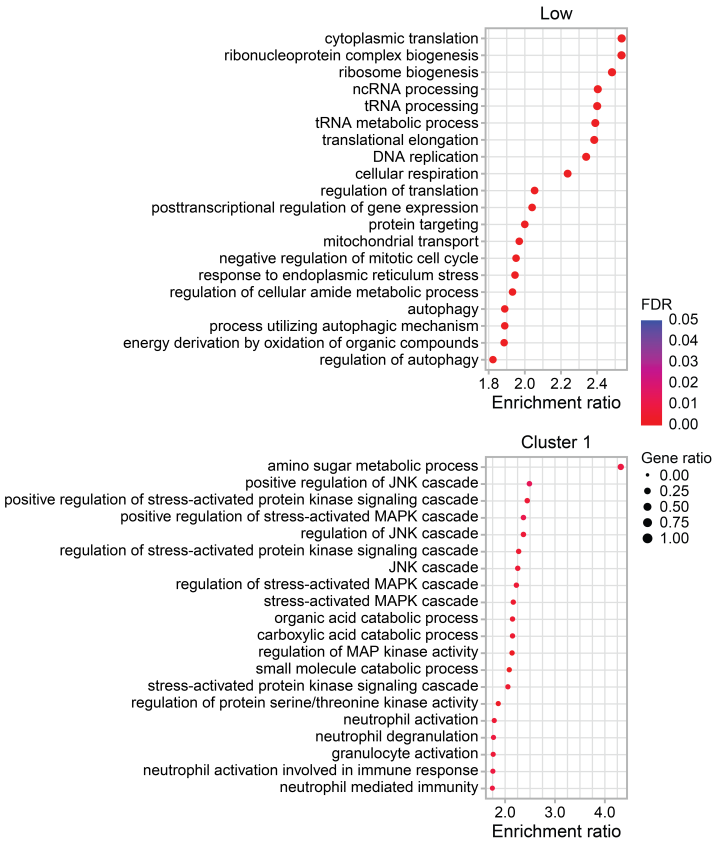

Figure 5.3.15: Overrepresentation analysis of genes associated with lowly methylated CGIs (top) and CGIs of cluster 1 (bottom).

We additionally performed clustering of T-ALL patients based on the top 500 variably expressed genes, which included known T-ALL marker genes, in order to investigate whether transcriptome- and methylation-based clustering would lead to similar groups of patients (Figures 5.3.19 and B.2.4). However, transcriptome-based clustering did not agree with the methylation-based subtypes defined using CGI methylation levels and instead grouped samples according to their genetic subtype.

Finally, we used methylation entropy to measure intra-tumor heterogeneity of DNA methylation to compare T-ALL patients from different methylation-based subtypes (Figure 5.3.20). Entropy was significantly higher in patients of the T-ALL$^{\text{LM}}$ and T-ALL$^{\text{IM}}$ groups compared to T-ALL$^{\text{HM}}$, suggesting more homogeneous methylation across reads in highly methylated T-ALL samples (Wilcoxon rank-sum test, $p = 2.3 * 10^{-6}$ and $p = 0.0002$ respectively). Previous studies in other cancer types, such as B cell lymphoma, found that higher intra-tumor heterogeneity is associated with poorer prognosis and survival [132, 275]. This would align with previous studies with much larger methylation array-based T-ALL cohorts showing that their defined CIMP-positive T-ALL cases demonstrated better overall survival than T-ALL samples with lower CGI methylation [254].

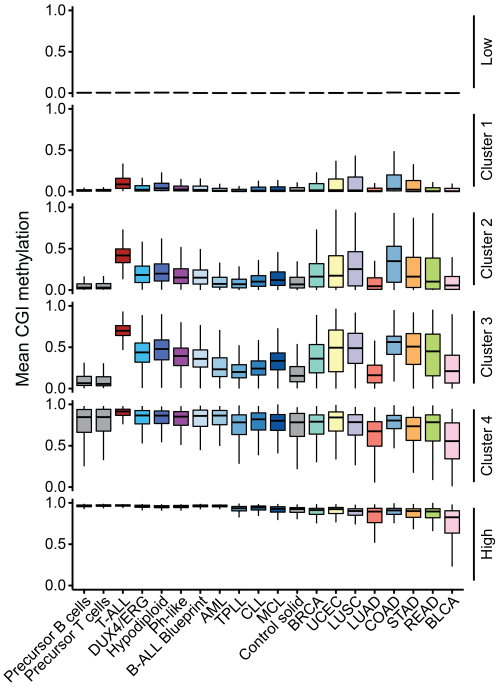

Figure 5.3.16: Methylation levels of CGIs per cluster across different hematopoietic and solid tumor types.

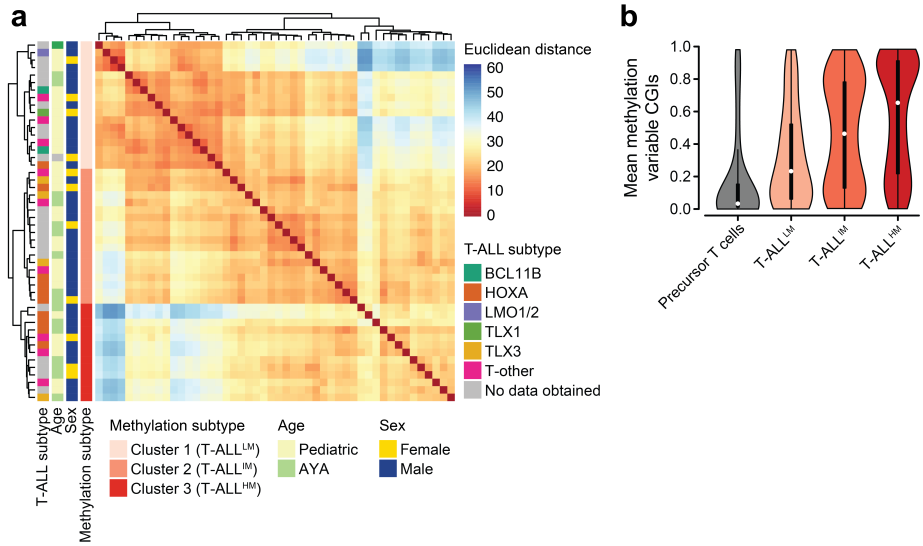

Figure 5.3.17: Hierarchical clustering of T-ALL patients identified three main clusters exhibiting different CGI methylation levels.

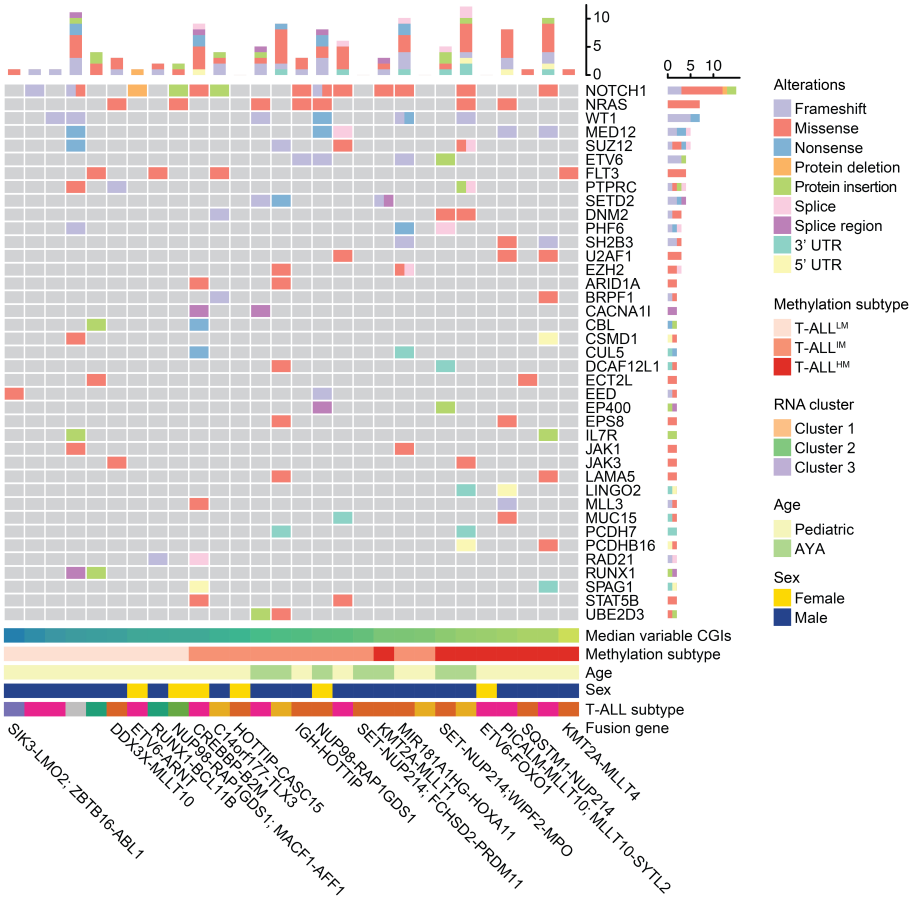

Figure 5.3.18: Frequently mutated genes and their respective mutation status across T-ALL patients.

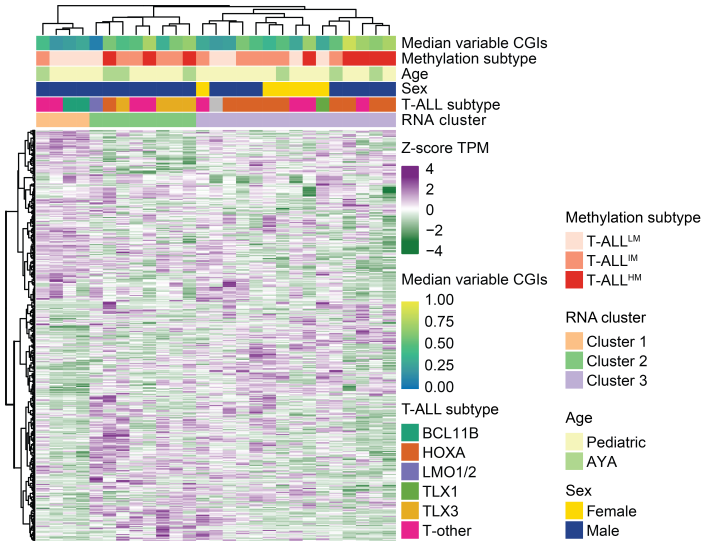

Figure 5.3.19: Hierarchical clustering of the 500 most variably expressed genes across T-ALL patients. Samples overall group based on the genetic or transcriptomic subtype such as TLX3 overexpression or HOXA subtype.

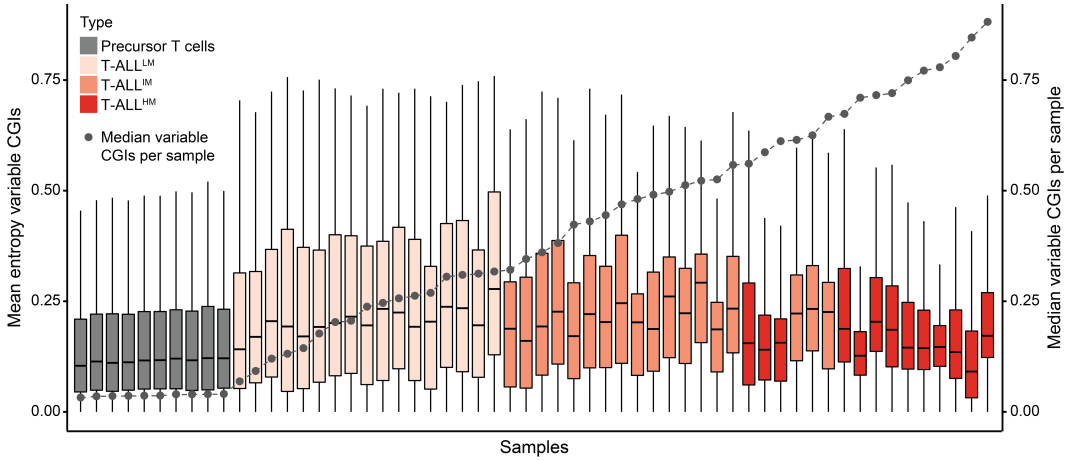

Figure 5.3.20: Methylation entropy of variable CGIs in T-ALL patients. The entropy is highest for patients of the T-ALL^{LM} and T-ALL^{IM} group.

5.3.4 Expression of epigenetic regulators associated with methylation levels

Given the generally higher methylation levels of CGIs and the complete genome in T-ALL compared to B-ALL but also the wide range of methylation levels across T-ALL patients, we aimed to identify genes - specifically epigenetic regulators - that could play a role in establishing and regulating these different landscapes. No recurrent mutations in epigenetic regulators have been detected in T-ALL patients that could explain the differences in this subtype (Figure 5.3.18). We, therefore, conducted a correlation test using the log2-transformed expression of each gene (measured in TPM) and the average global or CGI methylation across patients with T-ALL or B-ALL (subtypes DUX4/ERG and Ph-like). We excluded patients with hypodiploid B-ALL to avoid confounding expression effects due to the high aneuploidy. We detected 1,898 genes significantly correlated with average global DNA methylation and 1,833 significantly correlated with CGI mean methylation levels (1,390 genes detected in both analyses). Many of these genes were associated with B or T lymphocyte-specific pathways such as lymphocyte differentiation and B cell activation because of the generally higher methylation levels in T-ALL compared to B-ALL subtypes (Figure B.2.5).

When examining a panel of epigenetic regulators that are involved in the direct or indirect regulation of DNA methylation, we found that the *de novo* methyltransferase DNMT3B was among the significantly correlating genes with respect to both global and CGI methylation (Figure 5.3.21, selection of epigenetic regulators described in section 5.2.6). This includes the expression of catalytically active DNMT3B isoforms (DNMT3B-002 and, in some patients, DNMT3B-001). These isoforms are usually not expressed in adult tissues where instead, the catalytically inactive isoform DNMT3B-003 is expressed [276]. Additionally, the maintenance DNA methyltransferase DNMT1 significantly positively correlated with CGI methylation levels. Besides, TET1 and IDH2 were detected to correlate positively with both global and CGI methylation levels. TET1 actively removes DNA methylation, which could hint at a potential feedback loop between *de novo* methyltransferase and TET activity. IDH2 has been implicated in strong CGI hypermethylation and a CIMP subtyping in glioma; however, through mutation and not expression differences [143]. Previously reported epigenetic regulators in T-ALL model systems, such as MYC, did not correlate

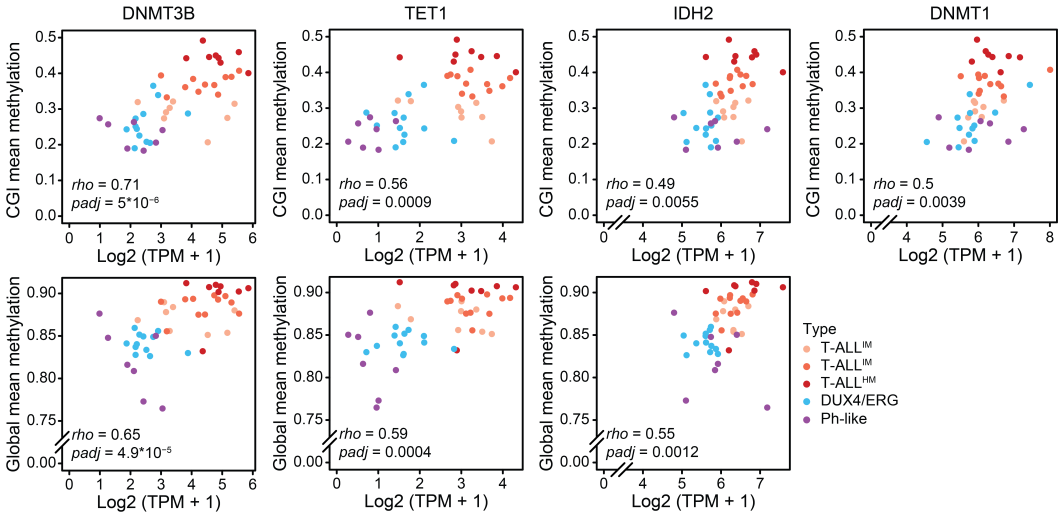

Figure 5.3.21: Epigenetic regulators significantly correlated with global or CGI methylation levels.

with the differences in methylation levels.

Although the correlation analysis revealed potential candidates that could be involved in regulating different methylation levels across ALL subtypes and within T-ALL, it might miss genes that are affected only in a small subset of patients. We, therefore, additionally inspected the promoter methylation status of our epigenetic regulator selection (Figure 5.3.22). Promoters were mostly unmethylated across B-ALL samples with sporadic exceptions. However, 26% of T-ALL patients showed hypermethylation of the *TET2* promoter (methylation > 0.2) accompanied by a decrease or complete loss of expression of the TET2 gene and associated with overall high CGI and global methylation levels (Figures 5.3.22, 5.3.23 and 5.3.24). Additionally, some T-ALL patients exhibited hypermethylation of the *TET1* promoter, which frequently coincided with *TET2* promoter hypermethylation (Figure 5.3.22). The tumor suppressor gene *WT1* also exhibited frequent promoter hypermethylation associated with decreased expression across T-ALL patients. Interestingly, hypermethylation of *WT1* was largely mutually exclusive to mutations in *WT1* but primarily affected patients with high CGI methylation levels and frequently also *TET2* promoter hypermethylation (Figure 5.3.22). Studies in AML showed that WT1 recruits TET2 to its target sites, suggesting a connection between the loss of the two genes in T-ALL patients that might lead to strong hypermethylation levels [268].

5.3.5 ALL cell lines as model systems

Promoter hypermethylation correlated with decreased expression of TET2 in a subset of T-ALL patients with generally high global and CGI methylation. To test whether the loss of TET2 affects the methylation levels, we aimed to use T-ALL cancer cell lines as model systems as these offer the possibility to manipulate the genome via CRISPR-mediated knockout and study the effects. However, cell lines have also been shown to deviate epigenetically from primary tumors frequently. More specifically, many cell lines have been shown to hypermethylate CGIs to greater extents

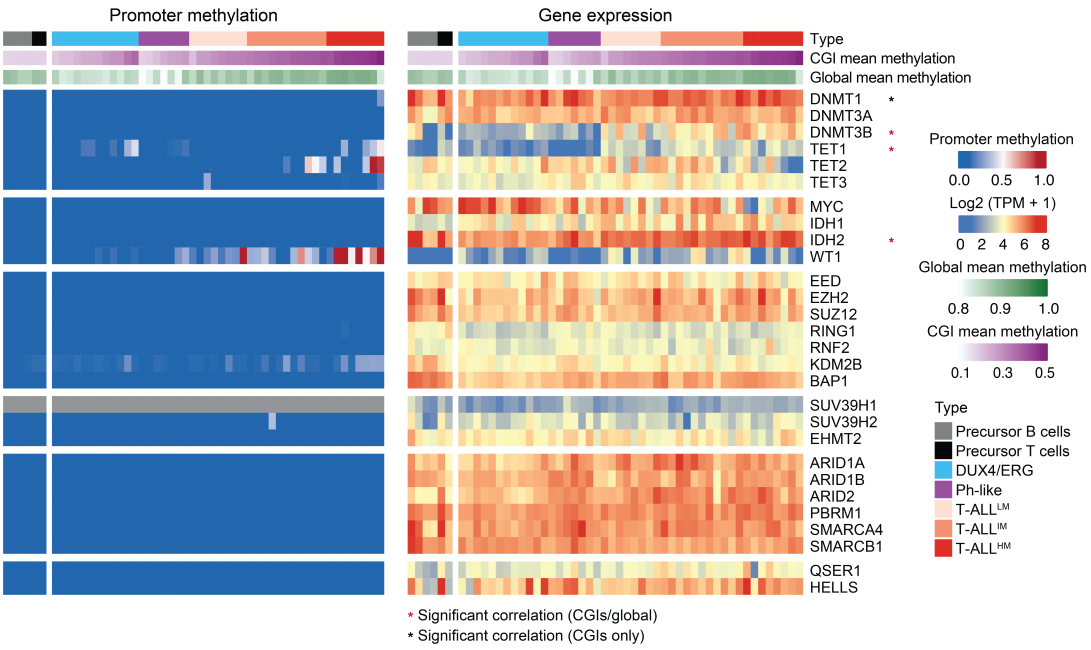

Figure 5.3.22: Promoter methylation and expression status of epigenetic regulators. TET1, TET2, WT1, and KDM2B exhibit elevated levels of promoter methylation in highly methylated T-ALL samples reaching up to 100% for TET2 and WT1.

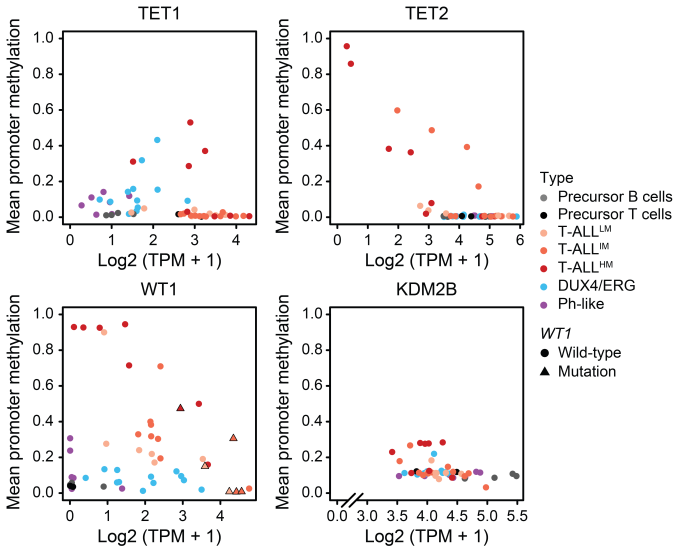

Figure 5.3.23: Correlation of promoter methylation and expression for TET1, TET2, WT1, and KDM2B. For TET2 and WT1, high promoter methylation is associated with low or no expression of the respective gene.

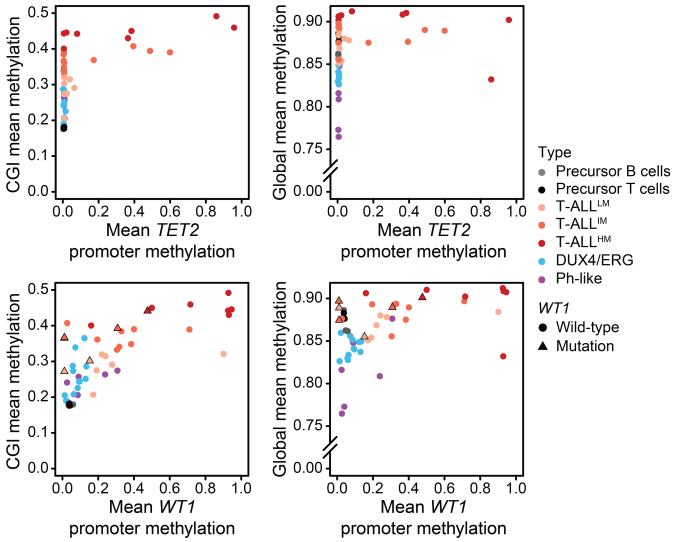

Figure 5.3.24: Correlation of TET2 (top) and WT1 (bottom) methylation status with sample-wise CGI (left) and global (right) methylation levels. High promoter methylation of these two genes seems to co-occur with high CGI and global methylation levels.

than the respective primary tumors [146, 147]. Additionally, long-term culture can affect global methylation levels and lead to the deepening of PMDs (see section 2.5.2) [72, 147]. Therefore, we first sequenced a selection of nine T-ALL and five B-ALL cell lines to assess whether their methylation landscape would reflect the respective primary patient samples. Clustering based on the 5% most variably methylated CpGs using healthy, tumor, and cell line samples showed that cell lines overall group with samples from their lineage of origin except for NALM-6, which groups together with T-ALL cell lines and highly methylated T-ALL cases (Figure 5.3.25). Three T-ALL cell lines resembled the intermediate CGI methylation of patients from the T-ALL^{IM} group (namely MOLT-16, Jurkat, and PEER). The remaining T-ALL cell lines (DND41, PER-117, RMPI-8402, LOUCY, TALL-1 and ALL-SIL) exhibited high CGI methylation levels comparable to T-ALL^{HM} patients (Figures 5.3.26 and B.2.7). In contrast, B-ALL cell lines frequently showed higher CGI methylation levels than their primary subtype, except for MHH-CALL-2, which more closely resembled the hypodiploid B-ALL primary methylome. NALM-6, a B-ALL cell line of the DUX4/ERG subtype, exhibited particularly high CGI methylation levels, which could explain why it clusters with highly methylated T-ALL cell lines. On the global level, cell lines frequently showed stronger hypomethylation, which might stem from culture-induced effects and clearly distinguishes the cell lines from the highly methylated genome of primary ALL samples (Figure 5.3.26).

The high CGI methylation levels of six T-ALL cell lines are difficult to interpret as they could either reflect the epigenetic regulation of T-ALL^{HM} patients or, similar to the global decrease in methylation, represent a culture-induced artifact. Therefore, we inspected the promoter methylation levels of the panel of epigenetic regulators previously examined for primary patients (Figure 5.3.27). The cell lines DND41, TALL-1, and LOUCY showed complete methylation of the *TET2* and *WT1* promoters similar to the effect observed in a subset of patients. Other T-ALL cell lines such as Jurkat, PEER, and MOLT-16 exhibited an unmethylated *TET2* promoter. In contrast, the *WT1* promoter was methylated in almost all cell lines, although to different extents with some-

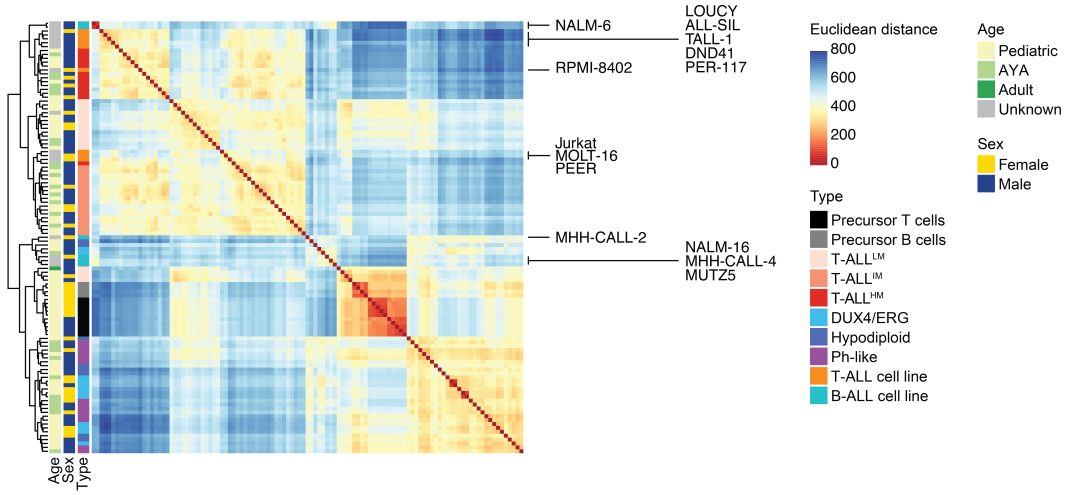

Figure 5.3.25: Hierarchical clustering of healthy lymphocytes, ALL patient samples, and ALL cell lines. Most cell lines group to primary patient samples of the same lineage, with the exception of NALM-6, which groups together with T-ALL cell lines and patients.

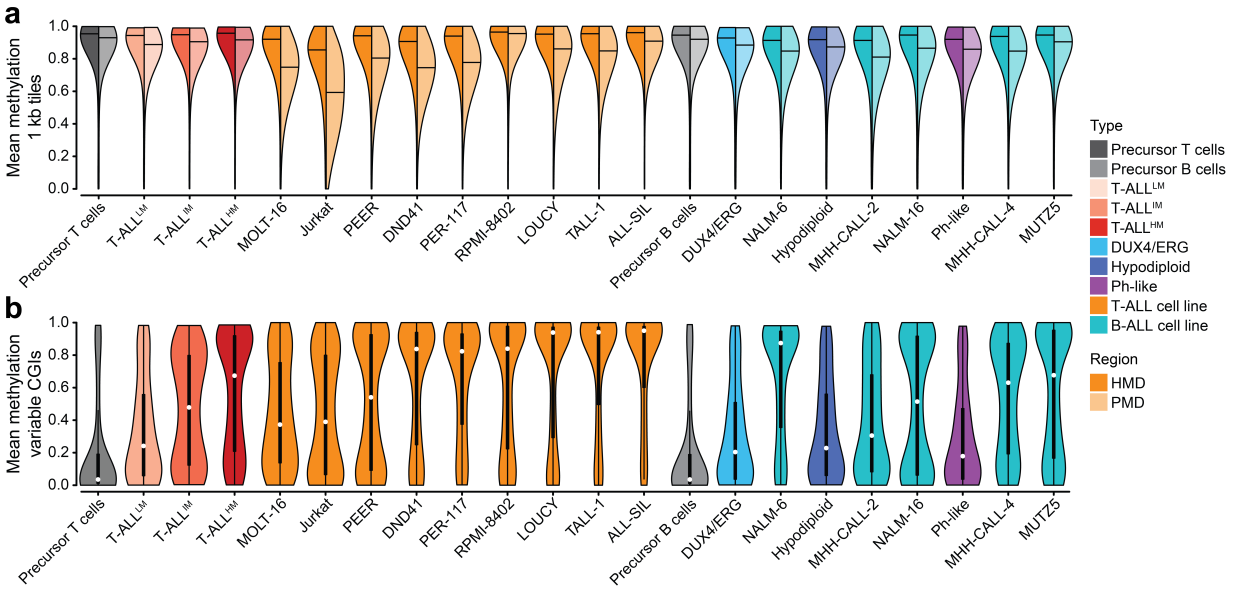

Figure 5.3.26: Violin plot of HMD and PMD (top) as well as variable CGI (bottom) methylation in ALL subtypes and cell lines. Cell lines frequently show a decrease in genome-wide methylation and more extreme CGI methylation levels than their corresponding primary ALL subtypes.

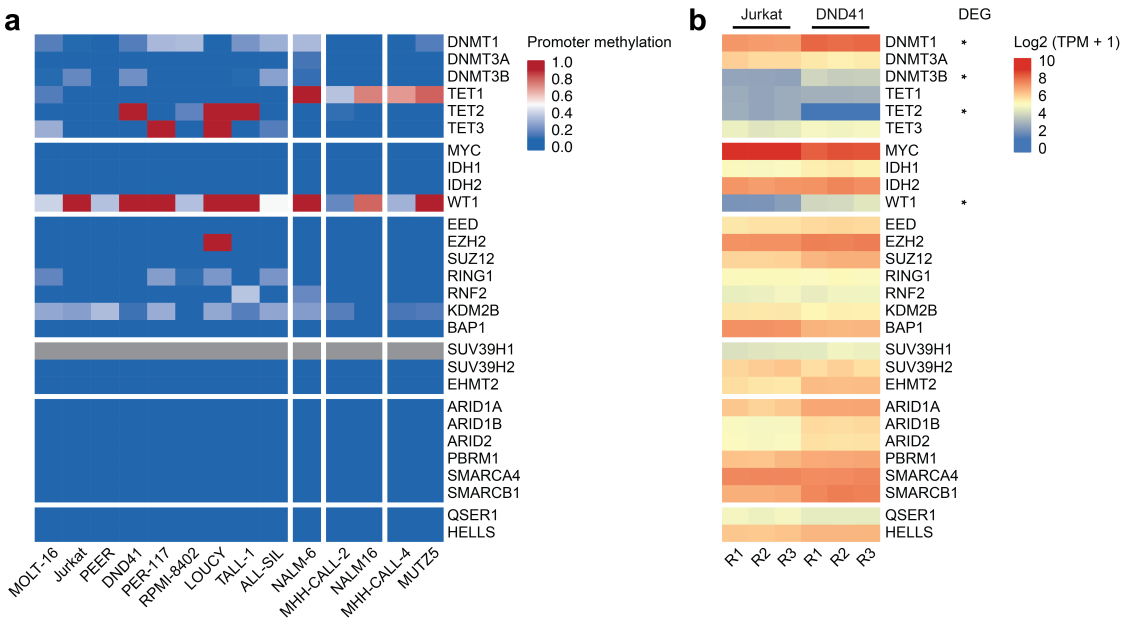

Figure 5.3.27: a) Promoter methylation status of epigenetic regulators in ALL cell lines. b) Epigenetic regulator expression in DND41 and Jurkat T-ALL cell lines.

times relatively low methylation levels. In line with our findings in the primary patients, B-ALL cell lines did not exhibit hypermethylation of the *TET2* promoter. Instead, *TET1* was methylated in all five cell lines. We then selected the cell line Jurkat resembling T-ALL^{IM} patients and the T-ALL^{HM}-like cell line DND41 and subjected them to expression profiling using RNA-Seq (Figure 5.3.27). Indeed we found that TET2 is expressed in Jurkat (unmethylated promoter) and silenced in DND41 (highly methylated promoter). Using differential expression analysis based on three replicates per cell line, we detected significant up-regulation of the DNA methyltransferases DNMT1 and DNMT3B in DND41 compared to Jurkat (Figure 5.3.27). These two enzymes are also positively correlated with increased CGI methylation levels across primary ALL patients (DNMT3B also correlated with global methylation levels).

We next knocked out *TET2* in Jurkat cells to characterize the effects of TET2 loss in a T-ALL^{IM}-like cell line (Figure B.2.8). After 20 days, cells were collected for WGBS and RNA-Seq profiling. When comparing the methylation levels of Jurkat cells with and without *TET2* knockout (KO), we observed hypermethylation of largely already highly methylated CpGs (Figure 5.3.28). However, methylation levels in Jurkat with *TET2* KO did not yet reach levels of the higher methylated cell lines PEER and DND41. Using sliding windows separated by HMDs and PMDs, we found that globally both HMDs and PMDs gain methylation in Jurkat KO cells compared to the wild type (WT), which is most pronounced in PMDs (Figure 5.3.29). However, potentially due to the low PMD methylation levels in Jurkat WT cells, PMDs still lag compared to cell lines with high methylation levels, while HMDs reach comparable levels. CGIs also gained methylation upon loss of TET2 and reached levels similar to the cell line PEER but remained less methylated than the extreme levels of DND41 (Figure 5.3.29). Notably, upon loss of TET2, we observed significant up-regulation of DNMT3B, which is also positively correlated with CGI and global methylation across ALL patients (Figure 5.3.29). We, therefore, concluded that TET2 seems to have a par-

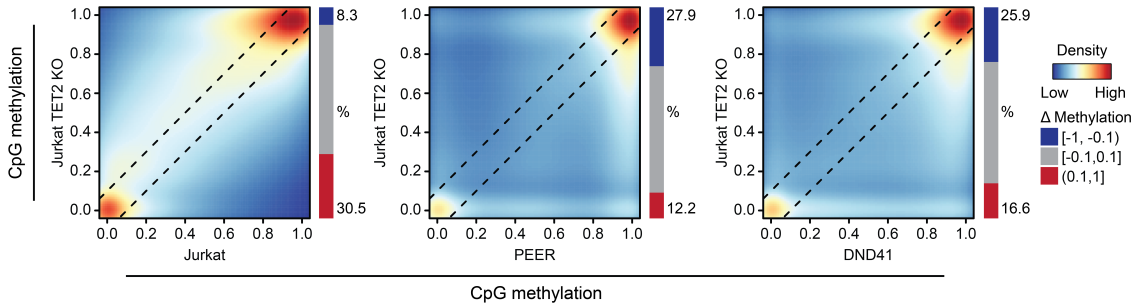

Figure 5.3.28: CpG-wise comparison of Jurkat with and without TET2 knockout (left), PEER, and Jurkat with TET2 knockout (middle), as well as DND41 and Jurkat with TET2 knockout (right).

tial contribution to the CGI and global methylation levels in Jurkat cells. However, it does not fully explain the methylation differences between intermediately and highly methylated T-ALL cell lines. Additionally, a feedback loop between TET2 and DNMT3B might exist as DNMT3B expression is triggered by the loss of TET2 and could add to the hypermethylation effect caused by the loss of a DNA demethylase enzyme by increased *de novo* methylation activity. Together our findings in T-ALL patients and cell lines highlight a role for both TET2 and DNMT3B in shaping the T-ALL methylome.

5.4 Discussion

This study provides extensive whole-genome methylation data sets of healthy lymphocyte progenitors, ALL patients, and cell lines and enables insights into the unique DNA methylation landscape of ALL. In contrast to most other hematopoietic and solid tumor types, ALL - specifically T-ALL - exhibits a stably highly methylated genome without the classic global hypomethylation previously described as a pan-cancer phenomenon. The highly methylated genome is present in pediatric, adolescent, and adult samples, contrasting previous hypotheses that a lack of global hypomethylation could be an exclusive feature of pediatric tumors. Besides ALL, this unusual tumor methylation landscape could only be observed in AML as reported previously [72]. Both ALL and AML are acute leukemias, in contrast to chronic leukemia types like CLL, and are characterized by the rapid accumulation of immature hematopoietic cell types (lymphoid precursors in ALL, myeloid precursors in AML). In contrast, other hematopoietic tumors like CLL and MCL arise from further differentiated cell types (germinal center or memory B cells). These cells are already less methylated genome-wide compared to precursor lymphocytes (Figure 5.3.3), which could imply a link of global hypomethylation to the cell-of-origin methylation levels or stage. On the other hand, TPLL - a different type of T cell leukemia - exhibits global loss of methylation despite the high background methylation in the original primary T cells. Additionally, although precursor B and T lymphocytes show very similar methylation levels and should arise from comparable differentiation stages, T-ALL and B-ALL subtypes exhibit mild differences in the global methylation levels. Although we observed positive correlations between the expression of DNA methyltransferases (DNMT3B, DNMT1) and the level of methylation across T- and B-ALL samples, we cannot causally explain this difference yet. Further - ideally, pan-cancer - studies would

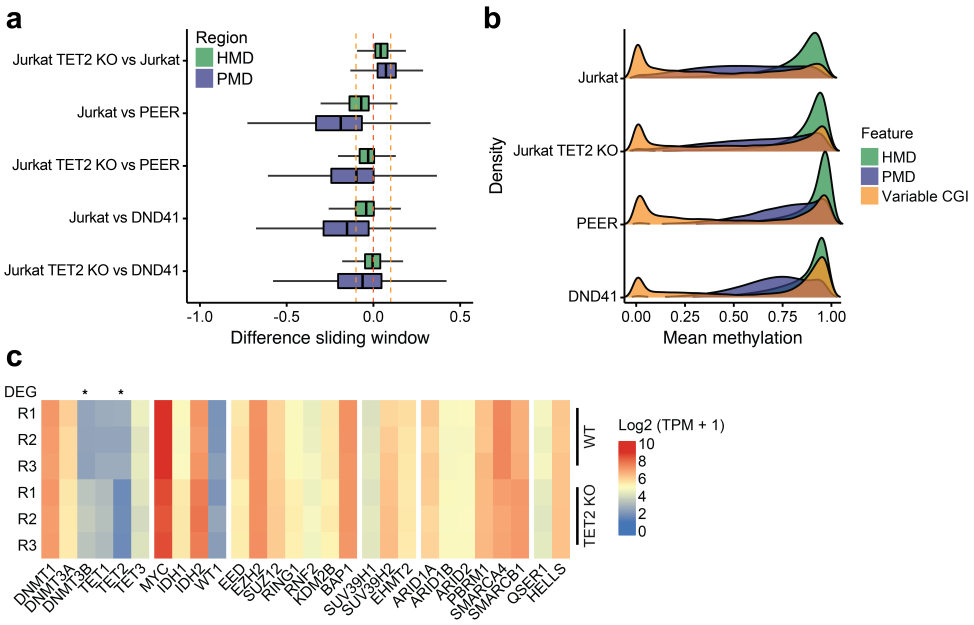

Figure 5.3.29: a) Boxplots showing the delta mean methylation of sliding windows in HMDs and PMDs for different T-ALL cell lines, including Jurkat with and without TET2 knockout. Upon loss of TET2 in Jurkat, methylation in both HMDs and PMDs rises and reaches HMD levels comparable to that of more highly methylated T-ALL cell lines such as DND41 and PEER, while PMDs still lack behind. b) Density of mean methylation in variable CGIs and one kb genomic tiles split into HMDs and PMDs. CGIs also gain methylation after the loss of TET2 in Jurkat cells but do not reach levels of DND41 or PEER. c) Epigenetic regulator expression in Jurkat with and without TET2 knockout.

be needed to determine whether differentiation stages are related to global hypomethylation levels.

As introduced in section 2.5.2, the prevailing model of global hypomethylation assumes that it progressively occurs in PMDs that are late in replication timing, and thus PMD methylation levels deepen in cancer due to the unusually high numbers of cell divisions. However, both ALL and AML are highly proliferative despite exhibiting little to no global loss of methylation. Another study based on colon cancer suggested that instead of a tumor-promoting feature, global hypomethylation is associated with chromatin reorganization and topological changes, which in turn could represent a defense mechanism of the cell based on their findings [122]. As acute leukemias progress rapidly, the absence or low degree of PMD methylation loss under this model might indicate that the cells were able to deactivate a potential defense mechanism or failed to induce such a response. Studies investigating and developing this model further will likely benefit from exceptional cases such as ALL and AML to fully understand the underlying regulatory mechanisms.

In this study, we additionally show that instead of clearly separating into previously defined CIMP groups, T-ALL patients exhibit a wide-spread range of CGI hypermethylation levels and targets. The concept of CIMP has been introduced in section 2.5.2 with the main drawback that it has been previously characterized based on a small selection of probes on the 450k array (single CpGs)

located in CGIs that are often uniquely defined for each tumor type or study (see section 2.5.2). Therefore, the actual CGI methylation levels for CIMP-positive cases of different tumor types might be variable and not comparable, leading to limited interpretability: CIMP-positive patients with breast cancer might have very different CGI methylation levels and targets compared to CIMP-positive colon cancer patients. CIMP-positive cases of one tumor type might even exhibit levels comparable to CIMP-negative samples of another. Therefore, taking all or most CGIs into account when investigating this feature in a given tumor type might help place the extent of hypermethylation in context with other malignancies while simultaneously allowing to compare patients within the indication.

The dynamic CGI methylation levels across T-ALL patients in this study led to a classification of CGIs into different target types defining CGIs that stay consistently unmethylated in the healthy and tumor context and others that seemed to be primed for different hypermethylation levels across T-ALL patients. Most of the frequent hypermethylation targets were associated with promoters whose corresponding genes were already silent in healthy cells. This aligns with previous findings that reported tumor-specific preferential hypermethylation of CGIs that are, for example, repressed by Polycomb complexes in normal tissues [127]. The preference for systematically lower or higher methylation levels of CGI groups defined based on T-ALL patients could be recapitulated in a pan-cancer comparison. This suggests that although CGI hypermethylation has been reported to be specific for different tumor types and even subtypes, a pan-cancer mechanism might exist that predisposes CGI groups with certain features to different hypermethylation levels. Combining these results with pan-cancer chromatin data and chromosomal architecture could give important insights into the underlying dynamics. Given the observed pan-cancer effect, our CGI groups can serve as a guideline in future studies, allowing more specific analyses tailored to CGIs of interest.

