INAUGURAL – DISSERTATION

for obtaining the doctoral degree
of the

Combined Faculty of Mathematics, Engineering and Natural Sciences
of
Ruprecht-Karls-University, Heidelberg

Presented by

Katherine Alexandra Kelly, M.Sc.
born in Galway, Ireland

Oral examination: 24th March 2025

Acute Myeloid Leukemia with Deletion 5q: an Epigenetic Perspective

Referees:
Prof. Dr. Christoph Plass
Prof. Dr. Karsten Rippe

Abstract

Acute myeloid leukemia (AML) is a hematological cancer characterised by a block in differentiation and accelerated proliferation of myeloid progenitor cells. Epigenetic regulators are among the most frequent targets for mutations and structural variations in AML, and the disruption of these genes can result in profound epigenetic heterogeneity between and within tumors. Deletion 5q [del(5q)] is the most common copy number alteration (CNA) in older AML patients and is associated with poor clinical outcome and therapy resistance, however the mechanisms linking del(5q) to leukemic development and progression are not understood.

I began this thesis with an analysis of DNA methylation profiles from 477 elderly AML patients using a DNA methylome deconvolution approach. Here I discovered that del(5q) AML constitutes an epigenetically distinct subgroup characterised by a unique signature of DNA hypermethylation. In an attempt to pinpoint the epigenetic disturbance leading this signature to arise, I investigated the 5q Minimally Deleted Region (MDR) for potential epigenetic regulators, and identified the H3K9me1/2 demethylase *KDM3B* as a promising target. Precise mapping of the MDR, together with differential transcriptional, protein and mutational analysis of 5q genes strengthened the argument that *KDM3B* is the most likely candidate for haploinsufficiency in del(5q) AML. I further linked the del(5q) methylation signature to dysregulation of other H3K9me1/2 regulators, and consistent overexpression of the *de novo* DNA methyltransferase and leukemic stem cell marker, *DNMT3B*. Moreover, I discovered that del(5q) and *MECOM*-overexpressing leukemias share a common DNA methylation signature, which in both subgroups coincides with increased expression of *DNMT3B*. These findings suggest that del(5q) AML deserves to be reappraised as an epigenetically-defined subgroup, which I suggest may be driven by haploinsufficiency of KDM3B.

Reduction in protein levels of KDM3B should result in an increase in H3K9me1/2. In addition, I hypothesised that haploinsufficiency of this enzyme may result in an imbalanced removal of H3K9me1/2, such that variable patterns of these histone marks may arise between cells. Such cell-to-cell epigenomic heterogeneity could provide a powerful driving force for leukemic progression by allowing selection of favorable phenotypes throughout cancer evolution and in response to therapy. To study this phenomenon, I developed a heterogeneity metric called epiCHAOS (epigenetic/Chromatin Heterogeneity Assessment Of Single cells). EpiCHAOS is the first tool that enables quantitative comparisons of epigenetic heterogeneity between single-cell groups/clusters within a biological sample. I

validated epiCHAOS *in silico* and demonstrated its functionality by applying the metric to a range of biological datasets from developmental systems, cancers, and aging, to investigate both genome-wide and region-specific differences in epigenetic heterogeneity.

Finally, to investigate the epigenetic consequences of KDM3B disruption in AML, I analyzed single-cell assay for transposase-accessible chromatin with sequencing (ATAC-seq) data generated from *KDM3B*-heterozygous OCI-AML3 cell lines, which were established to mimic haploinsufficiency of the enzyme. Heterozygous deletion of *KDM3B* resulted in the expected global chromatin compaction as well as epigenetic heterogeneity at H3K9me1/2-associated regions.

This thesis provides two important contributions for the research community. First, my findings shed light on the mechanisms driving one of the most aggressive forms of AML, which until now has not been studied from an epigenetic perspective, and where *KDM3B* has received very little attention as a putative target gene. Secondly, I provide the first computational strategy for quantitative single-cell analysis of epigenomic heterogeneity, which should offer a useful tool for biologists, especially those interested in stemness, plasticity and mechanisms of therapy resistance in cancer.

Zusammenfassung

Die Akute Myeloische Leukämie (AML) ist eine hämatologische Krebserkrankung, die durch eine Blockade der Differenzierung myeloider Vorläuferzellen entsteht. Epigenetische Regulatoren sind in AML häufig mutiert oder von strukturellen Variationen betroffen, und die Störung dieser Gene kann zu einer tiefgreifenden epigenetischen Heterogenität zwischen und innerhalb von Tumoren führen. Del(5q) ist die häufigste Kopienzahlveränderung bei älteren AML-Patienten und wird mit einem schlechten klinischen Ergebnis und Therapieresistenz in Verbindung gebracht. Die Mechanismen, die den del(5q) mit der Entwicklung und dem Fortschreiten von Leukämie in Verbindung bringen, sind bislang nicht verstanden.

Ich begann diese Arbeit mit einer Analyse von DNA-Methylierungsprofilen von 477 älteren AML-Patienten unter Verwendung eines Methylom-Dekonvolutionsansatzes. Dabei entdeckte ich, dass del(5q) AML eine einzigartige epigenetische Untergruppe darstellt, die durch DNA-Hypermethylierung gekennzeichnet ist. In einem Versuch, die epigenetische Störung zu bestimmen, die zur Entstehung dieser Signatur führt, untersuchte ich die minimal deletierte 5q-Region (MDR) auf potenzielle epigenetische Regulatoren und identifizierte die H3K9me1/2-Demethylase KDM3B als vielversprechendes Ziel. Eine genaue Kartierung der MDR sowie eine differenzielle Transkriptions-, Protein- und Mutationsanalyse von 5q-Genen untermauerten das Argument, dass KDM3B der wahrscheinlichste Kandidat für Haploinsuffizienz bei del(5q)-MDR ist. Ich brachte die del(5q)-Methylierungssignatur außerdem mit einer Dysregulation anderer H3K9me1/2-Regulatoren und einer konsistenten Überexpression der de novo-DNA-Methyltransferase und des leukämischen Stammzellmarkers DNMT3B in Verbindung. Darüber hinaus entdeckte ich, dass del(5q)- und MECOM-überexprimierende Leukämien eine gemeinsame DNA-Methylierungssignatur aufweisen, die in beiden Untergruppen mit einer erhöhten Expression von DNMT3B einhergeht. Diese Ergebnisse legen nahe, dass del(5q) AML als epigenetisch definierte Untergruppe neu bewertet werden sollte, die durch Haploinsuffizienz von KDM3B verursacht werden könnte.

Eine Verringerung des Proteinspiegels von KDM3B sollte zu einem Anstieg von H3K9me1/2 führen. Darüber hinaus habe ich die Hypothese aufgestellt, dass eine Haploinsuffizienz dieses Enzyms zu einer unausgewogenen Entfernung von H3K9me1/2 führt, sodass zwischen den Zellen unterschiedliche Muster dieser Histonmarkierungen auftreten könnten. Eine solche epigenomische Heterogenität von Zelle zu Zelle könnte eine treibende Kraft für die Entwicklung von Leukämie darstellen, indem sie die Auswahl günstiger Phänotypen

während der Krebsentwicklung und als Reaktion auf die Therapie ermöglicht. Um dieses Phänomen zu untersuchen, habe ich ein Heterogenitätsmaß namens epiCHAOS (epigenetic/Chromatin Heterogeneity Assessment Of Single cells) entwickelt, welches Vergleiche der epigenomischen Heterogenität in Einzelzelldatensätzen ermöglicht. Ich habe epiCHAOS in silico validiert und seine Funktionalität demonstriert, indem ich das Maß auf eine Reihe biologischer Datensätze aus Entwicklungssystemen, Krebs und Alterung angewendet habe, die zeigte, dass es sowohl genomweite als auch regionsspezifische Unterschiede in der epigenetischen Heterogenität gibt.

Um schließlich die epigenetischen Folgen der KDM3B-Störung bei AML zu untersuchen, habe ich Einzelzell-ATAC-seq-Daten analysiert, die aus KDM3B-heterozygoten OCI-AML3-Zelllinien generiert wurden. Diese Zellen wurden als Modell der Haploinsuffizenz von KDM3B entwickelt. Die heterozygote Deletion von KDM3B führte zur erwarteten globalen Chromatinverdichtung sowie zu epigenetischer Heterogenität in H3K9me1/2-assoziierten Regionen.

Diese Arbeit liefert zwei wichtige Beiträge für die wissenschaftliche Gemeinschaft. Erstens zeigen meine Ergebnisse, dass eine hochaggresive Form von AML eine epigenetisch definierte Untergruppe der Erkrankung darstellt. Diese Untergruppe ist durch Dysregulation von KDM3B definiert, welches bislang wenig untersucht wurde. Zweitens liefere ich die erste computerbasierte Strategie zur quantitativen Einzelzellanalyse der epigenomischen Heterogenität, die ein nützliches Werkzeug für Krebsbiologen darstellen wird. EpiCHAOS ist besonder relevant für Forschende im Bereich von Stammzellen, Plastizität und zur Untersuchung von Therapieresistenz.

Contents

1 Introduction **1**
1.1 Acute Myeloid Leukemia . . . 1
1.1.1 Genetic and epigenetic dysregulation in AML . . . 2
1.1.2 AML in the elderly . . . 3
1.1.3 AML treatment and clinical stratification . . . 3
1.1.4 Leukemic stem cells . . . 4
1.1.5 AML with a complex karyotype . . . 4
1.1.6 Past and present perspectives on deletion 5q AML . . . 5
1.2 Epigenetics . . . 7
1.2.1 DNA methylation . . . 8
1.2.2 Histone modifications and nucleosome positioning . . . 9
1.2.3 Epigenetics and epigenetic enzymes beyond the regulation of gene expression . . . 10
1.2.4 Epigenetic disorder in cancer . . . 11
1.2.5 Epigenetic strategies for cancer subgrouping . . . 13
1.2.6 Epigenetic heterogeneity and its implications in cancer . . . 13
1.2.7 Techniques for studying epigenetic modifications . . . 15
1.2.8 Techniques for studying epigenetic heterogeneity . . . 16

2 Motivation & Aims **19**
2.1 Characterising the DNA methylation landscape of AML in the elderly . . . 19
2.2 Investigating the epigenetic underpinnings of del(5q) AML . . . 20
2.3 Developing a computational strategy to quantify cell-to-cell epigenomic heterogeneity . . . 20
2.4 Investigating the epigenetic consequences of KDM3B haploinsufficiency . . . 21

3 DNA methylation-based characterisation of Acute Myeloid Leukemia **23**
3.1 Results . . . 23
3.1.1 Epigenetic characterisation of AML in older patients using methylome deconvolution . . . 23
3.1.2 Methylation signatures of normal hematopoietic cell types and HSPC stages . . . 24
3.1.3 Methylation signatures reflecting disruption of epigenetic regulators and transcription factors . . . 27

3.1.4 DNA methylation subgroups defined by unique molecular, cytogenetic and clinical features . . . 29
3.1.5 Del(5q) AML is an epigenetically distinct subgroup defined by a signature of DNA hypermethylation . . . 31
3.1.6 A hypermethylation signature enriched at developmental genes . . . 31
3.1.7 LMC3 carries prognostic significance independently of del(5q) . . . 34
3.1.8 LMC3 is present at low levels in a small subgroup of del(5q) MDS . 35
3.2 Discussion . . . 36

4 Investigating the epigenetic underpinnings of del(5q) AML 39
4.1 Results . . . 39
4.1.1 Mutations in epigenetic regulators are rare in del(5q) AML . . . 39
4.1.2 Narrowing down candidate genes in the minimally deleted region . . 41
4.1.3 Evidence for *KDM3B* as the target of the 5q deletion . . . 43
4.1.4 Mutual exclusivity of del(5q) and IDH mutations . . . 46
4.1.5 Linking the del(5q) methylation signature to DNMT3B . . . 47
4.1.6 Overexpression of *DNMT3B* in del(5q) AML may be regulated by DNA methylation in LSCs . . . 50
4.1.7 Linking the del(5q) methylation signature to H3K9me1/2 methylation 51
4.1.8 Epigenetic similarity of del(5q) and *MECOM*-overexpressing AML converge on overexpression of *DNMT3B* . . . 53
4.2 Discussion . . . 54

5 EpiCHAOS: a metric for quantifying epigenomic heterogeneity in single cell data 61
5.1 Results . . . 61
5.1.1 Development and *in silico* validation of epiCHAOS . . . 61
5.1.2 EpiCHAOS scores are minimally influenced by technical noise and choice of clustering parameters . . . 67
5.1.3 EpiCHAOS reflects epigenetic heterogeneity associated with developmental plasticity . . . 68
5.1.4 EpiCHAOS correlates with features of plasticity in malignant cells . 72
5.1.5 EpiCHAOS reveals increased epigenetic heterogeneity associated with hematopoietic aging . . . 74
5.1.6 EpiCHAOS reveals elevated epigenetic heterogeneity at PRC2 targeted regions and promoters of developmental genes . . . 74
5.1.7 EpiCHAOS is applicable to single-cell epigenomics data from any modality . . . 77
5.2 Discussion . . . 78

6 Investigating the epigenetic consequences of KDM3B haploinsufficiency in AML 83
6.1 Results . . . 83
6.1.1 Heterozygous deletion of *KDM3B* results in global chromatin compaction . . . 85

6.1.2 Heterozygous deletion of *KDM3B* results in cell-to-cell epigenetic heterogeneity at H3K9me1/2-associated regions . . . 88
6.2 Discussion . . . 89

7 Conclusions & Future Perspectives **91**

8 Materials and Methods **93**
8.1 deoxyribonucleic acid (DNA) methylation analyses . . . 93
8.1.1 Preprocessing of EPIC array data from the ASTRAL-1 cohort . . . 93
8.1.2 Methylome deconvolution . . . 93
8.1.3 Estimating sample purity . . . 94
8.1.4 Estimation of LMC proportions in external datasets . . . 94
8.1.5 Prediction of Leukemic Cell of Origin . . . 94
8.1.6 Biological interpretation of LMCs . . . 94
8.1.7 Defining LMC-based subgroups . . . 95
8.1.8 Differential methylation analysis . . . 95
8.1.9 Comparison of DNA methylation in del(5q) LSCs and blasts . . . 95
8.1.10 DNA methylation variation . . . 95
8.2 Gene expression analysis . . . 96
8.3 Copy number and mutational analyses . . . 96
8.3.1 Mutual exclusivity of mutations & CNA patterns . . . 96
8.3.2 Investigation of the 5q minimally deleted region . . . 96
8.3.3 Assessment of del(5q) mutations . . . 97
8.4 Protein expression analysis . . . 97
8.5 Clinical data analyses . . . 97
8.6 Single-cell epigenome and transcriptome analyses . . . 98
8.6.1 Calculating epiCHAOS scores . . . 98
8.6.2 Generating synthetic datasets with controlled heterogeneity . . . 99
8.6.3 Simulating scATAC-seq data with varying sequencing depth . . . 99
8.6.4 scATAC-seq data processing and analysis . . . 100
8.6.5 Differential heterogeneity analysis . . . 100
8.6.6 scRNA-seq analyses . . . 100
8.6.7 scChIP-seq analysis . . . 100
8.6.8 scNMT-seq and scTAM-seq analysis . . . 101
8.6.9 scATAC-seq data analysis from KDM3B deletion cells . . . 101

Contributions **101**
8.6.10 Publications . . . 103
8.6.11 Poster presentations . . . 104

Appendix **104**

Bibliography **122**

Acknowledgements **139**

Chapter 1

Introduction

1.1 Acute Myeloid Leukemia

AML is the most common form of acute leukemia in adults, and is also the most aggressive, with a five-year survival rate of 24% [1]. It is characterised by the clonal expansion of abnormally differentiated myeloid progenitor cells known as blasts, which results in impaired production of normal hematopoietic cells in the bone marrow and blood.

Normal hematopoeitic development begins with the hematopoietic stem cell (HSC)s – a quiescent population with a unique capacity for self-renewal and the ability to differentiate into different hematopoeitic progenitors [multipotent progenitor (MPP)s, lymphoid-primed multipotent progenitor (LMPP)s, common myeloid progenitor (CMP)s, common lymphoid progenitor (CLP)s, granuloyte/monocyte progenitor (GMP)s and megakaryocyte/erythroid progenitor (MEP)s]. These hematopoietic stem or progenitor cell (HSPC)s ultimately give rise to the range of committed cell types of the myeloid (monocytes and neutrophils), lymphoid [B, T and Natural Killer (NK) cells] and erythroid (red blood cells) lineages (Figure 1.1). Although there is controversy about whether these HSPC stages exist as distinct hierarchically organised cell states, or rather form a differentiation continuum [2], AML is thought usually to develop at a stage of hematopoietic development around the GMP, with fewer cases arising from pre-GMP stages of differentiation [3].

Depending on the type of genetic or cytogenetic alterations which are detected, AML is diagnosed clinically by the presence of either >10 or >20 percent of blast cells in the bone marrow or peripheral blood [4].