Although T-ALL patients have been previously reported to show different levels of CGI methylation (CIMP-positive and -negative), no recurrent mutations or expression differences in epigenetic regulators were identified that could be associated with the two groups. Instead, one study hypothesized that high CGI methylation levels might stem from differences in the mitotic age of the original thymocytes due to a preleukemic phase using data from mouse models and patients with T-ALL [255]. Patients spanning different age groups could serve as a model for some age-related DNA methylation effects. However, we could not observe a significant association of T-ALL methylation-based subtypes with the age of patients. On the other hand, the grouping into CIMP-positive and -negative cases is based on a limited amount of CpGs as well as extracted from a clustering approach, while we observed a wide range of CGI methylation levels. Therefore, comparing two rather simplistic groups might not identify associated expression changes. Our correlation test-based analysis showed a positive correlation of DNMT3B, DNMT1, and TET1 expression with global or CGI methylation levels across ALL subtypes. Additionally, we observed *TET2* promoter hypermethylation in a subset of T-ALL^{IM} and T-ALL^{HM} patients accompanied by decreased TET2 expression. The correlation between methylation levels and expression of both DNMTs and TET1 leads to the possibility that part of the observed CGI hypermethylation levels is, in fact, hydroxymethylation, which is not distinguishable from 5-methylcytosine by WGBS. Increased expression of DNMT3B and TET1 might increase methylation turnover at CGIs [277]. If TET2 is additionally lost in some patients, this equilibrium could be disturbed and shifted towards *de novo* methylation, increasing the overall DNA methylation levels.

Finally, we could recapitulate an effect of *TET2* hypermethylation and association between TETs and DNMTs in Jurkat cells by knocking out the *TET2* gene, which led to up-regulation of DNMT3B and was accompanied by an increase in global as well as CGI methylation levels. This shows that the loss of TET2 seems to have an effect on the overall methylation levels. However, we cannot distinguish whether the loss of *TET2* represents a driving event of high methylation levels or a side effect of already extensive hypermethylation in T-ALL patients. Although our selection of T-ALL cell lines seemed to mimic certain aspects of the DNA methylation dynamics in T-ALL patients, the methylome of ALL cell lines, in general, frequently deviated from that of the primary indication. This was most apparent on the genome-wide level, where cell lines presented with strong hypomethylation likely originating from long-term culture. These findings highlight that selected cancer cell lines can be used to investigate epigenetic dynamics and regulation in tumors. However, their ability to serve as a model for a specific indication or subtype needs to be assessed carefully. Additionally, results need to be interpreted in light of potential changes in the epigenetic landscape and machinery that can be induced through extensive culturing of cells.

Chapter 6

Redefining DNA methylation landscapes across tumors and cell lines

In this study, the DNA methylomes of healthy human tissues, primary tumors, and cancer cell lines were compared, and the main forms of DNA methylation landscapes across these data sets were defined. Our investigations showed that primary tumors are characterized mainly by intermediate methylation levels intrinsic to the underlying cells and cannot be explained by tumor purity alone. In contrast, cancer cell lines frequently converge to one of two different DNA methylation states rarely found in primary tumors but are associated with the tumor type of origin.

Genomic DNA of healthy tissues and primary tumors for six different solid tumor types was purchased from OriGene and profiled using WGBS at the Broad Institute by Dr. Kathleen Steinmann and Dr. Andreas Gnirke. Cancer cell lines were ordered from DSMZ, ATCC, the Korean Cell Bank, and the JCRB Cell Bank, cultured by Dr. Raha Weigert, and profiled using WGBS at the Max Planck Institute for Molecular Genetics by Dr. Nina Bailly (library preparation) and the Sequencing Core Facility.

6.1 Biological background

6.1.1 Cancer cell lines as model systems

In the previous chapter, a single tumor type, acute lymphoblastic leukemia, was analyzed in detail, and its unusually highly methylated genome was put into perspective regarding the observed degree of pan-cancer global hypo- and CGI hypermethylation. After epigenetic regulators were identified in patients that potentially play a role in establishing and maintaining the unusual ALL landscape, the immediate effect of TET2 loss was tested using a perturbation experiment in the T-ALL cancer cell line Jurkat. This is a commonly used strategy when the consequences of specific changes observed in cancer need to be investigated: Analyzing the effect of mutations, loss, or overexpression of a particular gene as well as drug screenings and many other experiments related to understanding the regulation of cancer cells are frequently carried out in cancer cell lines [145]. Primary tumors already established the genetic and epigenetic landscapes that can be

observed after sequencing and cannot be manipulated to understand the underlying regulations. Cancer cell lines, on the other hand, can be kept in culture indefinitely and offer the opportunity to easily test the effect of genetic perturbations or drug treatments. Consortia like the Genomics of Drug Sensitivity in Cancer project (GDSC) or the Cancer Cell Line Encyclopedia (CCLE) use large collections of cancer cell lines to conduct detailed studies on genetic dependency mapping, molecular profiling, and drug screenings that aim to link drug responses and genetic as well as epigenetic set-ups to advance therapeutic development [278–284].

6.1.2 DNA methylation in primary tumors and cancer cell lines

Gain of methylation at PRC2-targeted CGIs and genome-wide loss of methylation have been not only widely reported as a hallmark of tumors but also as an effect of aging and long-term culture [7, 72, 128]. Specifically, cancer cell lines have been shown to exhibit more extreme hypermethylation levels (as well as additional CGI targets) [146, 147] and more pronounced global hypomethylation compared to their primary counterparts [72, 147] (see section 2.5.2 for more details). This was also observed when profiling the methylome of Jurkat cells as a model of T-ALL described in section 5.3.5 where the genome-wide methylation landscape was characterized by hypomethylation in contrast to the highly methylated genome of primary T-ALL cases. On the other hand, B-ALL cell lines such as NALM-6 and MHH-CALL-4 exhibited higher CGI methylation levels than corresponding patients reaching methylation levels up to 100%.

Such observations have been previously attributed to the increase in the number of cell divisions in culture, which also applies to tumor progression compared to healthy, somatic cells: Progressive genome-wide loss of methylation could be linked to the potentially compromised fidelity of the maintenance DNA methyltransferase DNMT1 in late-replicating PMDs, an effect that accumulates over time [72]. Hypermethylation of CGIs is also thought to occur stochastically over cell divisions as a consequence of changes in chromatin state where previously PRC2-repressed CGIs transition to a more stable silencing by DNA methylation and H3K9me3 (epigenetic switch model) [4, 5, 126, 128]. These hypotheses essentially point to a proliferation- and, therefore, also a time-dependent model: Starting with the early stages of tumorigenesis, PMD hypo- and CGI hypermethylation become more pronounced the more the cells proliferate. This phenomenon then leads to more extreme methylation levels in culture as cancer cell lines can grow indefinitely in contrast to an actual primary tumor whose lifespan is ultimately linked to its removal or death of the patient.

In addition to this time-related model, the complexity of primary tumors compared to cancer cell lines represents a source that could contribute to the observed DNA methylation differences: Primary tumors are often heterogeneous, comprising different genetic subclones but also infiltration by immune cells, neighboring healthy tissue and stromal contribution [285–287]. Different cell types can also exhibit differential methylation at promoters or enhancers linked to their cellular identity and associated transcriptional differences [263]. In contrast, cancer cell lines are unaffected by somatic contamination, and long-term culture is known to reduce the population complexity [288, 289]. Together with the general prevalence of methylation changes across tumor types, this raises the question of whether a uniform “cancer DNA methylation landscape” with low PMD and high CGI methylation levels exists that tumors progress towards or already

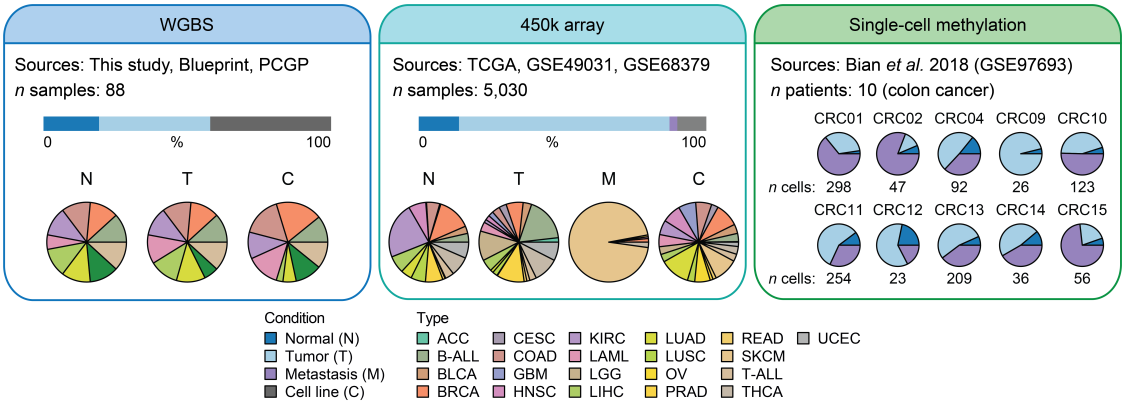

Figure 6.2.1: Overview of healthy, tumor, metastasis, and cell line samples analyzed in this study.

maintain, shadowed by clonal evolution and reduced purity.

6.1.3 Aims and scope of the study

Here, we aimed to investigate global DNA methylation dynamics during tumorigenesis and compare these to the genomic distribution of this modification within cancer cell lines. For this purpose, we generated a high-resolution reference of primary tissues, tumors, and cancer cell lines profiled with WGBS and subsequently integrated thousands of publicly available data sets. Our study reveals that intermediate DNA methylation is a defining feature of most tumor types and samples across both PMDs and CGIs. In contrast, cell lines largely maintain extreme states, including an “inverse bimodal” landscape where PMD methylation is exceptionally low and CGI methylation is strikingly high, as well as a state of extreme global hypermethylation affecting the entire genome. Investigation of read-level methylation across our cohort and of publicly available single tumor cell methylation profiles confirm that intermediate methylation is an intrinsic feature of most tumor cells *in vivo*, largely independent of tumor purity and can persist across aggressive population bottlenecks such as metastasis. Finally, we demonstrate that distinct tumor types are prone to acquire certain DNA methylation landscapes in culture, which is reflected by their mutational signatures as well as sensitivity to specific classes of small molecule inhibitors. Our study highlights the striking conservation and maintenance of specific properties of a pan-cancer epigenome as well as distinct types of major shifts that frequently occur *in vitro*.

6.2 Materials and methods

6.2.1 Cohort overview

For this study, a broad selection of publicly available methylation data sets of different types was used and complemented by newly generated WGBS data sets (Figure 6.2.1 and Table 6.2.1). In the following, the compiled cohorts are introduced and described.

Tumor type	Abbreviation	WGBS source	Array source
Adrenocortical carcinoma	ACC	-	TCGA
B cell acute lymphoblastic leukemia	B-ALL	EGAS00001005203	GSE49031
Bladder urothelial carcinoma	BLCA	-	TCGA
Breast adenocarcinoma	BRCA	This study	TCGA
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	-	TCGA
Colon adenocarcinoma	COAD	This study	TCGA
Glioblastoma multiforme	GBM	-	TCGA
Head and neck squamous cell carcinoma	HNSC	-	TCGA
Kidney renal clear cell carcinoma	KIRC	This study	TCGA
Acute myeloid leukemia	LAML	Blueprint	TCGA
Brain lower grade glioma	LGG	-	TCGA
Liver hepatocellular carcinoma	LIHC	This study	TCGA
Lung adenocarcinoma	LUAD	This study	TCGA
Lung squamous cell carcinoma	LUSC	-	TCGA
Ovarian serous cystadenocarcinoma	OV	-	TCGA
Pancreatic adenocarcinoma	PAAD	This study	-
Prostate adenocarcinoma	PRAD	-	TCGA
Rectum adenocarcinoma	READ	-	TCGA
Skin cutaneous melanoma	SKCM	-	TCGA
T cell acute lymphoblastic leukemia	T-ALL	EGAS00001005203	GSE49031
Thyroid carcinoma	THCA	-	TCGA
Uterine corpus endometrial carcinoma	UCEC	-	TCGA

Table 6.2.1: Tumor types that are part of the WGBS and 450k array cohort, their abbreviations used in the text and figures, and the public source the data sets were obtained from.

Whole-genome bisulfite sequencing

The WGBS cohort used in this study is comprised of nine different indications: breast adenocarcinoma (BRCA), colon adenocarcinoma (COAD), renal cell clear cell carcinoma (KIRC), hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), pancreatic adenocarcinoma (PAAD), acute myeloid leukemia (LAML) as well as B and T cell acute lymphoblastic leukemia (B-ALL, T-ALL, Table 6.2.1).

Genomic DNA for solid healthy and primary tumor samples (BRCA, COAD, KIRC, LIHC, LUAD, and PAAD) was obtained from OriGene. Only tumor samples with a purity $\geq 80\%$ were selected. Two healthy and four tumor samples were obtained for each indication, except for PAAD, where only two tumor samples of sufficient purity were available.

Four primary LAML samples and one multipotent hematopoietic progenitor sample as control were obtained from the Blueprint Epigenome project [257]. Precursor B and T cells (two samples each) as well as B- and T-ALL patients (four samples each) were obtained from the ALL study described in the previous chapter [238]. B-ALL patients were selected from the DUX4-rearranged/ERG-deregulated subtype, and T-ALL samples were selected from different methylation-based subtypes to span the wide range of T-ALL methylation levels.

The following cell lines were purchased from the German Collection of Microorganisms and Cell Cultures (DSMZ): CL-11, TALL-1, MOLT-16, SUP-B15, RCH-ACV, MHH-CALL-2, PL-21, AML-193, OCI-M1, SIG-M5, MONO-MAC-1, HCC827, T-47D, MDA-MB-231, MDA-MB-468, EFM-19, CL-14, CL-40, SW948, BXPC-3, HUP-T3, PA-TU-8988S, PANC-1, A549, Hep-G2, A498, NALM-6, Jurkat, and DND-41. The following cell lines were purchased from ATCC: HCT116, MCF7, BT-20 (HTB-19), HCC38, A704, and SW1417. The cell line SNU-1272 was purchased from the Korean cell bank, and the cell line KMRC-1 was purchased from the JCRB cell bank. All cancer cell lines were cultured according to the vendor's recommendations. To the best of our knowledge, we aimed to avoid cell lines known to be contaminated, and we prioritized cell lines grown out of primary tumors and not their respective metastasis, although this could not always be ensured.

In total, the WGBS cohort included 17 healthy samples, 34 primary tumor samples, and 37 cancer cell lines.

450k array

The Cancer Genome Atlas comprises the largest collection of methylation data sets for healthy and tumor samples profiled with the Infinium HumanMethylation450 BeadChip (450k array). Additionally, for some tumor types, also metastasis samples were profiled and available via TCGA. Similarly to the WGBS cohort, we aimed to only select samples from TCGA with high tumor purity ($\geq 80\%$) to minimize confounding effects due to contamination. A detailed description of our filtering process is described in the next section 6.2.2. To complement this cohort, we added additional samples from a study profiling B- and T-ALL patients [290] as well as a large cancer cell line cohort [279] that was reduced to the cell lines matching a tumor type for which primary tumor samples were available (Table 6.2.1).

In total, the 450k array cohort contained 5,030 samples, including 705 healthy samples, 3,681 primary tumor samples, 137 metastases, and 507 cancer cell lines.

Single-cell WGBS

A colon cancer cohort comprising 10 patients profiled using single-cell WGBS was obtained from Bian et al. [291]. For each patient, healthy and primary tumor cells from up to six different

sampling sites were available (Figure 6.2.1). Additionally, for most patients, cells from different types of metastasis were measured.

6.2.2 Initial data processing

Whole genome bisulfite sequencing

The experimental procedures of WGBS library preparation are described in the appendix C. The quality of the sequencing runs was inspected using FastQC (version 0.11.9) [171]. Reads were subjected to trimming using cutadapt (version 2.4, [172]): Bases with a quality score less than 20 were removed as well as adapter content, followed by the trimming of 10 and 5 nucleotides from the 5' and 3' end of the first read and 15 and 5 nucleotides from the 5' and 3' end of the second read respectively. The trimmed reads were then aligned to the human reference genome (hg19) using BSMAP (version 2.90; parameters: -v 0.1 -s 16 -q 20 -w 100 -S 1 -u -R) [183]. Following the alignment, PCR duplicates were removed using GATK with the 'MarkDuplicates' command (version 4.1.4.1) [187]. Afterwards, methylation rates were called with mcall (MOABS package, version 1.3.2) [188]. The resulting methylation rates were filtered such that only CpGs covered by at least 10 and at most 150 reads on autosomes were considered for downstream analyses.

450k array

Publicly available ALL data sets generated using the Illumina Infinium HumanMethylation450 BeadChip were processed using the Minfi R package (version 1.32.0) [292]. Data was loaded with the 'read.metharray.exp' function. Failed positions were identified using the function 'detectionP' (parameters: type = "m+u"). Data were normalized using the Noob normalization ('preprocessNoob', parameters: dyeMethod = "single). The methylation ratio was computed using the 'ratioConvert' function (parameters: what = "both", keepCN = TRUE). The following probes were excluded from downstream analyses:

- Probes overlapping known SNPs (Minfi: 'addSnpInfo', 'dropLociWithSnp', parameters: snps = c("Probe", "SBE", "CpG"), maf = 0.01) as these might no longer reflect CpGs in respective patients.
- Non-CpG positions as only CpGs were considered for the analyses.
- Probes locates on sex chromosomes in line with sequencing data analyses.
- Probes that failed the detection test (p-value ≥ 0.05). These can be considered failures during the experiment because methylated and unmethylated channels both report levels of background signal [293].
- Probes known to frequently cross-react, which means they align to multiple positions in the genome [294].

For samples from TCGA and Iorio et al. [279] processed beta values were downloaded, but positions were reduced to the probes passing all filtering steps during the processing of the ALL samples.

Tumor purity estimation for TCGA tumor samples was obtained from Aran et al. [295]. Different methods exist to estimate tumor purity from various data sources, including DNA methylation, gene expression, and mutations [296–298]. We decided to use purity estimates from a consensus prediction method instead of one of the many tools that estimate tumor purity based on DNA methylation alone, even though not all TCGA tumor types are covered by Aran et al. In addition to providing purity estimates based on more data types, we also found that consensus prediction may be better suited to address the confounding nature of intermediate DNA methylation in cancer. DNA methylation-based purity estimates rely on the assumption that intermediately methylated probes in primary tumors are mostly the result of cellular contamination between hypermethylated cancer and hypomethylated somatic cells [296]. However, our analyses confirm that primary tumors generally maintain CGIs in intermediately methylated states (see sections 6.2.5 and 6.3.3). As such, DNA methylation-based estimates appear to over-estimate the degree to which cellular contamination explains intermediate CGI levels, leading to consistently lower purity estimates than found using a consensus prediction method (Figure 6.2.2). Based on our data, purity estimates using DNA methylation alone might not be perfectly suited to estimate tumor purity, given the mixed effect of intrinsic intermediate methylation and additional contamination from healthy cells to varying degrees. To this point, we decided to additionally include the TCGA-LAML cohort, which based on the DNA methylation-based purity estimator InfiniumPurify exhibits purity levels of almost 100% [296]. This is in line with the fact that leukemic samples are usually flow-sorted or purified using density gradient centrifugation and, therefore, might generally exhibit higher purities than solid tumors [238,299]. The TCGA cohort was thus reduced according to the following criteria:

1. Only tumor and metastasis samples that were either part of the TCGA-LAML cohort or had an estimated purity $\geq 80\%$ were considered.
2. Only healthy, tumor and metastasis samples were selected that were associated with a type for which at least one cancer cell line was available from Iorio et al. [279]
3. For the LGG cohort specifically, only tumor samples with a known classification into IDH wild type or mutant were considered (see next section).

Similarly, the cancer cell line cohort obtained from Iorio et al. was reduced to cell lines from tumor types that survived the filtering criteria applied to TCGA or matched the ALL cohort (B-ALL and T-ALL).

Single-cell WGBS

CpG methylation measurements for single-cell WGBS colon cancer data sets were obtained from Bian et al. [291]. Matching CpGs on the plus and minus strands were combined, and only CpGs on autosomes were considered (matching the processing of WGBS methylation rates). Only cells with ≥ 3 million covered CpGs were selected for downstream analyses. For each patient, a pseudo bulk was generated at the level of CGIs and 100 kb genomic tiles (not per CpG due to the

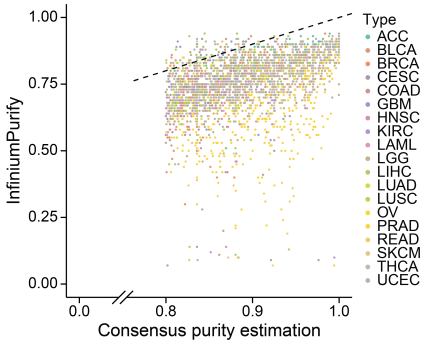

Figure 6.2.2: Comparison of tumor purity estimates using a consensus prediction method [295] and a methylation-based prediction method (InfiniumPurify) [296].

sparsity of single-cell WGBS data sets). Here, the average methylation of each CGI and tile was calculated per cell, and the respective methylation values were subsequently averaged across cells of the same patient and condition (healthy tissue, primary tumor, and different types of metastases) to mimic a bulk methylation sample similar to the standard WGBS data sets.

6.2.3 Overview of healthy, tumor, and cancer cell line methylomes

Genomic features

One and 100 kb genomic tiles were generated by segmenting the genome using bedtools makewindows (version 2.30.0; parameters: -w 1000 -s 1000 and -w 100000 -s 100000 respectively) [261]. Annotations of HMDs and PMDs, as well as solo-WCGW CpGs for hg19, were downloaded from https://zwdzwd.github.io/pmd [72]. The hg19 gene annotation was downloaded from GENCODE (V19). Promoters were defined as 1500 bp upstream and 500 bp downstream of the TSS. The annotation of tumor suppressor genes was downloaded from the TSGene database [300]. ChromHMM annotations for the human embryonic stem cell line HUES64 were downloaded from Roadmap (E016).

Annotations of CGIs were downloaded from UCSC. The chromatin state of CGIs was defined as the ChromHMM state with the largest overlap with each island, and a CGI was termed PRC2 target if the corresponding chromatin state was one of 10_TssBiv, 11_BivFlnk, 12_EnhBiv, 13_ReprPC, and 14_ReprPCWk. CGIs were defined as promoter CGI if at least 20% of the CGI or the promoter overlapped, as a gene body CGI if at least 20% of the CGI or the gene body overlapped but not with a promoter and intergenic for all remaining islands.

Feature-specific and genome-wide methylation measurements

For WGBS samples and single-cells from the single-cell WGBS colon cancer cohort, the arithmetic mean was calculated across features (tiles, CGIs). A feature was only considered if at least three CpGs were covered within a region. For samples profiled with the 450k array, the average

methylation of each CGI was computed using beta values of probes located within the respective CGI. For any analysis and technology, CGIs with an average methylation > 0.2 were termed methylated, otherwise unmethylated.

Compared to WGBS, the 450k array spans a limited number of CpGs enriched for regulatory features such as CGIs, promoters, enhancers, and gene bodies [92]. This represents a challenge for estimating PMD methylation levels: PMDs are typically characterized as large-scale late-replicating domains that are GC- and gene-poor [72], properties that can be assessed well using WGBS but less so using the biased sampling of the 450k array. As described in section 2.3, isolated solo-WCGW CpGs are most prone to hypomethylation and therefore provide an accurate measurement to determine the degree of hypomethylation in a sample. Of these CpGs, 6,214 are located within previously defined common pan-cancer PMDs and covered by the 450k array. Of these, 4,832 survive the filtering steps applied on the 450k array cohort as described above [72]. In order to enable and streamline analyses of WGBS and array samples alike, the methylation levels of solo-WCGW CpGs in common pan-cancer PMDs were therefore used to assess genome-wide PMD methylation levels across healthy, tumor, metastasis, and cell line samples. These CpGs were challenging to use within the single-cell data sets as within a single-cell, a CpG can, in theory, only reflect three different methylation measurements (0, 0.5, or 1 depending on the allelic methylation status), creating uninformative median values across solo-WCGW CpGs of either 0 or 1 in most cases. As these measurements do not enable a clear assessment of the degree to which individual cells maintain their PMDs, we performed an independent analysis of the single-cell data that differs from our analyses of WGBS or 450k array data. Specifically, we calculated the average methylation of 100 kb tiles in common pan-cancer PMDs as our genome-wide methylation measurement.

Genome browser tracks were generated using IGV (version 2.15.2) [232].

Definition of hypermethylated CGIs

Hypermethylated (hyper) CGIs per tumor, metastasis, and cell line sample (WGBS and 450k array) in comparison to healthy tissues were defined as follows: For each CGI, first, the median signature of healthy control samples profiled with the respective technology was defined (median over the average CGI methylation of each healthy sample). If available, the matching healthy tissue was used for this purpose. Otherwise, the median signature of all healthy array-based control samples was used (only applicable to some TCGA cohorts where no matching normal tissue was available, such as ACC or LGG). Then the average methylation of each CGI in the malignant sample was compared to the healthy signature. CGIs that were unmethylated in the healthy condition (methylation ≤ 0.2), methylated in the malignant sample (methylation > 0.2), and with a difference > 0.1 between malignant and median normal signature were termed hyper CGIs for the respective sample. For the single-cell WGBS patients, hyper CGIs per patient were defined analogously using the *in silico* generated pseudo bulk of normal and tumor cells for each patient specifically.

Hypermethylated CGIs per tumor type (WGBS and 450k array) were defined based on the sets of hypermethylated CGIs defined for each tumor sample (see above): CGIs that were termed hypermethylated in at least 75% of tumor samples of a specific type were considered as “common”

hyper CGIs for that type. Hypermethylation sets were defined separately for WGBS and array cohorts spanning the same tumor types due to differences in CpG coverage between the two technologies. For TCGA-THCA and TCGA-KIRC, 50% of the tumor samples were considered sufficient to select a CGI as hypermethylated due to the overall lower degree and more subtle trends of CGI hypermethylation across these cohorts. For the TCGA-LGG cohort, samples were split by IDH mutational status, and a separate set of hyper CGIs was defined for each subtype. We performed this additional subselection to ensure a fair comparison between primary tumors and cell lines: LGG tumors are mostly classified as IDH mutant, while all cell lines are reported as wild type, and IDH mutations appear to target a distinct CGI range (see section 2.5.2 on the glioma CIMP subtype) [143]. Performing our analyses without this stratification led to the selection of CGI sets that were biased towards the IDH mutant profile and confounded our ability to accurately infer the DNA methylation landscape of IDH wild type cell lines. Close inspection of all additional cancer types confirmed that this issue is substantially more prominent for LGG than other cohorts.

Saturation analysis of hypermethylated CGIs

In order to assess how uniformly the sets of common hyper CGIs are sampled across the tumor data sets of each type, a saturation analysis was performed. For this purpose, only tumor types with at least 100 samples were considered. For each tumor type, 100 iterations were performed, and within each iteration, 100 tumors were randomly sampled, followed by a calculation of the cumulative proportion of CGIs added with each subsequent sample. The results were then averaged across the 100 iterations to provide a more stable assessment of the degree to which individual samples contribute additional information to each tumor type-specific hyper CGI profile. The same analysis was performed for the entire sets of tumor samples and cancer cell lines, respectively.

6.2.4 Definition of DNA methylation states

A primary objective of this study was to broadly group WGBS and 450k array samples according to the methylation levels of hyper CGIs and solo-WCGW CpGs (in PMDs) instead of the genetic identity of target sequences. To accomplish this, we first clustered healthy, tumor, and cell line samples based on ECDF of each feature (hyper CGIs and solo-WCGW CpGs). This analysis was performed separately for each technology and feature as the distribution of methylation levels between WGBS and array samples can be affected by the differences in CpG sampling (the rather universal coverage of different genomic regions in WGBS compared to biased enrichment of regulatory features in 450k arrays). Technological differences are especially relevant for the solo-WCGW CpGs that are covered to a much larger extent in WGBS compared to array samples (on average 1.4 million per WGBS sample). The sample-to-sample distances based on the distribution of hyper CGI or solo-WCGW CpG methylation levels were computed using the Kolmogorov-Smirnov (KS) distance. Samples were then hierarchically clustered and visualized using the ComplexHeatmap R package (version 2.9.4) [266]. Four clusters for hyper CGIs and solo-WCGW CpGs each were extracted from the hierarchical clustering to examine the properties of broad sample groupings.

Ultimately, a simple scoring method was implemented that allowed simultaneous examination of both PMD and CGI compartments in a manner that was largely robust to the technology used. For this purpose, the median methylation of solo-WCGW CpGs in common PMDs, as well as the median methylation of tumor type-specific hyper CGIs, was used. WGBS samples were clustered based on both features using the ConsensusClusterPlus R package (version 1.48.0; function: ConsensusClusterPlus; parameters: maxK=12, reps=100, pItem=0.8, pFeature=1, clusterAlg="pam", distance = "euclidean", seed = 42) [265]. By assessing the change in the area under the ECDF curve as described in section 5.2.4 as well as the intra-cluster variability, the optimal number of clusters ($n = 5$) was determined. The clusters were annotated based on the overall PMD and hyper CGI methylation levels (low, intermediate, or high) and considered as different DNA methylation states (or landscapes). Specifically, the resulting methylation states were termed PMD^{high} CGI^{low} (somatic), PMD^{high} CGI^{int} (intermediate), PMD^{int} CGI^{int} (intermediate), PMD^{high} CGI^{high} (extreme hypermethylation) and PMD^{low} CGI^{high} (inverse bimodal). The array samples were subsequently assigned to the nearest WGBS-based cluster using a k nearest neighbor (k-NN) classification ($k = 10$). The WGBS samples were chosen as the reference even though the array cohort contains a larger number of samples because the CpG and, therefore, feature-wise coverage was considered more complete and better reflective of the actual genomic methylation distribution than those obtained from the array cohort. We found that clustering on the median methylation values of CGI and PMD methylation ECDFs was robust, and results conducted on more detailed parts of their methylation distributions (interquantile range or 10/90th percentiles) provided highly similar results.

6.2.5 Read-level analysis

The average methylation for each read for WGBS samples was obtained using RLM (version 1.0.0, see chapter 4) [222]. Read-wise average methylation measurements were aggregated per hypermethylated CGI to compare the cumulative distribution of read-level methylation across healthy, tumor, and cell line samples. In order to generate browser tracks, the read-wise methylation levels were aggregated in 500 bp sliding windows (step size 100 bp). Entropy per 4-mer of CpGs was calculated using RLM. Mean entropy per CGI was calculated using the arithmetic mean. Only 4-mers covered by at least 10 and at most 150 reads were considered.

In order to assess if higher methylation entropy in tumors is a consequence of cellular heterogeneity (such as contamination) versus per-molecule, stochastic methylation across cells (allelic heterogeneity), we performed an *in silico* mixing experiment. For this purpose, we selected high-purity T-ALL and matching T cell samples [238]. T-ALL allowed us to model the effects of cellular contamination with high-purity samples as well as to directly investigate the effects of fully hypermethylated tumor samples versus those with more intermediate methylation levels. Our selected samples included a T-ALL sample with extremely high CGI methylation levels as well as a sample with intermediate CGI methylation levels, and a healthy sample with somatic DNA methylation landscape. For each 4-mer and tumor sample, different *in silico* mixtures of epialleles were randomly sampled, reflecting different purities. Specifically, for each 4-mer, 20 epialleles were randomly selected, of which a specific fraction was sampled from the tumor, while the remaining fraction was sampled from the control. The sampling was restricted to 4-mers covered by at least 20 reads in each sample and located within a hyper CGI specific for

T-ALL. For each 4-mer, the methylation entropy, fraction of discordant epialleles, and average methylation were computed. This way, two scenarios could be modeled and compared with the entropy measurements obtained from primary samples:

1. Highly methylated CGIs in tumor cells contaminated by unmethylated somatic CGI methylation levels to different degrees (mixture of T-ALL tumor 1 with normal 1).

2. The effect of contamination with somatic cells on stochastic, intermediate methylation levels intrinsic to primary tumor cells (mixture of T-ALL tumor 4 with normal 1).

6.2.6 Correction of DNA methylation measurements by tumor purity

In order to correct methylation measurements of solid tumors based on the reported purity (pathology for WGBS, consensus prediction by Aran et al. for 450k arrays), an aggressive but simple strategy was employed. This strategy is very likely to overestimate the effect of cellular contamination with non-tumor cell types but allowed us to strictly assess the stability of our state assignments. For every sample and the associated fraction of actual tumor cells p (purity) with $0.8 \leq p \leq 1$ for samples within our cohort, it was assumed that the remaining fraction of cells $1-p$ stemmed from healthy, somatic cells with a perfect bimodal distribution where CGIs are completely unmethylated and the remaining genome is fully methylated (Figure 6.2.3). For a given feature (CGI or solo-WCGW CpG), the (average) methylation of the actual tumor cells Me_T together with the contaminating somatic cells and their assumed methylation Me_N lead to the combined methylation measurement Me_C (as observed in the primary data) as follows:

$$Me_C = p * Me_T + (1-p) * Me_N \tag{6.1}$$

According to the assumption of fully unmethylated CGIs in normal somatic cells ($Me_N = 0$), the following tumor-specific (corrected) CGI-specific methylation levels can be obtained:

$$Me_T = \frac{Me_C}{p} \tag{6.2}$$

When considering a fully methylated somatic genome outside of CGIs, including solo-WCGW CpGs with $Me_N = 1$, the actual tumor cell background methylation can be calculated as

$$Me_T = \frac{Me_C - (1-p)}{p} \tag{6.3}$$

This correction was applied to both WGBS and 450k array solid tumor samples, and the corrected hyper CGI and solo-WCGW CpG methylation measurements were used to re-assign samples to DNA methylation states as described above (k-NN approach).

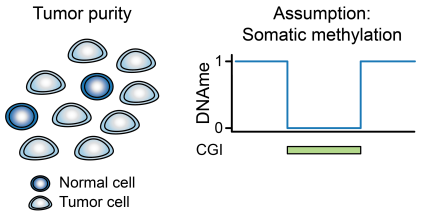

Figure 6.2.3: Schematic illustrating the correction method applied to WGBS and 450k array methylation estimates for CGIs and PMDs. A perfect bimodal distribution of somatic cells was assumed where CGIs are fully unmethylated while the genomic background is fully methylated. The methylation levels of solid tumors were adjusted using this model together with the tumor purity reported or estimated per sample.

6.2.7 Single-cell WGBS analysis

Euclidean distances between healthy and tumor cells and between cells within and across different tumor and metastasis sampling sites were computed per patient based on the average methylation of each hyper CGI or 100 kb tile in common PMDs. Heatmaps of hyper CGI and 100 kb tile methylation per patient were visualized using the ComplexHeatmap R package.

6.2.8 Association of cancer cell line DNA methylation states with tumor type

For all analyses related to the association of cancer cell line DNA methylation landscapes with tumor type, only cancer cell lines of the 450k array cohort were used that were assigned to their respective state with a probability ≥ 0.7 (based on the *k*-NN classification). Cell lines that lacked sufficient drug screenings were excluded (less than 90% of the overall screened drugs tested, see next paragraph). Additionally, only tumor types with at least eight available cell lines after the filtering steps were considered. The association of 1) the two main cancer cell line landscapes (PMD^{high} CGI^{high} and PMD^{low} CGI^{high}) with tumor type as well as 2) the intermediate (PMD^{high} CGI^{int}, PMD^{int} CGI^{int}) compared to the remaining states with tumor type were assessed using two-sided Fisher's exact test. *P*-values were corrected for multiple testing using FDR and adjusted *P*-values < 0.05 were considered significant.

Culture conditions, drug responses (measured by IC50, the molecular concentration of the drug to inhibit a biological process by 50% [301]), and mutations of the cancer cell line cohort were obtained from Iorio et al. The set of epigenetic regulator genes was obtained from dbEM [302], and the Cancer Gene Census, with common oncogenes and tumor suppressor genes, was obtained from COSMIC [303]. Epigenetic regulators and cancer driver genes were selected as recurrently mutated if a mutation was reported in 5% and 10% of all considered cancer cell lines, respectively (mutations in epigenetic regulators are overall rarer than in driver genes). Additionally, driver genes previously reported for each tumor type and their associated population frequency were obtained from Bailey et al. [304] (TCGA), Studd et al. [305] (B-ALL) and Liu et al. [306] (T-ALL). Heatmaps and oncoprint were visualized using the ComplexHeatmap R package. The association between mutation status per recurrently mutated gene and either the DNA methylation state or tumor type was assessed using a two-sided Fisher's exact test (significance was assessed as described above).

For the drug response analysis, only drugs tested in at least 90% of the considered samples were included. The analysis was limited to cell lines of the two most frequent states (PMD^{high} CGI^{high} and PMD^{low} CGI^{high}) due to the overall low sample size of cell lines in an intermediate methylation state. The association of DNA methylation landscape and response to a specific drug was assessed independently for each drug using a simple logistic regression model:

$$ln\left(\frac{Y_{ij}}{1-Y_{ij}}\right) = \beta_0 + \beta_1 * X_{1ij} \tag{6.4}$$

Here, Y_{ij} represents the DNA methylation landscape (two possible outcomes) of cell line i of tumor type j, X_{1ij} represents the drug response (centered, ln-transformed IC50) of cell line i of tumor type j, β_0 represents the intercept and β_1 represents the coefficient of X_1 with $i \in 1,...,n_j$ and $j \in 1,...,m$. m represents the total number of tumor types within the cohort. P-values derived from each logistic regression were corrected using FDR, and adjusted P-values < 0.05 were considered significant. The odds ratio for each drug was obtained as e^{β}. Additionally, a mixed effects model was built for every drug using the lme4 R package (version 1.1-23) [307] where the tumor type of the cell lines was added as a known random effect. For this purpose, a random intercept model was used based on the observation that different types have a higher likelihood of being associated with one of the two landscapes:

$$ln\left(\frac{Y_{ij}}{1-Y_{ij}}\right) = (\beta_0 + u_{0j}) + \beta_1 * X_{1ij} \tag{6.5}$$

Here, u_{0j} represents the random intercept for each tumor type j. Significance was assessed as described for the simple logistic regression models.

6.2.9 Copy number analysis

The copy number analysis for WGBS samples was performed using Control-FREEC (version 11.6) [308]. Due to the wide ranges of karyotypes being reported for the same cancer cell lines and because the overall ploidy of a cell line was not relevant for the analysis, a diploid genome was assumed as the baseline of all samples, including cell lines. Then, instead of the exact copy number indicated by the tool, only the copy number deviation status (amplification, deletion) was used to assess whether specific chromosomes or genomic regions have more or fewer copies compared to the baseline. For each sample, the total size of genomic regions that were amplified or deleted was visualized.

6.3 Results

6.3.1 Characterizing DNA methylation landscapes of tumors and cell lines

Given our aim to investigate the DNA methylation landscapes across primary tumors and cancer cell lines, we first verified that tumors of our WGBS cohort resemble the widely reported

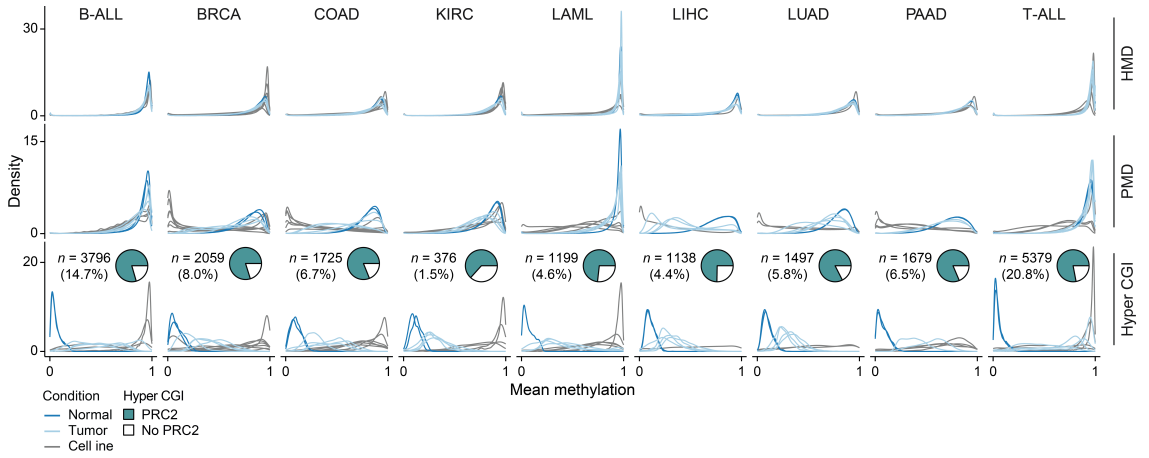

Figure 6.3.1: Mean methylation distribution of one kb tiles in HMD/PMDs and hypermethylated CGIs defined per tumor type profiled using WGBS. Pie charts show the fraction of hypermethylated CGIs targeted by PRC2 in hESCs.

DNA methylation hallmarks of tumorigenesis. We observed gain of methylation at previously unmethylated CGIs, ranging from 376 commonly hypermethylated CGIs in KIRC to 5,379 in T-ALL (see section 6.2.3 for the definition of common hyper CGIs per tumor type. Simultaneously, PMDs decrease genome-wide compared to matching healthy tissue (Figure 6.3.1). In line with previous studies, we also find that hypermethylated CGIs are highly enriched for regulation by PRC2 in stem cells, which includes nearly all tumor suppressor genes that become hypermethylated in primary tumors within our cohort (Figures 6.3.1 and C.2.2). As previously reported, cancer cell lines tend to exhibit more extreme methylation levels: In both solid and hematopoietic cancer cell lines, CGI methylation levels are frequently much higher than in primary tumors, often reaching methylation states close to 100% (Figure 6.3.1). Additionally, many cell lines exhibit extremely depleted PMD methylation levels in comparison to primary tumors. However, we also observe cell lines with PMD methylation levels comparable to healthy samples, a phenomenon previously only reported for primary acute leukemias (see chapter 5, Figure 6.3.1). Finally, some cell lines exhibit intermediate CGI and PMD methylation levels comparable with the distributions of primary tumors (Figure 6.3.1).

We inspected the same features across the larger 450k array-based cohort to generalize our findings. For this purpose, we generated a simple scoring method to examine PMDs and CGIs within both technologies. For this purpose, we defined each sample (WGBS and array) according to the median methylation status of solo-WCGW CpGs within a set of previously defined, pan-cancer PMDs [72] as well as for a tumor-specific subset of hypermethylated CGIs (separately defined for WGBS and array cancer types, see section 6.2.3, Figure 6.3.2). A combined investigation of both PMD and CGI methylation showed striking consistency across primary tumors of both cohorts (Figure 6.3.2). In general, primary tumors show only mild loss of PMD methylation alongside *de novo* methylation of CGIs, although PMD methylation levels can exhibit substantial variation. Nonetheless, cancer cell lines consistently reside in the periphery of primary tumor samples, suggesting one or multiple alternative epigenomes. In particular, CGI methylation levels are abnormally high, while PMD methylation levels exhibit a striking range from near complete to almost no methylation (Figure 6.3.2).