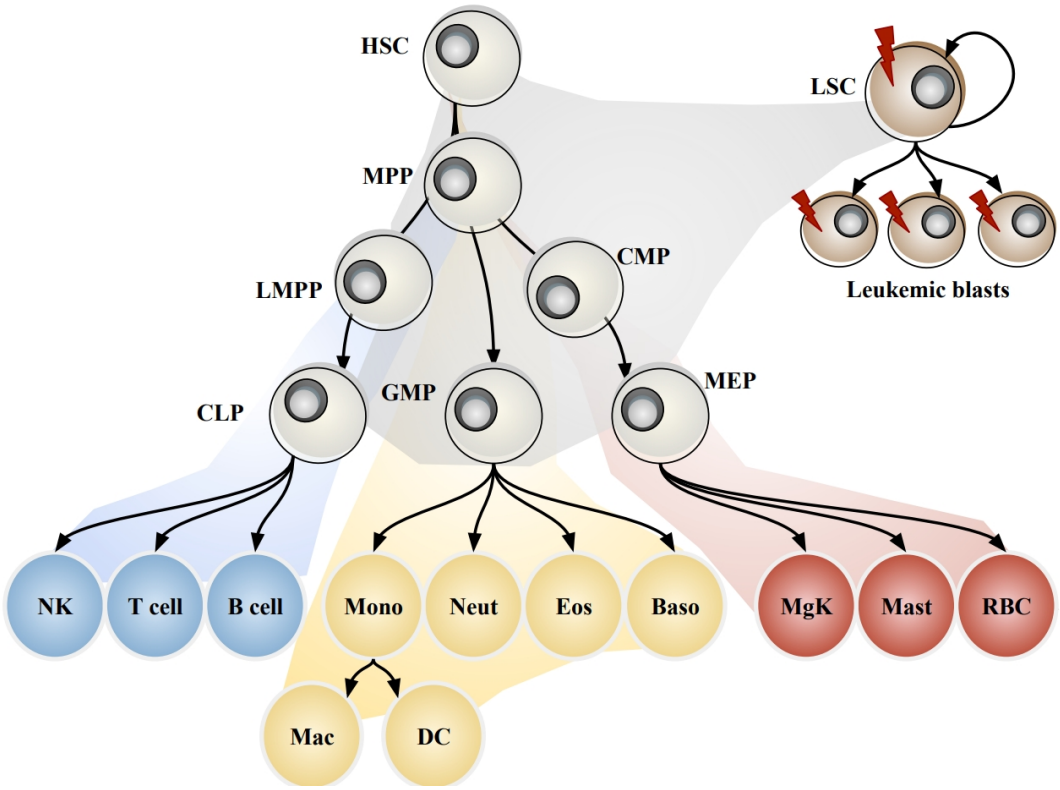

Figure 1.1: Healthy and malignant hematopoiesis. Schematic describing normal hematopoietic cell differentiation (left) and the emergence of AML (right). HSCs give rise to hematopoietic progenitor cells (MPPs, LMPPs, CMPs, GMPs, CLPs and MEPs), which in turn give rise to committed cell types of the myeloid [monocytes (Mono), neutrophils (Neut), eosinophils (Eos), basophils (Baso), macrophages (Mac) and dendritic cells (DC)], lymphoid (B-cell, T-cell and NK-cells) and megakaryocyte/erythroid lineage [red blood cells (RBC), mast cells (Mast), and megakaryocytes (MgK), which will give rise to platelets]. AML can arise from different stages of myeloid progenitor development. AML leukemic stem cell (LSC)s – which resemble HSCs in their self-renewing capacity – give rise to leukemic blasts which proliferate in the bone marrow and blood.

1.1.1 Genetic and epigenetic dysregulation in AML

By comparison to most solid tumors, the mutational burden and frequency of cytogenetic abnormalities in AML is relatively low [5]. Nevertheless, certain classes of genes are recurrently targeted for mutations; these include genes encoding hematopoietic transcription factor (TF)s such as RUNX1 and CEBPα, signalling molecules such as FLT3 and RAS, splicing factors such as SRSF2, cohesin complex components such as RAD21 and STAG2 – and most notably, epigenetic enzymes [6]. Structural genomic alterations including translocations, inversions, CNAs, and enhancer hijacking events are also detected in a considerable number of AML patients, many of which also disrupt hematopoietic TFs and epigenetic enzymes [7, 8]. Epigenetic dysregulation is considered to be a defining feature

of AML, wherein regulators of DNA methylation such as DNMT3A and TET2, histone modifiers such as KMT2A, EZH2, and other chromatin factors such as NPM1 and ASXL1 are among the earliest and most common targets for mutations, structural rearrangements and copy number alterations [9, 10, 11, 12, 13, 14, 15, 16]. These aberrations can disrupt global epigenetic regulation, leading to changes in DNA methylation and chromatin architecture. As a result, AML represents a heterogeneous class of tumors – not only genetically, phenotypically and clinically, but also at the epigenetic level [16, 17].

1.1.2 AML in the elderly

Like most cancers, AML is more common in older individuals, with a median age of diagnosis between 68 and 72 years [18]. In this thesis I will focus on AML in the elderly, in which the clinical and molecular characteristics differ from those in younger adults and in pediatric AML patients [19]. In particular, older patients more often present with complex cytogenetics and a high incidence of deletions on chromosome 5, 7 and 17. Elderly patients tend to have worse clinical outcome and a higher frequency of therapy resistance [19]. Recurrent translocations and inversions such as t(8;21) and inv(16), which are associated with favorable outcome, are by contrast more common in younger patients [19, 20], whereas fusion genes such as *RUNX1::RUNX1T1, CBFB::MYH11,* and *KMT2A::MLLT3*, are especially common in pediatric AML and rarely occur in adults [21]. It is also especially notable that AML in older patients is often preceded by myelodysplastic syndrome (MDS) – a premalignant hematological disorder, which often already harbor AML-associated mutations and cytogenetic changes such as del(5q), del(7q), trisomy 8, mutations in *DNMT3A, ASXL1, TET2, TP53* and others [19, 20]. Given these distinctions, and the fact that many of the above alterations influence epigenetic processes, it seems justifiable to separate age-specific cohorts for the purpose of epigenetic characterisation.

1.1.3 AML treatment and clinical stratification

Therapeutic management of AML in younger adults typically involves induction chemotherapy and hematopoietic stem cell transplantation [22]. Recently, targeted therapies such as FLT3 inhibitors and IDH inhibitors have also been incorporated into the clinical regimen for certain AML subgroups [22]. In older AML patients, and other cases in which intensive chemotherapy may not be tolerated, the current standard of care involves treatment with venetoclax – a BCL2 inhibitor – and a hypomethylating agent, azacitidine or decitabine [23]. While most patients respond well to these regimens, relapse is common in certain subgroups.

Frameworks for categorising AML in order to guide clinical decisions are continuously evolving. Traditionally, the French American British (FAB) nomenclature was used to characterise AML into subgroups based on the morphological appearance and differentiation stage of the leukemic cells [24]. Current protocols instead stratify AML based on

cytogenetic and genetic findings into three risk groups with favorable, intermediate and unfavorable prognosis, which were recently updated in the 2022 European LeukemiaNet (ELN) recommendations [4]. The favorable risk group includes patients with t(8;21) and inv(16), while the unfavorable risk group include patients with inv(3);t(3;3), t(6;9), t(9;22), t(8;16), del(5q), monosomy 7/5, or a complex karyotype (ckAML) [4]. Beyond these cytogenetic factors, mutations are also being incorporated into AML risk stratification, for example bZIP in-frame *CEBPA* mutations are indicative of a favorable outcome, whereas mutations in *TP53, ASXL1, BCOR, EZH2, RUNX1, SF3B1, SRSF2, STAG2, U2AF1,* and *ZRSR2* are associated with inferior outcome [25, 4]. The intermediate risk category includes patients with t(9;11), *FLT3-ITD* mutations, and other cytogenetic and molecular abnormalities which are neither classified as favorable or adverse [4].

1.1.4 Leukemic stem cells

According to the Cancer Stem Cell (CSC) hypothesis, cancers are organised as cellular hierarchies, at the apex of which resides a type of multipotent cancer cell which exhibits characteristics of normal stem cells – namely, the ability to self-renew and repopulate the tumor mass [26]. In AML, these cells are referred to as LSCs, and are distinguished from leukemic blasts – the transformed cells which make up the bulk of the tumor. In AML this stem cell capacity is usually proven based on the ability of cells to engraft and give rise to leukemia in immune-deficient mice – an ability that was linked to the CD34+CD38- immunophenotype [27]. In clinical settings, LSCs are believed to be a major contributor to therapy resistance and relapse [28]. This has been attributed to the fact that LSCs can exist in a state of cell-cycle quiescence, which allows them to avoid cell-cycle-dependent therapeutic insults [29]. A similar phenomenon is also well documented in other cancer types, where a population of so-called cancer persister cells can exist for years in a state of dormancy, giving rise to relapse if they later re-enter the cell cycle [30].

Based on this understanding, LSC gene expression signatures have been developed as predictive/prognostic biomarkers that could be used to identify AML patients at high risk of relapse after induction chemotherapy. Most notable is the LSC17 score, which is calculated as the weighted gene expression of 17 genes which were selected based on differential expression between LSC+ and LSC- cell fractions, as well as their expression associating with worse clinical outcome [31]. The LSC17 score is known to be increased in the more aggressive AML subgroups such as ckAML.

1.1.5 AML with a complex karyotype

CkAML is an aggressive AML subgroup defined by the presence of $\geq$3 unrelated chromosome abnormalities in the absence of other class-defining genetic lesions [32, 33]. CkAML comprises about 10-12 percent of AML patients, but is more common in the elderly, especially in patients progressing from the pre-malignant MDS, and in therapy-related AML. As mentioned above, ckAML falls into the category of unfavorable risk, however, unlike

some other events in this category, ckAML remains mechanistically a poorly understood subgroup. The majority of chromosomal aberrations in ckAML are unbalanced changes such as deletions and gains, rather than balanced changes such as translocations and inversions [34]. Attempts to further characterise ckAML have been limited, with the exception of the "typical" and "atypical" subclassification proposed by Mrozek *et al.* [33]. By this definition, "typical" ckAML are those which carry deletions on chromosome 5/7q and/or 17p, and are characterised by a higher frequency of *TP53* mutations and a worse clinical outcome compared to "atypical" ckAML lacking these alterations [33, 35]. However, with the exception of del(17p), which is linked to the famous tumor suppressor gene (TSG), *TP53*, the mechanisms linking the most common chromosome deletions [del(5q) and del(7q)] to the development and progression of aggressive leukemias are largely unknown.