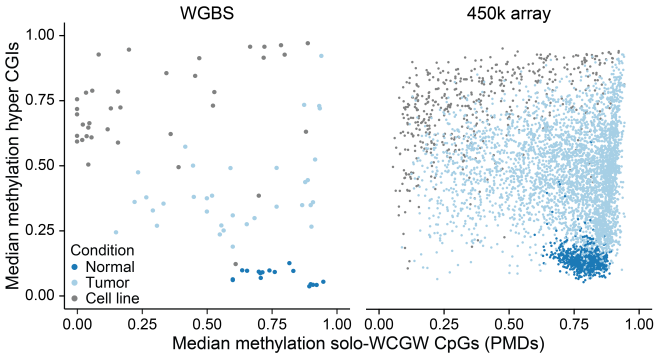

Figure 6.3.2: Comparison of the median methylation of solo-WCGW CpGs in common PMDs and hyper CGIs per WGBS and array sample.

In addition to our investigation of methylation levels, we wanted to understand how the identity of hypermethylated CGIs (which CGIs are targeted) relates to the tumor type. Here, we also aimed to address previous reports that cancer cell lines tend to methylate more CGIs than observed *in vivo*. We find that each sample hypermethylates an additional set of CGIs compared to the "common set" per tumor type, which is frequently larger in cancer cell lines than primary tumors (Figure 6.3.3). The additional CGIs are generally lower in methylation compared to the commonly hypermethylated CGIs, and the number of additional targets correlates with the methylation level of commonly hypermethylated CGIs. This observation seems consistent with the stochastic but continuous acquisition of *de novo* methylation over time and cell divisions.

Broadly, these trends can also be observed over the larger array cohort, with cell lines methylating a larger number of CGIs than their matched tumor type but with similar patterns (Figure C.2.3). However, we noticed a specific exception within the LGG cohort that could be explained by biased sampling between IDH mutant (IDH^{MT}) and wild type (IDH^{WT}) samples: While most of the LGG primary tumor samples carry an IDH mutation, all cell lines are classified as IDH^{WT} (Figure C.2.3). As IDH mutational status is known to affect the number and identity of CGI targets (see section 2.5.2 on the glioma CIMP subtype), we, therefore, split the LGG cohort into IDH^{MT} and IDH^{WT} and corrected our tumor to cell line comparison to only include IDH^{WT} samples (Figure C.2.3). After this correction, we found the same overall trends between LGG primary tumors and their cell lines as for other cancer types.

Despite the larger number of hypermethylated CGIs in individual cancer cell lines, these CGIs are generally also methylated elsewhere within our primary patient data: We observe a nearly complete overlap between hypermethylated CGIs observed across either primary tumors or cancer cell lines. This indicates that the overall set of CGIs that can potentially be methylated is restricted (Figure 6.3.4). Individual cancer types hypermethylate a limited and defined subset of these CGIs that characterize and distinguish them. We find that these cancer type-specific subsets are rapidly saturated across patient samples: every CGI that is commonly methylated within a given cancer type is observed after including a small number of individual patients (Figure 6.3.4). In contrast, cell lines tend to sample hypermethylated CGIs to a greater extent, which suggests that they are less constrained by their cell type of origin (Figure 6.3.4). These observations are overall in line with the CGI-centered analysis in the previous chapter, where we showed that

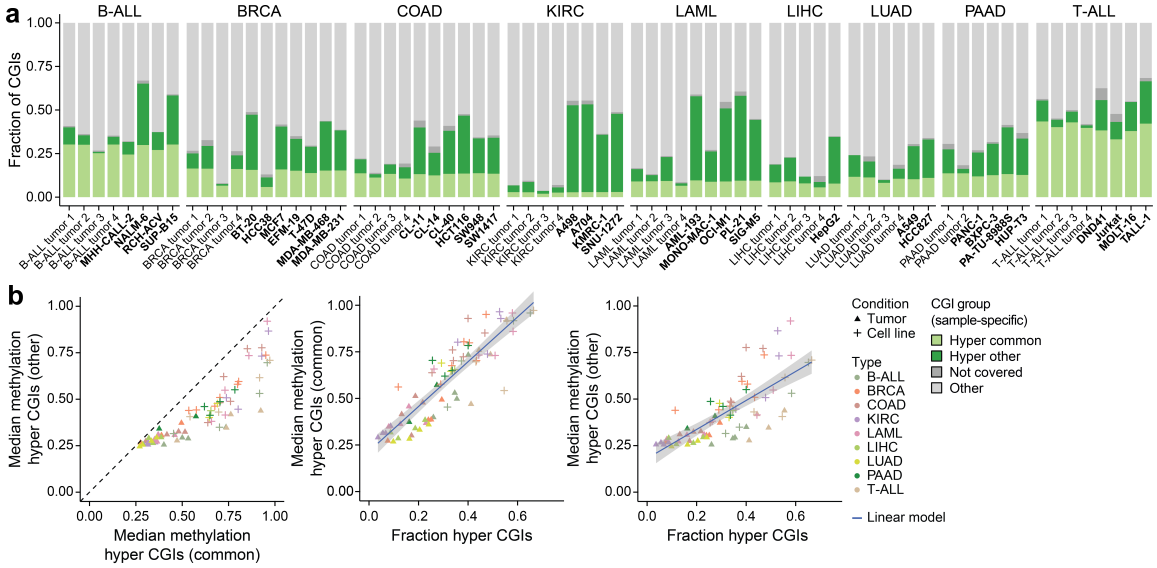

Figure 6.3.3: a) Hypermethylated CGIs per tumor and cell line sample (WGBS) separated by common targets across their respective tumor type and additional targets. Each sample hypermethylates other CGIs on top of the commonly hypermethylated CGIs per type, a number that is overall more extreme in cancer cell lines than primary tumors. b) Scatterplots comparing the methylation levels of commonly and additionally hypermethylated CGIs with the overall fraction of hypermethylated CGIs per WGBS sample.

specific sets of CGIs always remain unmethylated or consistently methylated across a pan-cancer cohort, while the remaining CGIs exhibit different levels of methylation across tumor types (see section 5.3.2).

6.3.2 Tumors and cell lines converge to distinct DNA methylation landscapes

Our initial observations suggested that the combination of PMD and CGI methylation levels could be used to broadly group cancer samples into "DNA methylation landscapes." We, therefore, aimed to develop a clustering approach that would reflect the patterns we observe in healthy, tumor, and cell line samples and could be applied to both cohorts (WGBS and 450k array). We would then use the separation into clusters to describe how consistently different DNA methylation states are observed across different types of cancers, tumor stages, metastasis, and adaptation to culture. First, we investigated the distribution (not identity) of DNA methylation levels at hypermethylated CGIs and PMDs (defined by solo-WCGW CpGs). For this purpose, we applied a clustering approach based on the ECDF for each feature and data type. For both cohorts (WGBS and 450k array) and features, we identified clusters of different ranges of DNA methylation distributions, reflecting low, intermediate, and high methylation. By CGI methylation, intermediately methylated clusters are comprised almost completely of primary tumor samples, while low and high CGI methylation are properties of normal and cell line samples, respectively (Figures 6.3.5 and C.2.4). By PMD methylation, normal as well as tumor samples are part of clusters with high or intermediate levels, while the clusters with low PMD methylation are mainly comprised of cell lines (Figures 6.3.5 and C.2.4). Clusters with extremely high PMD methylation contain not

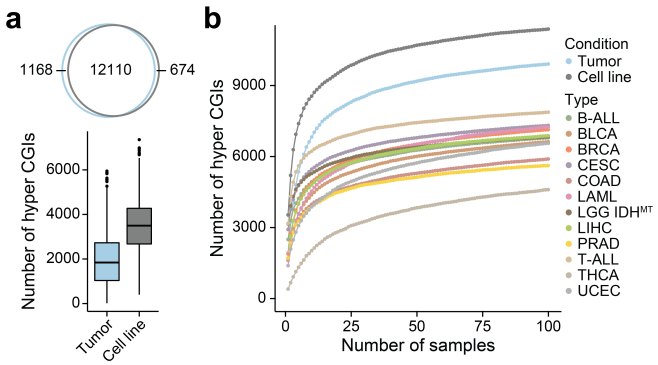

Figure 6.3.4: a) Top: Venn diagram showing the nearly complete overlap of CGIs called hypermethylated in any tumor or cell line sample (array cohort). Bottom: Number of CGIs hypermethylated per tumor and cell line sample. Cell lines hypermethylate a larger number of CGIs in comparison to primary tumors. b) Saturation analysis of the number of hyper CGIs cumulatively observed from random samples of 100 patient tumors per type, including across types or all cell lines (see section 6.2.3). The majority of CGIs hypermethylated per group is observed after only a few samples, suggesting that each cancer type is constrained in the overall number of possible targets, which also applies to the entirety of tumor and cell line samples when randomly sampling across types.

only healthy samples but also acute leukemia patients, which have been previously reported to exhibit essentially no to minimal genome-wide loss of methylation (see chapter 5).

Methylation levels of CGIs and PMDs consistently appeared to distinguish normal, primary, and cell line samples. Therefore, we aimed to develop a combined clustering approach that accounts for both features (CGIs and PMDs) to identify major forms of DNA methylation landscapes in cancer. For this purpose, we used a consensus clustering-based method based on the high-resolution WGBS cohort. As features, we considered the median methylation of hyper CGIs as well as of solo-CpGs within common PMDs per sample (see sections 5.2.4 and 6.2.4). This approach resulted in five basic DNA methylation states that generally separated samples according to their status as somatic, primary, or cell line (Figures C.2.5 and 6.3.6). As expected, healthy samples were primarily assigned to a "somatic landscape" (PMD^{high}, CGI^{low}), while primary cancers were generally found to be intermediately methylated, either in a PMD^{high}, CGI^{int} or PMD^{int}, CGI^{int} state (Figures 6.3.6 and C.2.6). The remaining two states are associated with properties observed across cancer cell lines: a state of "extreme hypermethylation" (PMD^{high}, CGI^{high}) that also includes primary ALL samples and an "inverse bimodal" (PMD^{low}, CGI^{high}) state that is enriched almost entirely for cell lines (Figures 6.3.6 and C.2.6). Only a minority of cancer cell lines are associated with intermediate states. Although the clustering was established based on the WGBS cohort, we found that we could also assign the larger 450k array cohort to these states using a *k*-nearest neighbor classification. The resulting assignments reflected the overall distribution of healthy, tumor, and cell line samples across the five states as observed based on the WGBS data (Figures 6.3.6, and C.2.6 and C.2.7).

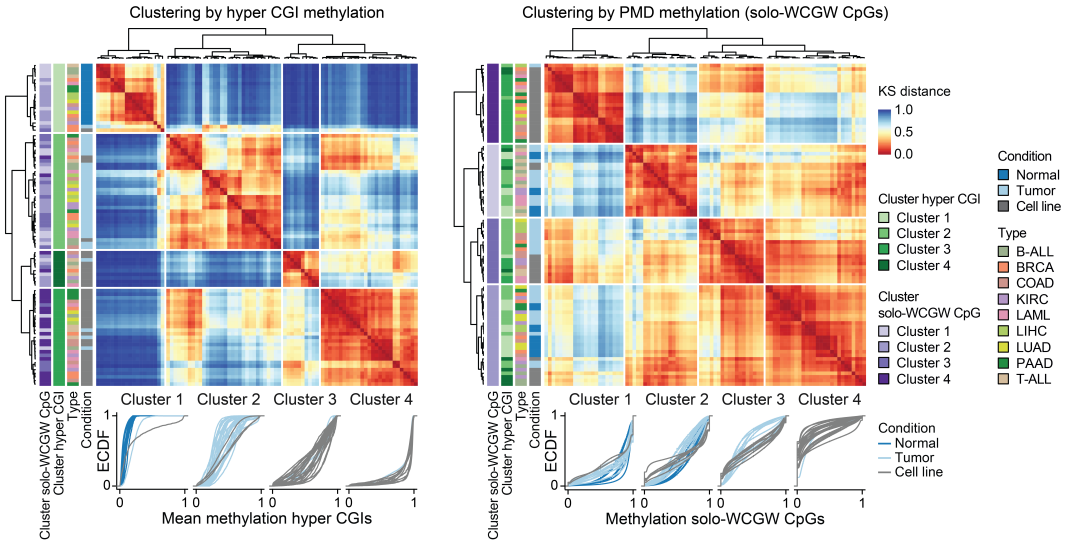

Figure 6.3.5: Hierarchical clustering of WGBS samples based on the ECDF of hyper CGIs (left) or solo-WCGW CpGs in common PMDs (right).

6.3.3 Allelic heterogeneity underlies intermediate DNA methylation in tumors

The intermediate methylation levels observed in primary tumors could be reflective of different phenomena: a cellular heterogeneity model where intermediate methylation is explained by a combination of tumor purity and epigenetic heterogeneity between cells (mixing of somatic and cancer cell line-associated states) or an allelic heterogeneity model where independent DNA molecules primarily exhibit a similar degree of intermediate methylation (intrinsic to tumor cells, Figure 6.3.7). To distinguish between these two scenarios, we used the single-read methylation information that reflects single molecules within the bulk population profiled by WGBS. Read methylation distributions across tumor type-specific hypermethylated CGIs demonstrated strong enrichment of intermediate-methylated molecules within primary cancer samples of different types (Figures 6.3.7, 6.3.8 and C.2.8). Somatic cells and cell lines also exhibit read-level methylation distributions consistent with their assignments, with the majority of reads either being fully unmethylated or fully methylated, respectively (Figures 6.3.7, 6.3.8 and C.2.8).

To further strengthen the observation that intermediate CGI methylation is an intrinsic form of regulation within primary tumors and not the consequence of cellular heterogeneity, we measured the methylation variation across independent molecules (sequencing reads) using DNA methylation entropy, which examines the configuration of methylated and unmethylated CpGs on the same read in a given sequence context (see chapter 4). Again, methylation entropy levels are highly consistent with our global landscape assignments, with the somatic-enriched CGI^{low} state showing low entropy and methylation, primary-enriched CGI^{int} states showing high entropy and cell line-enriched CGI^{high} states returning to a lower entropy but high methylation state (Figure 6.3.9).

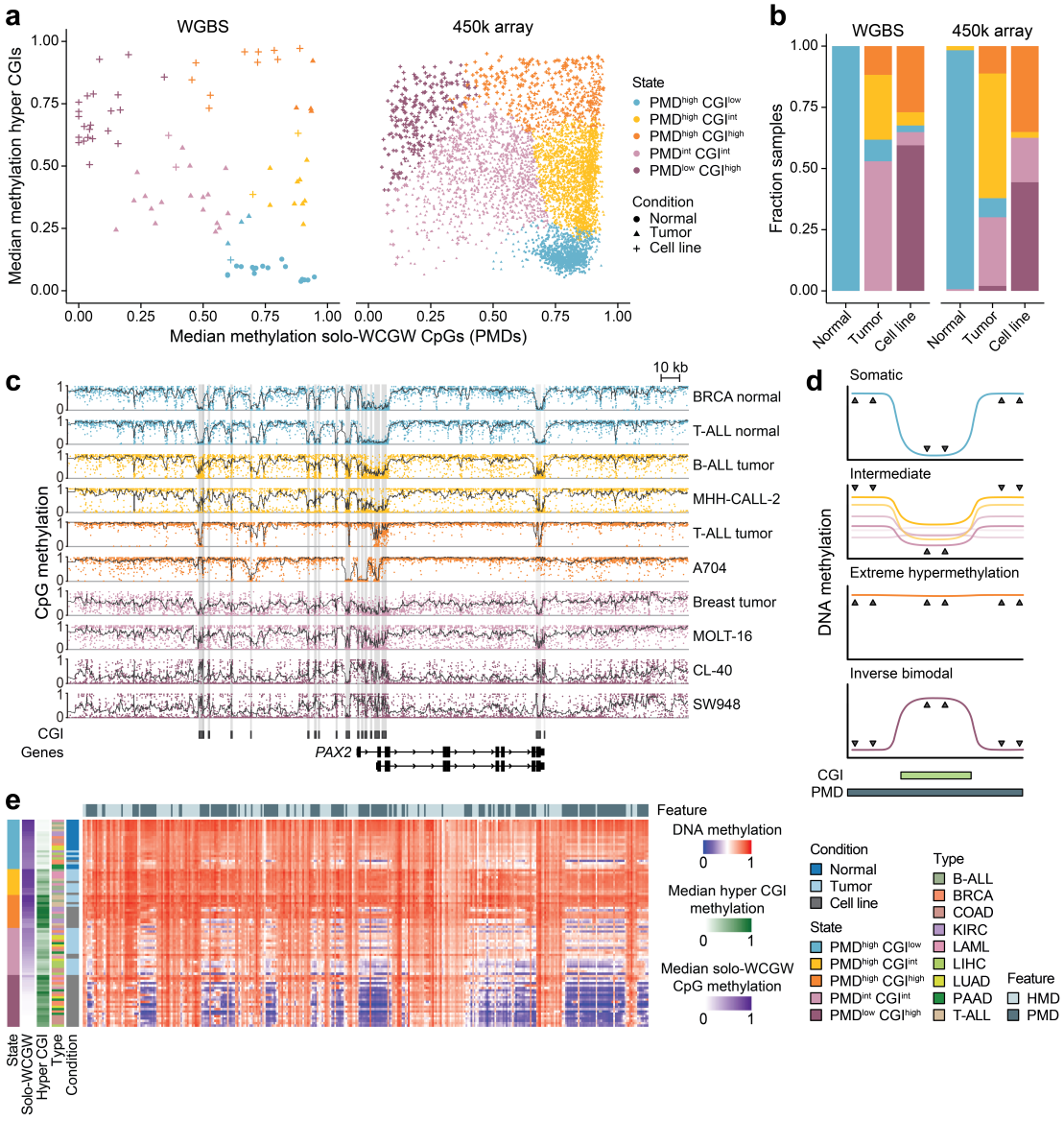

Figure 6.3.6: a) Relation of the median methylation of solo-WCGW CpGs in common PMDs and hyper CGIs per WGBS and array sample colored by DNA methylation state defined by consensus clustering. b) Distribution of DNA methylation states across healthy, tumor, and cell line samples for the WGBS and array cohort. c) Genome browser track of the *PAX2* locus showing exemplary WGBS samples for all five methylation states identified by consensus clustering. d) Schematic of the four main DNA methylation landscapes that can be observed across somatic, tumor, and cell line samples. e) Chromosome-scale heatmap of the average methylation of the WGBS cohort along chromosome 16p (100 kb tiles).

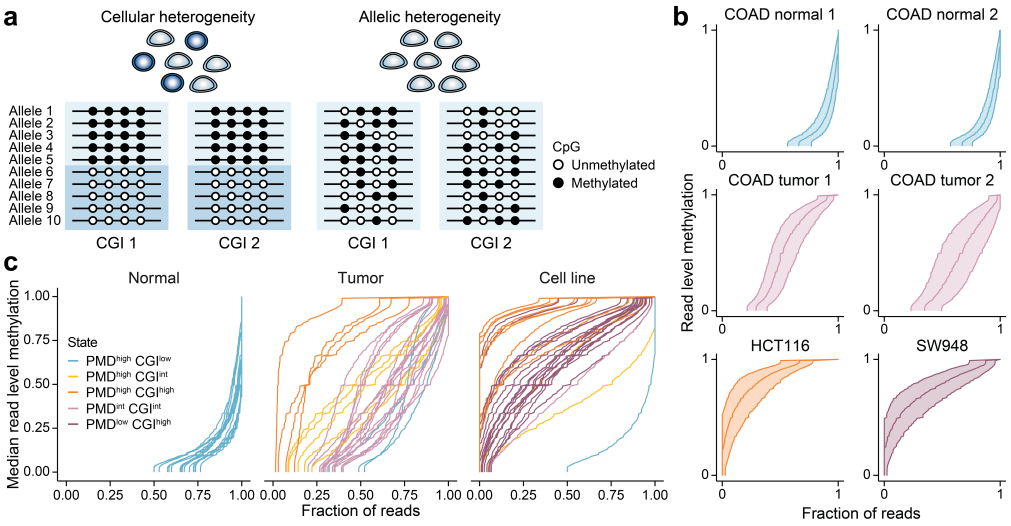

Figure 6.3.7: a) Schematic of cellular versus allelic forms of DNA methylation heterogeneity as they could contribute to intermediate DNA methylation observed in primary tumors. Cellular heterogeneity reflects the mixture (or contamination) of cells with distinct methylation patterns at individual CGIs, while allelic heterogeneity is defined by stochastic methylation that spans the majority of molecules within the population. b) Cumulative read-level methylation distributions across hyper CGIs within healthy colon tissue, primary tumors, and cell lines. Lines reflect the median, 25%, and 75% quantile across CGIs. c) Cumulative read-level methylation distributions across hyper CGIs for all healthy, tumor, and cell line samples (lines reflect the median).

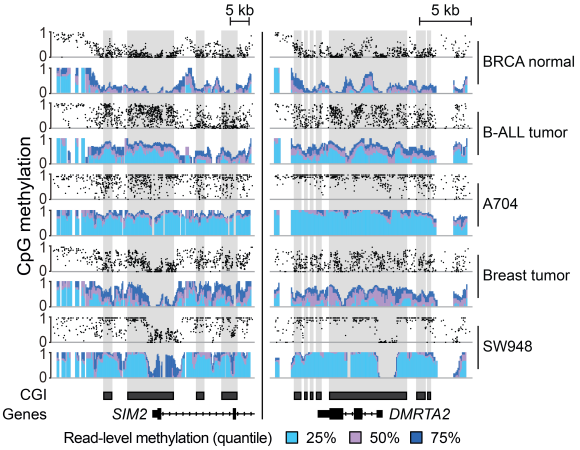

Figure 6.3.8: Genome browser tracks of read-level methylation across *SIM2* and *DMRTA2* loci for exemplary WGBS samples associated with different states. In addition to the population average CpG methylation rates (black), the methylation distributions of the underlying single reads are shown, confirming the substantial enrichment of intermediately methylated DNA molecules within primary tumor samples compared to either healthy samples or cell lines.

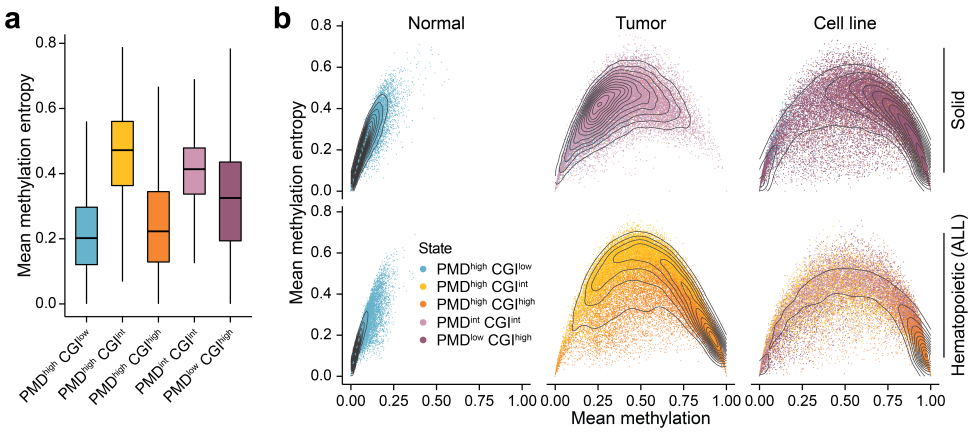

Figure 6.3.9: a) Boxplots of the average methylation entropy across hyper CGIs for samples associated with the five different DNA methylation states. The entropy is highest for samples associated with intermediate CGI methylation, which are primarily primary tumor samples. b) Scatterplots showing the relationship between mean CpG methylation and mean methylation entropy at hypermethylated CGIs for healthy tissues, corresponding cancer samples, and cell lines.

6.3.4 The effect of tumor purity on DNA methylation levels

All samples selected for this study were reported to have a tumor purity of $\geq$ 80% based on pathology (WGBS) or consensus purity prediction (450k array). Nonetheless, we wanted to address the effects of tumor purity on the assignment of individual primary samples to a given "DNA methylation landscape" in comparison to purer cancer cell lines. For this purpose, we made use of high-purity T-ALL and matching T cell samples and performed an *in silico* mixing experiment where sequencing reads from tumor and healthy cells were combined at different concentrations reflecting different artificial purities. Due to the large range of CGI methylation levels observed in T-ALL (see chapter 5), we were able to select one T-ALL sample with intermediate and one with extremely high, cancer cell line-like CGI methylation levels (Figure 6.3.10). Notably, even at purities far lower than reported for our tumor samples, our entropy-based results showed high consistency between our primary WGBS cohorts and the intermediate methylation spike-in. In contrast, spike-in with an extremely hypermethylated sample showed much lower entropy-based values than those observed for our primary data cohort. This suggests that despite potential contamination, intermediate methylation observed in our tumor samples does not stem from a cellular mixture of somatic and cancer cell line-like cells (Figure 6.3.10).

An analysis comparable to read level-based metrics cannot be performed for 450k arrays due to the nature of the technology (see section 2.6). Therefore, we employed an aggressive strategy to correct methylation levels according to the reported purity for both hyper CGIs and PMDs. As somatic cell contamination would cause the underestimation of hyper CGI and overestimation of solo-WCGW CpG (PMD) methylation levels, we corrected the methylation measurements for primary solid tumor samples by assuming that healthy cells would be completely unmethylated at CGIs and fully methylated elsewhere (Figure 6.2.3). Although very likely to overestimate the effects of cellular contamination, these corrections allowed us to evaluate how much closer pri-

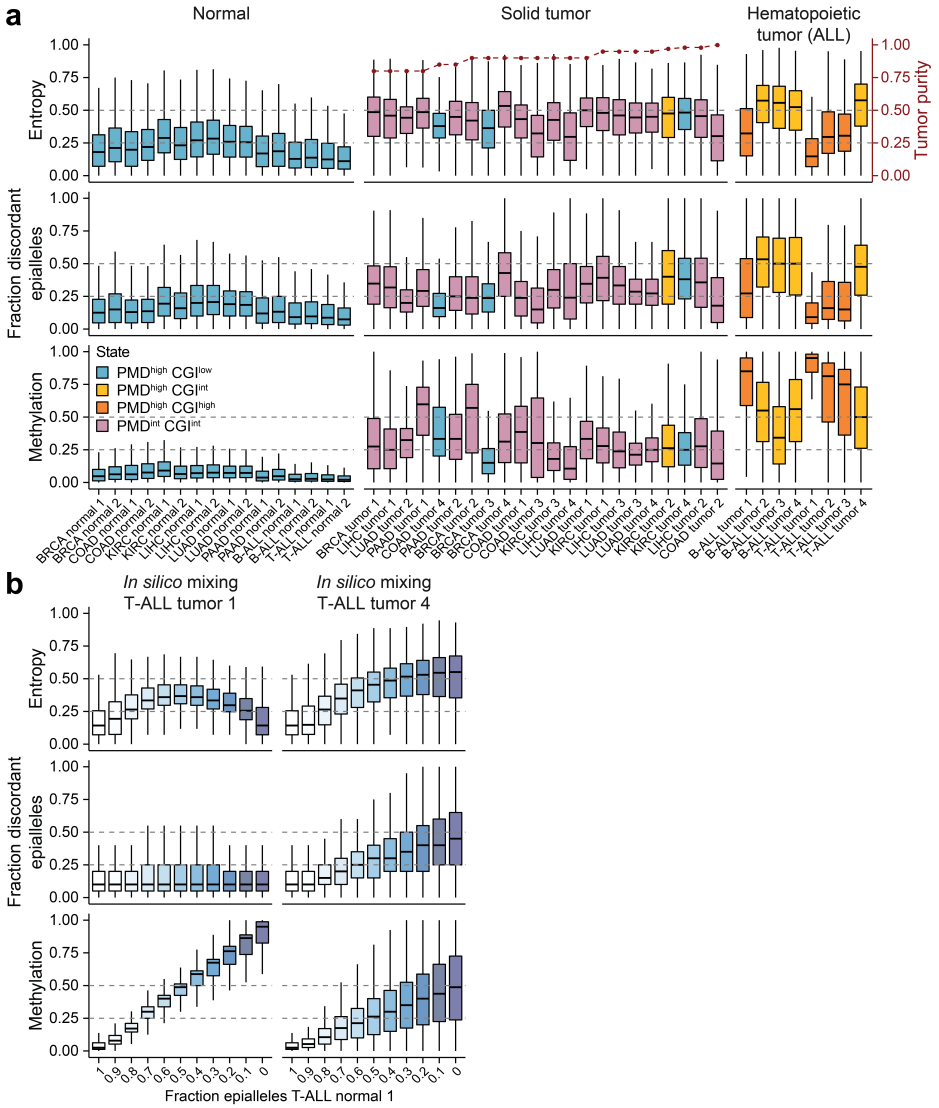

Figure 6.3.10: Boxplots showing the methylation entropy (top), the fraction of discordant reads (middle), and methylation (bottom) per 4-mer in hyper CGIs) for a) healthy and tumor samples as well as b) an *in silico* mixing experiment mimicking the effect of different purity levels on the three measurements (see section 6.2.5).

mary tumors could get to levels observed in cell lines (primarily the inverse bimodal PMD^{low}, CGI^{high} landscape). After correction, we reassigned our WGBS and array cohort samples using the revised methylation measurements for hyper CGIs and PMDs (Figures 6.3.11 and C.2.9). Only a minority of samples changed their state compared to their original assignment, and these were not always moved into regimes consistent with cancer cell line signatures (20% of array tumor samples were assigned to a different state, 7% were assigned to a PMD^{low} CGI^{high}, and 7% were assigned to a PMD^{high} CGI^{high} landscape). Moreover, samples that changed their assignment were generally characterized as being on the border between states and were assigned with lower probability pre- and post-correction (Figure C.2.9). Although the assignment of solid primary tumors to cancer cell line-enriched states increases slightly, the distribution of primary

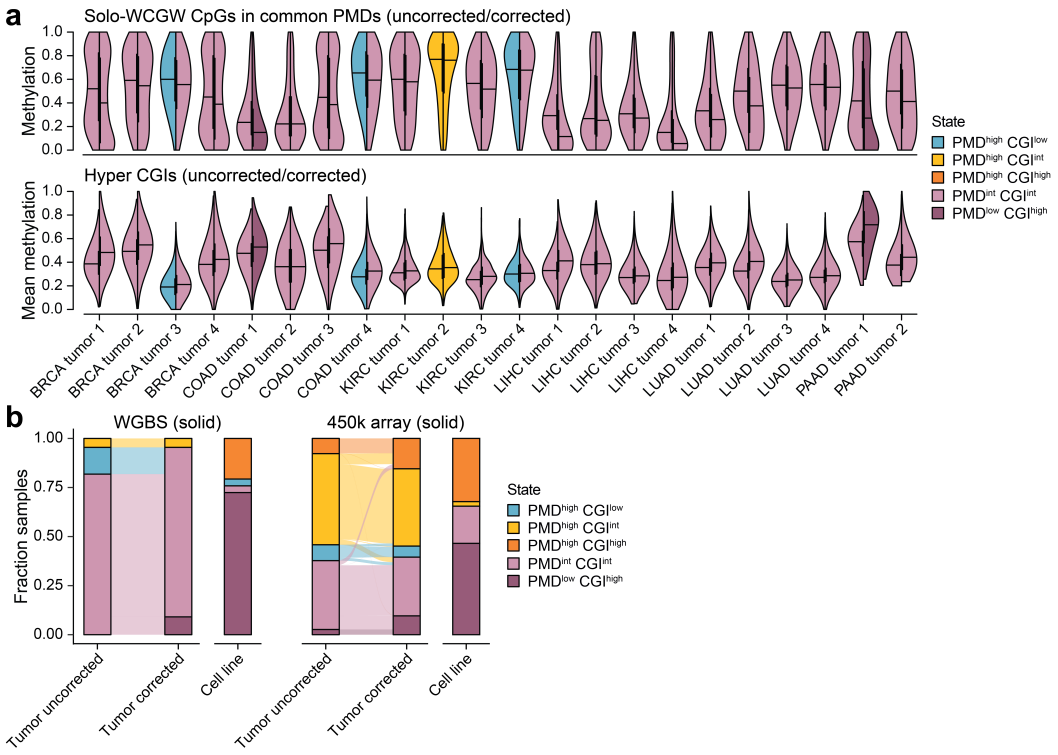

Figure 6.3.11: a) Distribution of solo-WCGW CpG and hyper CGI methylation before and after stringent tumor purity correction demonstrates minimal changes to the overall landscape (solid tumors, WGBS cohort). Violin plots are colored by the assigned DNA methylation state pre- and post-correction (left and right side of each violin plot, respectively). b) Distribution of DNA methylation states across solid tumor samples before and after tumor purity correction compared to cancer cell lines (WGBS and array cohort).

tumors across states still differs greatly in comparison to cancer cell lines 6.3.11. Together, these results highlight that stochastic, intermediate methylation is intrinsic to tumor cells and that the observed differences in DNA methylation landscapes between tumors and cell lines are not mainly an effect of reduced tumor purity.

6.3.5 Intermediate DNA methylation levels across single tumor cells

To consolidate our observations of intra-tumor DNA methylation heterogeneity, we examined a previously published cohort of 10 colorectal cancer patients where the methylation status of single cells was profiled at different positions along the tumor. Of these 10 tumors, a range of six to 142 cells were profiled from up to six distinct positions (ranging from one to 50 cells per tumor site). In order to mimic the detection of hypermethylated CGIs in bulk tumors, we generated “pseudo bulk” samples per patient using the average methylation across healthy or tumor cells per CGI. These pseudo bulk samples group closely to actual bulk methylomes from colon cancer patients (WGBS), verifying this approach (Figure 6.3.12). We identified between 702 and 1,795 hypermethylated CGIs per patient, which showed the same enrichment for PRC2 as found in bulk cohorts (Figure 6.3.13). When examining CGI and PMD methylations levels

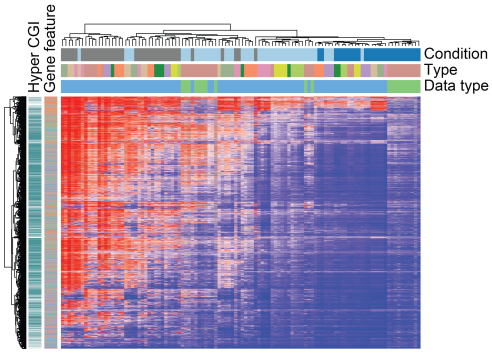

Figure 6.3.12: Heatmap showing the clustering of WGBS and single-cell WGBS (pseudo bulk) patients based on the union of hypermethylated CGIs defined for each colon cancer patient (single-cell WGBS cohort). Pseudo bulk samples group closely to the colon cancer patients profiled with WGBS.

of single cells from all 10 patients, we observed a similar departure from normal as seen in the pseudo bulk samples, including *de novo* methylation of CGIs and loss of PMD methylation that resulted in intermediate methylation levels (Figure 6.3.13). For the majority of patient tumors, single tumor cell methylomes group closely together by these metrics with minimal variation between one another (Figures 6.3.13 and 6.3.14). When inspecting the single-cell methylation levels across hyper CGIs defined on the pseudo-bulk, we observed that single cells consistently methylate targets to similar degrees with minor deviations, which indicates shared regulation at these sites across tumor cells (Figures 6.3.14 and C.2.10).

In order to measure the consistency of DNA methylation levels across the spatial tumor organization, we calculated the within-sampling site and across-sampling site distances of hyper CGI and PMD methylation for all 10 patients (Figure 6.3.15). Although this test confirms the consistency of methylation signatures across most cells, regardless of sampling site, we observed instances where different tumor regions show deviating hyper CGI or PMD methylation patterns (Figure 6.3.15). In patients 4 and 11, clonal heterogeneity at hyper CGIs can be detected, which nonetheless affects the minority of sampled cells within each tumor (9 and 8% of cells with a difference > 2 between within and across sampling site distance for the two patients, respectively). However, these clonal patterns remain in an intermediate methylation regime, reflecting the expected primary solid tumor methylation landscape (Figure C.2.11). Additionally, clonal heterogeneity at hyper CGIs does not seem to be coupled with heterogeneity across PMDs. Here, patients 1, 13, and 14 exhibit mostly sampling site-specific differences, where PMD methylation levels drop within a fraction of tumor cells approaching a PMD^{low} state (Figures 6.3.15, 6.3.16 and C.2.12). Notably, in these cases, the consistent genome-wide loss of PMD methylation across regions appears to be more reflective of a cellular adaptation rather than a gradual stochastic loss linked to cell divisions. For example, when we compare the global CpG methylation levels of PMD^{low} tumor fractions to other sampling sites, we see that the genome overall is comparatively hypomethylated (Figure 6.3.16). However, even within these PMD^{low} cell subpopulations, CGI methylation levels remain stable and resemble the remaining cells of the tumor (Figure 6.3.16). Based on most of our data, the maintenance of intermediate methylation, therefore, appears to be a primary event during tumorigenesis.

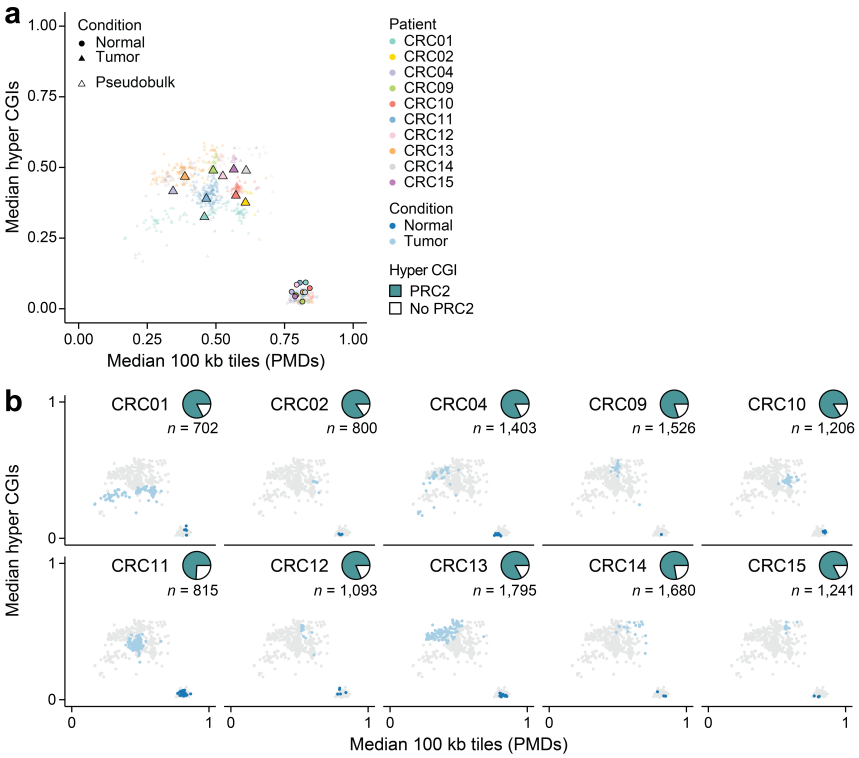

Figure 6.3.13: a) Comparison of the median methylation of 100 kb tiles in common PMDs and hyper CGIs per healthy and tumor cell measured for 10 colon cancer patients. Triangles mark the corresponding pseudo bulk measurements generated *in silico*. b) The same scatterplot as in a) colored specifically by cells of each patient. Similar to the WGBS cohort, hyper CGIs are frequently targeted by PRC2 in hESCs (pie charts).

6.3.6 DNA methylation across tumor stages and metastases *in vivo*

Our analyses highlighted the striking stability of an intermediately methylated epigenome within most primary tumors, a signature that appears to be commonly lost as cells adapt to culture. To proxy the effects of cell division and genetic bottlenecks, we next investigated the stability of intermediate DNA methylation levels across clinical stages and metastasis. For this purpose, we made use of the 450k array cohort, which includes samples from different tumor stages as well as metastases. When considering the methylation status of variable CGIs, we find healthy, tumor, metastasis, and cell line samples largely group closely according to their tissue-of-origin signature (Figures 6.3.17 and C.2.13, samples visualized using Uniform Manifold Approximation and Projection). Within these subclusters, samples are distributed with healthy and cell line samples at the endpoints, which primarily reflects the overall number of hypermethylated CGIs per sample (Figures 6.3.17 and C.2.13). This is consistent with our findings in section 6.3.1 that different tumor types have the potential to methylate a distinct subset of CGIs from the full set that can be methylated in cancer with varying levels and additional targets from primary samples to cell lines. When inspecting hyper CGI and PMD methylation split by clinical stage and metastasis, we observed that the overall levels are mostly stable across tumor progression and exhibit consistent differences compared to healthy somatic cells but also cancer cell lines (Figures

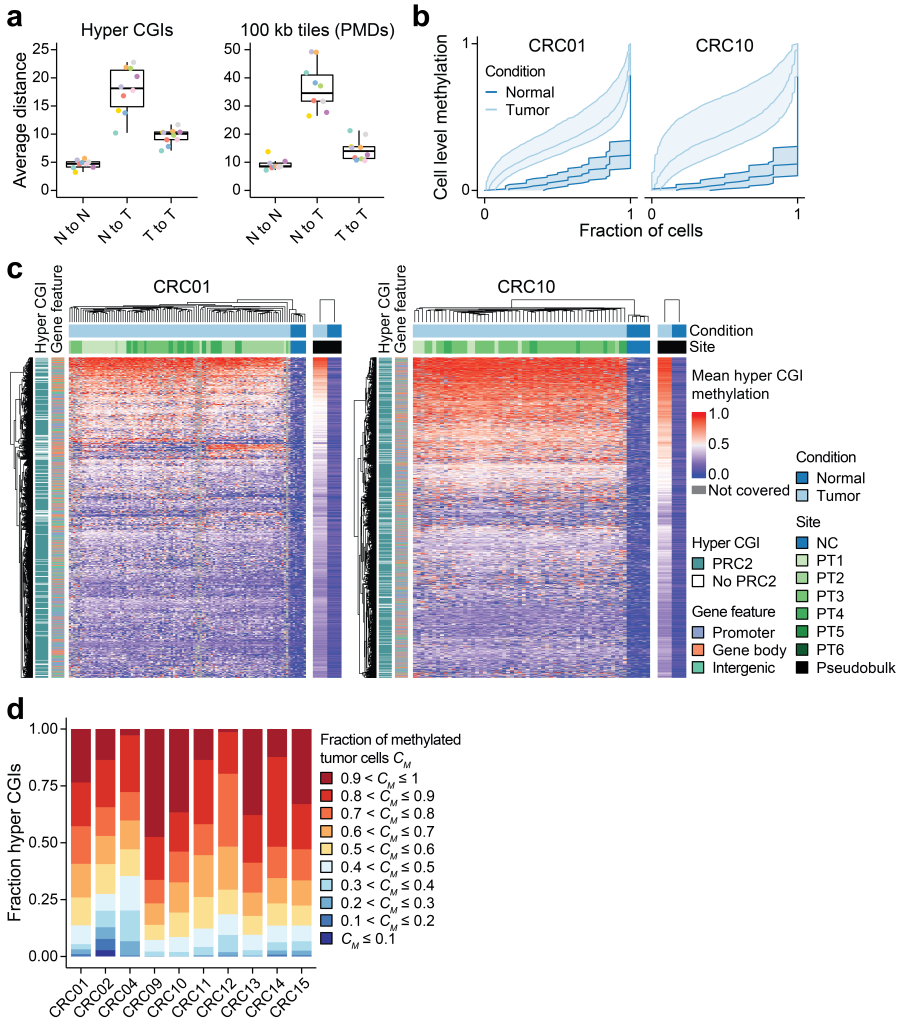

Figure 6.3.14: a) Pairwise Euclidean distance between healthy and tumor cells per patient. Within patients, the largest difference in methylation pattern exists between normal and tumor cells, whereas healthy or tumor cells show comparatively minimal differences between cells of the same type. b) Cumulative single-cell-level methylation distributions for hyper CGIs within two exemplary patients. Lines reflect the median, 25%, and 75% quantile across CGIs. c) Heatmaps of hyper CGI methylation across healthy and tumor cells for two exemplary patients (1 and 10). The methylation status of CGIs within single cells is largely similar to one another and maintained at intermediate levels. d) For all 10 patients, the fraction of cells for which hypermethylated CGIs are called methylated (restricted to cells for which the CGI is covered).