1.1.6 Past and present perspectives on deletion 5q AML

A deletion in the long arm of chromosome 5 is the most common copy number alteration in older AML patients, occurring in about 80% of patients with ckAML, and in rare cases, as a sole abnormality [33]. Del(5q) is an early event in leukemogenesis. This is evidenced firstly by the fact that del(5q) is typically a clonal event, and secondly by its frequent occurrence in premalignant MDS, which can progress to ckAML over time [36, 37, 38]. Although the deleted regions are typically large, spanning several tens of megabase (MB)s, analyses of overlapping deleted segments across large numbers of patients have resulted in the definition of a minimally deleted region (MDR) within 5q31.2, where the search for del(5q) TSGs has focused [39, 40, 41, 42]. A conservative definition of the MDR in del(5q) AML and high risk MDS is a region of around 1MB containing 50 genes, flanked by *IL9* and *UBE2D2* [43, 44], however regions as small as four genes have been proposed [42]. While this 5q31.2 MDR is common to high-risk MDS and AML, it is important to distinguish this region from the more telomeric 5q32-33 locus associated with low-risk MDS or so-called "5q syndrome", which has been linked to haploinsufficiency of the ribosomal gene *RPS14*, located on 5q33. Ebert *et al.* identified *RPS14* as the 5q-syndrome haploinsufficiency target through ribonucleic acid (RNA) interference screening, whereby each of the 40 genes in the 5q32-33 MDR were systematically inhibited, and the effects on HSPC function were evaluated. Here only shRNAs targeting the RPS14 gene resulted in a block in erythroid differentiation, which recapitulated clinical symptoms of the 5q syndrome [45, 46].

Haploinsufficiency is a gene dosage effect, whereby loss of only one allele is sufficient to reduce protein levels of its product in such a way that impairs normal cellular function and contributes to tumorigenesis [47]. This stands in contrast to the classic definition of a TSG, which was originally proposed by Knudson, where two "hits" (inactivating mutations or deletions) are required in order for a malignant phenotype to be fully propagated [48]. Since no genes within the 5q31.2 locus were identified to be recurrently mutationally

inactivated on the second allele, the prevailing hypothesis is that the del(5q) phenotype in high-risk AML is also driven by haploinsufficiency of one or more genes in the MDR. Several genes within and around the MDR have been investigated as haploinsufficiency candidates, including *CTNNA1* [44], *EGR1* [49, 50], *CSNK1A1* [51], *KDM3B* [52], *DELE1* [37] and *ETF1* [53]. However, attempts to pinpoint the target of del(5q), and to delineate the mechanisms linking the loss of these genes to leukemic progression in AML have been inconclusive.

CTNNA1, which encodes for the alpha-catenin protein, has been highly regarded among putative del(5q) targets since its proposition in a 2007 study by Liu *et al.* [44]. Here, the promoter of the *CTNNA1* gene was found to be epigenetically silenced, resulting in biallelic inactivation of the gene in a subset of del(5q) samples. This evidence for a "second hit" led the authors to speculate that CTNNA1 may be a tumor suppressor. Moreover, they showed that restoration of the gene's expression in a del(5q) cell line resulted in enhanced cell death. Considering the role of catenin family proteins in regulating cytoskeletal organisation, the authors suggested a model whereby the loss of CTNNA1 in leukemia-initiating/stem cells may disrupt the balance between symmetric (a stem cell giving rise to two identical stem cells, thereby replenishing the stem cell pool) versus asymmetric (a stem cell giving rise to one stem and one differentiated cell) cell fate decisions. They hypothesised that an increase in symmetric cell divisions resulting from CTNNA1 loss could confer a selective advantage to the malignant stem cell pool [44]. However, this remained speculation and was not experimentally validated. The authors also noted inconsistencies in the loss of CTNNA1 expression among del(5q) patients, with 4 out of the 12 del(5q) cases studied retaining normal levels of CTNNA1 expression. Later, Mikhalkovitch *et al.* carried out a range of functional experiments *in vivo* to follow up on the role of *CTNNA1* in AML, and did not detect any signs of aberrant HSC function or the development of myeloid neoplasms in *CTNNA1*-deficient mice [54]. It is also important to note that *CTNNA1* lies slightly outside most restricted definitions of the MDR [55, 39, 56, 57, 42], making it unlikely to be the key leukemogenic factor.

The same year that CTNNA1 emerged as a promising candidate, Joslin *et al.* reported on the haploinsufficiency of the *EGR1* gene [49]. *EGR1* encodes a zinc-finger TF with diverse roles in regulating proliferation, differentiation and apoptosis. They showed that haploinsufficiency of EGR1 led to disrupted myeloid development in an AML mouse model treated with an alkylating agent [49]. EGR1 was considered a promising candidate because it had already been identified as a classical tumor suppressor in other cancer types [58] and was known to play a role in regulating important tumor suppressors including TP53 and PTEN [59, 60]. It was suggested that haploinsufficiency of EGR1 confers a fitness advantage to HSCs in conditions of stress, such as following chemotherapy, potentially explaining why del(5q) is more common in therapy-related AML and in the setting of chronic inflammation associated with aging [50]. *EGR1* remains today one of the most plausible del(5q) target genes, however, some researchers have been cautious to accept

this tempting explanation. Independent studies reported that *EGR1* was inconsistently downregulated in AML patients with del(5q) [61], and that its expression did not correlate with the del(5q) clonal burden [62]. Moreover, the fact that EGR1-deficient cells in Joslin *et al.'s* study did not show evidence of hematopoietic disruption in the absence of a DNA damaging agent led to some scepticism [49]. Later studies suggested that the ability of EGR1-haploinsufficient HSCs to clonally expand was highly variable among recipients. This led to the proposition that another gene – *CSNK1A1* – deletion of which led to a more consistent and stronger clonal advantage in competitive transplantation assays – may be a more likely target [63, 64]. Schneider *et al.* showed that haploinsufficiency of CSNK1A1 induced HSC expansion, with homozygous deletion conferring HSC failure [51]. Moreover, *CSNK1A1* was found to be targeted by mutations in some cases of MDS [51]. Nevertheless this gene lies somewhat distal to the high-risk MDR, making it unlikely a sole del(5q) target in AML.

To date, the smallest interval of the 5q MDR proposed by Mackinnon *et al.* comprises only *EGR1, REEP1, KDM3B* and *ETF1* (the latter being on the border of the region and included tentatively) [42]. Of these genes, only *KDM3B* gene expression was found to be significantly reduced in del(5q) samples [61]. Later it was shown that introducing *KDM3B* into a del(5q) leukemic cell line suppressed clonogenic growth, suggesting that *KDM3B*, which encodes for a H3K9me1/2 demethylase enzyme, might act as a haploinsufficient TSG [52]. Nevertheless *KDM3B* received very little attention compared to other putative targets.

Evidently, the identity of the pivotal del(5q) target gene(s) remains controversial, and the quest to discover such genes is ongoing. For example, a recent study by Spinella *et al.* suggested that haploinsufficiency of the *DELE1* gene, which encodes a mitochondrial protein, could contribute to leukemia progression by inhibiting mitochondrial stress-induced apoptosis [37].

1.2 Epigenetics

The term "epigenetics" refers to any mitotically/meiotically heritable regulatory elements of the genome which can influence gene expression but which do not involve modifications of the underlying DNA sequence. While the DNA sequence remains stable throughout organismal development – being largely identical between different differentiated cell types and the zygote from which they derived – epigenetic marks are dynamic and reversible, being continuously remodeled throughout the cell cycle, and influenced by cell-cell signaling and environmental factors including aging and diseases [65, 66, 67]. The most recognised – but not only – function of epigenetic modifications is in the control of gene expression [68].

The proper organisation of the DNA strand within the nucleus is provided by the chro-

matin structure. Chromatin can be categorised as either heterochromatin, which exists in a highly compact and transcriptionally inactive state, or euchromatin, which is loosely packaged, thus allowing access by TFs and other elements required to permit transcriptional activity [68]. Heterochromatin can be further subdivided into the permanently repressed "constitutive heterochromatin" – which encompasses repetitive regions such as the centromeres and telomeres – and the more dynamic "facultative heterochromatin" – which may be active or inactive depending on the cellular context [69]. Chromatin accessibility and epigenetic control of gene expression is mediated by an array of chemical and protein modifications which occur either on the DNA itself, or on the tails of histone proteins which form the nucleosomes around which the DNA are wrapped (Figure 1.2).

1.2.1 DNA methylation

The primary epigenetic modification of DNA is methylation – the covalent addition of a methyl group (CH3) to the 5' Carbon of the cytosine base within a CpG dinucleotide. CpG dinucleotides are positions of the genome where a cytosine is followed by a guanine. Such sites occur throughout the genome but are enriched in clusters of close proximity – known as CpG islands – which are present at the promoters of most active genes [70]. While most CpG sites throughout the genome are stably methylated, CpG islands in the promoters of active genes are typically unmethylated, and methylation of such promoters is associated with gene silencing [70]. Outside of CpG islands, the link between DNA methylation and gene expression is more variable; for example, methylation of gene bodies has been shown to correlate with active gene expression and may have an impact on gene splicing [71], whereas intergenic methylation can influence gene activity through the modulation of enhancers, insulators, and other regulatory elements [71].