6.3.18 and C.2.14). These observations do not seem to support a model in which CGI hyper- and PMD hypomethylation accumulate as a function of increasing numbers of cell divisions that can be associated with later tumor stages and metastases.

To further deepen our understanding of DNA methylation distributions in primary tumors and metastases, we re-examined the eight colorectal cancer patients for which single-cell measurements included one or several distinct well-sampled metastases. Consistent with our primary

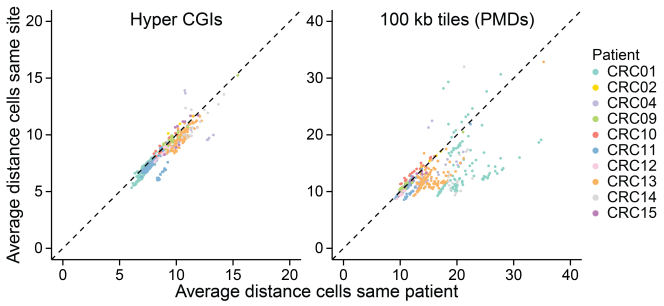

Figure 6.3.15: Pairwise Euclidean distance between tumor cells within or across independent sampling sites for each patient calculated based on hyper CGI (left) and PMD (right) methylation. CGIs within and across sampling sites are generally homogeneously methylated across sites, while some patients exhibit heterogeneity in PMD methylation that corresponds to an adaptive loss of methylation within specific subclones.

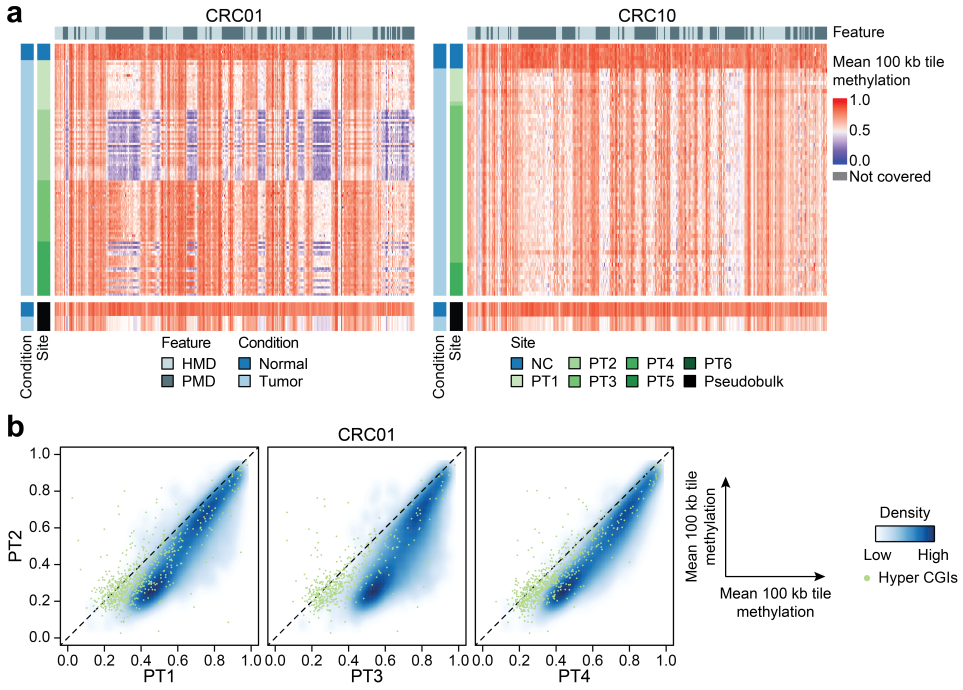

Figure 6.3.16: a) Chromosome-scale heatmaps of the average methylation across cells of two exemplary patients (patients 1 and 10) along chromosome 16p (100 kb tiles). For patient 1, a subset of cells from a specific sampling site exhibits a sharp drop in PMD methylation levels. b) 100 kb tile-wise density plot comparing the methylation of different sampling sites in patient 1. The loss of methylation across cells from the PT2 site seems to affect the whole genome, with the exception of most hyper CGIs that remain consistently intermediate.

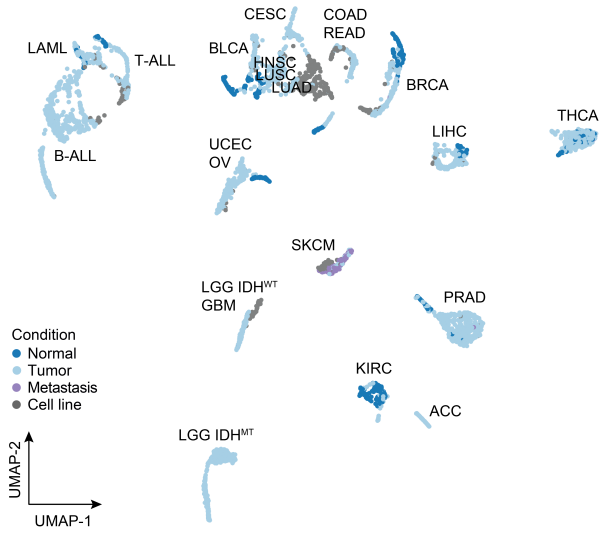

Figure 6.3.17: Uniform manifold approximation and projection (UMAP) plot of 705 healthy, 3,681 tumor, 137 metastasis, and 507 cell line samples based on the binary methylation status of 23,345 commonly covered CGIs (≤ 0.2 unmethylated, > 0.2 methylated, array cohort).

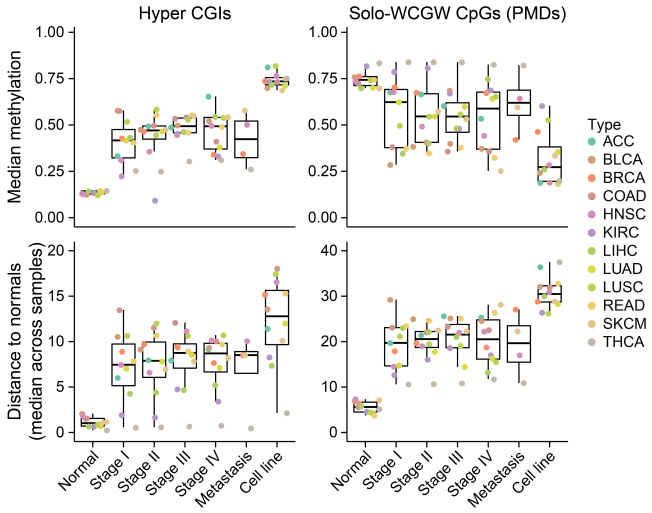

Figure 6.3.18: Top: Median hyper CGI and solo-WCGW CpG methylation across samples per type and separated by tumor stage or condition. Bottom: Distance of tumor, metastasis, and cell line to healthy samples based on hyper CGI and solo-WCGW CpG methylation.

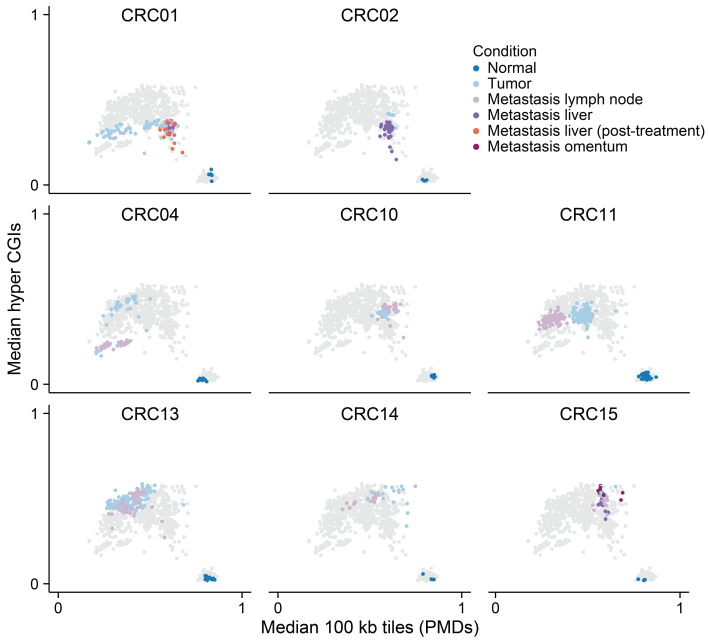

Figure 6.3.19: Relation of the median methylation of 100 kb tiles in common PMDs and hyper CGIs across single-cells as in Figure 6.3.13 including profiled metastases.

tumor analysis (Figure 6.3.13), we find that metastatic cells reside within the same intermediate state as the primary tumor population average (Figure 6.3.19). Moreover, when examining hyper CGI and PMD methylation independently, we mostly see minimal deviation between cells at either the PMD or CGI level, which also extends across multiple independent metastases of the same patient from distinct tissues (Figures 6.3.20 and C.2.15). Across these eight patients, we find a single example where a metastasis displays a CGI methylation pattern consistent with cellular heterogeneity: a lymph node metastasis of patient 4 closely matches a subclonal pattern found within the primary tumor (Figures 6.3.20 and C.2.15). More generally, metastatic adaptations generally reflect changes that can also be found in the corresponding primary tumors, including a loss of PMD methylation that does not affect intermediate CGI methylation levels (see patient 11, Figures 6.3.20 and C.2.15). Taken together, our results confirm the stability of intermediately methylated DNA methylation landscapes over the duration of tumorigenesis, including across aggressive cellular bottlenecks such as metastases. Our investigation of single-cell methylation data confirms that predominantly similar methylation levels are maintained across many cells and large spans of tumor development with only minimal cellular heterogeneity. A key exception appears to be the loss of methylation in PMDs in a subset of cells of some tumors, which does not seem to reflect the accumulation of stochastic, gradual loss over time but instead is reminiscent of a global adaptation as it does not extend to intermediately methylated CGIs.

6.3.7 Associating cell line methylation landscape with additional features

In contrast to primary tumors, cancer cell lines frequently present with one of two alternative DNA methylation landscapes that are rarely found *in vivo*. We aimed to identify factors associated

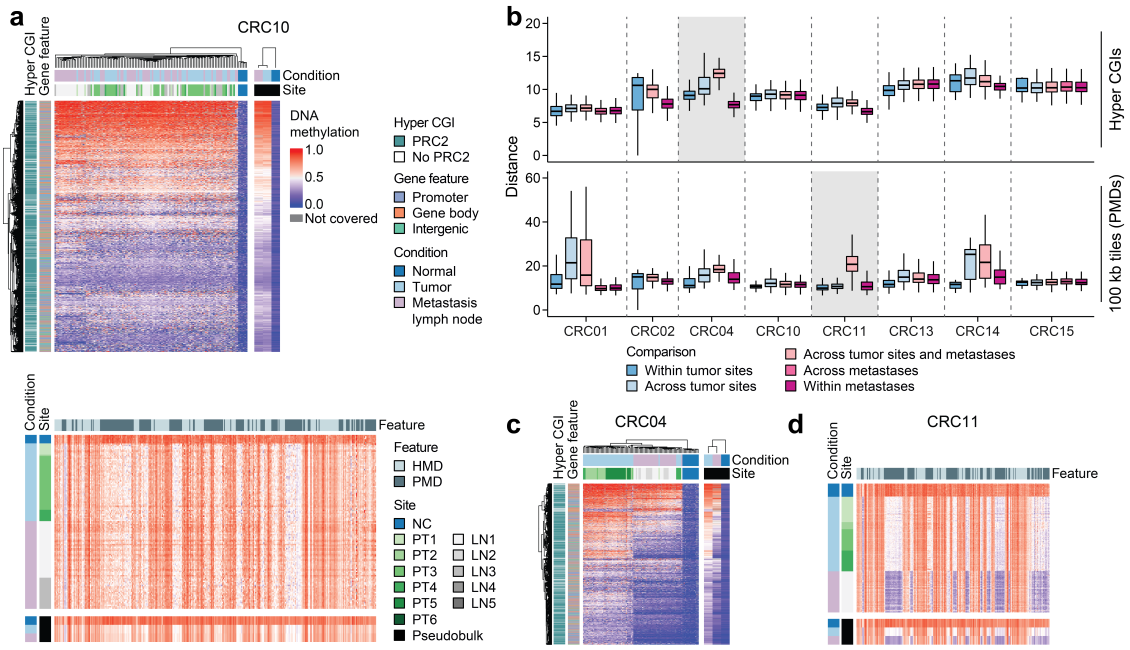

Figure 6.3.20: a) For patient 10, heatmaps of hyper CGI and chromosome-scale methylation across primary tumor cells and cells of a lymph node metastasis. Different sampling sites are indicated. b) Pairwise Euclidean distances between primary tumor cells and their associated metastases for all eight colon cancer patients for which sufficient metastatic cells were collected. Overall, single tumor cells display remarkably similar methylation profiles both within and across sampling sites, and these trends also hold when multiple independent metastases are considered (see patients 1 and 15). Only a single sufficiently sampled tumor (patient 4, see c) demonstrates a degree of cellular heterogeneity that gives rise to a metastatic subclone. Pairwise distances for PMD methylation (bottom) highlight three instances where global methylation levels drop, including two contributing to metastases (see an example in d).

with either of these landscapes to understand what defines the epigenetic trajectory of a cancer cell line. For this purpose, we focused on the larger 450k array cohort and cell lines that were assigned to their respective state with a high probability (≥ 0.7, Figure 6.3.21). We first investigated the enrichment of different cancer types in DNA methylation landscapes and found that many types seem biased towards one of the two main states (Figure 6.3.21). Specifically, T-ALL and KIRC cell lines are significantly associated with the extreme hypermethylation state, while GBM and LIHC are significantly associated with the inverse bimodal state (two-sided Fisher's exact test, Figure 6.3.21, Table C.2.1). When comparing the fraction of cell lines that are extremely hypermethylated with the fraction of primary tumors in a PMD^{high} state per cancer type, we observe a positive correlation with the exception of LAML, brain and thyroid tumors (Figure 6.3.21). This indicates that cell lines derived from tumor types with relatively high PMD methylation levels tend to exhibit high genome-wide methylation levels and less frequently transition to an inverse bimodal state. Different culture conditions do not seem to influence the separation of cell lines into different states but are most strikingly associated with tumor type (leukemia vs solid tumor, Figure C.2.16).

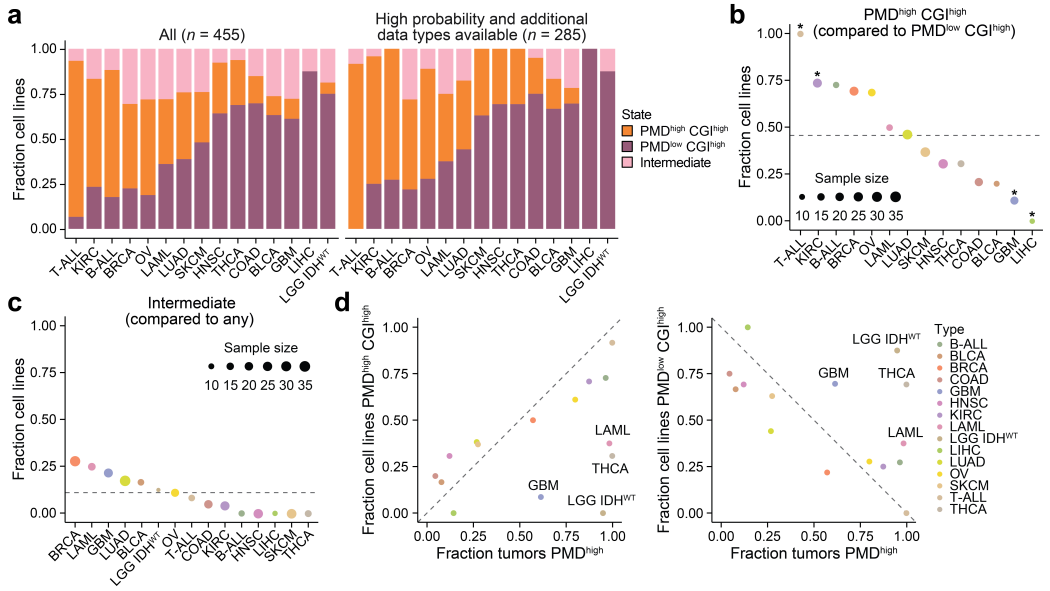

Figure 6.3.21: a) Fraction of cell lines in different DNA methylation states considering all samples and samples with high k-NN probability (≥ 0.7). Both panels are limited to the tumor types with at least eight cell lines after selection by k-NN probability. b) Fraction of cell lines in the extreme hypermethylation compared to the inverse bimodal state as the two dominant *in vitro* DNA methylation landscapes (excluding intermediate cell lines). T-ALL and KIRC are significantly associated with the extreme hypermethylation state, while GBM and LIHC are significantly associated with the inverse bimodal state (two-sided Fisher's exact test). The horizontal line indicates the overall fraction of cell lines in an extreme hypermethylation state across the cohort when only considering PMD^{high} CGI^{high} and PMD^{low} CGI^{high} cell lines. c) Fraction of cell lines in an intermediate state compared to the extreme hypermethylation and inverse bimodal state. No tumor type is significantly associated with the intermediate state compared to the two other states, which could also be impacted by the overall low number of intermediate cell lines (two-sided Fisher's exact test). The horizontal line indicates the overall fraction of intermediate cell lines across the cohort. d) Scatterplot comparing the fraction of tumors in a PMD^{high} state with the fraction of cell lines in a PMD^{high} CGI^{high} (left) and PMD^{low} CGI^{high} (right) state per tumor type (considering all cell line samples in comparison to b). Except for LAML, brain, and thyroid tumors, the fraction of tumor samples in a PMD^{high} state correlates well with the fraction of cell lines in an extreme hypermethylation state (Pearson's $r = 0.49$ with and 0.94 without the exceptional types) and anti-correlates with the fraction of cell lines in an inverse bimodal state (Pearson's $r =$ -0.52 with and -0.88 without the exceptional types).

We hypothesized that commonly mutated genes might be linked to the establishment of different DNA methylation landscapes in cancer cell lines. Therefore, we compared the frequency of mutations in known cancer drivers (the Cancer Gene Census defined by COSMIC) and epigenetic regulators across cell lines associated with different methylation states (data from Iorio et al. [279], Figures 6.3.22 and C.2.18). We observed that some genes are more frequently mutated in either the extreme hypermethylation or the inverse bimodal state. This includes the PRC2 component EZH2 as well as H3K4me3 demethylases that are more frequently mutated in cell lines with extreme hypermethylation, while nuclear receptor coactivators seem to be rather targeted in inverse bimodal cell lines. However, none of these are significantly associated with landscape. Instead, many are linked to the underlying tumor types, which include known driver genes such as *TP53*, *APC* and *KRAS* (two-sided Fisher's exact test, 6.3.22). More generally, we found that extremely hypermethylated and inverse bimodal lines are prone to an increase in genetic mutations per line compared to the few intermediate cell lines, potentially indicating a tendency towards higher mutational load (Figure C.2.17). However, these trends were subtle. More strikingly, we investigated our WGBS cohort for changes in copy number variation and found that cell lines in an inverse bimodal state show greater variance in chromosomal amplifications and deletions compared to either primary tumors or other cell lines (WGBS cohort, Figure 6.3.23). DNA methylation plays a role in genome stability, suggesting that the drastic decrease in PMD methylation levels might expose the genome to chromosomal aberrations.

Lastly, we used the extensive drug screens of cell lines provided by the GDSC to identify drugs that might have a differential effect on cell lines with specific methylation states. We identified 47 significant associations of methylation state with drug response (measured by IC50) using logistic regressions (Figures 6.3.24 and C.2.19). Of these, most drugs displayed an increased IC50 (and therefore relative resistance) related to the inverse bimodal state. Only for three drugs an increase in resistance was associated with the extreme hypermethylation state. However, in line with our previous findings, these effects seem to be mainly linked to the underlying tumor types that are associated with the different states: no significant associations between DNA methylation state and drug response were identified when type was added as a random effect to the regression models (see section 6.2.8). Together our findings highlight an association between *in vitro* DNA methylation landscapes and the tumor type of origin, which is reflected by mutational signatures and drug response.

6.4 Discussion

Cancer cell lines have long been known to exhibit more extreme methylation levels at selected CGIs and PMDs compared to primary tumors [146, 147]. However, a clear genome-wide assessment of the consistency and genomic nature of these alterations across commonly established cancer cell lines and compared to a large set of primary tumors has not been undertaken. Here, we showed that while most tumors display intermediate CGI and PMD methylation levels, cancer cell lines converge not to one but two main alternate DNA methylation landscapes: an inverse bimodal landscape with high CGI and low PMD methylation levels as well as an extreme hypermethylation state that affects CGIs and PMDs alike (Figure 6.4.1). Only a minority of cell lines seem to be able to maintain intermediate DNA methylation levels. Using read-level analysis and

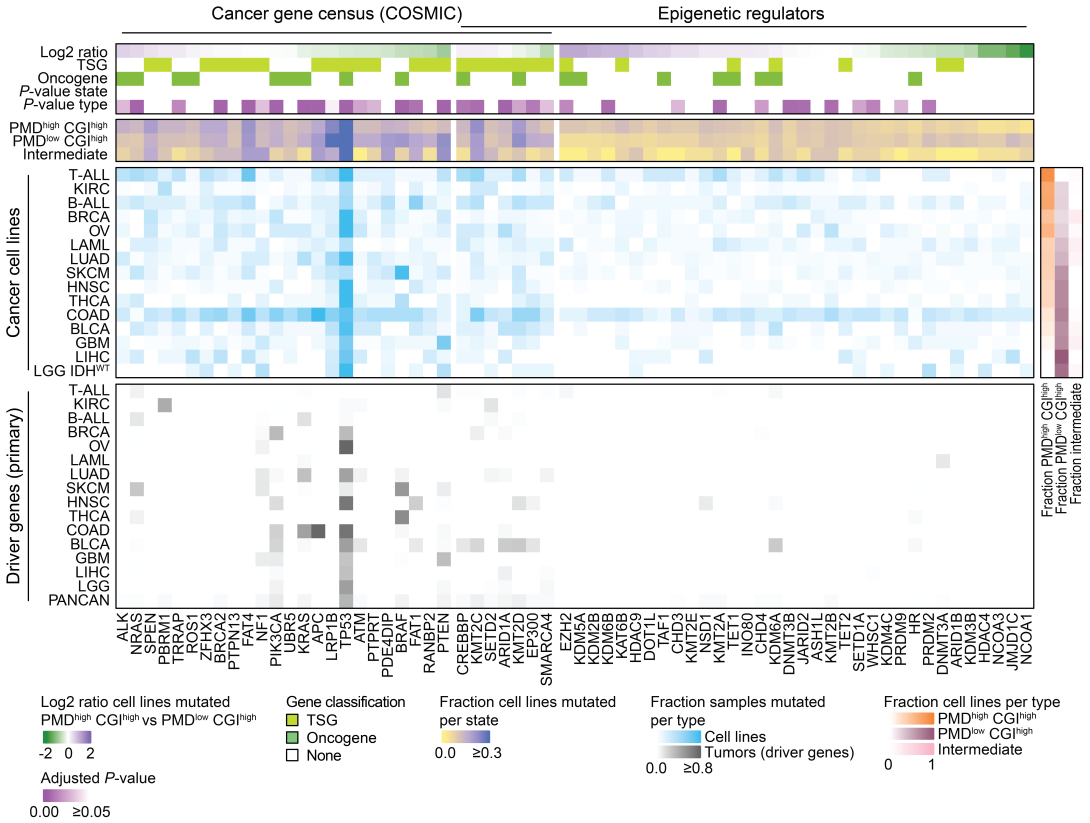

Figure 6.3.22: Overview of the mutational landscape across cancer cell lines showcasing commonly mutated cancer driver genes and epigenetic regulators (see section 6.2.8). Top: The log2 ratio indicates the enrichment of mutations in cell lines with an extreme hypermethylation state over cell lines from an inverse bimodal state. Adjusted *P*-values were obtained using two-sided Fisher's exact tests. Middle: Fraction of cell lines mutated per state and type for each gene. Bottom: Mutation frequency of cancer driver genes previously reported across primary tumor cohorts and pan-cancer [304–306].

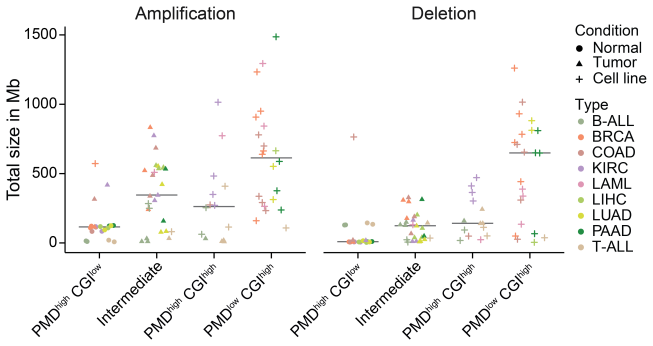

Figure 6.3.23: The total size of chromosomal aberrations separated into amplifications and deletions across WGBS healthy, tumor, and cell line samples separated by state. Cell lines in an inverse bimodal state show a greater variance in the amount/size of chromosomal aberrations.

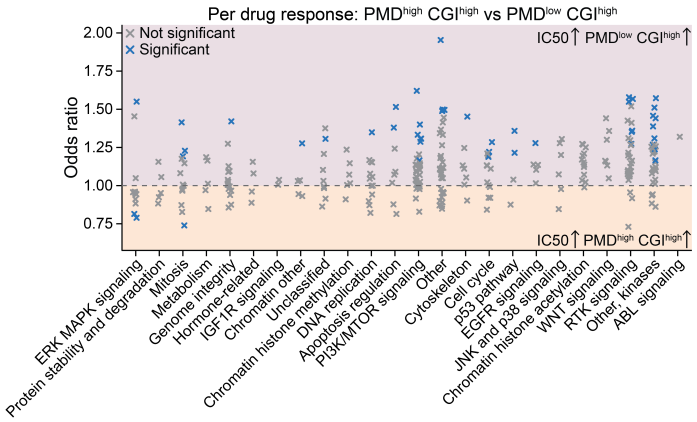

Figure 6.3.24: Odds ratio per drug with respect to the DNA methylation landscape (see section 6.2.8). Values > 1 indicate that cell lines are more likely to be in an inverse bimodal state with growing IC50 (increased drug resistance), while values < 1 indicate that cell lines are more likely to be in an extreme hypermethylation state with increasing IC50.

single-cell WGBS data sets, we showed that intermediate methylation is an intrinsic property of most primary tumors that is reflective of the majority of tumor cells and cannot be explained by diminished tumor purity. This genome-wide intermediate methylation is propagated through advanced cancer stages and metastasis, suggesting that extended proliferation *in vivo* generally does not drive DNA methylation to the extreme levels observed in culture. Finally, we uncovered a type-dependent relationship between cell line states and PMD methylation levels in primary tumors, which was verified by mutational profiles and drug sensitivity patterns.

Although the predominant cell line states are rarely observed *in vivo*, some exceptions exist where primary tumors display similar methylation levels. T-ALL has been reported to exhibit extreme hypermethylation of CGIs in a subset of patients, while PMDs lose almost no methylation compared to healthy precursor T cells (see chapter 5) [238]. In line with these findings, T-ALL patients in our cohort are frequently associated with the extreme hypermethylation state, similar to their respective cell lines. Additionally, some tumor types, such as LIHC, frequently exhibit relatively low PMD methylation levels; however, CGI methylation remains intermediate in contrast to the inverse bimodal state observed in many cell lines. This suggests that extreme hypermethylation of CGIs coupled with almost complete loss of methylation in PMDs mainly represents an artifact of cell culture as suggested previously [147]. It remains to be examined whether the extreme hypermethylation state in a subset of cancer cell lines is linked to the observed landscapes in T-ALL patients or whether these represent distinct forms of epigenetic regulation with a shared phenotype.

Using a colon cancer cohort profiled with single-cell WGBS, we showed that intermediate DNA methylation in primary tumors is largely consistent per cell across CGIs and PMDs. In addition to the relatively sparse single-cell data, our read-level analysis of contiguous CpGs on the same molecule verifies the intrinsic nature of intermediate methylation at CGIs that is characterized by a highly entropic state. High DNA methylation entropy reflects the presence of many distinct epialleles within the population. The stability of these genome-wide intermediate methylation

landscapes through cancer progression and metastasis suggests a different type of regulation compared to previously suggested models of stochastic methylation gain or loss linked to mitotic cell divisions. The propagation of stochastic methylation patterns could be facilitated by ongoing methylation turnover, which might be impaired over time in culture for cell lines with an inverse bimodal landscape.

The near universality of intermediate methylation at CGIs and PMDs across cells in primary tumors seems striking given the clonal evolutions within primary tumors [309]. Within the single-cell WGBS cohort, we observed few cases of seeming clonal adaptation at CGIs. However, the CGIs targeted for hypermethylation in each clone and the associated cells remained in an intermediate regime. This suggests that while the adaptation of clones for selective advantage is possible, this does not interfere with the general nature of DNA methylation across primary tumor cells. Additionally, we observed rare cases where cells of a tumor sampled at a specific site transitioned to a PMD^{low} state in comparison to other sampling sites. The loss of methylation in comparison to other measured sites appeared genome-wide with the exception of hyper CGIs that remained intermediately methylated. Additionally, this landscape was not necessarily propagated to metastasis derived from the respective tumors. Together, these observations rather point to a global adaptation of a specific clone within the tumor in contrast to a purely proliferation-related loss of methylation (Figure 6.4.1). It also suggests that intermediate DNA methylation likely occurs early in tumorigenesis as it is largely shared across cells, with global adaptations only affecting a subset of cells if present.

Given the data at hand, we were not able to investigate whether primary tumors and cancer cell lines are subject to different types of DNA methylation regulation, how the two main *in vitro* landscapes are established, and what potential consequences could arise with respect to targeted therapies that rely on established or newly generated cell lines for drug screens. Additionally, given the nature of the cohorts and the bias of certain tumor types towards one of the two landscapes, a much larger sample size and different sampling strategy would be needed to investigate molecular differences between cell lines of the same type but different DNA methylation states. This could help to identify molecular markers that might be linked to one of the two main cell line landscapes independent of the tumor type-specific bias that limits this pan-cancer study. In this line, we also cannot link our observations of different DNA methylation landscapes across tumors and cell lines to previous studies reporting changes in chromatin modifications during tumorigenesis (H3K27me3, H3K9me3) [4, 5, 126, 128]. Future studies with more comprehensive and evenly sampled cohorts would be needed to uncover the underlying regulatory principles of the different DNA methylation landscapes and potential effects when using cancer cell lines as model systems.

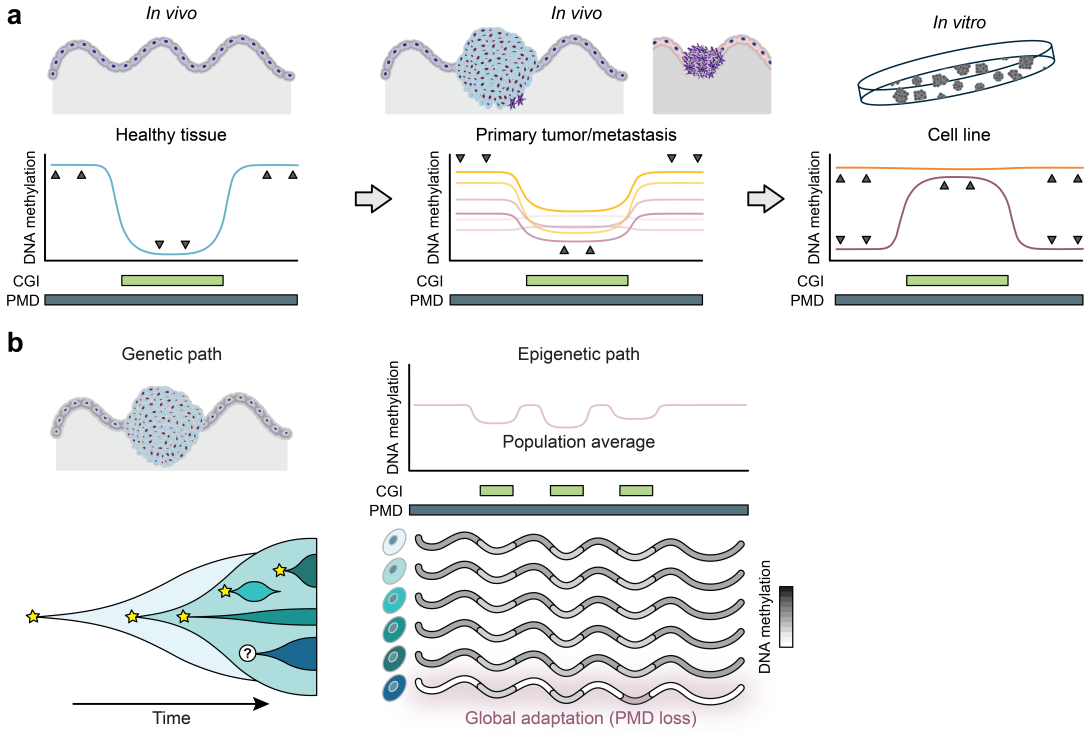

Figure 6.4.1: a) Single-cell and read-level analyses indicate that primary tumors are largely characterized by the maintenance of intermediate hyper CGI and PMD methylation levels. The transition to culture, however, directs cells towards either an extreme inverse bimodal or globally hypermethylated state, which seems to be tightly linked to the tumor type of origin. b) The intermediate DNA methylation levels in most primary tumors are intrinsic to the underlying cell population and cannot be explained by decreased tumor purity alone. The prevalence of highly similar CGI methylation patterns across cells corresponds to minimal cellular heterogeneity for most targets. This indicates that this pattern represents an ancestral state that precedes subclonally acquired mutations. Cells may acquire an additional epigenetic adaptation that drives global methylation levels to lower values, but the conservation of CGI targets would indicate that this represents a secondary event. It remains to be seen if this adaptation stems from genetic mutations within a specific pathway. This figure was generated with the help of Dr. Zachary D. Smith.

Chapter 7

Concluding remarks

This thesis highlights the importance of understanding the underlying heterogeneity of population-wide DNA methylation measurements. For this purpose, chapter 4 presented a new application to make read-level methylation metrics easily accessible from bisulfite sequencing data sets. Chapter 5 and 6 showcased what we can learn by considering fragment- or cell-wise methylation patterns, including the behavior of tumors or cells in culture. Average methylation rates remain a robust measure and our most important tool when defining DNA methylation landscapes or identifying differentially methylated regions between samples, such as patients or cell types. Average methylation rates are also robust across scales and can describe features from single base pairs to megabase-level domains. Nonetheless, average methylation on its own can be misleading without additional interpretation, in particular using methods such as those described in this thesis. This is most notably the case with respect to relatively intermediate DNA methylation levels, as the distinction between cellular and allelic heterogeneity could have implications regarding the underlying form of genome regulation or the complexity of the sample. Thus, considering the methylation levels per molecule in addition to the average across the population could lead to important insights that might be missed otherwise.

The key advantage of read-level methylation metrics, as presented in this thesis, is that they can be extracted from already existing data sets. Whenever a bulk population is profiled using whole-genome or reduced representation bisulfite sequencing, these measurements can be computed and integrated with the analyses of the average methylation rates. It is, therefore, easily accessible, and its use should be generally considered by sequencing-based DNA methylation studies. However, as discussed in chapter 4, these metrics are limited to CpG-rich regions, and fragments across different loci cannot be connected to each other as the cellular identity of each fragment is unknown. Therefore, in the future, it will become more and more important and valuable to generate high-quality single-cell methylation data sets. Currently, the generation of high-quality single-cell data sets is still costly and has comparatively low throughput if high-coverage methylomes should be produced. Additionally, single-cell data sets cover fewer CpGs per cell than what would be expected from a bulk experiment due to the more problematic effect of degradation after bisulfite conversion (more material is available for bulk experiments to compensate for this) [170]. Due to the high costs, many studies pivot to single-cell RRBS, which again enriches for CpG-dense regions and therefore covers even fewer CpGs compared to the respective bulk sequencing. However, chapter 6 showed that a substantial amount of in-

formation could be extracted already from higher-coverage single-cell WGBS data sets, which includes regions with relatively low CpG density, such as PMDs. Recent studies have started to address the need for cheaper, high-coverage single-cell DNA methylation solutions, which will lead to better opportunities to generate such cohorts in the future [310].

Long-read sequencing represents an additional branch that could be of interest to future studies analyzing read-level methylation. Although, similar to short reads, it is not possible to connect different reads to the same cell, the length of the reads themselves enables the analysis of read-wise methylation on a different level compared to classic next-generation sequencing data sets. Reads that span from multiple kilobases up to megabases of the genome can be used to link the methylation of different regulatory elements within a single allele and also span large parts of CpG-poor partially methylated domains. For this purpose, new read-level methylation metrics would need to be developed that can account for the longer reads and the possibility of drop-outs, which reflect a low likelihood during methylation calling and might therefore lead to missing CpG measurements within a read [311]. Excitingly, in addition to cytosine methylation, the long-read technology developed by Oxford Nanopore can capture other base modifications, which could be integrated into future studies. However, improved base calling algorithms would be required for this purpose as currently only a few modifications, such as 5hmC and N6-methyladenosine (m6A), can be reliably detected in addition to regular 5mC [312]. At the same time, Nanopore sequencing can already be coupled with chromosome conformation capture techniques. This could open interesting avenues also for read-wise methylation measurements, such as in the context of enhancer-promoter interactions and the accompanying methylation patterns [313].

Using read-level and single-cell methylation measurements, chapter 6 introduced that intermediate, stochastic DNA methylation is intrinsic to most tumor cells *in vivo*, a feature that is commonly lost in culture. Because this landscape seems to be almost ubiquitously observed across cells of the same tumor and robustly maintained *in vivo*, the question arises whether this represents a distinct form of genome regulation, and if so, why it is so faithfully propagated. The stochastic methylation observed in tumor cells could be a footprint of consistent DNA methylation turnover, which would require a constant targeting of the whole genome by *de novo* methylation enzymes. From the perspective of potentially desired genomic instability or stable silencing of tumor suppressor or other genes, a truly demethylated genome or extremely high methylation of promoter CGIs, such as those observed in culture, seem more beneficial and easier to maintain. Therefore, further investigation of the differences between primary tumors with exceptional methylomes, such as ALL (chapter 5), other tumor types, and cancer cell lines could help disentangle why most tumors maintain their genome in an intermediately methylated state. Additionally, exploring the parallels and similarities of the primary tumor DNA methylation landscape with different physiological processes, such as aging and extraembryonic development, could help elucidate the underlying regulation and potential purpose of this unusual methylome. Especially in light of more and more epigenetic therapies that are tested and admitted to treating tumors in patients, including inhibitors to DNA methylation transferases, it seems crucial to consistently improve our understanding of the regulation and function of the cancer epigenome [314].

Appendix A

Lambda3

This appendix contains additional figures and tables referenced in chapter 3.

A.1 Query data sets

The query data sets used to test and evaluate the performance of Lambda3' nucleotide mode were sampled as follows: Query data sets were downloaded from their respective sources. The data sets q1 and q2 were obtained from https://frl.publisso.de/data/frl:6425521/plant_associated/short_read/rhimgCAMI2_sample_0_reads.tar.gz and https://frl.publisso.de/data/frl:6425521/strain/short_read/strmgCAMI2_sample_0_reads.tar.gz, respectively. The data sets q3 and q4 were obtained from the Sequencing Read Archive (accession numbers ERR1877758 and SRR6043351). Only the first read files were used to mimic a single-end sequencing experiment. For q5, we obtained multiple bisulfite sequencing experiments of different fungi species from the Gene Expression Omnibus (accession numbers GSM3074692, GSM3074693, GSM3074694, GSM3074695, GSM3074696, GSM3074705, GSM3074706, GSM3074711, GSM3074716, and GSM3074718), which were combined and randomly samples. The data sets q1-q3 were *in silico* bisulfite converted as described in section 3.4. All FASTQ files were subsequently converted to FASTA format and the first 200 MB were extracted as test data sets.

A.2 Supplementary figures and tables

Parameter	Nucleotide mode			Bisulfite mode		
	Default	Sensitive	Fast	Default	Sensitive	Fast
seed length (search0)	14	14	-	17	16	-
seed offset (search0)	9	3	-	10	8	-
seed delta (search0)	0	0	-	0	0	-
seed length	14	14	14	17	15	17
seed offset	7	3	9	10	10	10
seed delta	1	1	0	1	1	0
pre-scoring threshold	1.4			1.5		
bit score threshold	46, 46, 47 (q1, q2, q3)			68, 68, 68 (q1, q2, q3)		

Table A.2.1: Parameters selected for the bisulfite mode in comparison to the regular nucleotide search. The keyword "search0" indicates that these were parameters selected for the first round of the iterative search described in section 3.2. For the fast modes, only one iteration of the search is performed.

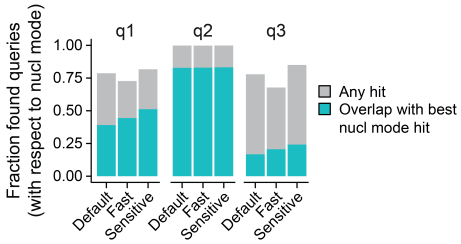

Figure A.2.1: Bar plots visualizing the fraction of queries found by Lambda3's nucleotide mode that were also found by the bisulfite mode separated by profile (default, fast, and sensitive). Additionally, the fraction of queries is visualized for which the best hit overlaps with the best hit located by the respective profile of Lambda3's nucleotide mode.

Appendix B

Acute lymphoblastic leukemia

This appendix contains additional methods, figures, and tables referenced in chapter 5.

B.1 Supplementary methods

B.1.1 Library preparation

Whole-genome bisulfite sequencing

WGBS libraries from patient samples and ALL cell lines (except Jurkat and DND41) were generated at the St. Jude Children's Research Hospital. 200 ng of genomic DNA was bisulfite converted, including 0.2% of DNA from the Lambda phage. Libraries were generated according to the manufacturer's instructions using the TruSeq DNA Methylation kit. Four independent libraries per sample were prepared and sequenced on a HiSeq 2000, generating 101 base pair paired-end reads. WGBS libraries for Jurkat and DND41 were generated at the Max Planck Institute for Molecular Genetics by Dr. Alexandra L. Mattei and the Sequencing Core Facility. Genomic DNA was extracted using the PureLink Genomic DNA Mini Kit, sheared, and bisulfite converted using the EZ DNA Methylation-Gold Kit. Libraries were generated according to the manufacturer's instructions using the Accel-NGS Methyl-seq DNA library kit. Per sample, one library was prepared and sequenced on a NovaSeq 6000, generating 150 bp paired-end reads. On average, 536 million fragments were generated per sample.

RNA sequencing

RNA sequencing data sets of the cell lines Jurkat and DND41 were generated at the Max Planck Institute for Molecular Genetics by Dr. Alexandra L. Mattei and the Sequencing Core Facility. RNA was extracted using the Qiagen RNeasy Mini Kit, and RNA-Seq libraries were prepared using the KAPA Stranded mRNA-seq Kit according to the manufacturer's instructions. Libraries were sequenced on a NovaSeq 6000, generating 100 bp paired-end reads. Transcriptome sequencing of patients was carried out at the St. Jude Children's Research Hospital as described previously, and most data sets were already published in earlier studies (see section 5.2.1).

B.1.2 Cell line experiments

Cell culture

Jurkat (DSMZ ACC 282) and DND41 (DSMZ ACC 525) cells were cultured at the MPIMG in RPMI 1640 medium (Thermo Fisher 61870044) with 10% FBS. PEER (ACC6, DSMZ), PER-117 (Gift from Ursula Kees, Perth), MOLT-16 (ACC29, DSMZ), RPMI-8402 (ACC290, DSMZ), LOUCY (ACC394, DSMZ), TALL-1 (ACC521, DSMZ), ALL-SIL (ACC511, DSMZ), NALM-6 (ACC128, DSMZ), NALM-16 (ACC680, DSMZ), MHH-CALL-2 (ACC341, DSMZ), MHH-CALL-4 (ACC337, DSMZ), and MUTZ5 (ACC490, DSMZ) were cultured at the St. Jude Children's Research Hospital in RPMI 1640 medium containing 10% or 20% fetal bovine serum (HyClone), penicillin/streptomycin (100 U/mL), and glutamine (100 μM). Cell identity was confirmed by short tandem repeat (STR) profiling using a PowerPlex Fusion System (Promega). All of the cell lines were confirmed as Mycoplasma spp. free using the Universal Mycoplasma Detection Kit (American Type Culture Collection, Manassas, VA).

TET2 knockout in Jurkat cells

Jurkat cells were transfected with px458 (Addgene plasmid no. 48138) containing a guide RNA (target sequence: CTTATGGTCAAATAACGACT [315]) targeting exon 3 of the *TET2* gene and expressing a GFP reporter. The transfection was carried out using the Amaxa 4D nucleofector X-Unit (Lonza) following the manufacturer's recommendations. GFP-positive cells were sorted by FACS as single cells into a 96 well plate for clonal expansion and screening. Percentages of sorted cells were analyzed using FlowJo (version 10.3). Disruption of the targeted locus was verified by genotyping PCR and sanger sequencing (primer pair: forward GTCTGGTCAACAAGCTGCGC, reverse AAAGCTGGGGTGTGGCTATC).

B.2 Supplementary figures and tables

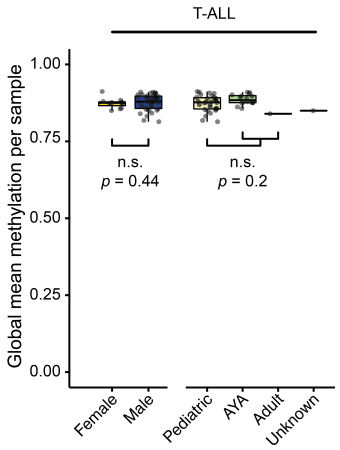

Figure B.2.1: Global average methylation per sample measured excluding CpGs in CGIs for T-ALL samples split by sex (left) and age (right) of the patients.

CGI group/cluster	***p*-value**	**Effect size (Cramer's V)**
Low	$< 2.2 * 10^{-16}$	0.23
Cluster 1	$< 2.2 * 10^{-16}$	0.38
Cluster 2	$< 2.2 * 10^{-16}$	0.67
Cluster 3	$< 2.2 * 10^{-16}$	0.69
Cluster 4	$< 2.2 * 10^{-16}$	0.18
High	$1.69 * 10^{-10}$	0.1
All covered	$< 2.2 * 10^{-16}$	0.25

Table B.2.1: Change in chromatin state proportions of CGI clusters (Chi-squared test). The *p*-value is shown as a measure of significance. Cramer's V is shown as a measure of the effect size.

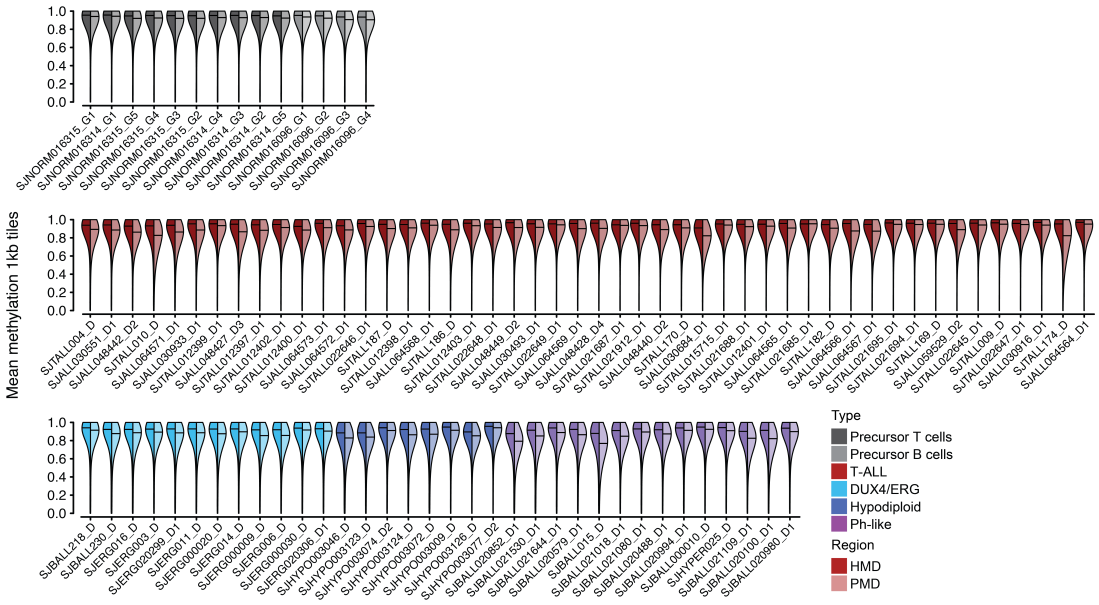

Figure B.2.2: Violin plots of HMDs and PMDs for each healthy and ALL sample.

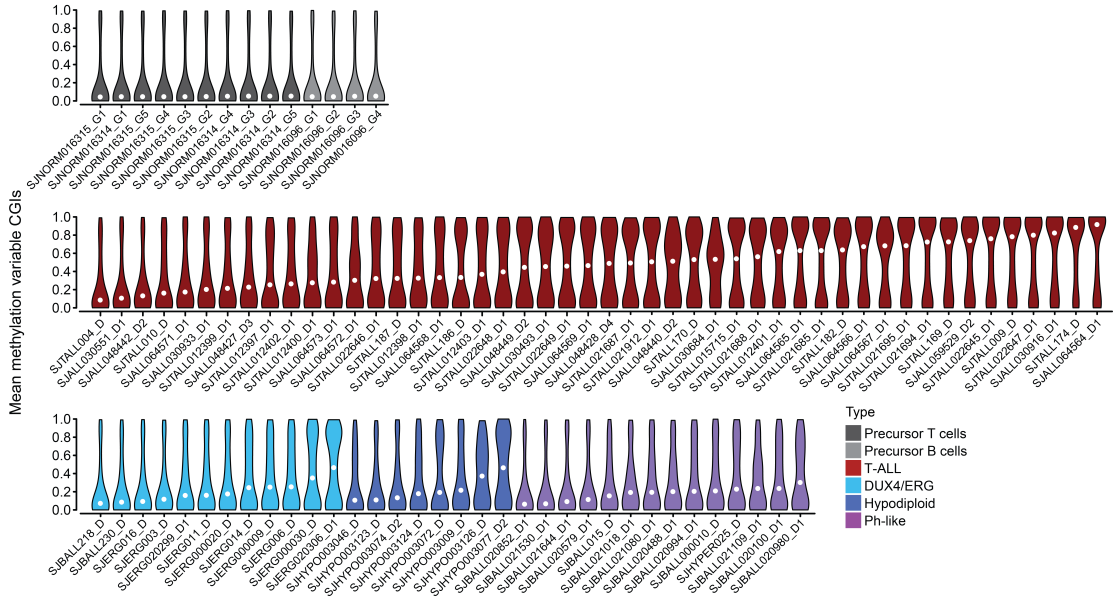

Figure B.2.3: Violin plots of variable CGIs for each healthy and ALL sample.

Variable row	**Variable column**	***p*-value**
Sex	T-ALL group (LM, IM, HM)	1
Age group	T-ALL group (LM, IM, HM)	0.152
HOXA subtype (yes/no)	T-ALL group (LM, IM, HM)	0.324
TLX3 subtype (yes/no)	T-ALL group (LM, IM, HM)	0.343
HOXA or TLX3 subtype (yes/no)	T-ALL group (LM, IM, HM)	0.03
NOTCH1 mutation (yes/no)	T-ALL group (LM, IM, HM)	0.891
NRAS mutation (yes/no)	T-ALL group (LM, IM, HM)	1
WT1 mutation (yes/no)	T-ALL group (LM, IM, HM)	0.853
MED12 mutation (yes/no)	T-ALL group (LM, IM, HM)	0.839
SUZ12 mutation (yes/no)	T-ALL group (LM, IM, HM)	1
ETV6 mutation (yes/no)	T-ALL group (LM, IM, HM)	0.272
FLT3 mutation (yes/no)	T-ALL group (LM, IM, HM)	0.807
Sex	T-ALL group (LM, HM)	1
Age group	T-ALL group (LM, HM)	0.064
HOXA subtype (yes/no)	T-ALL group (LM, HM)	0.282
TLX3 subtype (yes/no)	T-ALL group (LM, HM)	1
HOXA or TLX3 subtype (yes/no)	T-ALL group (LM, HM)	0.119
NOTCH1 mutation (yes/no)	T-ALL group (LM, HM)	1
NRAS mutation (yes/no)	T-ALL group (LM, HM)	1
WT1 mutation (yes/no)	T-ALL group (LM, HM)	1
MED12 mutation (yes/no)	T-ALL group (LM, HM)	0.576
SUZ12 mutation (yes/no)	T-ALL group (LM, HM)	1
ETV6 mutation (yes/no)	T-ALL group (LM, HM)	0.471
FLT3 mutation (yes/no)	T-ALL group (LM, HM)	1

Table B.2.2: Association of T-ALL subtypes and covariates (Fisher's exact test). The first two columns indicate the variables tested. The *p*-value is shown as a measure of significance.

Figure B.2.4: Standardized expression of T-ALL marker genes across T-ALL samples.

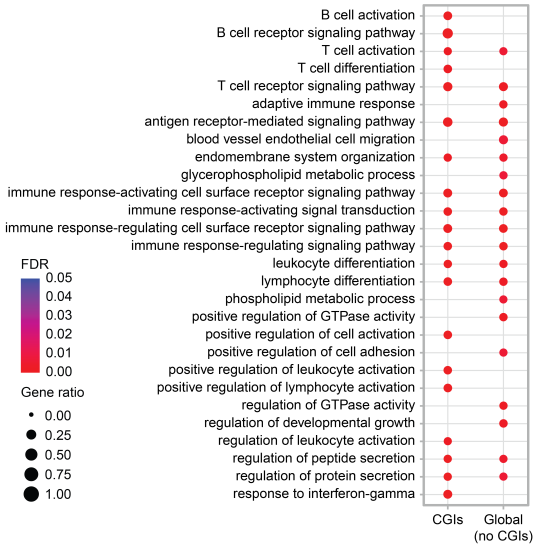

Figure B.2.5: Overrepresentation analysis of genes significantly correlated with global or CGI methylation levels. Enriched GO terms are associated with B and T lymphocyte development because of the overall higher methylation levels in T-ALL compared to B-ALL.

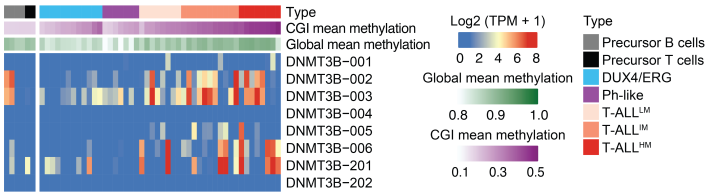

Figure B.2.6: Expression of DNMT3B isoforms across ALL patients. T-ALL patients express the catalytically active isoforms DNMT3B-001 and DNMT3B-002 in addition to the catalytically inactive isoform DNMT3B-003.

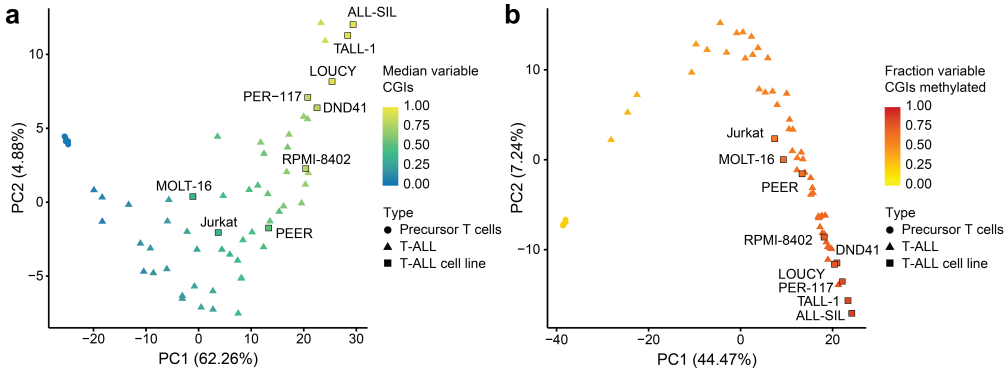

Figure B.2.7: a) PCA based on the mean methylation of variable CGIs of precursor T cell and T-ALL samples. T-ALL cancer cell lines are projected onto the PCA based on the same features. b) PCA based on the methylation status (methylated/unmethylated) of variable CGIs of precursor T cell and T-ALL samples. T-ALL cancer cell lines are projected onto the PCA based on the same features.

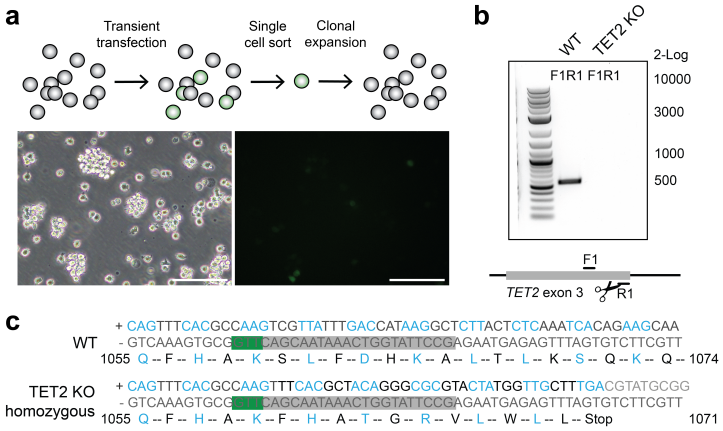

Figure B.2.8: a) TET2 was targeted at exon 3 in Jurkat cells using a GFP-expressing Cas9 plasmid with a single guide RNA. Transfected Jurkat cells were sorted by GFP expression using FACS, and single clones were picked, expanded, and screened by genotyping. b) cDNA genotyping PCR results amplifying a region close to the cut site of the guide RNA shows a product in wild type cells but not in the knockout. c) 7 kb insertion in Jurkat cells at the guide RNA cut sites after TET2 knockout, which results in a premature stop codon in exon 3.

Appendix C

Cancer cell lines

This appendix contains additional methods, figures, and tables referenced in chapter 6.

C.1 Supplementary methods

C.1.1 Library preparation

Whole-genome bisulfite sequencing

The DNA was sheared in Covaris micro TUBE AFA Fiber Pre-Slit Snap-Cap tubes (SKU: 520045) and cleaned up with the Zymo DNA Clean & Concentrator-5 kit (#D4013) following the manufacturer's guidelines. Sheared gDNA was bisulfite converted following manufacturer's guidelines with the EZ DNA Methylation-Gold Kit (Zymo #D5005), and libraries were prepared using the Accel-NGS Methyl-seq DNA library kit (Swift Biosciences, #30024-SWI). Libraries were cleaned using Agencourt AMPure XP beads (Beckman Coulter, #A63881), and the absence of adapters was confirmed on the Agilent TapeStation HS D5000. The final libraries were sequenced on a NovaSeq platform (Illumina), yielding 150 bp paired-end reads.

C.2 Supplementary figures and tables

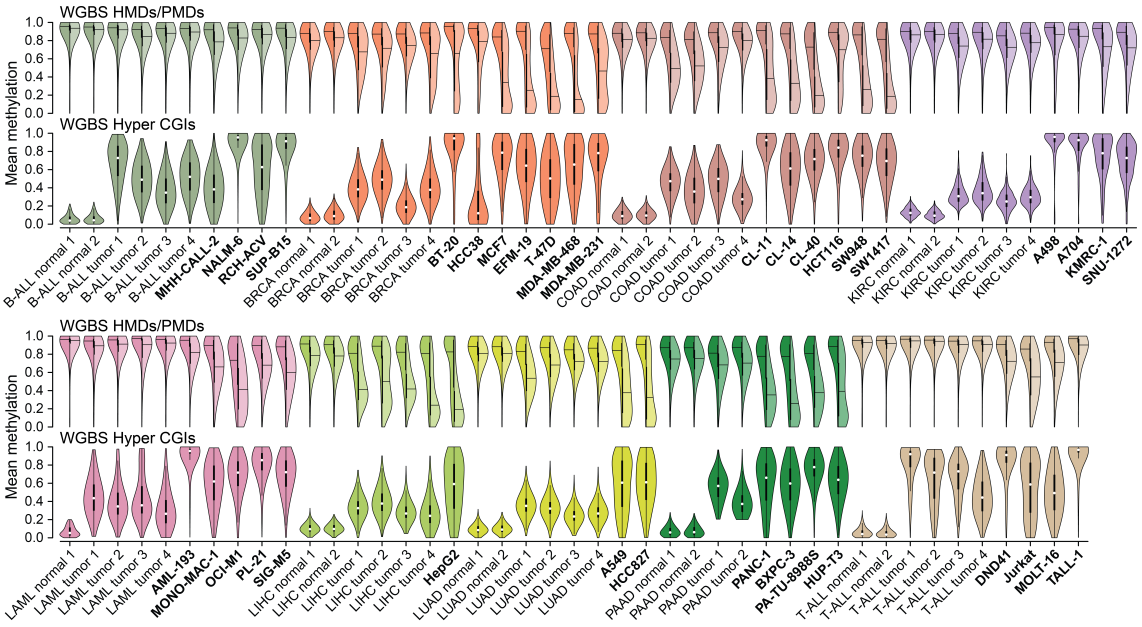

Figure C.2.1: Violin plots of HMD, PMD, and hyper CGI methylation per healthy, tumor, and cell line sample.

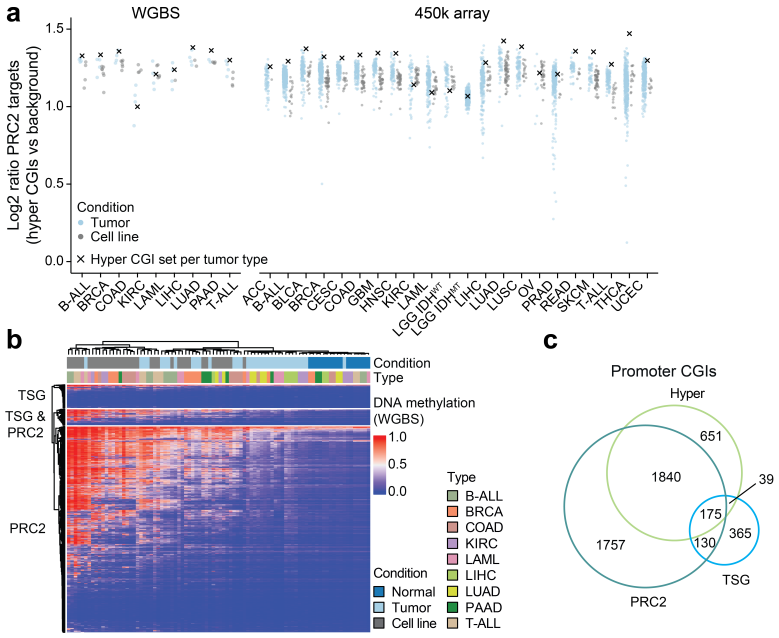

Figure C.2.2: a) Enrichment of hypermethylated CGIs per tumor and cell line sample (dots) as well as tumor type (black crosses) for PRC2 targets for both the WGBS (left) and 450k array (right) cohort. b) Heatmap showing the mean CGI methylation for those associated with tumor suppressor gene promoters (TSGs), PRC2-based regulation, or both across our initial WGBS cohort. Notably, TSGs are prone to hypermethylation if they are canonically repressed by PRC2. c) Overlap of promoter CGIs hypermethylated in any tumor types as measured by WGBS, CGIs targeted by PRC2 in hESCs, and CGIs associated with TSG promoters.

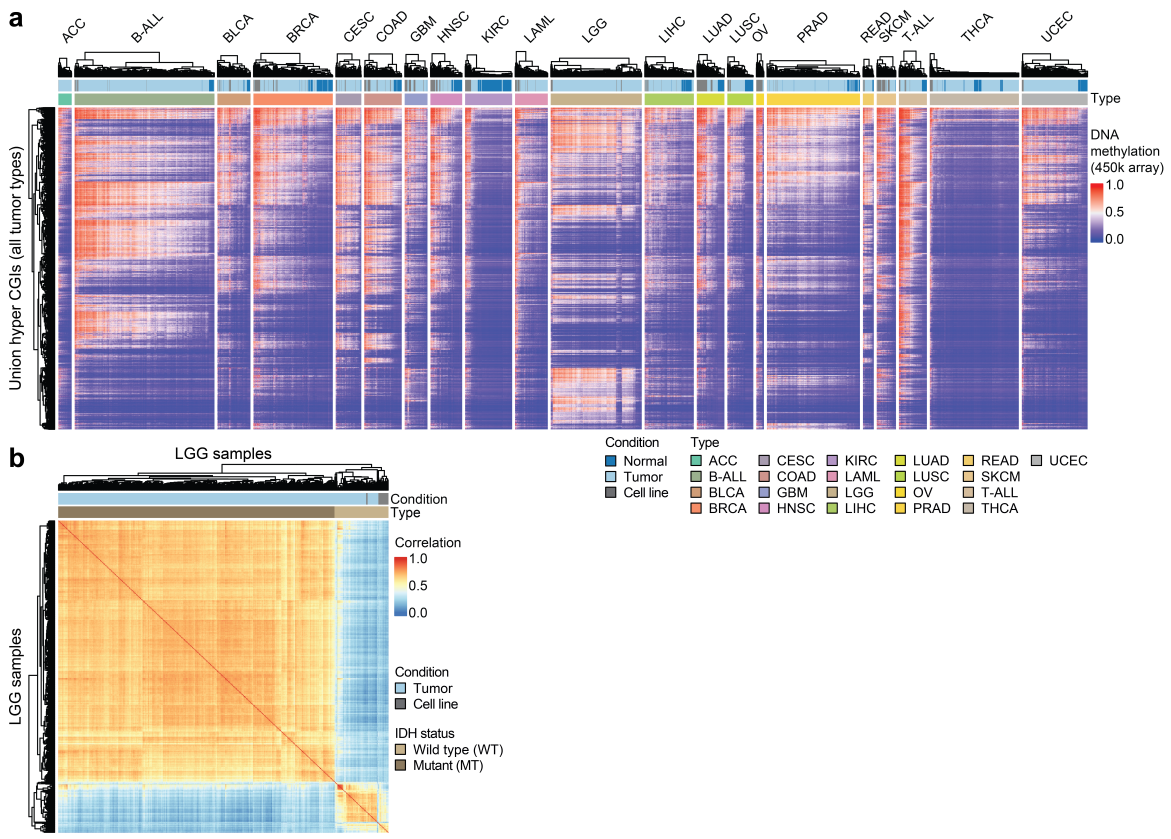

Figure C.2.3: a) Heatmap showing the union of commonly hypermethylated CGIs across all tumor types (array cohort). Cell lines hypermethylate similar targets to their corresponding primary tumors but with a greater frequency. They also frequently methylate additional targets in comparison to the primary tumor samples. Primary tumors overall show similar hypermethylation patterns to one another, with the exception of LGG that stratify into two distinct patterns. b) Correlation heatmap of LGG tumor and cell line samples with IDH mutational status indicated on the top.

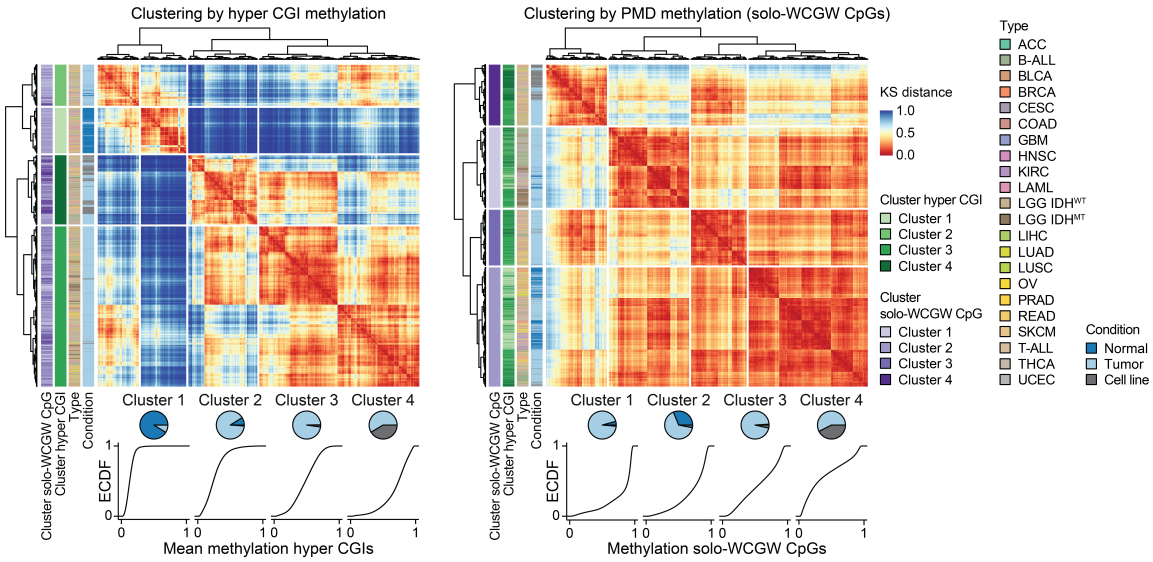

Figure C.2.4: Hierarchical clustering of 450k array samples based on the ECDF of hyper CGIs (left) or solo-WCGW CpGs in common PMDs (right).

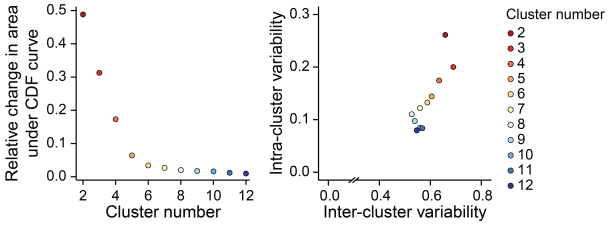

Figure C.2.5: Relative change in the area under the ECDF curve and intra-cluster variability associated with different numbers of clusters during consensus clustering of the median hyper CGI and solo-WCGW CpG methylation levels (see section 6.2.4). Five clusters were chosen as optimum (six clusters only separated the intermediate tumor-enriched states further).

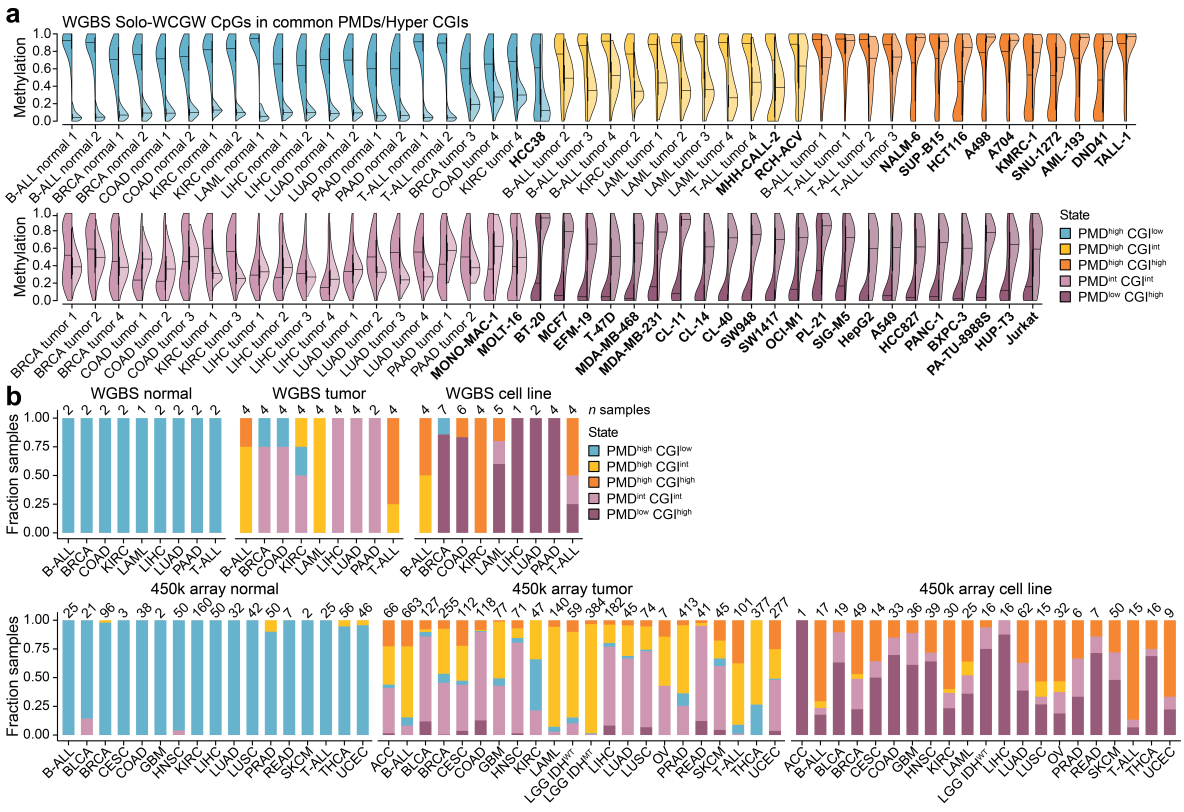

Figure C.2.6: a) Methylation of solo-WCGW CpGs (left) and hypermethylated CGIs (right) for single healthy, tumor, and cancer cell line samples (profiled with WGBS) colored by their methylation state assignment. b) Distribution of DNA methylation states across healthy, tumor, and cell line samples shown for both WGBS and array cohorts and separated by tumor types.

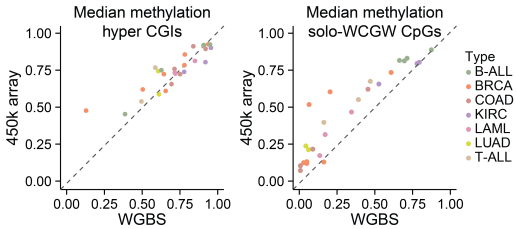

Figure C.2.7: Comparison of hyper CGI and solo-WCGW CpG methylation levels of cell lines profiled by WGBS and 450k array. Hyper CGI methylation levels are relatively consistent between WGBS and arrays, whereas solo-WCGW CpG methylation tends to be higher in arrays than in WGBS. This trend could be linked to the bias in the 450k array probe selection towards functional regulatory elements.

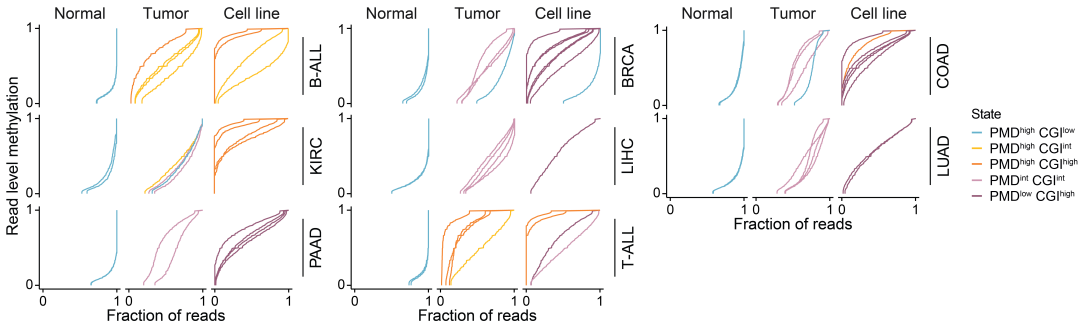

Figure C.2.8: Cumulative read-level methylation distributions across hyper CGIs for all healthy, tumor, and cell line samples separated by tumor type (lines reflect the median).

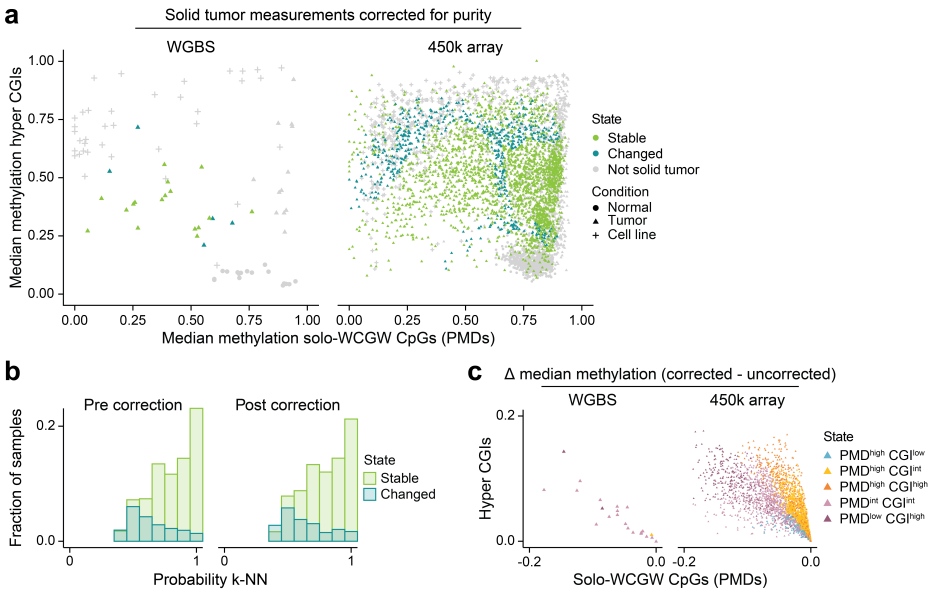

Figure C.2.9: a) Relation of the median methylation of solo-WCGW CpGs in common PMDs and hyper CGIs per WGBS and 450k array sample colored by solid tumor samples that either retain or change their DNA methylation state assignment after correction. Tumor samples that change states after correction are primarily located at the borders between two or more states. b) Probability of solid tumor samples (array cohort) with which they were assigned to their state according to the *k*-NN classification before and after purity correction. Tumor samples that changed methylation states after correction showed overall lower assignment probabilities before and after correction compared to the tumor samples whose status remained unchanged. c) Scatterplots showing the difference in methylation levels between the corrected and uncorrected average methylation values of CGIs and PMDs demonstrate the minor effect of this correction on the overall landscape.

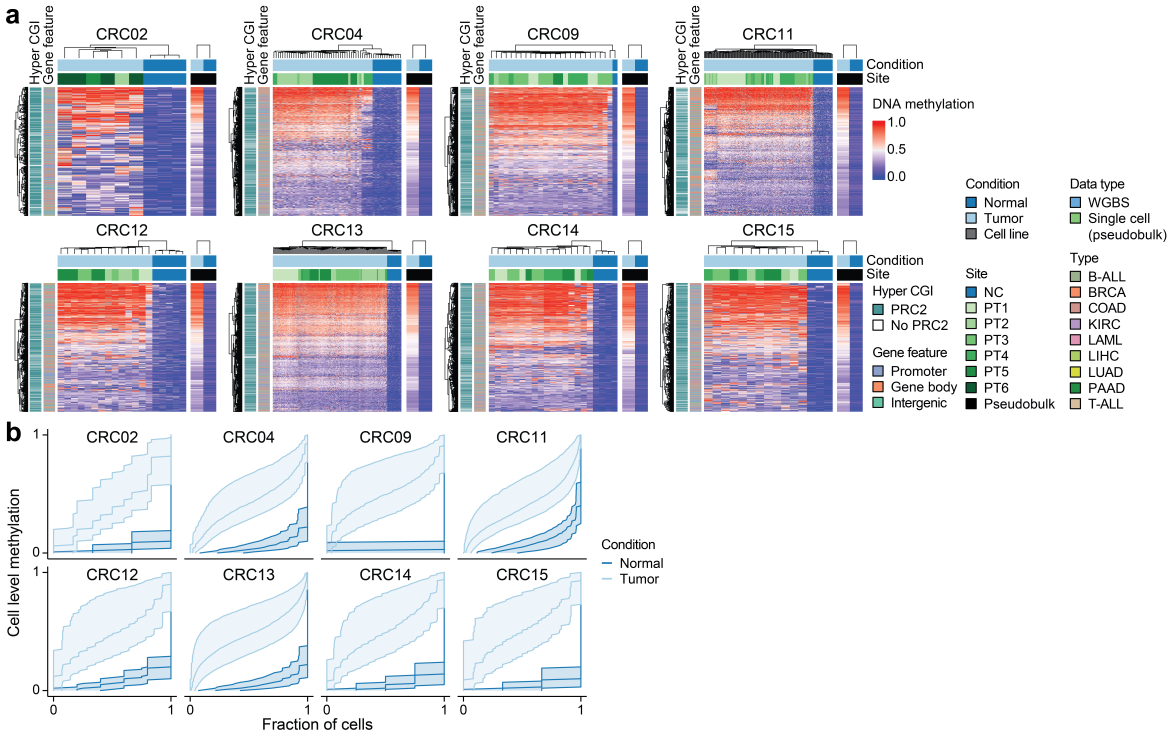

Figure C.2.10: a) Heatmaps of hyper CGI methylation across healthy and tumor cells for the remaining patients not shown in Figure 6.3.14. b) Cumulative single cell-level methylation distributions for hyper CGIs within the remaining patients not shown in Figure 6.3.14. Lines reflect the median, 25%, and 75% quantile across CGIs.

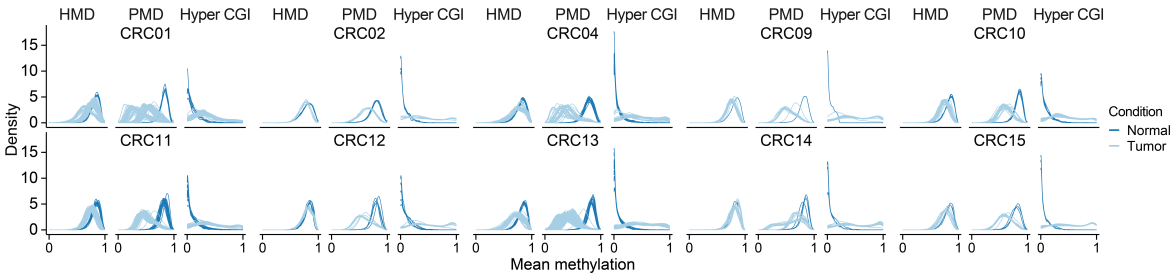

Figure C.2.11: Density of HMD, PMD, and hyper CGI methylation levels for each patient shown for healthy and tumor cells.

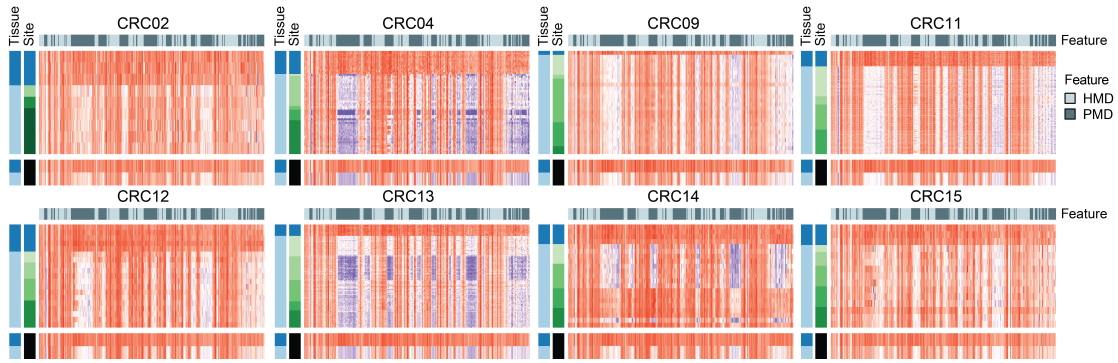

Figure C.2.12: Chromosome-scale heatmaps of the average methylation across cells of the remaining patients not shown in Figure 6.3.16 along chromosome 16p (100 kb tiles).