DNA methylation is mediated by a family of enzymes called DNA methyltransferase (DNMT)s, which catalyze the transfer of a methyl group from S-adenosylmethionine (SAM) to DNA. This family includes the so-called maintenance methyltransferase, DNMT1, and the *de novo* methyltransferases, DNMT3A and DNMT3B [72]. The latter are predominantly active during embryogenesis where they are responsible for establishing the patterns of DNA methylation in developing tissues [72]. Meanwhile, DNMT1 is stably expressed in all adult tissues where it copies the already established methylation pattern to daughter cells as part of every cell division [73]. DNA demethyation can be a passive process, whereby dividing cells fail to remethylate newly synthesised daughter strands, for example, if DNMT1 is inhibited. However, DNA can also be actively demethylated – a mechanism mediated by the family of so-called ten-eleven translocation (TET) enzymes [74]. These enzymes add a hydroxyl group onto methylated cytosines, which are subsequently converted back to unmethylated cytosines. Each different cell type harbors a unique signature of DNA methylation which instructs cell-type-specific gene expression and is thus essential for maintaining cell identity throughout mitotic cell divisions [75].

1.2.2 Histone modifications and nucleosome positioning

Beyond DNA methylation, various epigenetic modifications of histone proteins also play a role in regulating gene expression. Four core histone proteins – H2A, H2B, H3 and H4 – each present in two units – form an octamer, around which approximately 147 base pairs of DNA is wrapped. Each such unit is called a nucleosome, and the positioning of such nucleosomes has a critical impact on the regulation of gene expression, by influencing the accessibility of chromatin to TFs and other regulatory proteins [76]. Nucleosome-free, open chromatin regions are associated with active transcription, whereas the positioning of nucleosomes over promoters or enhancers can inhibit the binding of such factors and thereby repress transcriptional activity [76]. Within the nucleosome, each histone protein consists of a globular domain and an N-terminal tail which protrudes from the nucleosome, where it is accessible for placement of various epigenetic modifications. Chemical modifications to the histone tail affect chromatin structure and ultimately gene expression by altering the electrostatic interactions between histones and DNA, and by providing a platform for the recruitment of other chromatin modifying proteins and complexes that modulate transcription [77].

The best studied of such modifications are histone methylation and acetylation, but others exist such as ubiquitination and phosphorylation. The effect of such modifications on gene expression depends not only on the chemical group added, but also at which location within the histone complex and which amino acid is modified. For example, H3K9me3 (trimethylation of the ninth position Lysine on histone H3) is a repressive chromatin mark associated with gene silencing, whereas H3K4me3 (trimethylation of the 4th position Lysine on histone H3) is associated with gene activation and is typically found at the promoters of expressed genes [77]. The overall effect on the chromatin conformation and gene expression is determined by the combination of histone modifications on different residues as well as the number of moieties which are attached (for example H3K9 may be mono, di, or tri-methylated, with different consequences). A large number of enzymes participate in the modification of histones, for example histone lysine methyltransferase (KMT)s and lysine demethylase (KDM)s mediate the addition and removal of methyl groups to histone lysine residues, while histone acetyltransferase (HAT)s and histone deacetylase (HDAC)s control the placement of acetyl groups. Each of these families contain numerous enzymes which exhibit overlapping but distinct patterns of activity and tissue/context-specificity [77].

DNA and histone modifications are closely interconnected. For example, repressive histone marks can direct the establishment of DNA methylation [78], and *de novo* DNA methyltransferases, DNMT3A and DNMT3B are known to interact with several histone methyltransferases including SUV39H1, EZH2 and G9a [79, 80].

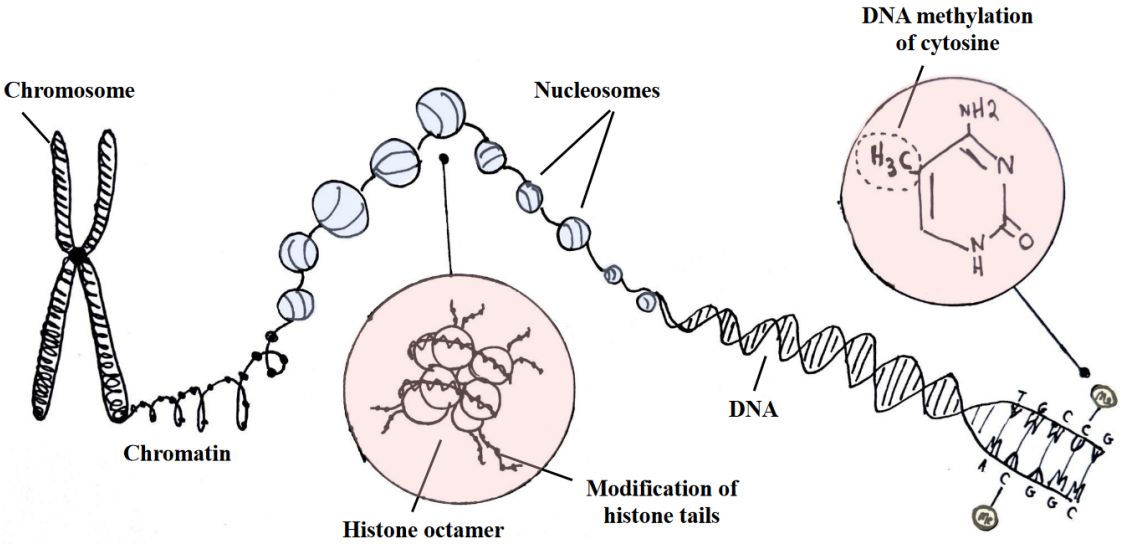

Figure 1.2: Epigenetic modifications of the DNA and histones. Schematic outlining the structure of chromosomes and the major sites of epigenetic modifications in the genome. Histone modifications occur on the tails of histone proteins, which make up the nucleosomes around which DNA is wrapped to form chromatin. The second main epigenetic modification – DNA methlation – involves the addition of a methyl group (CH3) to the 5' carbon of the cytosine ring at CpG dinucleotides.

1.2.3 Epigenetics and epigenetic enzymes beyond the regulation of gene expression

While epigenetics is most often studied as a system overseeing the regulation of gene expression, the importance of epigenetics extends beyond transcriptional control. For example, epigenetic processes are tightly linked to genome stability and DNA repair [81]. In fact, a prevailing theory is that DNA methylation evolutionarily evolved as a means of protecting against genomic disruption by parasitic insertions [82].

The loosening of chromatin structure facilitated by epigenetic modifications is a critical step in the early DNA damage response, which is needed to improve the accessibility of chromatin for the DNA repair machinery [83]. Various chromatin remodelling complexes and regulators of histone modifications are also recruited to sites of DNA damage where they have been implicated in DNA damage detection and repair processes such as non-homologous end-joining (NHEJ) and homologous recombination (HR). Most notably, phosphorylation of the histone H2A is a crucial event in response to DNA damage, and is thought to facilitate the accumulation of DNA repair proteins at damaged sites [84].

Enzymes which mediate histone (de-)/methylation and (de-)/acetylation can also participate in post-translational modification of other non-histone proteins, thereby functioning as regulators of protein activity. An important example is the regulation of the tumor-

suppressor TP53 by lysine methylation and acetylation. Acetylation of various lysine residues in the TP53 protein were shown to increase during DNA damage, serving to induce its activity as a transcription factor. Lysine methylation – mediated by enzymes such as SET9, LSD1, G9a and others – has been shown to have both activating and repressive effects on TP53 activity, depending on the amino acid position and number of methyl groups [85].

1.2.4 Epigenetic disorder in cancer

Cancer has historically been studied as a genetic disease, in which mutations and structural alterations in the DNA give rise to uncontrolled proliferation and evasion of growth signals and thereby drive malignant progression [86]. Nowadays, epigenetic reprogramming is appreciated as a hallmark of cancer, and it is increasingly recognised that epigenetic disturbances can themselves represent cancer driver events, by influencing the transcriptional state of cancer-associated genes [87].

Changes in DNA methylation are the most widely studied epigenetic aberrations in cancer, and are a notable feature of most tumors [88]. Typically, abnormal hypermethylation affects CpG-rich rich regions of the genome such as promoters. Promoter CpG hypermethylation has been recognised as a mechanism of silencing TSGs, and has been described for several well known tumor suppressors including *VHL* in clear cell renal carcinoma and *BRCA1* in breast and ovarian cancers [89, 90]. Beyond these notable targets, a striking example of the widespread CpG island hypermethylation in cancer is the so-called "CpG Island Methylator Phenotype" or "CIMP", which has been most famously recognised in a subset of colorectal cancers, though it has also been observed in other cancer types [91]. CIMP colorectal cancers display exceptionally high levels of CpG island hypermethylation, resulting in the inactivation of several TSGs, most notably the DNA mismatch repair gene, *MLH1*, suppression of which gives rise to genomic instability [92]. CpG hypermethylation in cancer is also associated with an increased rate of cytosine to thymine mutations, which include hotspot mutations for critical TSGs such as *TP53* [93]. Epigenetic alterations in other regulatory regions such as enhancers and CTCF binding sites can also lead to altered gene expression by influencing promoter/enhancer interactions and disrupting the structure of topologically associated domains [94, 95]. In contrast to the localised hypermethylation of CpG islands, cancer cells are characterised by global hypomethylation throughout repetitive DNA sequences and late replicating regions of the genome, which in normal cells remain extensively methylated [96]. This global hypomethylation is thought to be linked to the rate of cell proliferation, which is reflected in its association with late replication timing [97]. In this scenario, the methylation maintenance machinery may fail to remethylate newly synthesised daughter strands during successive rounds of rapid cell division and replication, resulting in a progressive loss of methylation in regions of the cancer genome that replicate later in the S phase of the cell cycle [97]. Global hypomethylation is also believed to promote genomic instability in can-

cer by interfering with the protective function of heterochromatin, notably at telomeres, centromeres and other repetitive regions [98]. Loss of methylation was shown to destabilise peri-centromeric chromatin, thereby promoting large-scale chromosomal rearrangements [96]. Hypomethylation-mediated de-repression of transposable elements can also encourage genomic rearrangements [99]. Thus, epigenomic and genomic disorder in cancer cells need to be appreciated as closely interconnected developments.