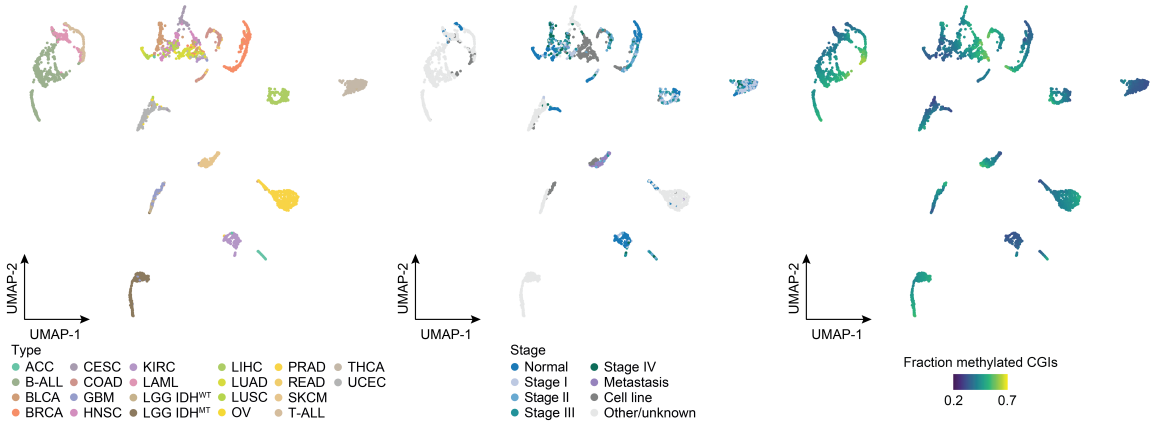

Figure C.2.13: UMAP as shown in Figure 6.3.17 colored by type (left), stage (middle), and fraction of methylated CGIs (right).

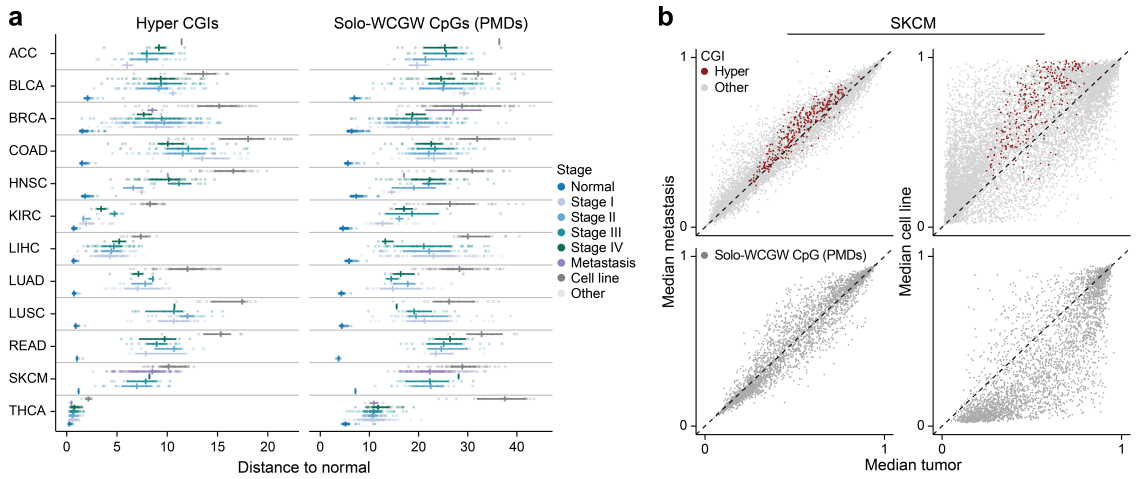

Figure C.2.14: a) Distribution of hyper CGI and PMD methylation levels across healthy samples, different tumor stages, metastasis, and cell lines (array cohort). Horizontal lines denote the IQR; vertical lines denote the median. Overall, methylation levels across different tumor stages and metastatic samples are more similar to each other than the methylation levels of corresponding cancer cell lines. b) Comparison of median hyper CGI (top) and PMD (bottom) methylation levels between primary tumors and metastases (left) as well as cell lines (right) for the SKCM array cohort. Metastases resemble the methylation levels of primary tumors more closely than corresponding cancer cell lines despite a comparable genetic bottleneck and more extensive proliferation than within primary tumors.

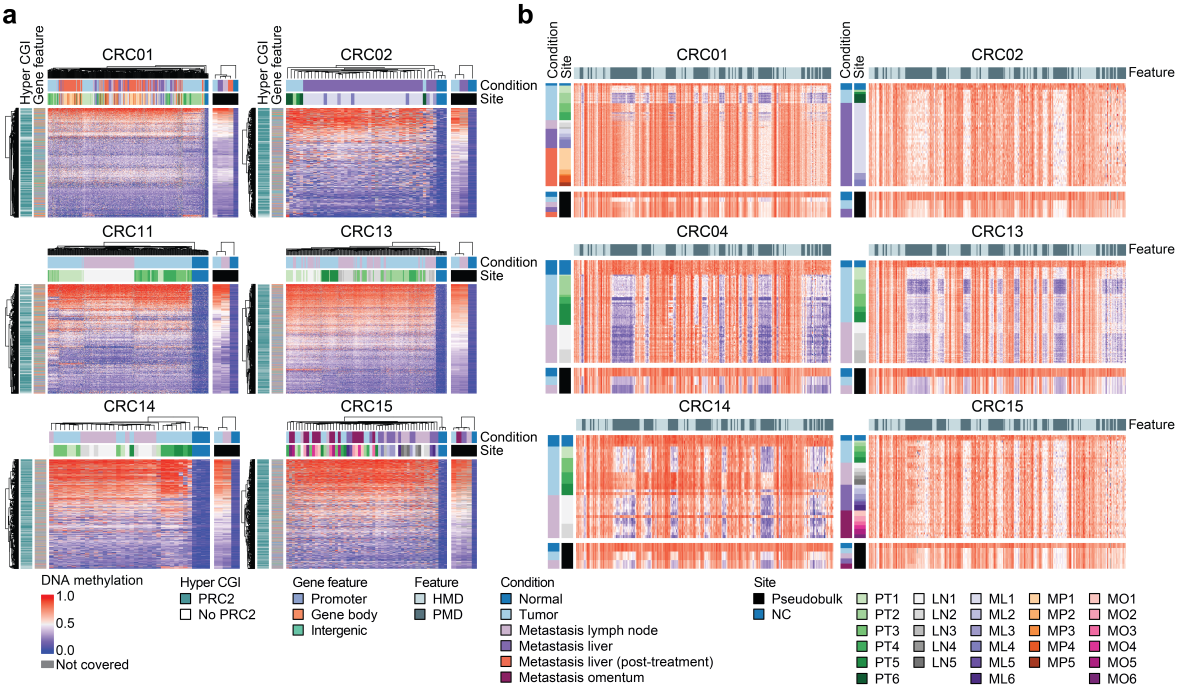

Figure C.2.15: a) Heatmaps of hyper CGI methylation across healthy, tumor, and metastasis cells for the remaining colon cancer patients with profiled metastases not shown in Figure 6.3.20. b) Chromosome-scale heatmaps of the average methylation across healthy, tumor and metastasis cells of the remaining patients not shown in Figure 6.3.20 along chromosome 16p (tile resolution = 100 kb).

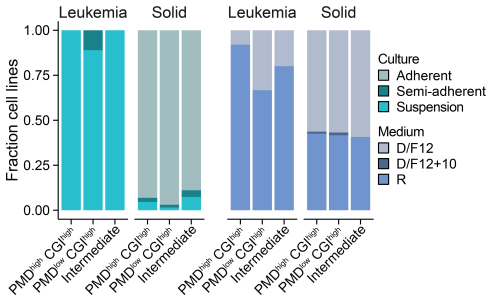

Figure C.2.16: Fraction of cell lines associated with different culture conditions (left) and media (right) separated by state and broad type (leukemia vs. solid tumor).

Type	Adjusted p-value high vs inverse	Adjusted p-value intermediate vs other
B-ALL	0.18	0.86
BLCA	0.18	0.86
BRCA	0.07	0.05
COAD	0.07	0.89
GBM	0.02	0.46
HNSC	0.2	0.35
KIRC	0.02	0.86
LAML	0.83	0.35
LGG IDH^{WT}	-	1
LIHC	0.02	0.86
LUAD	1	0.59
OV	0.14	1
SKCM	0.48	0.35
T-ALL	0.002	1
THCA	0.48	0.8

Table C.2.1: Association of methylation landscape and tumor type (Fisher's exact test). The p-value corrected by FDR is shown as a measure of significance for the comparison of extreme hypermethylation with inverse bimodal landscape and the comparison of intermediate with any other landscape.

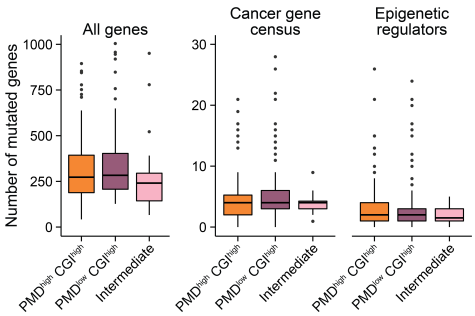

Figure C.2.17: Number of mutated genes (any, recurrently mutated driver, or epigenetic regulator genes) across cell lines associated with the extreme hypermethylation, inverse bimodal, or an intermediate methylation state. While different cell line states do not show substantial enrichment for specific driver or epigenetic regulator mutations, the overall number of mutated genes per line is subtly higher for PMD^{high} CGI^{high} and PMD^{low} CGI^{high} lines than for intermediately methylated lines that more closely resemble primary tumors.

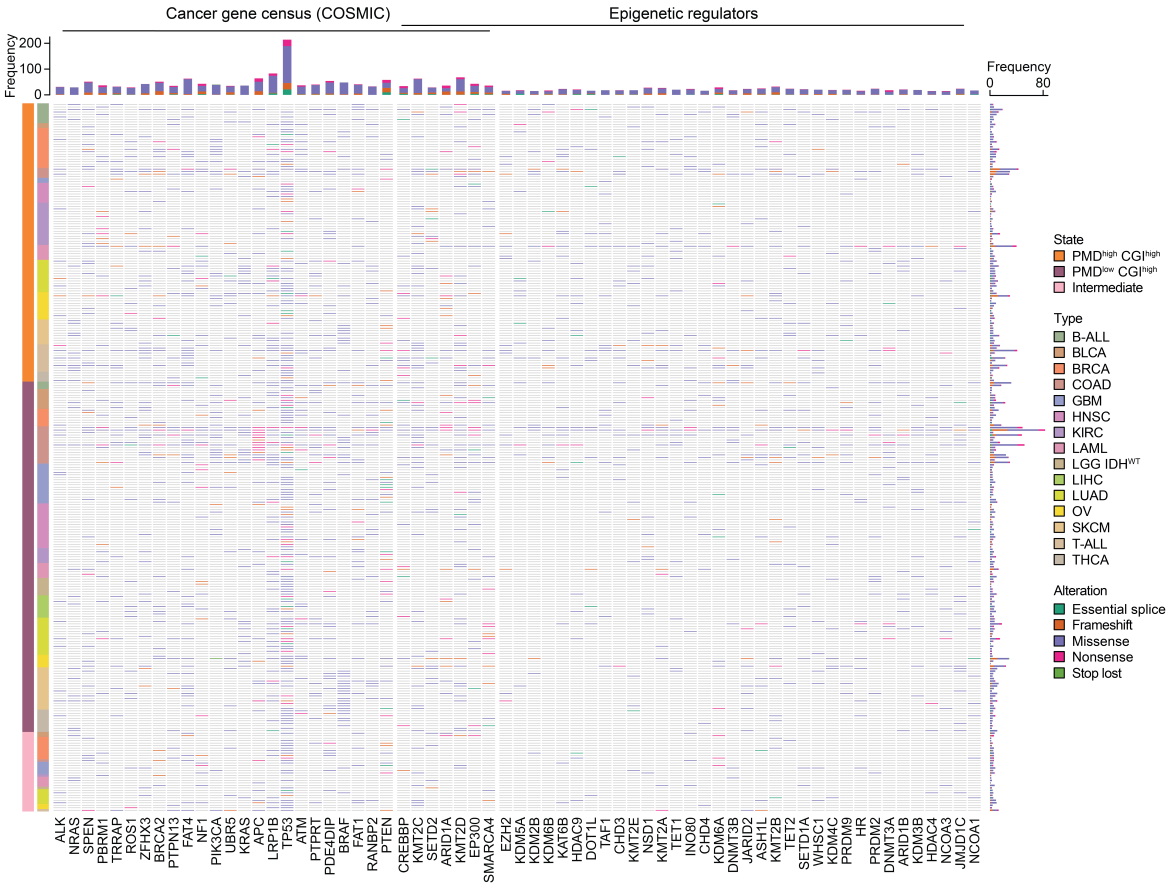

Figure C.2.18: Oncoprint showing mutations in recurrently mutated driver genes and epigenetic regulators across all cell lines ordered by state and type.

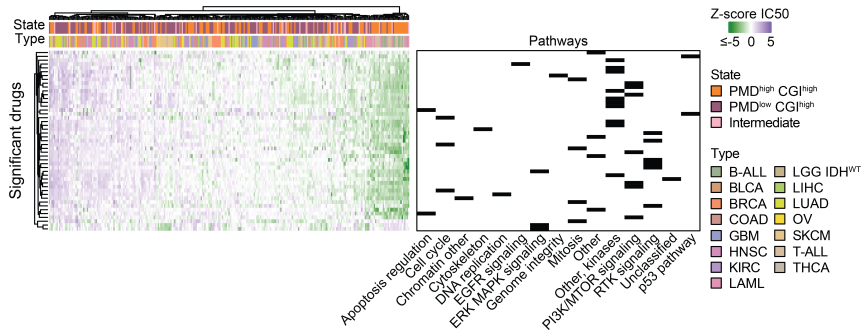

Figure C.2.19: Standardized IC50 for drugs significantly associated with extreme hypermethylation or inverse bimodal state. Differential response to drugs between the PMD^{high} CGI^{high} and PMD^{low} CGI^{high} state seems to be predominantly defined by leukemia compared to solid tumor samples.

Bibliography

[1] D. Hanahan and R. A. Weinberg, "The hallmarks of cancer," *Cell*, vol. 100, pp. 57–70, Jan 2000.

[2] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *Cell*, vol. 144, pp. 646–674, Mar 2011.

[3] D. Hanahan, "Hallmarks of Cancer: New Dimensions," *Cancer Discov*, vol. 12, pp. 31–46, Jan 2022.

[4] S. B. Baylin and P. A. Jones, "Epigenetic Determinants of Cancer," *Cold Spring Harb Perspect Biol*, vol. 8, Sep 2016.

[5] S. B. Baylin and P. A. Jones, "A decade of exploring the cancer epigenome - biological and translational implications," *Nat Rev Cancer*, vol. 11, pp. 726–734, Sep 2011.

[6] J. S. You and P. A. Jones, "Cancer genetics and epigenetics: two sides of the same coin?," *Cancer Cell*, vol. 22, pp. 9–20, Jul 2012.

[7] V. K. Rakyan, T. A. Down, S. Maslau, T. Andrew, T. P. Yang, H. Beyan, P. Whittaker, O. T. McCann, S. Finer, A. M. Valdes, R. D. Leslie, P. Deloukas, and T. D. Spector, "Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains," *Genome Res*, vol. 20, pp. 434–439, Apr 2010.

[8] S. E. Johnstone, V. N. Gladyshev, M. J. Aryee, and B. E. Bernstein, "Epigenetic clocks, aging, and cancer," *Science*, vol. 378, pp. 1276–1277, Dec 2022.

[9] Z. D. Smith, J. Shi, H. Gu, J. Donaghey, K. Clement, D. Cacchiarelli, A. Gnirke, F. Michor, and A. Meissner, "Epigenetic restriction of extraembryonic lineages mirrors the somatic transition to cancer," *Nature*, vol. 549, pp. 543–547, Sep 2017.

[10] R. Lister, R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar, and J. R. Ecker, "Highly integrated single-base resolution maps of the epigenome in Arabidopsis," *Cell*, vol. 133, pp. 523–536, May 2008.

[11] S. J. Cokus, S. Feng, X. Zhang, Z. Chen, B. Merriman, C. D. Haudenschild, S. Pradhan, S. F. Nelson, M. Pellegrini, and S. E. Jacobsen, "Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning," *Nature*, vol. 452, pp. 215–219, Mar 2008.

[12] M. Casado-Pelaez, A. Bueno-Costa, and M. Esteller, "Single cell cancer epigenetics," *Trends Cancer*, vol. 8, pp. 820–838, Oct 2022.

[13] C. B. Hug and J. M. Vaquerizas, "The Birth of the 3D Genome during Early Embryonic Development," *Trends Genet*, vol. 34, pp. 903–914, Dec 2018.

[14] D. U. Gorkin, D. Leung, and B. Ren, "The 3D genome in transcriptional regulation and pluripotency," *Cell Stem Cell*, vol. 14, pp. 762–775, Jun 2014.

[15] R. A. Beagrie, A. Scialdone, M. Schueler, D. C. Kraemer, M. Chotalia, S. Q. Xie, M. Barbieri, I. de Santiago, L. M. Lavitas, M. R. Branco, J. Fraser, J. Dostie, L. Game, N. Dillon, P. A. Edwards, M. Nicodemi, and A. Pombo, "Complex multi-enhancer contacts captured by genome architecture mapping," *Nature*, vol. 543, pp. 519–524, Mar 2017.

[16] T. Cremer and M. Cremer, "Chromosome territories," *Cold Spring Harb Perspect Biol*, vol. 2, p. a003889, Mar 2010.

[17] R. Zhao, M. S. Bodnar, and D. L. Spector, "Nuclear neighborhoods and gene expression," *Curr Opin Genet Dev*, vol. 19, pp. 172–179, Apr 2009.

[18] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, "Comprehensive mapping of long-range interactions reveals folding principles of the human genome," *Science*, vol. 326, pp. 289–293, Oct 2009.

[19] B. van Steensel and A. S. Belmont, "Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression," *Cell*, vol. 169, pp. 780–791, May 2017.

[20] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren, "Topological domains in mammalian genomes identified by analysis of chromatin interactions," *Nature*, vol. 485, pp. 376–380, Apr 2012.

[21] J. Zuin, J. R. Dixon, M. I. van der Reijden, Z. Ye, P. Kolovos, R. W. Brouwer, M. P. van de Corput, H. J. van de Werken, T. A. Knoch, W. F. van IJcken, F. G. Grosveld, B. Ren, and K. S. Wendt, "Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells," *Proc Natl Acad Sci U S A*, vol. 111, pp. 996–1001, Jan 2014.

[22] S. S. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden, "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping," *Cell*, vol. 159, pp. 1665–1680, Dec 2014.

[23] J. R. Dixon, I. Jung, S. Selvaraj, Y. Shen, J. E. Antosiewicz-Bourget, A. Y. Lee, Z. Ye, A. Kim, N. Rajagopal, W. Xie, Y. Diao, J. Liang, H. Zhao, V. V. Lobanenkov, J. R. Ecker, J. A. Thomson, and B. Ren, "Chromatin architecture reorganization during stem cell differentiation," *Nature*, vol. 518, pp. 331–336, Feb 2015.

[24] M. Vietri Rudan, C. Barrington, S. Henderson, C. Ernst, D. T. Odom, A. Tanay, and S. Hadjur, "Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture," *Cell Rep*, vol. 10, pp. 1297–1309, Mar 2015.

[25] O. Symmons, V. V. Uslu, T. Tsujimura, S. Ruf, S. Nassari, W. Schwarzer, L. Ettwiller, and F. Spitz, "Functional and topological characteristics of mammalian regulatory domains," *Genome Res*, vol. 24, pp. 390–400, Mar 2014.

[26] Y. Shen, F. Yue, D. F. McCleary, Z. Ye, L. Edsall, S. Kuan, U. Wagner, J. Dixon, L. Lee, V. V. Lobanenkov, and B. Ren, "A map of the cis-regulatory sequences in the mouse genome," *Nature*, vol. 488, pp. 116–120, Aug 2012.

[27] H. Zheng and W. Xie, "The role of 3D genome organization in development and cell differentiation," *Nat Rev Mol Cell Biol*, vol. 20, pp. 535–550, Sep 2019.

[28] D. G. Lupiáñez, K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J. M. Opitz, R. Laxova, F. Santos-Simarro, B. Gilbert-Dussardier, L. Wittler, M. Borschiwer, S. A. Haas, M. Osterwalder, M. Franke, B. Timmermann, J. Hecht, M. Spielmann, A. Visel, and S. Mundlos, "Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions," *Cell*, vol. 161, pp. 1012–1025, May 2015.

[29] E. Giorgio, D. Robyr, M. Spielmann, E. Ferrero, E. Di Gregorio, D. Imperiale, G. Vaula, G. Stamoulis, F. Santoni, C. Atzori, L. Gasparini, D. Ferrera, C. Canale, M. Guipponi, L. A. Pennacchio, S. E. Antonarakis, A. Brussino, and A. Brusco, "A large genomic deletion leads to enhancer adoption by the lamin B1 gene: a second path to autosomal dominant adult-onset demyelinating leukodystrophy (ADLD)," *Hum Mol Genet*, vol. 24, pp. 3143–3154, Jun 2015.

[30] A. L. Valton and J. Dekker, "TAD disruption as oncogenic driver," *Curr Opin Genet Dev*, vol. 36, pp. 34–40, Feb 2016.

[31] V. W. Zhou, A. Goren, and B. E. Bernstein, "Charting histone modifications and the functional organization of mammalian genomes," *Nat Rev Genet*, vol. 12, pp. 7–18, Jan 2011.

[32] H. Woo, S. Dam Ha, S. B. Lee, S. Buratowski, and T. Kim, "Modulation of gene expression dynamics by co-transcriptional histone methylations," *Exp Mol Med*, vol. 49, p. e326, Apr 2017.

[33] C. A. Musselman, M. E. Lalonde, J. Côté, and T. G. Kutateladze, "Perceiving the epigenetic landscape through histone readers," *Nat Struct Mol Biol*, vol. 19, pp. 1218–1227, Dec 2012.

[34] T. Kouzarides, "Chromatin modifications and their function," *Cell*, vol. 128, pp. 693–705, Feb 2007.

[35] C. D. Allis and T. Jenuwein, "The molecular hallmarks of epigenetic control," *Nat Rev Genet*, vol. 17, pp. 487–500, Aug 2016.

[36] G. G. Wang and C. D. Allis, ""Misinterpretation" of a histone mark is linked to aberrant stem cells and cancer development," *Cell Cycle*, vol. 8, pp. 1982–1983, Jul 2009.

[37] P. Chi, C. D. Allis, and G. G. Wang, "Covalent histone modifications–miswritten, misinterpreted and mis-erased in human cancers," *Nat Rev Cancer*, vol. 10, pp. 457–469, Jul 2010.

[38] B. M. Turner, "Reading signals on the nucleosome with a new nomenclature for modified histones," *Nat Struct Mol Biol*, vol. 12, pp. 110–112, Feb 2005.

[39] B. E. Bernstein, M. Kamal, K. Lindblad-Toh, S. Bekiranov, D. K. Bailey, D. J. Huebert, S. McMahon, E. K. Karlsson, E. J. Kulbokas, T. R. Gingeras, S. L. Schreiber, and E. S. Lander, "Genomic maps and comparative analysis of histone modifications in human and mouse," *Cell*, vol. 120, pp. 169–181, Jan 2005.

[40] T. H. Kim, L. O. Barrera, M. Zheng, C. Qu, M. A. Singer, T. A. Richmond, Y. Wu, R. D. Green, and B. Ren, "A high-resolution map of active promoters in the human genome," *Nature*, vol. 436, pp. 876–880, Aug 2005.

[41] A. Shilatifard, "The COMPASS family of histone H3K4 methylases: mechanisms of regulation in development and disease pathogenesis," *Annu Rev Biochem*, vol. 81, pp. 65–95, Jul 2012.

[42] N. D. Heintzman, R. K. Stuart, G. Hon, Y. Fu, C. W. Ching, R. D. Hawkins, L. O. Barrera, S. Van Calcar, C. Qu, K. A. Ching, W. Wang, Z. Weng, R. D. Green, G. E. Crawford, and B. Ren, "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome," *Nat Genet*, vol. 39, pp. 311–318, Mar 2007.

[43] A. Local, H. Huang, C. P. Albuquerque, N. Singh, A. Y. Lee, W. Wang, C. Wang, J. E. Hsia, A. K. Shiau, K. Ge, K. D. Corbett, D. Wang, H. Zhou, and B. Ren, "Identification of H3K4me1-associated proteins at mammalian enhancers," *Nat Genet*, vol. 50, pp. 73–82, Jan 2018.

[44] S. A. Shinsky, K. E. Monteith, S. Viggiano, and M. S. Cosgrove, "Biochemical reconstitution and phylogenetic comparison of human SET1 family core complexes involved in histone methylation," *J Biol Chem*, vol. 290, pp. 6361–6375, Mar 2015.

[45] M. P. Creyghton, A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp, L. A. Boyer, R. A. Young, and R. Jaenisch, "Histone H3K27ac separates active from poised enhancers and predicts developmental state," *Proc Natl Acad Sci U S A*, vol. 107, pp. 21931–21936, Dec 2010.

[46] R. Raisner, S. Kharbanda, L. Jin, E. Jeng, E. Chan, M. Merchant, P. M. Haverty, R. Bainer, T. Cheung, D. Arnott, E. M. Flynn, F. A. Romero, S. Magnuson, and K. E. Gascoigne, "Enhancer Activity Requires CBP/P300 Bromodomain-Dependent Histone H3K27 Acetylation," *Cell Rep*, vol. 24, pp. 1722–1729, Aug 2018.

[47] Z. Sun, Y. Zhang, J. Jia, Y. Fang, Y. Tang, H. Wu, and D. Fang, "H3K36me3, message from chromatin to DNA damage repair," *Cell Biosci*, vol. 10, p. 9, Jan 2020.

[48] C. Huang and B. Zhu, "Roles of H3K36-specific histone methyltransferases in transcription: antagonizing silencing and safeguarding transcription fidelity," *Biophys Rep*, vol. 4, no. 4, pp. 170–177, 2018.

[49] J. A. Simon and R. E. Kingston, "Mechanisms of polycomb gene silencing: knowns and unknowns," *Nat Rev Mol Cell Biol*, vol. 10, pp. 697–708, Oct 2009.

[50] K. Hyun, J. Jeon, K. Park, and J. Kim, "Writing, erasing and reading histone lysine methylations," *Exp Mol Med*, vol. 49, p. e324, Apr 2017.

[51] R. Margueron and D. Reinberg, "The Polycomb complex PRC2 and its mark in life," *Nature*, vol. 469, pp. 343–349, Jan 2011.

[52] Z. Chen and Y. Zhang, "Maternal H3K27me3-dependent autosomal and X chromosome imprinting," *Nat Rev Genet*, vol. 21, pp. 555–571, Sep 2020.

[53] J. S. Becker, D. Nicetto, and K. S. Zaret, "H3K9me3-Dependent Heterochromatin: Barrier to Cell Fate Changes," *Trends Genet*, vol. 32, pp. 29–41, Jan 2016.

[54] P. Voigt, W. W. Tee, and D. Reinberg, "A double take on bivalent promoters," *Genes Dev*, vol. 27, pp. 1318–1338, Jun 2013.

[55] G. Mas, E. Blanco, C. Ballaré, M. Sansó, Y. G. Spill, D. Hu, Y. Aoi, F. Le Dily, A. Shilatifard, M. A. Marti-Renom, and L. Di Croce, "Promoter bivalency favors an open chromatin architecture in embryonic stem cells," *Nat Genet*, vol. 50, pp. 1452–1462, Oct 2018.

[56] Z. D. Smith and A. Meissner, "DNA methylation: roles in mammalian development," *Nat Rev Genet*, vol. 14, pp. 204–220, Mar 2013.

[57] M. V. C. Greenberg and D. Bourc'his, "The diverse roles of DNA methylation in mammalian development and disease," *Nat Rev Mol Cell Biol*, vol. 20, pp. 590–607, Oct 2019.

[58] P. A. Jones, "Functions of DNA methylation: islands, start sites, gene bodies and beyond," *Nat Rev Genet*, vol. 13, pp. 484–492, May 2012.

[59] X. Yang, H. Han, D. D. De Carvalho, F. D. Lay, P. A. Jones, and G. Liang, "Gene body methylation can alter gene expression and is a therapeutic target in cancer," *Cancer Cell*, vol. 26, pp. 577–590, Oct 2014.

[60] A. K. Maunakea, R. P. Nagarajan, M. Bilenky, T. J. Ballinger, C. D'Souza, S. D. Fouse, B. E. Johnson, C. Hong, C. Nielsen, Y. Zhao, G. Turecki, A. Delaney, R. Varhol, N. Thiessen, K. Shchors, V. M. Heine, D. H. Rowitch, X. Xing, C. Fiore, M. Schillebeeckx, S. J. Jones, D. Haussler, M. A. Marra, M. Hirst, T. Wang, and J. F. Costello, "Conserved role of intragenic DNA methylation in regulating alternative promoters," *Nature*, vol. 466, pp. 253–257, Jul 2010.

[61] S. Choufani, J. S. Shapiro, M. Susiarjo, D. T. Butcher, D. Grafodatskaya, Y. Lou, J. C. Ferreira, D. Pinto, S. W. Scherer, L. G. Shaffer, P. Coullin, I. Caniggia, J. Beyene, R. Slim, M. S. Bartolomei, and R. Weksberg, "A novel approach identifies new differentially methylated regions (DMRs) associated with imprinted genes," *Genome Res*, vol. 21, pp. 465–476, Mar 2011.

[62] M. Ehrlich and M. Lacey, "DNA methylation and differentiation: silencing, upregulation and modulation of gene expression," *Epigenomics*, vol. 5, pp. 553–568, Sep 2013.

[63] Z. Jin and Y. Liu, "DNA methylation in human diseases," *Genes Dis*, vol. 5, pp. 1–8, Mar 2018.

[64] H. Wu, J. Tao, and Y. E. Sun, "Regulation and function of mammalian DNA methylation patterns: a genomic perspective," *Brief Funct Genomics*, vol. 11, pp. 240–250, May 2012.

[65] C. Haggerty, H. Kretzmer, C. Riemenschneider, A. S. Kumar, A. L. Mattei, N. Bailly, J. Gottfreund, P. Giesselmann, R. Weigert, B. Brändl, P. Giehr, R. Buschow, C. Galonska, F. von Meyenn, M. B. Pappalardi, M. T. McCabe, L. Wittler, C. Giesecke-Thiel, T. Mielke, D. Meierhofer, B. Timmermann, F. J. Müller, J. Walter, and A. Meissner, "Dnmt1 has de novo activity targeted to transposable elements," *Nat Struct Mol Biol*, vol. 28, pp. 594–603, Jul 2021.

[66] J. Wang, S. Hevi, J. K. Kurash, H. Lei, F. Gay, J. Bajko, H. Su, W. Sun, H. Chang, G. Xu, F. Gaudet, E. Li, and T. Chen, "The lysine demethylase LSD1 (KDM1) is required for maintenance of global DNA methylation," *Nat Genet*, vol. 41, pp. 125–129, Jan 2009.

[67] R. M. Kohli and Y. Zhang, "TET enzymes, TDG and the dynamics of DNA demethylation," *Nature*, vol. 502, pp. 472–479, Oct 2013.

[68] H. Gowher, K. Liebert, A. Hermann, G. Xu, and A. Jeltsch, "Mechanism of stimulation of catalytic activity of Dnmt3A and Dnmt3B DNA-(cytosine-C5)-methyltransferases by Dnmt3L," *J Biol Chem*, vol. 280, pp. 13341–13348, Apr 2005.

[69] N. Veland, Y. Lu, S. Hardikar, S. Gaddis, Y. Zeng, B. Liu, M. R. Estecio, Y. Takata, K. Lin, M. W. Tomida, J. Shen, D. Saha, H. Gowher, H. Zhao, and T. Chen, "DNMT3L facilitates DNA methylation partly by maintaining DNMT3A stability in mouse embryonic stem cells," *Nucleic Acids Res*, vol. 47, pp. 152–167, Jan 2019.

[70] H. Kretzmer, *Methods for DNA Methylation Sequencing Analysis and their Application on Cancer Data*. Dissertation, Universität Leipzig, May 2016.

[71] X. Wu and Y. Zhang, "TET-mediated active DNA demethylation: mechanism, function and beyond," *Nat Rev Genet*, vol. 18, pp. 517–534, Sep 2017.

[72] W. Zhou, H. Q. Dinh, Z. Ramjan, D. J. Weisenberger, C. M. Nicolet, H. Shen, P. W. Laird, and B. P. Berman, "DNA methylation loss in late-replicating domains is linked to mitotic cell division," *Nat Genet*, vol. 50, pp. 591–602, Apr 2018.

[73] A. Salhab, K. Nordström, G. Gasparoni, K. Kattler, P. Ebert, F. Ramirez, L. Arrigoni, F. Müller, J. K. Polansky, C. Cadenas, J. G Hengstler, T. Lengauer, T. Manke, and J. Walter, "A comprehensive analysis of 195 DNA methylomes reveals shared and cell-specific features of partially methylated domains," *Genome Biol*, vol. 19, p. 150, Sep 2018.

[74] B. E. Decato, J. Qu, X. Ji, E. Wagenblast, S. R. V. Knott, G. J. Hannon, and A. D. Smith, "Characterization of universal features of partially methylated domains across tissues and species," *Epigenetics Chromatin*, vol. 13, p. 39, Oct 2020.

[75] T. Pachano, V. Sánchez-Gaya, T. Ealo, M. Mariner-Faulí, T. Bleckwehl, H. G. Asenjo, P. Respuela, S. Cruz-Molina, M. Muñoz-San Martín, E. Haro, W. F. J. van IJcken, D. Landeira, and A. Rada-Iglesias, "Orphan CpG islands amplify poised enhancer regulatory activity and determine target gene responsiveness," *Nat Genet*, vol. 53, pp. 1036–1049, Jul 2021.

[76] M. Gardiner-Garden and M. Frommer, "CpG islands in vertebrate genomes," *J Mol Biol*, vol. 196, pp. 261–282, Jul 1987.

[77] D. Takai and P. A. Jones, "Comprehensive analysis of CpG islands in human chromosomes 21 and 22," *Proc Natl Acad Sci U S A*, vol. 99, pp. 3740–3745, Mar 2002.

[78] J. L. Glass, R. F. Thompson, B. Khulan, M. E. Figueroa, E. N. Olivier, E. J. Oakley, G. Van Zant, E. E. Bouhassira, A. Melnick, A. Golden, M. J. Fazzari, and J. M. Greally, "CG dinucleotide clustering is a species-specific property of the genome," *Nucleic Acids Res*, vol. 35, pp. 6798–6807, Nov 2007.

[79] H. Wu, B. Caffo, H. A. Jaffee, R. A. Irizarry, and A. P. Feinberg, "Redefining CpG islands using hidden Markov models," *Biostatistics*, vol. 11, pp. 499–514, Jul 2010.

[80] D. M. Jeziorska, R. J. S. Murray, M. De Gobbi, R. Gaentzsch, D. Garrick, H. Ayyub, T. Chen, E. Li, J. Telenius, M. Lynch, B. Graham, A. J. H. Smith, J. N. Lund, J. R. Hughes, D. R. Higgs, and C. Tufarelli, "DNA methylation of intragenic CpG islands depends on their transcriptional activity during differentiation and disease," *Proc Natl Acad Sci U S A*, vol. 114, pp. E7526–E7535, Sep 2017.

[81] E. Schilling and M. Rehli, "Global, comparative analysis of tissue-specific promoter CpG methylation," *Genomics*, vol. 90, pp. 314–323, Sep 2007.

[82] L. Shen, Y. Kondo, Y. Guo, J. Zhang, L. Zhang, S. Ahmed, J. Shu, X. Chen, R. A. Waterland, and J. P. Issa, "Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters," *PLoS Genet*, vol. 3, pp. 2023–2036, Oct 2007.

[83] M. Weber, I. Hellmann, M. B. Stadler, L. Ramos, S. Pääbo, M. Rebhan, and D. Schübeler, "Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome," *Nat Genet*, vol. 39, pp. 457–466, Apr 2007.

[84] F. Mohn, M. Weber, M. Rebhan, T. C. Roloff, J. Richter, M. B. Stadler, M. Bibel, and D. Schübeler, "Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors," *Mol Cell*, vol. 30, pp. 755–766, Jun 2008.

[85] I. Okamoto and E. Heard, "Lessons from comparative analysis of X-chromosome inactivation in mammals," *Chromosome Res*, vol. 17, pp. 659–669, Oct 2009.

[86] A. Meissner, T. S. Mikkelsen, H. Gu, M. Wernig, J. Hanna, A. Sivachenko, X. Zhang, B. E. Bernstein, C. Nusbaum, D. B. Jaffe, A. Gnirke, R. Jaenisch, and E. S. Lander, "Genome-scale DNA methylation maps of pluripotent and differentiated cells," *Nature*, vol. 454, pp. 766–770, Aug 2008.

[87] D. Macleod, J. Charlton, J. Mullins, and A. P. Bird, "Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island," *Genes Dev*, vol. 8, pp. 2282–2292, Oct 1994.

[88] J. P. Thomson, P. J. Skene, J. Selfridge, T. Clouaire, J. Guy, S. Webb, A. R. Kerr, A. Deaton, R. Andrews, K. D. James, D. J. Turner, R. Illingworth, and A. Bird, "CpG islands influence chromatin structure via the CpG-binding protein Cfp1," *Nature*, vol. 464, pp. 1082–1086, Apr 2010.

[89] J. Otani, T. Nankumo, K. Arita, S. Inamoto, M. Ariyoshi, and M. Shirakawa, "Structural basis for recognition of H3K4 methylation status by the DNA methyltransferase 3A ATRX-DNMT3-DNMT3L domain," *EMBO Rep*, vol. 10, pp. 1235–1241, Nov 2009.

[90] S. K. Ooi, C. Qiu, E. Bernstein, K. Li, D. Jia, Z. Yang, H. Erdjument-Bromage, P. Tempst, S. P. Lin, C. D. Allis, X. Cheng, and T. H. Bestor, "DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA," *Nature*, vol. 448, pp. 714–717, Aug 2007.

[91] K. D. Rasmussen and K. Helin, "Role of TET enzymes in DNA methylation, development, and cancer," *Genes Dev*, vol. 30, pp. 733–750, Apr 2016.

[92] M. Bibikova, B. Barnes, C. Tsan, V. Ho, B. Klotzle, J. M. Le, D. Delano, L. Zhang, G. P. Schroth, K. L. Gunderson, J. B. Fan, and R. Shen, "High density DNA methylation array with single CpG site resolution," *Genomics*, vol. 98, pp. 288–295, Oct 2011.

[93] R. A. Irizarry, C. Ladd-Acosta, B. Wen, Z. Wu, C. Montano, P. Onyango, H. Cui, K. Gabo, M. Rongione, M. Webster, H. Ji, J. Potash, S. Sabunciyan, and A. P. Feinberg, "The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores," *Nat Genet*, vol. 41, pp. 178–186, Feb 2009.

[94] A. Doi, I. H. Park, B. Wen, P. Murakami, M. J. Aryee, R. Irizarry, B. Herb, C. Ladd-Acosta, J. Rho, S. Loewer, J. Miller, T. Schlaeger, G. Q. Daley, and A. P. Feinberg, "Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts," *Nat Genet*, vol. 41, pp. 1350–1353, Dec 2009.

[95] M. Moarii, V. Boeva, J. P. Vert, and F. Reyal, "Changes in correlation between promoter methylation and gene expression in cancer," *BMC Genomics*, vol. 16, p. 873, Oct 2015.

[96] W. Xie, M. D. Schultz, R. Lister, Z. Hou, N. Rajagopal, P. Ray, J. W. Whitaker, S. Tian, R. D. Hawkins, D. Leung, H. Yang, T. Wang, A. Y. Lee, S. A. Swanson, J. Zhang, Y. Zhu, A. Kim, J. R. Nery, M. A. Urich, S. Kuan, C. A. Yen, S. Klugman, P. Yu, K. Suknuntha, N. E. Propson, H. Chen, L. E. Edsall, U. Wagner, Y. Li, Z. Ye, A. Kulkarni, Z. Xuan, W. Y. Chung, N. C. Chi, J. E. Antosiewicz-Bourget, I. Slukvin, R. Stewart, M. Q. Zhang, W. Wang, J. A. Thomson, J. R. Ecker, and B. Ren, "Epigenomic analysis of multilineage differentiation of human embryonic stem cells," *Cell*, vol. 153, pp. 1134–1148, May 2013.

[97] Y. Li, H. Zheng, Q. Wang, C. Zhou, L. Wei, X. Liu, W. Zhang, Y. Zhang, Z. Du, X. Wang, and W. Xie, "Genome-wide analyses reveal a role of Polycomb in promoting hypomethylation of DNA methylation valleys," *Genome Biol*, vol. 19, p. 18, Feb 2018.

[98] M. Jeong, D. Sun, M. Luo, Y. Huang, G. A. Challen, B. Rodriguez, X. Zhang, L. Chavez, H. Wang, R. Hannah, S. B. Kim, L. Yang, M. Ko, R. Chen, B. Göttgens, J. S. Lee, P. Gunaratne, L. A. Godley, G. J. Darlington, A. Rao, W. Li, and M. A. Goodell, "Large conserved domains of low DNA methylation maintained by Dnmt3a," *Nat Genet*, vol. 46, pp. 17–23, Jan 2014.

[99] S. Grosswendt, H. Kretzmer, Z. D. Smith, A. S. Kumar, S. Hetzel, L. Wittler, S. Klages, B. Timmermann, S. Mukherji, and A. Meissner, "Epigenetic regulator function through mouse gastrulation," *Nature*, vol. 584, pp. 102–108, Aug 2020.

[100] W. Ren, H. Fan, S. A. Grimm, Y. Guo, J. J. Kim, J. Yin, L. Li, C. J. Petell, X. F. Tan, Z. M. Zhang, J. P. Coan, L. Gao, L. Cai, B. Detrick, B. Çetin, Q. Cui, B. D. Strahl, O. Gozani, Y. Wang, K. M. Miller, S. E. O'Leary, P. A. Wade, D. J. Patel, G. G. Wang, and J. Song, "Direct readout of heterochromatic H3K9me3 regulates DNMT1-mediated maintenance DNA methylation," *Proc Natl Acad Sci U S A*, vol. 117, pp. 18439–18447, Aug 2020.

[101] W. Ren, H. Fan, S. A. Grimm, J. J. Kim, L. Li, Y. Guo, C. J. Petell, X. F. Tan, Z. M. Zhang, J. P. Coan, J. Yin, D. I. Kim, L. Gao, L. Cai, N. Khudaverdyan, B. Çetin, D. J. Patel, Y. Wang, Q. Cui, B. D. Strahl, O. Gozani, K. M. Miller, S. E. O'Leary, P. A. Wade, G. G. Wang, and J. Song, "DNMT1 reads heterochromatic H4K20me3 to reinforce LINE-1 DNA methylation," *Nat Commun*, vol. 12, p. 2490, May 2021.

[102] G. Rondelet, T. Dal Maso, L. Willems, and J. Wouters, "Structural basis for recognition of histone H3K36me3 nucleosome by human de novo DNA methyltransferases 3A and 3B," *J Struct Biol*, vol. 194, pp. 357–367, Jun 2016.

[103] T. Baubec, D. F. Colombo, C. Wirbelauer, J. Schmidt, L. Burger, A. R. Krebs, A. Akalin, and D. Schübeler, "Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation," *Nature*, vol. 520, pp. 243–247, Apr 2015.

[104] J. He, L. Shen, M. Wan, O. Taranova, H. Wu, and Y. Zhang, "Kdm2b maintains murine embryonic stem cell status by recruiting PRC1 complex to CpG islands of developmental genes," *Nat Cell Biol*, vol. 15, pp. 373–384, Apr 2013.

[105] N. P. Blackledge, A. M. Farcas, T. Kondo, H. W. King, J. F. McGouran, L. L. P. Hanssen, S. Ito, S. Cooper, K. Kondo, Y. Koseki, T. Ishikura, H. K. Long, T. W. Sheahan, N. Brockdorff, B. M. Kessler, H. Koseki, and R. J. Klose, "Variant PRC1 complex-dependent H2A ubiquitylation drives PRC2 recruitment and polycomb domain formation," *Cell*, vol. 157, pp. 1445–1459, Jun 2014.

[106] I. van Kruijsbergen, S. Hontelez, and G. J. Veenstra, "Recruiting polycomb to chromatin," *Int J Biochem Cell Biol*, vol. 67, pp. 177–187, Oct 2015.

[107] V. Parreno, A. M. Martinez, and G. Cavalli, "Mechanisms of Polycomb group protein function in cancer," *Cell Res*, vol. 32, pp. 231–253, Mar 2022.

[108] S. S. Nair and R. Kumar, "Chromatin remodeling in cancer: a gateway to regulate gene transcription," *Mol Oncol*, vol. 6, pp. 611–619, Dec 2012.

[109] J. A. Biegel, T. M. Busse, and B. E. Weissman, "SWI/SNF chromatin remodeling complexes and cancer," *Am J Med Genet C Semin Med Genet*, vol. 166C, pp. 350–366, Sep 2014.

[110] J. E. Audia and R. M. Campbell, "Histone Modifications and Cancer," *Cold Spring Harb Perspect Biol*, vol. 8, p. a019521, Apr 2016.

[111] Z. Zhao and A. Shilatifard, "Epigenetic modifications of histones in cancer," *Genome Biol*, vol. 20, p. 245, Nov 2019.

[112] S. Zhao, C. D. Allis, and G. G. Wang, "The language of chromatin modification in human cancers," *Nat Rev Cancer*, vol. 21, pp. 413–430, Jul 2021.

[113] D. N. Weinberg, C. D. Allis, and C. Lu, "Oncogenic Mechanisms of Histone H3 Mutations," *Cold Spring Harb Perspect Med*, vol. 7, Jan 2017.

[114] P. W. Lewis, M. M. Müller, M. S. Koletsky, F. Cordero, S. Lin, L. A. Banaszynski, B. A. Garcia, T. W. Muir, O. J. Becher, and C. D. Allis, "Inhibition of PRC2 activity by a gain-of-function H3 mutation found in pediatric glioblastoma," *Science*, vol. 340, pp. 857–861, May 2013.

[115] M. Ehrlich and R. Y. Wang, "5-Methylcytosine in eukaryotic DNA," *Science*, vol. 212, pp. 1350–1357, Jun 1981.

[116] A. P. Feinberg and B. Vogelstein, "Hypomethylation distinguishes genes of some human cancers from their normal counterparts," *Nature*, vol. 301, pp. 89–92, Jan 1983.

[117] M. A. Gama-Sosa, V. A. Slagel, R. W. Trewyn, R. Oxenhandler, K. C. Kuo, C. W. Gehrke, and M. Ehrlich, "The 5-methylcytosine content of DNA from human tumors," *Nucleic Acids Res*, vol. 11, pp. 6883–6894, Oct 1983.

[118] S. E. Goelz, B. Vogelstein, S. R. Hamilton, and A. P. Feinberg, "Hypomethylation of DNA from benign and malignant human colon neoplasms," *Science*, vol. 228, pp. 187–190, Apr 1985.

[119] K. D. Hansen, W. Timp, H. C. Bravo, S. Sabunciyan, B. Langmead, O. G. McDonald, B. Wen, H. Wu, Y. Liu, D. Diep, E. Briem, K. Zhang, R. A. Irizarry, and A. P. Feinberg, "Increased methylation variation in epigenetic domains across cancer types," *Nat Genet*, vol. 43, pp. 768–775, Jun 2011.

[120] B. P. Berman, D. J. Weisenberger, J. F. Aman, T. Hinoue, Z. Ramjan, Y. Liu, H. Noushmehr, C. P. Lange, C. M. van Dijk, R. A. Tollenaar, D. Van Den Berg, and P. W. Laird, "Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains," *Nat Genet*, vol. 44, pp. 40–46, Nov 2011.

[121] R. Lister, M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q. M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A. H. Millar, J. A. Thomson, B. Ren, and J. R. Ecker, "Human DNA methylomes at base resolution show widespread epigenomic differences," *Nature*, vol. 462, pp. 315–322, Nov 2009.

[122] S. E. Johnstone, A. Reyes, Y. Qi, C. Adriaens, E. Hegazi, K. Pelka, J. H. Chen, L. S. Zou, Y. Drier, V. Hecht, N. Shoresh, M. K. Selig, C. A. Lareau, S. Iyer, S. C. Nguyen, E. F. Joyce, N. Hacohen, R. A. Irizarry, B. Zhang, M. J. Aryee, and B. E. Bernstein, "Large-Scale Topological Changes Restrain Malignant Progression in Colorectal Cancer," *Cell*, vol. 182, pp. 1474–1489, Sep 2020.

[123] A. de Bustros, B. D. Nelkin, A. Silverman, G. Ehrlich, B. Poiesz, and S. B. Baylin, "The short arm of chromosome 11 is a "hot spot" for hypermethylation in human neoplasia," *Proc Natl Acad Sci U S A*, vol. 85, pp. 5693–5697, Aug 1988.

[124] M. Esteller, "CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future," *Oncogene*, vol. 21, pp. 5427–5440, Aug 2002.

[125] M. Widschwendter, H. Fiegl, D. Egle, E. Mueller-Holzner, G. Spizzo, C. Marth, D. J. Weisenberger, M. Campan, J. Young, I. Jacobs, and P. W. Laird, "Epigenetic stem cell signature in cancer," *Nat Genet*, vol. 39, pp. 157–158, Feb 2007.

[126] J. E. Ohm, K. M. McGarvey, X. Yu, L. Cheng, K. E. Schuebel, L. Cope, H. P. Mohammad, W. Chen, V. C. Daniel, W. Yu, D. M. Berman, T. Jenuwein, K. Pruitt, S. J. Sharkis, D. N. Watkins, J. G. Herman, and S. B. Baylin, "A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing," *Nat Genet*, vol. 39, pp. 237–242, Feb 2007.

[127] Y. Schlesinger, R. Straussman, I. Keshet, S. Farkash, M. Hecht, J. Zimmerman, E. Eden, Z. Yakhini, E. Ben-Shushan, B. E. Reubinoff, Y. Bergman, I. Simon, and H. Cedar, "Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer," *Nat Genet*, vol. 39, pp. 232–236, Feb 2007.

[128] D. Sproul and R. R. Meehan, "Genomic insights into cancer-associated aberrant CpG island hypermethylation," *Brief Funct Genomics*, vol. 12, pp. 174–190, May 2013.

[129] C. J. Lee, H. Ahn, D. Jeong, M. Pak, J. H. Moon, and S. Kim, "Impact of mutations in DNA methylation modification genes on genome-wide methylation landscapes and downstream gene activations in pan-cancer," *BMC Med Genomics*, vol. 13, p. 27, Feb 2020.

[130] G. Landan, N. M. Cohen, Z. Mukamel, A. Bar, A. Molchadsky, R. Brosh, S. Horn-Saban, D. A. Zalcenstein, N. Goldfinger, A. Zundelevich, E. N. Gal-Yam, V. Rotter, and A. Tanay, "Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues," *Nat Genet*, vol. 44, pp. 1207–1214, Nov 2012.

[131] D. A. Landau, K. Clement, M. J. Ziller, P. Boyle, J. Fan, H. Gu, K. Stevenson, C. Sougnez, L. Wang, S. Li, D. Kotliar, W. Zhang, M. Ghandi, L. Garraway, S. M. Fernandes, K. J. Livak, S. Gabriel, A. Gnirke, E. S. Lander, J. R. Brown, D. Neuberg, P. V. Kharchenko, N. Hacohen, G. Getz, A. Meissner, and C. J. Wu, "Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia," *Cancer Cell*, vol. 26, pp. 813–825, Dec 2014.

[132] H. Pan, Y. Jiang, M. Boi, F. Tabbò, D. Redmond, K. Nie, M. Ladetto, A. Chiappella, L. Cerchietti, R. Shaknovich, A. M. Melnick, G. G. Inghirami, W. Tam, and O. Elemento, "Epigenomic evolution in diffuse large B-cell lymphomas," *Nat Commun*, vol. 6, p. 6921, Apr 2015.

[133] Z. Shipony, Z. Mukamel, N. M. Cohen, G. Landan, E. Chomsky, S. R. Zeliger, Y. C. Fried, E. Ainbinder, N. Friedman, and A. Tanay, "Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells," *Nature*, vol. 513, pp. 115–119, Sep 2014.

[134] M. Toyota, N. Ahuja, M. Ohe-Toyota, J. G. Herman, S. B. Baylin, and J. P. Issa, "CpG island methylator phenotype in colorectal cancer," *Proc Natl Acad Sci U S A*, vol. 96, pp. 8681–8686, Jul 1999.

[135] D. J. Weisenberger, K. D. Siegmund, M. Campan, J. Young, T. I. Long, M. A. Faasse, G. H. Kang, M. Widschwendter, D. Weener, D. Buchanan, H. Koh, L. Simms, M. Barker, B. Leggett, J. Levine, M. Kim, A. J. French, S. N. Thibodeau, J. Jass, R. Haile, and P. W. Laird, "CpG island methylator phenotype underlies sporadic microsatellite instability and

is tightly associated with BRAF mutation in colorectal cancer," *Nat Genet*, vol. 38, pp. 787–793, Jul 2006.

[136] H. Noushmehr, D. J. Weisenberger, K. Diefes, H. S. Phillips, K. Pujara, B. P. Berman, F. Pan, C. E. Pelloski, E. P. Sulman, K. P. Bhat, R. G. Verhaak, K. A. Hoadley, D. N. Hayes, C. M. Perou, H. K. Schmidt, L. Ding, R. K. Wilson, D. Van Den Berg, H. Shen, H. Bengtsson, P. Neuvial, L. M. Cope, J. Buckley, J. G. Herman, S. B. Baylin, P. W. Laird, and K. Aldape, "Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma," *Cancer Cell*, vol. 17, pp. 510–522, May 2010.

[137] G. Garcia-Manero, J. Daniel, T. L. Smith, S. M. Kornblau, M. S. Lee, H. M. Kantarjian, and J. P. Issa, "DNA methylation of multiple promoter-associated CpG islands in adult acute lymphocytic leukemia," *Clin Cancer Res*, vol. 8, pp. 2217–2224, Jul 2002.

[138] A. D. Kelly, H. Kroeger, J. Yamazaki, R. Taby, F. Neumann, S. Yu, J. T. Lee, B. Patel, Y. Li, R. He, S. Liang, Y. Lu, M. Cesaroni, S. A. Pierce, S. M. Kornblau, C. E. Bueso-Ramos, F. Ravandi, H. M. Kantarjian, J. Jelinek, and J. P. Issa, "A CpG island methylator phenotype in acute myeloid leukemia independent of IDH mutations and associated with a favorable outcome," *Leukemia*, vol. 31, pp. 2011–2019, Oct 2017.

[139] A. Tanemura, A. M. Terando, M. S. Sim, A. Q. van Hoesel, M. F. de Maat, D. L. Morton, and D. S. Hoon, "CpG island methylator phenotype predicts progression of malignant melanoma," *Clin Cancer Res*, vol. 15, pp. 1801–1807, Mar 2009.

[140] B. P. Whitcomb, D. G. Mutch, T. J. Herzog, J. S. Rader, R. K. Gibb, and P. J. Goodfellow, "Frequent HOXA11 and THBS2 promoter methylation, and a methylator phenotype in endometrial adenocarcinoma," *Clin Cancer Res*, vol. 9, pp. 2277–2287, Jun 2003.

[141] F. Fang, S. Turcan, A. Rimner, A. Kaufman, D. Giri, L. G. Morris, R. Shen, V. Seshan, Q. Mo, A. Heguy, S. B. Baylin, N. Ahuja, A. Viale, J. Massague, L. Norton, L. T. Vahdat, M. E. Moynahan, and T. A. Chan, "Breast cancer methylomes establish an epigenomic foundation for metastasis," *Sci Transl Med*, vol. 3, p. 75ra25, Mar 2011.

[142] L. A. Hughes, V. Melotte, J. de Schrijver, M. de Maat, V. T. Smit, J. V. Bovée, P. J. French, P. A. van den Brandt, L. J. Schouten, T. de Meyer, W. van Criekinge, N. Ahuja, J. G. Herman, M. P. Weijenberg, and M. van Engeland, "The CpG island methylator phenotype: what's in a name?," *Cancer Res*, vol. 73, pp. 5858–5868, Oct 2013.

[143] S. Turcan, D. Rohle, A. Goenka, L. A. Walsh, F. Fang, E. Yilmaz, C. Campos, A. W. Fabius, C. Lu, P. S. Ward, C. B. Thompson, A. Kaufman, O. Guryanova, R. Levine, A. Heguy, A. Viale, L. G. Morris, J. T. Huse, I. K. Mellinghoff, and T. A. Chan, "IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype," *Nature*, vol. 483, pp. 479–483, Feb 2012.

[144] J. Su, Y. H. Huang, X. Cui, X. Wang, X. Zhang, Y. Lei, J. Xu, X. Lin, K. Chen, J. Lv, M. A. Goodell, and W. Li, "Homeobox oncogene activation by pan-cancer DNA hypermethylation," *Genome Biol*, vol. 19, p. 108, Aug 2018.

[145] P. Mirabelli, L. Coppola, and M. Salvatore, "Cancer Cell Lines Are Useful Model Systems for Medical Research," *Cancers (Basel)*, vol. 11, Aug 2019.

[146] D. J. Smiraglia, L. J. Rush, M. C. Frühwald, Z. Dai, W. A. Held, J. F. Costello, J. C. Lang, C. Eng, B. Li, F. A. Wright, M. A. Caligiuri, and C. Plass, "Excessive CpG island hypermethylation in cancer cell lines versus primary human malignancies," *Hum Mol Genet*, vol. 10, pp. 1413–1419, Jun 2001.

[147] M. F. Paz, M. F. Fraga, S. Avila, M. Guo, M. Pollan, J. G. Herman, and M. Esteller, "A systematic profile of DNA methylation in human cancer cell lines," *Cancer Res*, vol. 63, pp. 1114–1121, Mar 2003.

[148] R. Kempfer and A. Pombo, "Methods for mapping 3D chromosome architecture," *Nat Rev Genet*, vol. 21, pp. 207–226, Apr 2020.

[149] S. A. Quinodoz, N. Ollikainen, B. Tabak, A. Palla, J. M. Schmidt, E. Detmar, M. M. Lai, A. A. Shishkin, P. Bhat, Y. Takei, V. Trinh, E. Aznauryan, P. Russell, C. Cheng, M. Jovanovic, A. Chow, L. Cai, P. McDonel, M. Garber, and M. Guttman, "Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus," *Cell*, vol. 174, pp. 744–757, Jul 2018.

[150] P. J. Park, "ChIP-seq: advantages and challenges of a maturing technology," *Nat Rev Genet*, vol. 10, pp. 669–680, Oct 2009.

[151] H. S. Kaya-Okur, S. J. Wu, C. A. Codomo, E. S. Pledger, T. D. Bryson, J. G. Henikoff, K. Ahmad, and S. Henikoff, "CUT&Tag for efficient epigenomic profiling of small samples and single cells," *Nat Commun*, vol. 10, p. 1930, Apr 2019.

[152] P. J. Skene and S. Henikoff, "An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites," *Elife*, vol. 6, Jan 2017.

[153] S. P. Thomas, S. A. Haws, L. E. Borth, and J. M. Denu, "A practical guide for analysis of histone post-translational modifications by mass spectrometry: Best practices and pitfalls," *Methods*, vol. 184, pp. 53–60, Dec 2020.

[154] B. D. Singer, "A Practical Guide to the Measurement and Analysis of DNA Methylation," *Am J Respir Cell Mol Biol*, vol. 61, pp. 417–428, Oct 2019.

[155] B. Khulan, R. F. Thompson, K. Ye, M. J. Fazzari, M. Suzuki, E. Stasiek, M. E. Figueroa, J. L. Glass, Q. Chen, C. Montagna, E. Hatchwell, R. R. Selzer, T. A. Richmond, R. D. Green, A. Melnick, and J. M. Greally, "Comparative isoschizomer profiling of cytosine methylation: the HELP assay," *Genome Res*, vol. 16, pp. 1046–1055, Aug 2006.

[156] T. Zuo, B. Tycko, T. M. Liu, J. J. Lin, and T. H. Huang, "Methods in DNA methylation profiling," *Epigenomics*, vol. 1, pp. 331–345, Dec 2009.

[157] M. Frommer, L. E. McDonald, D. S. Millar, C. M. Collis, F. Watt, G. W. Grigg, P. L. Molloy, and C. L. Paul, "A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands," *Proc Natl Acad Sci U S A*, vol. 89, pp. 1827–1831, Mar 1992.

[158] J. T. Simpson, R. E. Workman, P. C. Zuzarte, M. David, L. J. Dursi, and W. Timp, "Detecting DNA cytosine methylation using nanopore sequencing," *Nat Methods*, vol. 14, pp. 407–410, Apr 2017.

[159] M. Jain, S. Koren, K. H. Miga, J. Quick, A. C. Rand, T. A. Sasani, J. R. Tyson, A. D. Beggs, A. T. Dilthey, I. T. Fiddes, S. Malla, H. Marriott, T. Nieto, J. O'Grady, H. E. Olsen, B. S. Pedersen, A. Rhie, H. Richardson, A. R. Quinlan, T. P. Snutch, L. Tee, B. Paten, A. M. Phillippy, J. T. Simpson, N. J. Loman, and M. Loose, "Nanopore sequencing and assembly of a human genome with ultra-long reads," *Nat Biotechnol*, vol. 36, pp. 338–345, Apr 2018.

[160] W. Zhou, P. W. Laird, and H. Shen, "Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes," *Nucleic Acids Res*, vol. 45, p. e22, Feb 2017.

[161] F. Eckhardt, J. Lewin, R. Cortese, V. K. Rakyan, J. Attwood, M. Burger, J. Burton, T. V. Cox, R. Davies, T. A. Down, C. Haefliger, R. Horton, K. Howe, D. K. Jackson, J. Kunde, C. Koenig, J. Liddle, D. Niblett, T. Otto, R. Pettett, S. Seemann, C. Thompson, T. West, J. Rogers, A. Olek, K. Berlin, and S. Beck, "DNA methylation profiling of human chromosomes 6, 20 and 22," *Nat Genet*, vol. 38, pp. 1378–1385, Dec 2006.

[162] F. Miura, Y. Enomoto, R. Dairiki, and T. Ito, "Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging," *Nucleic Acids Res*, vol. 40, p. e136, Sep 2012.

[163] N. Olova, F. Krueger, S. Andrews, D. Oxley, R. V. Berrens, M. R. Branco, and W. Reik, "Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data," *Genome Biol*, vol. 19, p. 33, Mar 2018.

[164] A. Raine, U. Liljedahl, and J. Nordlund, "Data quality of whole genome bisulfite sequencing on Illumina platforms," *PLoS One*, vol. 13, p. e0195972, Apr 2018.

[165] S. S. Nair, P. L. Luu, W. Qu, M. Maddugoda, L. Huschtscha, R. Reddel, G. Chenevix-Trench, M. Toso, J. G. Kench, L. G. Horvath, V. M. Hayes, P. D. Stricker, T. P. Hughes, D. L. White, J. E. J. Rasko, J. J. Wong, and S. J. Clark, "Guidelines for whole genome bisulphite sequencing of intact and FFPET DNA on the Illumina HiSeq X Ten," *Epigenetics Chromatin*, vol. 11, p. 24, May 2018.

[166] Z. Sun, J. Cunningham, S. Slager, and J. P. Kocher, "Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis," *Epigenomics*, vol. 7, pp. 813–828, Aug 2015.

[167] Babraham Bioinformatics, "Reduced Representation Bisulfite-Seq - A brief guide to RRBS." https://www.bioinformatics.babraham.ac.uk/projects/bismark/RRBS_Guide.pdf, 2013. Accessed: 2022-07-02.

[168] NuGEN, "Analysis Guide for NuGEN Ovation RRBS Methyl-Seq." https://github.com/nugentechnologies/NuMetRRBS, 2022.

[169] R. Lister and J. R. Ecker, "Finding the fifth base: genome-wide sequencing of cytosine methylation," *Genome Res*, vol. 19, pp. 959–966, Jun 2009.

[170] J. Ahn, S. Heo, J. Lee, and D. Bang, "Introduction to Single-Cell DNA Methylation Profiling Methods," *Biomolecules*, vol. 11, Jul 2021.

[171] S. Andrews, "FastQC: A Quality Control Tool for High Throughput Sequence Data." https://www.bioinformatics.babraham.ac.uk/projects/fastqc, 2010.

[172] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet.journal*, vol. 17, pp. 10–12, May 2011.

[173] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," *Bioinformatics*, vol. 30, pp. 2114–2120, Aug 2014.

[174] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nat Methods*, vol. 9, pp. 357–359, Mar 2012.

[175] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, pp. 1754–1760, Jul 2009.

[176] R. Li, Y. Li, K. Kristiansen, and J. Wang, "SOAP: short oligonucleotide alignment program," *Bioinformatics*, vol. 24, pp. 713–714, Mar 2008.

[177] C. Grehl, M. Wagner, I. Lemnian, B. Glaser, and I. Grosse, "Performance of Mapping Approaches for Whole-Genome Bisulfite Sequencing Data in Crop Plants," *Front Plant Sci*, vol. 11, p. 176, Feb 2020.

[178] G. Kunde-Ramamoorthy, C. Coarfa, E. Laritsky, N. J. Kessler, R. A. Harris, M. Xu, R. Chen, L. Shen, A. Milosavljevic, and R. A. Waterland, "Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing," *Nucleic Acids Res*, vol. 42, p. e43, Apr 2014.

[179] C. Otto, P. F. Stadler, and S. Hoffmann, "Fast and sensitive mapping of bisulfite-treated sequencing data," *Bioinformatics*, vol. 28, pp. 1698–1704, Jul 2012.

[180] F. Krueger and S. R. Andrews, "Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications," *Bioinformatics*, vol. 27, pp. 1571–1572, Jun 2011.

[181] D. Kim, B. Langmead, and S. L. Salzberg, "HISAT: a fast spliced aligner with low memory requirements," *Nat Methods*, vol. 12, pp. 357–360, Apr 2015.

[182] W. Guo, P. Fiziev, W. Yan, S. Cokus, X. Sun, M. Q. Zhang, P. Y. Chen, and M. Pellegrini, "BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data," *BMC Genomics*, vol. 14, p. 774, Nov 2013.

[183] Y. Xi and W. Li, "BSMAP:molecules whole genome bisulfite sequence MAPping program," *BMC Bioinformatics*, vol. 10, p. 232, Jul 2009.

[184] B. S. Pedersen, K. Eyring, S. De, I. V. Yang, and D. A. Schwartz, "Fast and accurate alignment of long bisulfite-seq reads," May 2014.

[185] S. Marco-Sola, M. Sammeth, R. Guigó, and P. Ribeca, "The GEM mapper: fast, accurate and versatile alignment by filtration," *Nat Methods*, vol. 9, pp. 1185–1188, Dec 2012.

[186] A. Merkel, M. Fernández-Callejo, E. Casals, S. Marco-Sola, R. Schuyler, I. G. Gut, and S. C. Heath, "gemBS: high throughput processing for DNA methylation data from bisulfite sequencing," *Bioinformatics*, vol. 35, pp. 737–742, Mar 2019.

[187] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo, "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data," *Genome Res*, vol. 20, pp. 1297–1303, Sep 2010.

[188] D. Sun, Y. Xi, B. Rodriguez, H. J. Park, P. Tong, M. Meong, M. A. Goodell, and W. Li, "MOABS: model based analysis of bisulfite sequencing data," *Genome Biol*, vol. 15, p. R38, Feb 2014.

[189] D. Ryan, "MethylDackel." https://github.com/dpryan79/MethylDackel, 2022.

[190] H. Hauswedell, J. Singer, and K. Reinert, "Lambda: the local aligner for massive biological data," *Bioinformatics*, vol. 30, pp. i349–355, Sep 2014.

[191] H. P. Hauswedell, *SeqAn3 – Sequence Analysis and Modern C++*. Dissertation, Freie Universität Berlin, Jun 2021.

[192] K. Reinert, B. Langmead, D. Weese, and D. J. Evers, "Alignment of Next-Generation Sequencing Reads," *Annu Rev Genomics Hum Genet*, vol. 16, pp. 133–151, May 2015.

[193] W. R. Pearson, "An introduction to sequence similarity ("homology") searching," *Curr Protoc Bioinformatics*, vol. Chapter 3, pp. 1–3, Jun 2013.

[194] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J Mol Biol*, vol. 215, pp. 403–410, Oct 1990.

[195] E. Project Consortium, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, pp. 57–74, Sep 2012.

[196] A. J. Vågene, A. Herbig, M. G. Campana, N. M. Robles García, C. Warinner, S. Sabin, M. A. Spyrou, A. Andrades Valtueña, D. Huson, N. Tuross, K. I. Bos, and J. Krause, "Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico," *Nat Ecol Evol*, vol. 2, pp. 520–528, Mar 2018.

[197] B. Buchfink, C. Xie, and D. H. Huson, "Fast and sensitive protein alignment using DIAMOND," *Nat Methods*, vol. 12, pp. 59–60, Jan 2015.

[198] B. Buchfink, K. Reuter, and H. G. Drost, "Sensitive protein alignments at tree-of-life scale using DIAMOND," *Nat Methods*, vol. 18, pp. 366–368, Apr 2021.

[199] C. Legendre, G. C. Gooden, K. Johnson, R. A. Martinez, W. S. Liang, and B. Salhia, "Whole-genome bisulfite sequencing of cell-free DNA identifies signature associated with metastatic breast cancer," *Clin Epigenetics*, vol. 7, p. 100, Sep 2015.

[200] M. L. Stackpole, W. Zeng, S. Li, C. C. Liu, Y. Zhou, S. He, A. Yeh, Z. Wang, F. Sun, Q. Li, Z. Yuan, A. Yildirim, P. J. Chen, P. Winograd, B. Tran, Y. T. Lee, P. S. Li, Z. Noor, M. Yokomizo, P. Ahuja, Y. Zhu, H. R. Tseng, J. S. Tomlinson, E. Garon, S. French, C. E. Magyar, S. Dry, C. Lajonchere, D. Geschwind, G. Choi, S. Saab, F. Alber, W. H. Wong, S. M. Dubinett, D. R. Aberle, V. Agopian, S. B. Han, X. Ni, W. Li, and X. J. Zhou, "Cost-effective methylome sequencing of cell-free DNA for accurately detecting and locating cancer," *Nat Commun*, vol. 13, p. 5566, Sep 2022.

[201] S. L. R. Barrett, E. A. Holmes, D. R. Long, R. C. Shean, G. E. Bautista, S. Ravishankar, V. Peddu, B. T. Cookson, P. K. Singh, A. L. Greninger, and S. J. Salipante, "Cell free DNA from respiratory pathogens is detectable in the blood plasma of Cystic Fibrosis patients," *Sci Rep*, vol. 10, p. 6903, Apr 2020.

[202] M. Kowarsky, J. Camunas-Soler, M. Kertesz, I. De Vlaminck, W. Koh, W. Pan, L. Martin, N. F. Neff, J. Okamoto, R. J. Wong, S. Kharbanda, Y. El-Sayed, Y. Blumenfeld, D. K. Stevenson, G. M. Shaw, N. D. Wolfe, and S. R. Quake, "Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA," *Proc Natl Acad Sci U S A*, vol. 114, pp. 9623–9628, Sep 2017.

[203] A. P. Cheng, P. Burnham, J. R. Lee, M. P. Cheng, M. Suthanthiran, D. Dadhania, and I. De Vlaminck, "A cell-free DNA metagenomic sequencing assay that integrates the host injury response to infection," *Proc Natl Acad Sci U S A*, vol. 116, pp. 18738–18744, Sep 2019.

[204] T. Li, K. Fan, J. Wang, and W. Wang, "Reduction of protein sequence complexity by residue grouping," *Protein Eng*, vol. 16, pp. 323–330, May 2003.

[205] L. R. Murphy, A. Wallqvist, and R. M. Levy, "Simplified amino acid alphabets for protein fold recognition and implications for folding," *Protein Eng*, vol. 13, pp. 149–152, Mar 2000.

[206] H. Hauswedell, "BioC++." https://github.com/biocpp, 2023.

[207] S. Gottlieb, "FMIndex-Collection." https://github.com/SGSSGene/fmindex-collection, 2023.

[208] W. S. Grant and R. Voorhies, "cereal - A C++11 library for serialization." https://github.com/USCiLab/cereal, 2023.

[209] K. Reinert, T. H. Dadi, M. Ehrhardt, H. Hauswedell, S. Mehringer, R. Rahn, J. Kim, C. Pockrandt, J. Winkler, E. Siragusa, G. Urgese, and D. Weese, "The SeqAn C++ template library for efficient sequence analysis: A resource for programmers," *J Biotechnol*, vol. 261, pp. 157–168, Nov 2017.

[210] R. Rahn, S. Budach, P. Costanza, M. Ehrhardt, J. Hancox, and K. Reinert, "Generic accelerated sequence alignment in SeqAn using vectorization and multi-threading," *Bioinformatics*, vol. 34, pp. 3437–3445, Oct 2018.

[211] S. Karlin and S. F. Altschul, "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes," *Proc Natl Acad Sci U S A*, vol. 87, pp. 2264–2268, Mar 1990.

[212] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, pp. 2078–2079, Aug 2009.

[213] F. Meyer, A. Fritz, Z. L. Deng, D. Koslicki, T. R. Lesker, A. Gurevich, G. Robertson, M. Alser, D. Antipov, F. Beghini, D. Bertrand, J. J. Brito, C. T. Brown, J. Buchmann, A. ç, B. Chen, R. Chikhi, P. T. L. C. Clausen, A. Cristian, P. W. Dabrowski, A. E. Darling, R. Egan, E. Es-

kin, E. Georganas, E. Goltsman, M. A. Gray, L. H. Hansen, S. Hofmeyr, P. Huang, L. Irber, H. Jia, T. S. rgensen, S. D. Kieser, T. Klemetsen, A. Kola, M. Kolmogorov, A. Korobeynikov, J. Kwan, N. LaPierre, C. Lemaitre, C. Li, A. Limasset, F. Malcher-Miranda, S. Mangul, V. R. Marcelino, C. Marchet, P. Marijon, D. Meleshko, D. R. Mende, A. Milanese, N. Nagarajan, J. Nissen, S. Nurk, L. Oliker, L. Paoli, P. Peterlongo, V. C. Piro, J. S. Porter, S. Rasmussen, E. R. Rees, K. Reinert, B. Renard, E. M. Robertsen, G. L. Rosen, H. J. Ruscheweyh, V. Sarwal, N. Segata, E. Seiler, L. Shi, F. Sun, S. Sunagawa, S. J. rensen, A. Thomas, C. Tong, M. Trajkovski, J. Tremblay, G. Uritskiy, R. Vicedomini, Z. Wang, Z. Wang, Z. Wang, A. Warren, N. P. Willassen, K. Yelick, R. You, G. Zeller, Z. Zhao, S. Zhu, J. Zhu, R. Garrido-Oter, P. Gastmeier, S. Hacquard, S. ler, A. Khaledi, F. Maechler, F. Mesny, S. Radutoiu, P. Schulze-Lefert, N. Smit, T. Strowig, A. Bremges, A. Sczyrba, and A. C. McHardy, "Critical Assessment of Metagenome Interpretation: the second round of challenges," *Nat Methods*, vol. 19, pp. 429–440, Apr 2022.

[214] M. Bahram, F. Hildebrand, S. K. Forslund, J. L. Anderson, N. A. Soudzilovskaia, P. M. Bodegom, J. Bengtsson-Palme, S. Anslan, L. P. Coelho, H. Harend, J. Huerta-Cepas, M. H. Medema, M. R. Maltz, S. Mundra, P. A. Olsson, M. Pent, S. lme, S. Sunagawa, M. Ryberg, L. Tedersoo, and P. Bork, "Structure and function of the global topsoil microbiome," *Nature*, vol. 560, pp. 233–237, Aug 2018.

[215] A. Nunn, C. Otto, P. F. Stadler, and D. Langenberger, "Comprehensive benchmarking of software for mapping whole genome bisulfite data: from read alignment to DNA methylation analysis," *Brief Bioinform*, vol. 22, Sep 2021.

[216] L. Liu, J. Feng, J. Polimeni, M. Zhang, H. Nguyen, U. Das, X. Zhang, H. Singh, X.-J. Yao, E. Leygue, S. K. P. Kung, and J. Xie, "Characterization of cell free plasma methyl-DNA from xenografted tumors to guide the selection of diagnostic markers for early-stage cancers," *Frontiers in Oncology*, vol. 11, Mar. 2021.

[217] A. J. Bewick, B. T. Hofmeister, R. A. Powers, S. J. Mondo, I. V. Grigoriev, T. Y. James, J. E. Stajich, and R. J. Schmitz, "Diversity of cytosine methylation across the fungal tree of life," *Nat Ecol Evol*, vol. 3, pp. 479–490, Mar 2019.

[218] Human Microbiome Project Consortium, "A framework for human microbiome research," *Nature*, vol. 486, pp. 215–221, Jun 2012.

[219] D. H. Huson and C. Xie, "A poor man's BLASTX–high-throughput metagenomic protein database search using PAUDA," *Bioinformatics*, vol. 30, pp. 38–39, Jan 2014.

[220] Y. Ye, J. H. Choi, and H. Tang, "RAPSearch: a fast protein similarity search tool for short reads," *BMC Bioinformatics*, vol. 12, p. 159, May 2011.

[221] Y. T. Huang, P. Y. Liu, and P. W. Shih, "Homopolish: a method for the removal of systematic errors in nanopore sequencing by homologous polishing," *Genome Biol*, vol. 22, p. 95, Mar 2021.

[222] S. Hetzel, P. Giesselmann, K. Reinert, A. Meissner, and H. Kretzmer, "RLM: Fast and simplified extraction of Read-Level Methylation metrics from bisulfite sequencing data," *Bioinformatics*, Oct 2021.

[223] M. Scherer, A. Nebel, A. Franke, J. Walter, T. Lengauer, C. Bock, F. Müller, and M. List, "Quantitative comparison of within-sample heterogeneity scores for DNA methylation data," *Nucleic Acids Res*, vol. 48, p. e46, May 2020.

[224] G. Elliott, C. Hong, X. Xing, X. Zhou, D. Li, C. Coarfa, R. J. Bell, C. L. Maire, K. L. Ligon, M. Sigaroudinia, P. Gascard, T. D. Tlsty, R. A. Harris, L. C. Schalkwyk, M. Bilenky, J. Mill, P. J. Farnham, M. Kellis, M. A. Marra, A. Milosavljevic, M. Hirst, G. D. Stormo, T. Wang, and J. F. Costello, "Intermediate DNA methylation is a conserved signature of genome regulation," *Nat Commun*, vol. 6, p. 6363, Feb 2015.

[225] E. A. Houseman, W. P. Accomando, D. C. Koestler, B. C. Christensen, C. J. Marsit, H. H. Nelson, J. K. Wiencke, and K. T. Kelsey, "DNA methylation arrays as surrogate measures of cell mixture distribution," *BMC Bioinformatics*, vol. 13, p. 86, May 2012.

[226] E. A. Houseman, M. L. Kile, D. C. Christiani, T. A. Ince, K. T. Kelsey, and C. J. Marsit, "Reference-free deconvolution of DNA methylation data and mediation by cell composition effects," *BMC Bioinformatics*, vol. 17, p. 259, Jun 2016.

[227] J. Charlton, T. L. Downing, Z. D. Smith, H. Gu, K. Clement, R. Pop, V. Akopian, S. Klages, D. P. Santos, A. M. Tsankov, B. Timmermann, M. J. Ziller, E. Kiskinis, A. Gnirke, and A. Meissner, "Global delay in nascent strand DNA methylation," *Nat Struct Mol Biol*, vol. 25, pp. 327–332, Apr 2018.

[228] H. Xie, M. Wang, A. de Andrade, M. d. e. F. Bonaldo, V. Galat, K. Arndt, V. Rajaram, S. Goldman, T. Tomita, and M. B. Soares, "Genome-wide quantitative assessment of variation in DNA methylation patterns," *Nucleic Acids Res*, vol. 39, pp. 4099–4108, May 2011.

[229] A. Jeltsch and R. Z. Jurkowska, "Allosteric control of mammalian DNA methyltransferases - a new regulatory paradigm," *Nucleic Acids Res*, vol. 44, pp. 8556–8575, Oct 2016.

[230] S. Guo, D. Diep, N. Plongthongkum, H. L. Fung, K. Zhang, and K. Zhang, "Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA," *Nat Genet*, vol. 49, pp. 635–642, Apr 2017.

[231] "SeqAn3 – the modern C++ library for sequence analysis." https://github.com/seqan/seqan3, 2021.

[232] H. Thórvaldsdottir, J. T. Robinson, and J. P. Mesirov, "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration," *Brief Bioinform*, vol. 14, pp. 178–192, Mar 2013.

[233] W. J. Kent, A. S. Zweig, G. Barber, A. S. Hinrichs, and D. Karolchik, "BigWig and BigBed: enabling browsing of large distributed datasets," *Bioinformatics*, vol. 26, pp. 2204–2207, Sep 2010.

[234] J. He, X. Sun, X. Shao, L. Liang, and H. Xie, "DMEAS: DNA methylation entropy analysis software," *Bioinformatics*, vol. 29, pp. 2044–2045, Aug 2013.

[235] C. A. Scott, J. D. Duryea, H. MacKay, M. S. Baker, E. Laritsky, C. J. Gunasekara, C. Coarfa, and R. A. Waterland, "Identification of cell type-specific methylation signals in bulk whole genome bisulfite sequencing data," *Genome Biol*, vol. 21, p. 156, Jul 2020.

[236] M. Scherer, "WSH package vignette." https://github.com/MPIIComputationalEpigenetics/WSHPackage/blob/master/vignettes/WSH.md, 2021.

[237] J. Charlton, E. J. Jung, A. L. Mattei, N. Bailly, J. Liao, E. J. Martin, P. Giesselmann, B. Brändl, E. K. Stamenova, F. J. Müller, E. Kiskinis, A. Gnirke, Z. D. Smith, and A. Meissner, "TETs compete with DNMT3 activity in pluripotent cells at thousands of methylated somatic enhancers," *Nat Genet*, vol. 52, pp. 819–827, Aug 2020.

[238] S. Hetzel, A. L. Mattei, H. Kretzmer, C. Qu, X. Chen, Y. Fan, G. Wu, K. G. Roberts, S. Luger, M. Litzow, J. Rowe, E. Paietta, W. Stock, E. R. Mardis, R. K. Wilson, J. R. Downing, C. G. Mullighan, and A. Meissner, "Acute lymphoblastic leukemia displays a distinct highly methylated genome," *Nat Cancer*, May 2022.

[239] T. H. Tran and S. P. Hunger, "The genomic landscape of pediatric acute lymphoblastic leukemia and precision medicine opportunities," *Semin Cancer Biol*, Nov 2020.

[240] M. L. Loh and C. G. Mullighan, "Advances in the genetics of high-risk childhood B-progenitor acute lymphoblastic leukemia and juvenile myelomonocytic leukemia: implications for therapy," *Clin Cancer Res*, vol. 18, pp. 2754–2767, May 2012.

[241] F. Malard and M. Mohty, "Acute lymphoblastic leukaemia," *Lancet*, vol. 395, pp. 1146–1162, Apr 2020.

[242] National Cancer Institute, "Childhood Acute Lymphoblastic Leukemia Treatment." https://www.cancer.gov/types/leukemia/patient/child-all-treatment-pdq. Accessed: 2022-07-18.

[243] Z. Gu, M. L. Churchman, K. G. Roberts, I. Moore, X. Zhou, J. Nakitandwe, K. Hagiwara, S. Pelletier, S. Gingras, H. Berns, D. Payne-Turner, A. Hill, I. Iacobucci, L. Shi, S. Pounds, C. Cheng, D. Pei, C. Qu, S. Newman, M. Devidas, Y. Dai, S. C. Reshmi, J. Gastier-Foster, E. A. Raetz, M. J. Borowitz, B. L. Wood, W. L. Carroll, P. A. Zweidler-McKay, K. R. Rabin, L. A. Mattano, K. W. Maloney, A. Rambaldi, O. Spinelli, J. P. Radich, M. D. Minden, J. M. Rowe, S. Luger, M. R. Litzow, M. S. Tallman, J. Racevskis, Y. Zhang, R. Bhatia, J. Kohlschmidt, K. Mrózek, C. D. Bloomfield, W. Stock, S. Kornblau, H. M. Kantarjian, M. Konopleva, W. E. Evans, S. Jeha, C. H. Pui, J. Yang, E. Paietta, J. R. Downing, M. V. Relling, J. Zhang, M. L. Loh, S. P. Hunger, and C. G. Mullighan, "PAX5-driven subtypes of B-progenitor acute lymphoblastic leukemia," *Nat Genet*, vol. 51, pp. 296–307, Feb 2019.

[244] L. Belver and A. Ferrando, "The genetics and mechanisms of T cell acute lymphoblastic leukaemia," *Nat Rev Cancer*, vol. 16, pp. 494–507, Jul 2016.

[245] J. Nordlund and A. C. Syvänen, "Epigenetics in pediatric acute lymphoblastic leukemia," *Semin Cancer Biol*, vol. 51, pp. 129–138, Aug 2018.

[246] K. H. Taylor, K. E. Pena-Hernandez, J. W. Davis, G. L. Arthur, D. J. Duff, H. Shi, F. B. Rahmatpanah, O. Sjahputera, and C. W. Caldwell, "Large-scale CpG methylation analysis

identifies novel candidate genes and reveals methylation hotspots in acute lymphoblastic leukemia," *Cancer Res*, vol. 67, pp. 2617–2625, Mar 2007.

[247] L. Milani, A. Lundmark, J. Nordlund, A. Kiialainen, T. Flaegstad, G. Jonmundsson, J. Kanerva, K. Schmiegelow, K. L. Gunderson, G. Lönnerholm, and A. C. Syvänen, "Allele-specific gene expression patterns in primary leukemic cells reveal regulation of gene expression by CpG site methylation," *Genome Res*, vol. 19, pp. 1–11, Jan 2009.

[248] J. Davidsson, H. Lilljebjörn, A. Andersson, S. Veerla, J. Heldrup, M. Behrendtz, T. Fioretos, and B. Johansson, "The DNA methylome of pediatric acute lymphoblastic leukemia," *Hum Mol Genet*, vol. 18, pp. 4054–4065, Nov 2009.