The importance of epigenetic dysregulation in human cancer is highlighted by the high frequency of mutations in epigenetic regulators – especially writers and erasers (enzymes which add or remove epigenetic modifications, respectively) of histone methylation. For example, mutations in *EZH2* – which encodes a component of the polycomb repressive complex 2 (polycomb repressive complex (PRC)2) that mediates H3K27me3 methylation – are common in lymphoma (activating mutations) and leukemia (inactivating mutations) [100, 101] and *EZH2* overexpression has been linked to worse clinical outcome in several other cancer types [102]. Mutations in the H3K27me3 demethylase *KDM6A/UTX* are also prevalent in several cancer types including bladder, lung and breast cancer [103]. The gene encoding the H3K4me1 methyltransferase, *KMT2D/MLL2*, is mutated in about 10 percent of all cancers, being among the most frequently mutated genes in lung, bladder and head and neck cancer [103]. These are just a few examples which highlight the pervasive involvement of epigenetic regulators in the mutational landscape of human cancers.

Metabolic disruptions can also manifest in epigenetic changes. A typical example is in AML with mutations in IDH1/2, which results in the accumulation of the oncometabolite 2-Hydroxyglutarate (2-HG). This oncometabolite serves as a competitive inhibitor of various α-ketoglutarate-dependent epigenetic enzymes, including the TET family of DNA demethylases as well as various histone lysine demethylases. IDH mutations therefore result in a signature of DNA hypermethylation [12].

Coupled with the fact that epigenetic alterations are, in principle, reversible events, this background has precipitated a growing interest in epigenetic strategies for cancer therapy [104]. Some such therapeutic compounds are already in clinical use; most notably, the DNA hypomethylating agents, azacitidine and decitabine, which are part of the standard clinical regimen for older patients with AML [105]. Other epigenetic drugs have recently entered clinical development, including for example EZH2 inhibitors, which are undergoing clinical trials for several tumor types including lymphomas and prostate cancer [106].

Epigenetic alterations have also attracted attention from a clinical perspective beyond therapeutic targeting. For example, since epigenetic alterations in cancer are thought to be early events which might precede malignant transformation, they have also gained interest as potential biomarkers for early detection, diagnosis and clinical decision-making. For example, hypermethylation of *MGMT*, which encodes the DNA repair enzyme O6-methylguanine-DNA methyltransferase, was identified as a key prognostic and predictive biomarker in glioma, where it indicates a favorable response to alkylating agent

chemotherapy [107]. In the future, DNA-methylation-based liquid biopsies to detect cancer-associated alterations in circulating tumor DNA or cell-free DNA might also offer promising avenues for detection and monitoring of tumor progression [108, 109].

A deeper understanding of the causes and consequences of epigenetic dysregulation in cancer could therefore have valuable translational implications.

1.2.5 Epigenetic strategies for cancer subgrouping

Patterns of epigenetic dysregulation – in particular DNA methylation – have become of interest for cancer subgrouping. Defining subgroups within heterogeneous tumors can aid diagnosis, risk stratification and therapeutic decision-making, as well as improving how researchers can approach the study of disease. A particularly successful example is the DNA methylation-based characterisation of tumors of the central nervous system proposed by Capper *et al.* [110], which is now used routinely in diagnostic settings. A similar approach was also later proposed for the classification of sarcomas [111].

In this thesis I will focus on AML, a salient example of epigenetic dysregulation and heterogeneity in cancer. Several previous attempts were undertaken to define epigenetic subgroups in AML using DNA methylation profiles. For example Figueroa *et al.* identified 16 methylation based AML subgroups using the HpaII tiny fragment enrichment by ligation-mediated polymerase chain reaction (HELP) assay [112], which focuses on methylation at promoters. These categories were later redefined to 14 subgroups using Enhanced Reduced Representation Bisulfite Sequencing (ERRBS) [113]. More recently, Giacopelli *et al.* defined a set of 13 methylation-based subgroups, so-called "epitypes" using EPIC and 450K array data from the BEAT and the cancer genome atlas (TCGA) AML cohorts [114]. These studies and others have succeeded in linking methylation signatures to many of the most common mutations and structural rearrangements including mutations in *DNMT3A, NPM1*, and *CEBPA*, as well as inv(16), t(15;17) and t(8;21) (rearrangements disrupting CBFβ, PML/RARA and RUNX1/RUNX1T, respectively). Other studies have identified DNA methylation signatures associated with overexpression of the oncogene, *MECOM/EVI-1* [115], as well as rearrangements of the *KMT2A/MLL* locus [32].

1.2.6 Epigenetic heterogeneity and its implications in cancer

Complex biological systems exhibit heterogeneity at multiple levels – from the genetic diversity distinguishing individual organisms from one another, to the epigenetically-regulated functional diversity of tissues and cell types within an organism, and more subtle cell-to-cell variations between cells of a defined type [116].

Cell-to-cell epigenetic heterogeneity is essential for the proper functioning of biological organisms, where it is thought to contribute to the maintainance of pluripotency and the control of differentiation and cell fate decisions [117]. The property of stemness is

associated with an elevated stochastic molecular variation, which endows cells with the potential to follow any of a range of differentiation trajectories. Such molecular "noise" is believed to decrease as cells become progressively more differentiated, during which they commit to a designated cell fate, and acquire the stable transcriptional program which that lineage demands [118]. Such heterogeneity can however have detrimental implications in the context of malignancy [118].

Cancers are in many respects heterogeneous entities. Inter-tumor heterogeneity – referring to the differences between tumors from any two individuals – implies that any two tumors of a given type can evolve in very different directions – exhibiting variations in genetic profiles, phenotypic behaviours (e.g. growth, metabolic and immune dynamics) as well as potentially harboring different epigenetic patterns. This inter-individual heterogeneity entails that patients with the same tumor type may not respond to treatment in a "one-size-fits-all" manner – an anomaly that has precipitated a growing interest in personalised oncology [119]. However, the promise of personalised cancer therapy is complicated by an additional layer of heterogeneity – that which exists between cells within any individual tumor, i.e. intra-tumor or inter-cellular heterogeneity.

Cancer formation begins with the clonal expansion of a single transformed cell, but a tumor can grow to become a heterogeneous mass, exhibiting a range of phenotypic characteristics [120, 121]. Intra-tumor heterogeneity emerges in part through the acquisition of different mutations in different cells, resulting in genetically distinct subclones. However, as the classical genetic view of cancer has expanded to one encompassing epigenetic disorder, it is also increasingly appreciated that phenotypic heterogeneity within tumors can stem from cell-to-cell differences in the patterns of DNA and histone modifications and chromatin accessibility [122].

In recent years this phenomenon of inter-cellular heterogeneity – closely linked to the concept of cancer plasticity – has received particular attention as a potential driver of tumor evolution and therapy resistance [120, 123, 124]. In this scenario, the capacity of a single tumor to exhibit a variety of epigenetic, transcriptional and ultimately functional states, creates a fitness advantage, enabling that tumor to adapt and respond to a variety of external and internal stresses, including microenvironmental pressures and therapeutic insults. This represents a major clinical challenge [121]. The higher the heterogeneity, the greater the chances that a part of the tumor will be intrinsically tolerant to a certain stressor, such that this subset of cells can persist in the face of therapy, and ultimately outgrow to form a resistant tumor.

Another advantage of such cell-to-cell heterogeneity in cancer evolution is in facilitating a "division of labor" that allows a tumor to function efficiently as a system, wherein different cells can take on different roles for the benefit of the whole. One can consider for example the functional diversity underlying the process of cancer metastasis. The metastatic cascade begins with degradation of the extracellular matrix and invasion of

local tissues, followed by extravasation and dissemination into the bloodstream. There, cells must survive passage through the circulation before ultimately settling and colonising a distant organ, where they will need to adapt to the foreign conditions of a new microenvironment. This undertaking requires a diversity of sometimes conflicting "skills" that cannot be shared by a homogeneous group of cells, but more likely distributed across a phenotypically heterogeneous and/or plastic cell population [125, 126]. In epithelial cancers, this process is also believed to be facilitated by a so-called epithelial/mesenchymal plasticity, whereby epithelial cells of the primary tumor acquire a mesenchymal phenotype (epithelial-to-mesenchymal transition) that endows them with invasive characteristics, and later reacquire epithelial traits (mesenchymal-to-epithelial transition) to allow attachment and colonisation in the metastatic site [127].

1.2.7 Techniques for studying epigenetic modifications

Various modern technologies have been developed to facilitate epigenetic studies, including methods for measuring DNA methylation, histone modification, chromatin accessibility and conformation. A range of techniques for studying DNA methylation have been established, most of which rely on the principle of bisulfite conversion. Here, DNA is treated with sodium bisulfite, resulting in the conversion of unmethylated cytosine residues to uracils, while methylated cytosines remain unchanged. Methylated and unmethylated cytosines can thus be distinguished. Genome-wide profiles of DNA methylation can be obtained using sequencing techniques such as whole genome bisulfite sequencing (WGBS), or reduced representation bisulifte sequencing (RRBS), the latter being restricted to CpG-dense regions of the genome [128, 129]. Array-based technologies have also been developed such as the Illumina Infinium HumanMethylation450 (450K) and HumanMethylationEPIC (EPIC) arrays, which provide methylation measurements at 450,000 and 850,000 CpG sites throughout the genome, respectively, with most probes covering promoters, enhancers and other regulatory regions of the genome [130, 131]. These technologies use specialised probes that distinguish between methylated and unmethylated cytosines based on the ratio of fluorescence signal produced during single-base extension and hybridisation with bisulfite-converted DNA [130].

Various techniques have also been developed for studying genome-wide patterns of histone modifications, as well as measuring DNA binding sites for TFs and other proteins. The most widely used is Chromatin Immunoprecipitation with sequencing (ChIP-seq), which relies on the principle of immunoprecipitation; using an antibody specific to the protein (or histone modification) of interest, to capture and enrich DNA segments to which the protein is attached [132]. Later, the purified DNA can be sequenced to determine precisely which regions in the genome the protein (or histone modification) has bound. Other techniques for studying histone modifications include Cleavage Under Targets & Release Using Nuclease (CUT&RUN) [133] and Antibody-guided Chromatin Tagmentation (ACT)-seq [134], which offer advantages over ChIP-seq, such as requiring a lower number of input

cells.

Genome-wide patterns of chromatin accessibility can be studied using the assay for ATAC-seq [135]. This can provide a broader insight into the gene regulatory landcape, since it does not require *a priori* knowledge about any specific epigenetic modification. ATAC-seq relies on the use of an enzyme called Tn5 transposase, which recognises regions of accessible chromatin, cuts the DNA at those regions, and tags it for sequencing [135].

In recent years, studies of DNA methylation, histone modifications and chromatin accessibility have been transformed by developments in single-cell sequencing technologies [136, 137, 138]. Single-cell ATAC-seq is currently the most widely used method for studying epigenetic patterns at the single cell level [136].

1.2.8 Techniques for studying epigenetic heterogeneity

Understanding intratumor heterogeneity is essential to the study of cancer, however it has historically been challenging to study this phenomenon due to technological limitations. Until recently, most studies of cancer have relied on traditional "bulk" sequencing approaches, by which subtle variations within cell populations are easily blurred. Since epigenetic signals are highly cell-type-specific, epigenetic data derived from tumor biopsies can be expected to retain signals from both tumor cells, and various non-cancerous cell types of the tumor microenvironment, with the added complexity that the tumor cell component likely itself represents a heterogeneous population. This cell-to-cell heterogeneity generally complicates the downstream analyses and interpretation of epigenomic data, making it difficult to decipher the causes and consequences of cancer-cell-specific epigenetic disturbances [139].

To address this challenge, various technical and computational strategies have been developed which facilitate studies of epigenetic heterogeneity at different scales. Most notably, the emergence and continuous evolution of single-cell sequencing technologies have made it possible to disentangle intratumor and microenvironmental heterogeneity at unprecedented resolution, at the genomic, transcriptomic and epigenomic levels. Using technologies such as single-cell RNA sequencing (scRNA-seq) and single-cell ATAC-seq (scATAC-seq), the presence of transcriptionally and epigenetically distinct clusters within tumors – the simplest layer of epigenetic heterogeneity – has been demonstrated across a broad range of cancer types [140, 141, 142, 143].

Several computational strategies were also developed to study transcriptional and epigenetic heterogeneity using "bulk" data. Methylome deconvolution provides one strategy for studying the epigenetic heterogeneity within a tumor and its microenvironment. This is a computational method designed to extract cell–type-specific information and other recurring sources of epigenetic variation in DNA methylation data from bulk tissue samples such as tumors [144]. A few recent studies have incorporated methods to quantify cell-to-cell

heterogeneity at the transcriptional level using distance-based or entropy-based metrics, or to quantify transcriptional "noise" using metrics such as the coefficient of variation [145, 146, 147, 148]. Others have taken advantage of read-level DNA methylation data to devise metrics of epigenetic heterogeneity that can be applied to bulk sequencing data [149, 150, 151, 152, 153]. However, such strategies are limited in discerning heterogeneity *between* from heterogeneity *within* cell states, and there have been no attempts so far to develop quantitative metrics of epigenetic heterogeneity in single-cell data.

Chapter 2

Motivation & Aims

AML with a complex karyotype remains a poorly understood subgroup, for which improved characterisation will be needed to decipher the molecular events leading to profound genomic instability and chemoresistance, and ultimately to point towards novel therapeutic avenues for AML patients. Perhaps most pressing is the need to understand the recurrent chromosomal aberrations which exemplify the most aggressive of these tumors, and for which the critical genes and underlying mechanisms remain elusive – the most common of which is del(5q). In this thesis I set out to improve our molecular understanding of AML in older patients through a DNA methylation-based characterisation. My initial findings led me to focus on del(5q) AML, and to investigate the H3K9me1/2 demethylase KDM3B as a haploinsufficiency candidate in this subgroup. In the later part of the thesis, I investigate the hypothesis that haploinsufficiency of KDM3B gives rise to cell-to-cell epigenetic heterogeneity. This hypothesis also motivated the development and validation of a computational strategy to quantify epigenetic heterogeneity in single-cell data. I therefore begin this thesis with a search for order in AML methylomes, which eventually leads me to a search for "chaos". The aims of the thesis are summarised below:

2.1 Characterising the DNA methylation landscape of AML in the elderly

The initial aim of this thesis was to characterise the DNA methylation landscape in elderly AML patients. Towards this, I focused on a DNA methylation dataset from 477 elderly AML patients from the ASTRAL-1 clinical trial [14, 154]. My specific aims were:

- Identify the major DNA methylation signatures in AML in the elderly using a methylome deconvolution approach
- Analyze and interpret the biological sources of each of these methylation signatures

by comparison to known hematopoietic cell types/states, and mutational, cytogenetic and clinical features of tumors

- Place elderly AML patients into epigenetic subgroups and describe each subgroup in terms of its molecular, clinical and cytogenetic characteristics

2.2 Investigating the epigenetic underpinnings of del(5q) AML

My findings from this initial analysis suggested that AML with a deletion on chromosome 5q [del(5q)] may represent an epigenetically distinct subgroup, which was previously unrecognised. This led me to seek an explanation for why del(5q) patients behave differently at the epigenetic level. Here, my specific aims were as follows:

- Investigate the MDR on chromosome 5 in a search for candidate target genes

- Investigate the transcriptional or mutational correlates of the del(5q) methylation signature with epigenetic regulators functionally related to the identified target gene, *KDM3B*

2.3 Developing a computational strategy to quantify cell-to-cell epigenomic heterogeneity

In the latter part of this thesis, I will investigate the hypothesis that haploinsufficiency of KDM3B gives rise to cell-to-cell epigenetic heterogeneity, which may provide a fitness advantage throughout leukemic progression. In order to investigate this, I designed epiCHAOS (epigenetic/Chromatin Heterogeneity Assessment of Single cells) - a computational strategy to quantify cell-to-cell epigenetic heterogeneity in single-cell epigenomic data. Here my aims were as follows:

- Conceive of a suitable metric for quantifying cell-to-cell heterogeneity using single-cell epigenomics data

- Validate the performance of the above metric using synthetic datasets where heterogeneity can be controlled

- Explore the functionality of epiCHAOS by applying it to a wide range of biological datasets from development, cancer and aging

- Use epiCHAOS to investigate differences in epigenetic heterogeneity between groups of cells and across different genomic regions

2.4 Investigating the epigenetic consequences of KDM3B haploinsufficiency

In the final chapter of this thesis, I aimed to investigate the epigenetic consequences of KDM3B disruption in AML. For this I analyzed single-cell ATAC-seq data derived from an OCI-AML3 cell line in which *KDM3B* was heterozygously deleted to mimic haploinsufficiency. Specifically, my aims were as follows:

- Identify the epigenetic differences in *KDM3B* deletion cells
- Apply epiCHAOS to test whether haploinsufficiency of KDM3B results in cell-to-cell epigenetic heterogeneity

Chapter 3

DNA methylation-based characterisation of Acute Myeloid Leukemia

3.1 Results

3.1.1 Epigenetic characterisation of AML in older patients using methylome deconvolution

Note: *The deconvolution pipeline described in this section, including CpG selection, K and Lambda selection and independent component analysis (ICA) was performed together with Linda Welte – a Masters student, under my supervision. The main findings presented within this and the following chapter also appear in the results of Kelly et al. [155] (manuscript under preparation).*

I began this doctoral thesis with the aim of characterising the DNA methylation landscape of AML in the elderly. Towards this, I analysed DNA methylation EPIC array profiles previously generated from the bone marrow or peripheral blood of 477 AML patients who had participated in the ASTRAL-1 clinical trial [154, 14]. This cohort was comprised of diagnostic samples from treatment naive AML patients, who were selected based on ineligibility for intensive chemotherapy, and was therefore enriched for elderly patients, with a median age of 77 (range 59-94 years).