[249] D. J. Stumpel, P. Schneider, E. H. van Roon, J. M. Boer, P. de Lorenzo, M. G. Valsecchi, R. X. de Menezes, R. Pieters, and R. W. Stam, "Specific promoter methylation identifies different subgroups of MLL-rearranged infant acute lymphoblastic leukemia, influences clinical outcome, and provides therapeutic options," *Blood*, vol. 114, pp. 5490–5498, Dec 2009.

[250] M. E. Figueroa, S. C. Chen, A. K. Andersson, L. A. Phillips, Y. Li, J. Sotzen, M. Kundu, J. R. Downing, A. Melnick, and C. G. Mullighan, "Integrated genetic and epigenetic analysis of childhood acute lymphoblastic leukemia," *J Clin Invest*, vol. 123, pp. 3099–3111, Jul 2013.

[251] P. Wahlberg, A. Lundmark, J. Nordlund, S. Busche, A. Raine, K. Tandre, L. Rönnblom, D. Sinnett, E. Forestier, T. Pastinen, G. Lönnerholm, and A. C. Syvänen, "DNA methylome analysis of acute lymphoblastic leukemia cells reveals stochastic de novo DNA methylation in CpG islands," *Epigenomics*, vol. 8, pp. 1367–1387, Oct 2016.

[252] S. T. Lee, M. O. Muench, M. E. Fomin, J. Xiao, M. Zhou, A. de Smith, J. I. Martín-Subero, S. Heath, E. A. Houseman, R. Roy, M. Wrensch, J. Wiencke, C. Metayer, and J. L. Wiemels, "Epigenetic remodeling in B-cell acute lymphoblastic leukemia occurs in two tracks and employs embryonic stem cell-like signatures," *Nucleic Acids Res*, vol. 43, pp. 2590–2602, Mar 2015.

[253] M. Almamun, B. T. Levinson, A. C. van Swaay, N. T. Johnson, S. D. McKay, G. L. Arthur, J. W. Davis, and K. H. Taylor, "Integrated methylome and transcriptome analysis reveals novel regulatory elements in pediatric acute lymphoblastic leukemia," *Epigenetics*, vol. 10, pp. 882–890, Aug 2015.

[254] Z. Haider, P. Larsson, M. Landfors, L. Köhn, K. Schmiegelow, T. Flaegstad, J. Kanerva, M. Heyman, M. Hultdin, and S. Degerman, "An integrated transcriptome analysis in T-cell acute lymphoblastic leukemia links DNA methylation subgroups to dysregulated TAL1 and ANTP homeobox gene expression," *Cancer Med*, vol. 8, pp. 311–324, Jan 2019.

[255] J. Roels, M. Thénoz, B. Szarzyńska, M. Landfors, S. De Coninck, L. Demoen, L. Provez, A. Kuchmiy, S. Strubbe, L. Reunes, T. Pieters, F. Matthijssens, W. Van Loocke, B. Erarslan-Uysal, P. Richter-Pechańska, K. Declerck, T. Lammens, B. De Moerloose, D. Deforce, F. Van Nieuwerburgh, L. C. Cheung, R. S. Kotecha, M. R. Mansour, B. Ghesquière, G. Van Camp, W. V. Berghe, J. R. Kowalczyk, T. Szczepański, U. P. Davé, A. E. Kulozik,

S. Goossens, D. J. Curtis, T. Taghon, M. Dawidowska, S. Degerman, and P. Van Vlierberghe, "Aging of preleukemic thymocytes drives CpG island hypermethylation in T-cell acute lymphoblastic leukemia," *Blood Cancer Discov*, vol. 1, pp. 274–289, Nov 2020.

[256] C. J. Poole, A. Lodh, J. H. Choi, and J. van Riggelen, "MYC deregulates TET1 and TET2 expression to control global DNA (hydroxy)methylation and gene expression to maintain a neoplastic phenotype in T-ALL," *Epigenetics Chromatin*, vol. 12, p. 41, Jul 2019.

[257] "The BLUEPRINT DCC Portal." http://dcc.blueprint-epigenome.eu/#/home, 2016.

[258] F. Krueger, F. James, P. Ewels, E. Afyounian, and B. Schuster-Boeckler, "Felixkrueger/trimgalore: v0.6.7 - doi via zenodo," July 2021.

[259] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, pp. 15–21, Jan 2013.

[260] M. Pertea, G. M. Pertea, C. M. Antonescu, T. C. Chang, J. T. Mendell, and S. L. Salzberg, "StringTie enables improved reconstruction of a transcriptome from RNA-seq reads," *Nat Biotechnol*, vol. 33, pp. 290–295, Mar 2015.

[261] A. R. Quinlan and I. M. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, pp. 841–842, Mar 2010.

[262] F. Jühling, H. Kretzmer, S. H. Bernhart, C. Otto, P. F. Stadler, and S. Hoffmann, "metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data," *Genome Res*, vol. 26, pp. 256–262, Feb 2016.

[263] A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y. C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K. H. Farh, S. Feizi, R. Karlic, A. R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L. H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, M. Kellis, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y. C. Wu, A. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K. H. Farh, S. Feizi, R. Karlic, A. R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall,

N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, N. Abdennur, M. Adli, M. Akerman, L. Barrera, J. Antosiewicz-Bourget, T. Ballinger, M. J. Barnes, D. Bates, R. J. Bell, D. A. Bennett, K. Bianco, C. Bock, P. Boyle, J. Brinchmann, P. Caballero-Campo, R. Camahort, M. J. Carrasco-Alfonso, T. Charnecki, H. Chen, Z. Chen, J. B. Cheng, S. Cho, A. Chu, W. Y. Chung, C. Cowan, Q. Athena Deng, V. Deshpande, M. Diegel, B. Ding, T. Durham, L. Echipare, L. Edsall, D. Flowers, O. Genbacev-Krtolica, C. Gifford, S. Gillespie, E. Giste, I. A. Glass, A. Gnirke, M. Gormley, H. Gu, J. Gu, D. A. Hafler, M. J. Hangauer, M. Hariharan, M. Hatan, E. Haugen, Y. He, S. Heimfeld, S. Herlofsen, Z. Hou, R. Humbert, R. Issner, A. R. Jackson, H. Jia, P. Jiang, A. K. Johnson, T. Kadlecek, B. Kamoh, M. Kapidzic, J. Kent, A. Kim, M. Kleinewietfeld, S. Klugman, J. Krishnan, S. Kuan, T. Kutyavin, A. Y. Lee, K. Lee, J. Li, N. Li, Y. Li, K. L. Ligon, S. Lin, Y. Lin, J. Liu, Y. Liu, C. J. Luckey, Y. P. Ma, C. Maire, A. Marson, J. S. Mattick, M. Mayo, M. McMaster, H. Metsky, T. Mikkelsen, D. Miller, M. Miri, E. Mukamel, R. P. Nagarajan, F. Neri, J. Nery, T. Nguyen, H. O'Geen, S. Paithankar, T. Papayannopoulou, M. Pelizzola, P. Plettner, N. E. Propson, S. Raghuraman, B. J. Raney, A. Raubitschek, A. P. Reynolds, H. Richards, K. Riehle, P. Rinaudo, J. F. Robinson, N. B. Rockweiler, E. Rosen, E. Rynes, J. Schein, R. Sears, T. Sejnowski, A. Shafer, L. Shen, R. Shoemaker, M. Sigaroudinia, I. Slukvin, S. Stehling-Sun, R. Stewart, S. L. Subramanian, K. Suknuntha, S. Swanson, S. Tian, H. Tilden, L. Tsai, M. Urich, I. Vaughn, J. Vierstra, S. Vong, U. Wagner, H. Wang, T. Wang, Y. Wang, A. Weiss, H. Whitton, A. Wildberg, H. Witt, K. J. Won, M. Xie, X. Xing, I. Xu, Z. Xuan, Z. Ye, C. A. Yen, P. Yu, X. Zhang, X. Zhang, J. Zhao, Y. Zhou, J. Zhu, Y. Zhu, S. Ziegler, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L. H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, M. Kellis, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, and M. Kellis, "Integrative analysis of 111 reference human epigenomes," *Nature*, vol. 518, pp. 317–330, Feb 2015.

[264] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data," *Machine learning*, vol. 52, pp. 91–118, Jul 2003.

[265] M. D. Wilkerson and D. N. Hayes, "ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking," *Bioinformatics*, vol. 26, pp. 1572–1573, Jun 2010.

[266] Z. Gu, R. Eils, and M. Schlesner, "Complex heatmaps reveal patterns and correlations in multidimensional genomic data," *Bioinformatics*, vol. 32, pp. 2847–2849, Sep 2016.

[267] D. Anastasiadi, A. Esteve-Codina, and F. Piferrer, "Consistent inverse correlation between DNA methylation of the first intron and gene expression across tissues and species," *Epigenetics Chromatin*, vol. 11, p. 37, Jun 2018.

[268] Y. Wang, M. Xiao, X. Chen, L. Chen, Y. Xu, L. Lv, P. Wang, H. Yang, S. Ma, H. Lin, B. Jiao, R. Ren, D. Ye, K. L. Guan, and Y. Xiong, "WT1 recruits TET2 to regulate its target gene expression and suppress leukemia cell proliferation," *Mol Cell*, vol. 57, pp. 662–673, Feb 2015.

[269] G. Dixon, H. Pan, D. Yang, B. P. Rosen, T. Jashari, N. Verma, J. Pulecio, I. Caspi, K. Lee, S. Stransky, A. Glezer, C. Liu, M. Rivas, R. Kumar, Y. Lan, I. Torregroza, C. He, S. Sidoli, T. Evans, O. Elemento, and D. Huangfu, "QSER1 protects DNA methylation valleys from de novo methylation," *Science*, vol. 372, Apr 2021.

[270] W. Yu, C. McIntosh, R. Lister, I. Zhu, Y. Han, J. Ren, D. Landsman, E. Lee, V. Briones, M. Terashima, R. Leighty, J. R. Ecker, and K. Muegge, "Genome-wide DNA methylation patterns in LSH mutant reveals de-repression of repeat elements and redundant epigenetic silencing pathways," *Genome Res*, vol. 24, pp. 1613–1623, Oct 2014.

[271] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biol*, vol. 15, p. 550, Dec 2014.

[272] J. I. Martin-Subero and C. C. Oakes, "Charting the dynamic epigenome during B-cell development," *Semin Cancer Biol*, vol. 51, pp. 139–148, Aug 2018.

[273] A. B. Brinkman, S. Nik-Zainal, F. Simmer, F. G. Rodríguez-González, M. Smid, L. B. Alexandrov, A. Butler, S. Martin, H. Davies, D. Glodzik, X. Zou, M. Ramakrishna, J. Staaf, M. Ringnér, A. Sieuwerts, A. Ferrari, S. Morganella, T. Fleischer, V. Kristensen, M. Gut, M. J. van de Vijver, A. L. Børresen-Dale, A. L. Richardson, G. Thomas, I. G. Gut, J. W. M. Martens, J. A. Foekens, M. R. Stratton, and H. G. Stunnenberg, "Partially methylated domains are hypervariable in breast cancer and fuel widespread CpG island hypermethylation," *Nat Commun*, vol. 10, p. 1749, Apr 2019.

[274] E. F. Wagner and A. R. Nebreda, "Signal integration by JNK and p38 MAPK pathways in cancer development," *Nat Rev Cancer*, vol. 9, pp. 537–549, Aug 2009.

[275] S. De, R. Shaknovich, M. Riester, O. Elemento, H. Geng, M. Kormaksson, Y. Jiang, B. Woolcock, N. Johnson, J. M. Polo, L. Cerchietti, R. D. Gascoyne, A. Melnick, and F. Michor, "Aberration in DNA methylation in B-cell lymphomas has a complex origin and increases with disease severity," *PLoS Genet*, vol. 9, p. e1003137, Jan 2013.

[276] D. J. Weisenberger, M. Velicescu, J. C. Cheng, F. A. Gonzales, G. Liang, and P. A. Jones, "Role of the DNA methyltransferase variant DNMT3b3 in DNA methylation," *Mol Cancer Res*, vol. 2, pp. 62–72, Jan 2004.

[277] A. Parry, S. Rulands, and W. Reik, "Active turnover of DNA methylation during cell fate decisions," *Nat Rev Genet*, vol. 22, pp. 59–66, Jan 2021.

[278] M. J. Garnett, E. J. Edelman, S. J. Heidorn, C. D. Greenman, A. Dastur, K. W. Lau, P. Greninger, I. R. Thompson, X. Luo, J. Soares, Q. Liu, F. Iorio, D. Surdez, L. Chen, R. J. Milano, G. R. Bignell, A. T. Tam, H. Davies, J. A. Stevenson, S. Barthorpe, S. R. Lutz, F. Kogera, K. Lawrence, A. McLaren-Douglas, X. Mitropoulos, T. Mironenko, H. Thi, L. Richardson, W. Zhou, F. Jewitt, T. Zhang, P. O'Brien, J. L. Boisvert, S. Price, W. Hur, W. Yang, X. Deng, A. Butler, H. G. Choi, J. W. Chang, J. Baselga, I. Stamenkovic, J. A. Engelman, S. V. Sharma, O. Delattre, J. Saez-Rodriguez, N. S. Gray, J. Settleman, P. A. Futreal, D. A. Haber, M. R. Stratton, S. Ramaswamy, U. McDermott, and C. H. Benes, "Systematic identification of genomic markers of drug sensitivity in cancer cells," *Nature*, vol. 483, pp. 570–575, Mar 2012.

[279] F. Iorio, T. A. Knijnenburg, D. J. Vis, G. R. Bignell, M. P. Menden, M. Schubert, N. Aben, E. alves, S. Barthorpe, H. Lightfoot, T. Cokelaer, P. Greninger, E. van Dyk, H. Chang, H. de Silva, H. Heyn, X. Deng, R. K. Egan, Q. Liu, T. Mironenko, X. Mitropoulos, L. Richardson, J. Wang, T. Zhang, S. Moran, S. Sayols, M. Soleimani, D. Tamborero, N. Lopez-Bigas, P. Ross-Macdonald, M. Esteller, N. S. Gray, D. A. Haber, M. R. Stratton, C. H. Benes, L. F. A. Wessels, J. Saez-Rodriguez, U. McDermott, and M. J. Garnett, "A Landscape of Pharmacogenomic Interactions in Cancer," *Cell*, vol. 166, pp. 740–754, Jul 2016.

[280] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin, A. Reddy, M. Liu, L. Murray, M. F. Berger, J. E. Monahan, P. Morais, J. Meltzer, A. Korejwa, J. Jané-Valbuena, F. A. Mapa, J. Thibault, E. Bric-Furlong, P. Raman, A. Shipway, I. H. Engels, J. Cheng, G. K. Yu, J. Yu, P. Aspesi, M. de Silva, K. Jagtap, M. D. Jones, L. Wang, C. Hatton, E. Palescandolo, S. Gupta, S. Mahan, C. Sougnez, R. C. Onofrio, T. Liefeld, L. MacConaill, W. Winckler, M. Reich, N. Li, J. P. Mesirov, S. B. Gabriel, G. Getz, K. Ardlie, V. Chan, V. E. Myer, B. L. Weber, J. Porter, M. Warmuth, P. Finan, J. L. Harris, M. Meyerson, T. R. Golub, M. P. Morrissey, W. R. Sellers, R. Schlegel, and L. A. Garraway, "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, pp. 603–607, Mar 2012.

[281] M. Ghandi, F. W. Huang, J. Jané-Valbuena, G. V. Kryukov, C. C. Lo, E. R. McDonald, J. Barretina, E. T. Gelfand, C. M. Bielski, H. Li, K. Hu, A. Y. Andreev-Drakhlin, J. Kim, J. M. Hess, B. J. Haas, F. Aguet, B. A. Weir, M. V. Rothberg, B. R. Paolella, M. S. Lawrence, R. Akbani, Y. Lu, H. L. Tiv, P. C. Gokhale, A. de Weck, A. A. Mansour, C. Oh, J. Shih, K. Hadi, Y. Rosen, J. Bistline, K. Venkatesan, A. Reddy, D. Sonkin, M. Liu, J. Lehar, J. M. Korn, D. A. Porter, M. D. Jones, J. Golji, G. Caponigro, J. E. Taylor, C. M. Dunning, A. L. Creech, A. C. Warren, J. M. McFarland, M. Zamanighomi, A. Kauffmann, N. Stransky, M. Imielinski, Y. E. Maruvka, A. D. Cherniack, A. Tsherniak, F. Vazquez, J. D. Jaffe, A. A. Lane, D. M. Weinstock, C. M. Johannessen, M. P. Morrissey, F. Stegmeier, R. Schlegel, W. C. Hahn, G. Getz, G. B. Mills, J. S. Boehm, T. R. Golub, L. A. Garraway, and W. R. Sellers, "Next-generation characterization of the Cancer Cell Line Encyclopedia," *Nature*, vol. 569, pp. 503–508, May 2019.

[282] A. Tsherniak, F. Vazquez, P. G. Montgomery, B. A. Weir, G. Kryukov, G. S. Cowley, S. Gill, W. F. Harrington, S. Pantel, J. M. Krill-Burger, R. M. Meyers, L. Ali, A. Goodale, Y. Lee, G. Jiang, J. Hsiao, W. F. J. Gerath, S. Howell, E. Merkel, M. Ghandi, L. A. Garraway, D. E. Root, T. R. Golub, J. S. Boehm, and W. C. Hahn, "Defining a Cancer Dependency Map," *Cell*, vol. 170, pp. 564–576, Jul 2017.

[283] E. Gonçalves, A. Segura-Cabrera, C. Pacini, G. Picco, F. M. Behan, P. Jaaks, E. A. Coker, D. van der Meer, A. Barthorpe, H. Lightfoot, T. Mironenko, A. Beck, L. Richardson, W. Yang, E. Lleshi, J. Hall, C. Tolley, C. Hall, I. Mali, F. Thomas, J. Morris, A. R. Leach, J. T. Lynch, B. Sidders, C. Crafter, F. Iorio, S. Fawell, and M. J. Garnett, "Drug mechanism-of-action discovery through the integration of pharmacological and CRISPR screens," *Mol Syst Biol*, vol. 16, p. e9405, Jul 2020.

[284] F. Vazquez and W. R. Sellers, "Are CRISPR Screens Providing the Next Generation of Therapeutic Targets?," *Cancer Res*, vol. 81, pp. 5806–5809, Dec 2021.

[285] J. Galon, A. Costes, F. Sanchez-Cabo, A. Kirilovsky, B. Mlecnik, C. s, M. Tosolini, M. Camus, A. Berger, P. Wind, F. é, P. Bruneval, P. H. Cugnenc, Z. Trajanoski, W. H. Fridman, and F. s, "Type, density, and location of immune cells within human colorectal tumors predict clinical outcome," *Science*, vol. 313, pp. 1960–1964, Sep 2006.

[286] H. Garner and K. E. de Visser, "Immune crosstalk in cancer progression and metastatic spread: a complex conversation," *Nat Rev Immunol*, vol. 20, pp. 483–497, Aug 2020.

[287] E. Sahai, I. Astsaturov, E. Cukierman, D. G. DeNardo, M. Egeblad, R. M. Evans, D. Fearon, F. R. Greten, S. R. Hingorani, T. Hunter, R. O. Hynes, R. K. Jain, T. Janowitz, C. Jorgensen, A. C. Kimmelman, M. G. Kolonin, R. G. Maki, R. S. Powers, E. é, D. C. Ramirez, R. Scherz-Shouval, M. H. Sherman, S. Stewart, T. D. Tlsty, D. A. Tuveson, F. M. Watt, V. Weaver, A. T. Weeraratna, and Z. Werb, "A framework for advancing our understanding of cancer-associated fibroblasts," *Nat Rev Cancer*, vol. 20, pp. 174–186, Mar 2020.

[288] S. N. Porter, L. C. Baker, D. Mittelman, and M. H. Porteus, "Lentiviral and targeted cellular barcoding reveals ongoing clonal dynamics of cell lines in vitro and in vivo," *Genome Biol*, vol. 15, p. R75, May 2014.

[289] S. R. Romanov, B. K. Kozakiewicz, C. R. Holst, M. R. Stampfer, L. M. Haupt, T. D. Tlsty, and T. D. Tlsty, "Normal human mammary epithelial cells spontaneously escape senescence and acquire genomic changes," *Nature*, vol. 409, pp. 633–637, Feb 2001.

[290] J. Nordlund, C. L. cklin, P. Wahlberg, S. Busche, E. C. Berglund, M. L. Eloranta, T. Flaegstad, E. Forestier, B. M. Frost, A. Harila-Saari, M. Heyman, O. G. nsson, R. Larsson, J. Palle, L. nnblom, K. Schmiegelow, D. Sinnett, S. ll, T. Pastinen, M. G. Gustafsson, G. nnerholm, and A. C. nen, "Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia," *Genome Biol*, vol. 14, p. r105, Sep 2013.

[291] S. Bian, Y. Hou, X. Zhou, X. Li, J. Yong, Y. Wang, W. Wang, J. Yan, B. Hu, H. Guo, J. Wang, S. Gao, Y. Mao, J. Dong, P. Zhu, D. Xiu, L. Yan, L. Wen, J. Qiao, F. Tang, and W. Fu, "Single-cell multiomics sequencing and analyses of human colorectal cancer," *Science*, vol. 362, pp. 1060–1063, Nov 2018.

[292] M. J. Aryee, A. E. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A. P. Feinberg, K. D. Hansen, and R. A. Irizarry, "Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays," *Bioinformatics*, vol. 30, pp. 1363–1369, May 2014.

[293] J. A. Heiss and A. C. Just, "Improved filtering of DNA methylation microarray data by detection p values and its impact on downstream analyses," *Clin Epigenetics*, vol. 11, p. 15, Jan 2019.

[294] Y. A. Chen, M. Lemire, S. Choufani, D. T. Butcher, D. Grafodatskaya, B. W. Zanke, S. Gallinger, T. J. Hudson, and R. Weksberg, "Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray," *Epigenetics*, vol. 8, pp. 203–209, Feb 2013.

[295] D. Aran, M. Sirota, and A. J. Butte, "Systematic pan-cancer analysis of tumour purity," *Nat Commun*, vol. 6, p. 8971, Dec 2015.

[296] X. Zheng, N. Zhang, H. J. Wu, and H. Wu, "Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies," *Genome Biol*, vol. 18, p. 17, Jan 2017.

[297] K. Yoshihara, M. Shahmoradgoli, E. nez, R. Vegesna, H. Kim, W. Torres-Garcia, V. o, H. Shen, P. W. Laird, D. A. Levine, S. L. Carter, G. Getz, K. Stemke-Hale, G. B. Mills, and R. G. Verhaak, "Inferring tumour purity and stromal and immune cell admixture from expression data," *Nat Commun*, vol. 4, p. 2612, Oct 2013.

[298] S. L. Carter, K. Cibulskis, E. Helman, A. McKenna, H. Shen, T. Zack, P. W. Laird, R. C. Onofrio, W. Winckler, B. A. Weir, R. Beroukhim, D. Pellman, D. A. Levine, E. S. Lander, M. Meyerson, and G. Getz, "Absolute quantification of somatic DNA alterations in human cancer," *Nat Biotechnol*, vol. 30, pp. 413–421, May 2012.

[299] L. Velten, B. A. Story, P. ndez Malmierca, S. Raffel, D. R. Leonce, J. Milbank, M. Paulsen, A. Demir, C. Szu-Tu, R. mel, C. Lutz, D. Nowak, J. C. Jann, C. Pabst, T. Boch, W. K. Hofmann, C. ller Tidow, A. Trumpp, S. Haas, and L. M. Steinmetz, "Identification of leukemic and pre-leukemic stem cells by clonal tracking from single-cell transcriptomics," *Nat Commun*, vol. 12, p. 1366, Mar 2021.

[300] M. Zhao, J. Sun, and Z. Zhao, "TSGene: a web resource for tumor suppressor genes," *Nucleic Acids Res*, vol. 41, pp. D970–976, Jan 2013.

[301] S. Aykul and E. Martinez-Hackert, "Determination of half-maximal inhibitory concentration using biosensor-based protein interaction analysis," *Anal Biochem*, vol. 508, pp. 97–103, Sep 2016.

[302] J. Singh Nanda, R. Kumar, and G. P. Raghava, "dbEM: A database of epigenetic modifiers curated from cancerous and normal genomes," *Sci Rep*, vol. 6, p. 19340, Jan 2016.

[303] Z. Sondka, S. Bamford, C. G. Cole, S. A. Ward, I. Dunham, and S. A. Forbes, "The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers," *Nat Rev Cancer*, vol. 18, pp. 696–705, Nov 2018.

[304] M. H. Bailey, C. Tokheim, E. Porta-Pardo, S. Sengupta, D. Bertrand, A. Weerasinghe, A. Colaprico, M. C. Wendl, J. Kim, B. Reardon, P. K. Ng, K. J. Jeong, S. Cao, Z. Wang, J. Gao, Q. Gao, F. Wang, E. M. Liu, L. Mularoni, C. Rubio-Perez, N. Nagarajan, I. s Ciriano, D. C. Zhou, W. W. Liang, J. M. Hess, V. D. Yellapantula, D. Tamborero, A. Gonzalez-Perez, C. Suphavilai, J. Y. Ko, E. Khurana, P. J. Park, E. M. Van Allen, H. Liang, M. S. Lawrence, A. Godzik, N. Lopez-Bigas, J. Stuart, D. Wheeler, G. Getz, K. Chen, A. J. Lazar, G. B. Mills, R. Karchin, L. Ding, S. J. Caesar-Johnson, J. A. Demchok, I. Felau, M. Kasapi, M. L. Ferguson, C. M. Hutter, H. J. Sofia, R. Tarnuzzer, Z. Wang, L. Yang, J. C. Zenklusen, J. J. Zhang, S. Chudamani, J. Liu, L. Lolla, R. Naresh, T. Pihl, Q. Sun, Y. Wan, Y. Wu, J. Cho, T. DeFreitas, S. Frazer, N. Gehlenborg, G. Getz, D. I. Heiman, J. Kim, M. S. Lawrence, P. Lin, S. Meier, M. S. Noble, G. Saksena, D. Voet, H. Zhang, B. Bernard, N. Chambwe, V. Dhankani, T. Knijnenburg, R. Kramer, K. Leinonen, Y. Liu, M. Miller, S. Reynolds, I. Shmulevich, V. Thorsson, W. Zhang, R. Akbani, B. M. Broom, A. M. Hegde, Z. Ju, R. S. Kanchi, A. Korkut, J. Li, H. Liang, S. Ling, W. Liu, Y. Lu, G. B. Mills, K. S. Ng, A. Rao, M. Ryan, J. Wang, J. N. Weinstein, J. Zhang, A. Abeshouse, J. Armenia, D. Chakravarty, W. K. Chatila, I. de Bruijn,

J. Gao, B. E. Gross, Z. J. Heins, R. Kundra, K. La, M. Ladanyi, A. Luna, M. G. Nissan, A. Ochoa, S. M. Phillips, E. Reznik, F. Sanchez-Vega, C. Sander, N. Schultz, R. Sheridan, S. O. Sumer, Y. Sun, B. S. Taylor, J. Wang, H. Zhang, P. Anur, M. Peto, P. Spellman, C. Benz, J. M. Stuart, C. K. Wong, C. Yau, D. N. Hayes, J. S. Parker, M. D. Wilkerson, A. Ally, M. Balasundaram, R. Bowlby, D. Brooks, R. Carlsen, E. Chuah, N. Dhalla, R. Holt, S. J. M. Jones, K. Kasaian, D. Lee, Y. Ma, M. A. Marra, M. Mayo, R. A. Moore, A. J. Mungall, K. Mungall, A. G. Robertson, S. Sadeghi, J. E. Schein, P. Sipahimalani, A. Tam, N. Thiessen, K. Tse, T. Wong, A. C. Berger, R. Beroukhim, A. D. Cherniack, C. Cibulskis, S. B. Gabriel, G. F. Gao, G. Ha, M. Meyerson, S. E. Schumacher, J. Shih, M. H. Kucherlapati, R. S. Kucherlapati, S. Baylin, L. Cope, L. Danilova, M. S. Bootwalla, P. H. Lai, D. T. Maglinte, D. J. Van Den Berg, D. J. Weisenberger, J. T. Auman, S. Balu, T. Bodenheimer, C. Fan, K. A. Hoadley, A. P. Hoyle, S. R. Jefferys, C. D. Jones, S. Meng, P. A. Mieczkowski, L. E. Mose, A. H. Perou, C. M. Perou, J. Roach, Y. Shi, J. V. Simons, T. Skelly, M. G. Soloway, D. Tan, U. Veluvolu, H. Fan, T. Hinoue, P. W. Laird, H. Shen, W. Zhou, M. Bellair, K. Chang, K. Covington, C. J. Creighton, H. Dinh, H. Doddapaneni, L. A. Donehower, J. Drummond, R. A. Gibbs, R. Glenn, W. Hale, Y. Han, J. Hu, V. Korchina, S. Lee, L. Lewis, W. Li, X. Liu, M. Morgan, D. Morton, D. Muzny, J. Santibanez, M. Sheth, E. Shinbrot, L. Wang, M. Wang, D. A. Wheeler, L. Xi, F. Zhao, J. Hess, E. L. Appelbaum, M. Bailey, M. G. Cordes, L. Ding, C. C. Fronick, L. A. Fulton, R. S. Fulton, C. Kandoth, E. R. Mardis, M. D. McLellan, C. A. Miller, H. K. Schmidt, R. K. Wilson, D. Crain, E. Curley, J. Gardner, K. Lau, D. Mallery, S. Morris, J. Paulauskis, R. Penny, C. Shelton, T. Shelton, M. Sherman, E. Thompson, P. Yena, J. Bowen, J. M. Gastier-Foster, M. Gerken, K. M. Leraas, T. M. Lichtenberg, N. C. Ramirez, L. Wise, E. Zmuda, N. Corcoran, T. Costello, C. Hovens, A. L. Carvalho, A. C. de Carvalho, J. H. Fregnani, A. Longatto-Filho, R. M. Reis, C. Scapulatempo-Neto, H. C. S. Silveira, D. O. Vidal, A. Burnette, J. Eschbacher, B. Hermes, A. Noss, R. Singh, M. L. Anderson, P. D. Castro, M. Ittmann, D. Huntsman, B. Kohl, X. Le, R. Thorp, C. Andry, E. R. Duffy, V. Lyadov, O. Paklina, G. Setdikova, A. Shabunin, M. Tavobilov, C. McPherson, R. Warnick, R. Berkowitz, D. Cramer, C. Feltmate, N. Horowitz, A. Kibel, M. Muto, C. P. Raut, A. Malykh, J. S. Barnholtz-Sloan, W. Barrett, K. Devine, J. Fulop, Q. T. Ostrom, K. Shimmel, Y. Wolinsky, A. E. Sloan, A. De Rose, F. Giuliante, M. Goodman, B. Y. Karlan, C. H. Hagedorn, J. Eckman, J. Harr, J. Myers, K. Tucker, L. A. Zach, B. Deyarmin, H. Hu, L. Kvecher, C. Larson, R. J. Mural, S. Somiari, A. Vicha, T. Zelinka, J. Bennett, M. Iacocca, B. Rabeno, P. Swanson, M. Latour, L. Lacombe, B. tu, A. Bergeron, M. McGraw, S. M. Staugaitis, J. Chabot, H. Hibshoosh, A. Sepulveda, T. Su, T. Wang, O. Potapova, O. Voronina, L. Desjardins, O. Mariani, S. Roman-Roman, X. Sastre, M. H. Stern, F. Cheng, S. Signoretti, A. Berchuck, D. Bigner, E. Lipp, J. Marks, S. McCall, R. McLendon, A. Secord, A. Sharp, M. Behera, D. J. Brat, A. Chen, K. Delman, S. Force, F. Khuri, K. Magliocca, S. Maithel, J. J. Olson, T. Owonikoko, A. Pickens, S. Ramalingam, D. M. Shin, G. Sica, E. G. Van Meir, H. Zhang, W. Eijckenboom, A. Gillis, E. Korpershoek, L. Looijenga, W. Oosterhuis, H. Stoop, K. E. van Kessel, E. C. Zwarthoff, C. Calatozzolo, L. Cuppini, S. Cuzzubbo, F. DiMeco, G. Finocchiaro, L. Mattei, A. Perin, B. Pollo, C. Chen, J. Houck, P. Lohavanichbutr, A. Hartmann, C. Stoehr, R. Stoehr, H. Taubert, S. Wach, B. Wullich, W. Kycler, D. Murawa, M. Wiznerowicz, K. Chung, W. J. Edenfield, J. Martin, E. Baudin, G. Bubley, R. Bueno, A. De Rienzo, W. G. Richards, S. Kalkanis, T. Mikkelsen, H. Noushmehr, L. Scarpace, N. Girard, M. Aymerich, E. Campo, E. é, A. L. Guillermo, N. Van Bang, P. T. Hanh, B. D. Phu, Y. Tang, H. Colman, K. Evason, P. R. Dottino,

J. A. Martignetti, H. Gabra, H. Juhl, T. Akeredolu, S. Stepa, D. Hoon, K. Ahn, K. J. Kang, F. Beuschlein, A. Breggia, M. Birrer, D. Bell, M. Borad, A. H. Bryce, E. Castle, V. Chandan, J. Cheville, J. A. Copland, M. Farnell, T. Flotte, N. Giama, T. Ho, M. Kendrick, J. P. Kocher, K. Kopp, C. Moser, D. Nagorney, D. O'Brien, B. P. O'Neill, T. Patel, G. Petersen, F. Que, M. Rivera, L. Roberts, R. Smallridge, T. Smyrk, M. Stanton, R. H. Thompson, M. Torbenson, J. D. Yang, L. Zhang, F. Brimo, J. A. Ajani, A. M. A. Gonzalez, C. Behrens, J. Bondaruk, R. Broaddus, B. Czerniak, B. Esmaeli, J. Fujimoto, J. Gershenwald, C. Guo, A. J. Lazar, C. Logothetis, F. Meric-Bernstam, C. Moran, L. Ramondetta, D. Rice, A. Sood, P. Tamboli, T. Thompson, P. Troncoso, A. Tsao, I. Wistuba, C. Carter, L. Haydu, P. Hersey, V. Jakrot, H. Kakavand, R. Kefford, K. Lee, G. Long, G. Mann, M. Quinn, R. Saw, R. Scolyer, K. Shannon, A. Spillane, J. Stretch, M. Synott, J. Thompson, J. Wilmott, H. Al-Ahmadie, T. A. Chan, R. Ghossein, A. Gopalan, D. A. Levine, V. Reuter, S. Singer, B. Singh, N. V. Tien, T. Broudy, C. Mirsaidi, P. Nair, P. Drwiega, J. Miller, J. Smith, H. Zaren, J. W. Park, N. P. Hung, E. Kebebew, W. M. Linehan, A. R. Metwalli, K. Pacak, P. A. Pinto, M. Schiffman, L. S. Schmidt, C. D. Vocke, N. Wentzensen, R. Worrell, H. Yang, M. Moncrieff, C. Goparaju, J. Melamed, H. Pass, N. Botnariuc, I. Caraman, M. Cernat, I. Chemencedji, A. Clipca, S. Doruc, G. Gorincioi, S. Mura, M. Pirtac, I. Stancul, D. Tcaciuc, M. Albert, I. Alexopoulou, A. Arnaout, J. Bartlett, J. Engel, S. Gilbert, J. Parfitt, H. Sekhon, G. Thomas, D. M. Rassl, R. C. Rintoul, C. Bifulco, R. Tamakawa, W. Urba, N. Hayward, H. Timmers, A. Antenucci, F. Facciolo, G. Grazi, M. Marino, R. Merola, R. de Krijger, A. P. Gimenez-Roqueplo, A. é, S. Chevalier, G. McKercher, K. Birsoy, G. Barnett, C. Brewer, C. Farver, T. Naska, N. A. Pennell, D. Raymond, C. Schilero, K. Smolenski, F. Williams, C. Morrison, J. A. Borgia, M. J. Liptay, M. Pool, C. W. Seder, K. Junker, L. Omberg, M. Dinkin, G. Manikhas, D. Alvaro, M. C. Bragazzi, V. Cardinale, G. Carpino, E. Gaudio, D. Chesla, S. Cottingham, M. Dubina, F. Moiseenko, R. Dhanasekaran, K. F. Becker, K. P. Janssen, J. Slotta-Huspenina, M. H. Abdel-Rahman, D. Aziz, S. Bell, C. M. Cebulla, A. Davis, R. Duell, J. B. Elder, J. Hilty, B. Kumar, J. Lang, N. L. Lehman, R. Mandt, P. Nguyen, R. Pilarski, K. Rai, L. Schoenfield, K. Senecal, P. Wakely, P. Hansen, R. Lechan, J. Powers, A. Tischler, W. E. Grizzle, K. C. Sexton, A. Kastl, J. Henderson, S. Porten, J. Waldmann, M. Fassnacht, S. L. Asa, D. Schadendorf, M. Couce, M. Graefen, H. Huland, G. Sauter, T. Schlomm, R. Simon, P. Tennstedt, O. Olabode, M. Nelson, O. Bathe, P. R. Carroll, J. M. Chan, P. Disaia, P. Glenn, R. K. Kelley, C. N. Landen, J. Phillips, M. Prados, J. Simko, K. Smith-McCune, S. VandenBerg, K. Roggin, A. Fehrenbach, A. Kendler, S. Sifri, R. Steele, A. Jimeno, F. Carey, I. Forgie, M. Mannelli, M. Carney, B. Hernandez, B. Campos, C. Herold-Mende, C. Jungk, A. Unterberg, A. von Deimling, A. Bossler, J. Galbraith, L. Jacobus, M. Knudson, T. Knutson, D. Ma, M. Milhem, R. Sigmund, A. K. Godwin, R. Madan, H. G. Rosenthal, C. Adebamowo, S. N. Adebamowo, A. Boussioutas, D. Beer, T. Giordano, A. M. Mes-Masson, F. Saad, T. Bocklage, L. Landrum, R. Mannel, K. Moore, K. Moxley, R. Postier, J. Walker, R. Zuna, M. Feldman, F. Valdivieso, R. Dhir, J. Luketich, E. M. M. Pinero, M. Quintero-Aguilo, C. G. Carlotti, J. S. Dos Santos, R. Kemp, A. Sankarankuty, D. Tirapelli, J. Catto, K. Agnew, E. Swisher, J. Creaney, B. Robinson, C. S. Shelley, E. M. Godwin, S. Kendall, C. Shipman, C. Bradford, T. Carey, A. Haddad, J. Moyer, L. Peterson, M. Prince, L. Rozek, G. Wolf, R. Bowman, K. M. Fong, I. Yang, R. Korst, W. K. Rathmell, J. L. Fantacone-Campbell, J. A. Hooke, A. J. Kovatich, C. D. Shriver, J. DiPersio, B. Drake, R. Govindan, S. Heath, T. Ley, B. Van Tine, P. Westervelt, M. A. Rubin, J. I. Lee, N. D. Aredes, and A. Mariamidze, "Comprehensive Characterization

of Cancer Driver Genes and Mutations," *Cell*, vol. 173, pp. 371–385, Apr 2018.

[305] J. B. Studd, A. J. Cornish, P. H. Hoang, P. Law, B. Kinnersley, and R. Houlston, "Cancer drivers and clonal dynamics in acute lymphoblastic leukaemia subtypes," *Blood Cancer J*, vol. 11, p. 177, Nov 2021.

[306] Y. Liu, J. Easton, Y. Shao, J. Maciaszek, Z. Wang, M. R. Wilkinson, K. McCastlain, M. Edmonson, S. B. Pounds, L. Shi, X. Zhou, X. Ma, E. Sioson, Y. Li, M. Rusch, P. Gupta, D. Pei, C. Cheng, M. A. Smith, J. G. Auvil, D. S. Gerhard, M. V. Relling, N. J. Winick, A. J. Carroll, N. A. Heerema, E. Raetz, M. Devidas, C. L. Willman, R. C. Harvey, W. L. Carroll, K. P. Dunsmore, S. S. Winter, B. L. Wood, B. P. Sorrentino, J. R. Downing, M. L. Loh, S. P. Hunger, J. Zhang, and C. G. Mullighan, "The genomic landscape of pediatric and young adult T-lineage acute lymphoblastic leukemia," *Nat Genet*, vol. 49, pp. 1211–1218, Aug 2017.

[307] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, Oct 2015.

[308] V. Boeva, T. Popova, K. Bleakley, P. Chiche, J. Cappo, G. Schleiermacher, I. Janoueix-Lerosey, O. Delattre, and E. Barillot, "Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data," *Bioinformatics*, vol. 28, pp. 423–425, Feb 2012.

[309] M. Greaves and C. C. Maley, "Clonal evolution in cancer," *Nature*, vol. 481, pp. 306–313, Jan 2012.

[310] R. V. Nichols, B. L. O'Connell, R. M. Mulqueen, J. Thomas, A. R. Woodfin, S. Acharya, G. Mandel, D. Pokholok, F. J. Steemers, and A. C. Adey, "High-throughput robust single-cell DNA methylation profiling with sciMETv2," *Nat Commun*, vol. 13, p. 7627, Dec 2022.

[311] Z. W. Yuen, A. Srivastava, R. Daniel, D. McNevin, C. Jack, and E. Eyras, "Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing," *Nat Commun*, vol. 12, p. 3438, Jun 2021.

[312] L. Xu and M. Seki, "Recent advances in the detection of base modifications using the Nanopore sequencer," *J Hum Genet*, vol. 65, pp. 25–33, Jan 2020.

[313] A. S. Deshpande, N. Ulahannan, M. Pendleton, X. Dai, L. Ly, J. M. Behr, S. Schwenk, W. Liao, M. A. Augello, C. Tyer, P. Rughani, S. Kudman, H. Tian, H. G. Otis, E. Adney, D. Wilkes, J. M. Mosquera, C. E. Barbieri, A. Melnick, D. Stoddart, D. J. Turner, S. Juul, E. Harrington, and M. ski, "Identifying synergistic high-order 3D chromatin conformations from genome-scale nanopore concatemer sequencing," *Nat Biotechnol*, vol. 40, pp. 1488–1499, Oct 2022.

[314] S. E. Bates, "Epigenetic Therapies for Cancer," *N Engl J Med*, vol. 383, pp. 650–663, Aug 2020.

[315] N. Verma, H. Pan, L. C. é, A. Shukla, Q. V. Li, B. Pelham-Webb, V. Teijeiro, F. lez, A. Krivtsov, C. J. Chang, E. P. Papapetrou, C. He, O. Elemento, and D. Huangfu, "TET proteins safeguard bivalent promoters from de novo methylation in human embryonic stem cells," *Nat Genet*, vol. 50, pp. 83–95, Jan 2018.

Statement of authorship

I declare to the Freie Universität Berlin that I have completed the submitted dissertation independently and without the use of sources and aids other than those indicated. The present thesis is free of plagiarism. I have marked as such all statements that are taken literally or in content from other writings. This dissertation has not been submitted in the same or similar form in any previous doctoral procedure. I agree to have my thesis examined by a plagiarism examination software.

Berlin, June 13, 2023 Sara Hetzel