I applied MeDeCom [156, 144] – an established reference-free DNA methylome deconvolution protocol based on constrained non-negative matrix factorisation, which allows decomposition of a set of bulk methylomes to recover their constituent latent methylation component (LMC)s and the proportions of LMCs across patients (Fig. 3.1A). For

deconvolution, I selected the 20,000 most variably methylated CpG sites across patients in order to retain the most informative features, while reducing computational burden. In this way, I partitioned the original "bulk" methylation matrix into a set of 11 LMCs, selected based on cross-validation (Appendix, Fig. 8.1). LMCs can be conceptualised as methylation "signatures" which summarise the major sources of DNA methylation variation within a dataset. In AML samples, they may represent for example different cell types present in the bone marrow or blood (signatures of non-leukemic origin) or other sources of epigenetic variation, such as the effects of mutations in epigenetic regulators (disease-specific sources).

By investigating the differences in LMC proportions across patients, I aimed to decipher the various cell-type-specific and disease-specific signatures, to link each disease-specific signature to distinct molecular and clinical features, and to use the LMCs to define a set of DNA methylation-based subgroups in elderly AML. Wherever specified in the proceeding sections, I also utilised publicly available DNA methylation data from two independent datasets to validate my interpretations – BEAT-OSU (EPIC array, n = 272) [114] and TCGA AML (450K array, n = 190). I separately estimated LMC proportions in these cohorts using a factor regression method provided by MeDeCom.

3.1.2 Methylation signatures of normal hematopoietic cell types and HSPC stages

Since DNA methylation is highly cell-type specific, I first wanted to identify which of the 11 components are derived from non-leukemic cell types present in the bone marrow or peripheral blood. For this I used a previously published DNA methylation dataset from a range of healthy hematopoietic cell types including monocytes, neutrophils, HSPCs, CD4+ and CD8+ T-cells, B-cells and natural killer (NK) cells [157]. I assessed the similarity between methylation profiles of each LMC and those of each normal hematopoietic cell type using Pearson correlations. I found that LMC1 strongly correlated with the methylomes of differentiated myeloid cells (neutrophils and monocytes), while LMC9 correlated strongly with lymphoid cell methylomes (Fig. 3.1B). In line with this, the monocyte/neutrophil-like LMC1 was enriched for hypomethylation of gene ontologies relating to neutrophil/-macrophage signaling processes, such as "granulocyte activation", "neutrophil activation", "neutrophil degranulation" and "neutrophil mediated immunity" (Appendix, Table 8.1). Meanwhile, LMC9-hypomethylated CpG sites were enriched for gene ontologies relating to lymphocyte activity, such as "T cell activation", "lymphocyte differentiation" and "antigen receptor-mediated signaling pathway" (Appendix, Table 8.2). As further confirmation, I estimated tumor purity in these samples using the established InfiniumPurify algorithm, and as expected, identified a strong correlation between LMC9 proportion and estimated sample purity [158] (Fig. 3.1C). Among the remaining LMCs, LMC8 showed a strong similarity to the methylomes of untransformed hematopoietic progenitor cells (Fig. 3.1B).

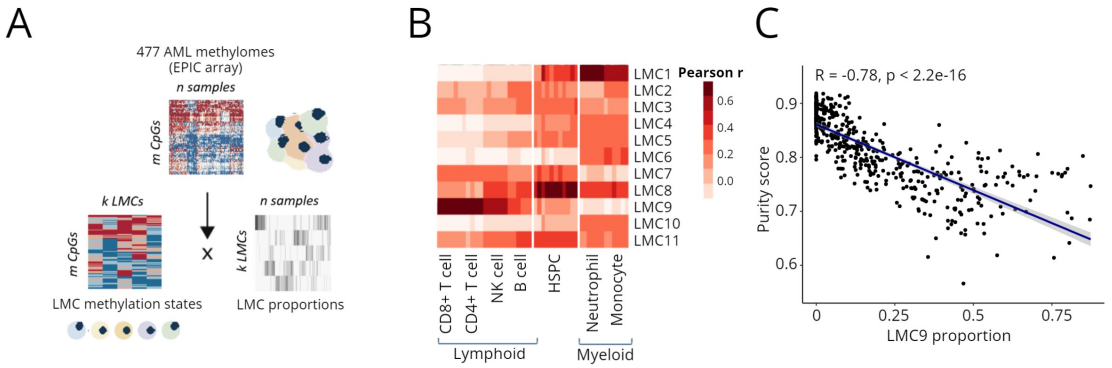

Figure 3.1: Linking LMCs to hematopoietic cell types. A. Schematic outlining the methylome deconvolution approach and its application to the EPIC array dataset. Methylation data measured from bulk samples is represented as a matrix with n columns (samples) and m rows (CpG sites). MeDeCom decomposes the matrix to derive two further matrices; one of LMCs (m CpGs x k LMCs) and one of LMC proportions (k LMCs x n samples). **B.** Heatmap showing correlation of AML LMCs to normal hematopoietic cell methylomes. Colour intensity represents Pearson correlation coefficient. **C.** Scatter plot correlating LMC9 proportion with sample purity estimated by InfiniumPurify. Pearson correlation coefficient and p-value shown.

As it is believed that AML can arise from different stages of myeloid progenitor development, which are epigenetically distinct, I hypothesised that some LMCs might reflect differences in the leukemic cell of origin. To test this I estimated the likely cell of origin of each AML methylome according to a previous strategy [159]: using 450K data from a range of sorted hematopoietic progenitor cells (HSCs, MPPs, LMPPs, CMPs, GMPs and MEPs) from healthy donors, I obtained a set of CpG sites that were differentially methylated between states. Based on hierarchical clustering on these CpG sites I identified three groups of GMP-like, MEP/CMP-like and MPP/LMPP-like AML samples (Fig. 3.2A). As previously described, the majority of AML samples resembled GMP stage methylomes, with smaller groups bearing similarity to the less differentiated progenitors. I tested for differences in LMC proportions across these three groups, and identified two components associated with less differentiated stages: LMC7 was strongly enriched in MPP/LMPP-like AML while LMC11 was enriched in MEP/CMP-like samples (Fig. 3.2B).

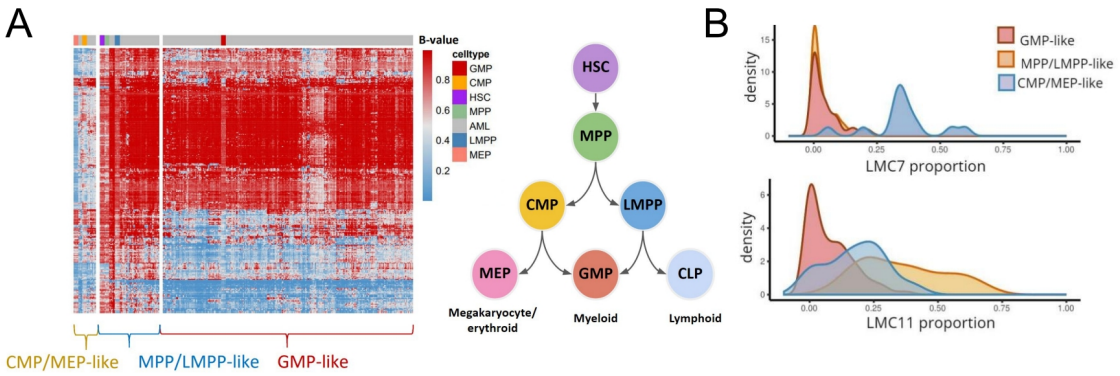

Figure 3.2: Linking LMCs to hematopoietic progenitor cell states. A. Heatmap obtained from hierarchical clustering of AML methylomes together with HSC, MPP, LMPP, CMP, MEP and GMP cells. Samples are clustered based on 216 regions of differential methylation between progenitor cell states. The resulting GMP-like (257), MPP/LMPP-like (47) and CMP/MEP-like (13) clusters serve as an estimate of the AML cells of origin. **B.** Density plots comparing LMC7 and LMC11 proportions in AML methylomes predicted to derive from GMP-like, MPP/LMPP-like, and CMP/MEP-like cells of origin as defined in (**A**).

3.1.3 Methylation signatures reflecting disruption of epigenetic regulators and transcription factors

Next, I decided to investigate which of the remaining LMCs might reflect epigenetic disruptions due to mutations in epigenetic regulators, which are common in AML. I tested for differences in LMC proportions in relation to patients' mutational status, and identified several LMCs enriched in patients carrying epigenetic regulator mutations, or their combinations (Fig. 3.3). I observed an enrichment of LMC4 in *NPM1*-mutated samples, and LMC5 in AML with *CEBPA* mutations. LMC10 was enriched in samples carrying co-occurring mutations in *DNMT3A* and *NPM1* LMC10 captured a signature of hypomethylation, with the lowest global methylation of all LMCs (median beta value = 0.29), consistent with previous descriptions of *DNMT3A*-mutated AML [160]. LMC10 was also increased in *FLT3-ITD* mutated AML and was enriched for hypomethylation at STAT transcription factor binding site (TFBS)s, as has been previously described for this subgroup [114] (Appendix, Table 8.3). In contrast, LMC11, which was enriched in *IDH1/2* mutated AML, had the highest methylation of all LMCs (median beta value = 0.81), capturing a global hypermethylation signature which is typical of this subgroup [161]. Methylomes predicted to derive from less differentiated cells (MEP/CMP-like and MPP/LMPP-like AML) also carried higher levels of LMC11 compared to the more differentiated GMP-like AML. Interestingly, patients with co-occurring *DNMT3A/IDH* and *DNMT3A/TET2* mutations, rather than carrying either global hyper/hypomethylation signatures, were enriched for the normal HSPC-like LMC8. This pattern is reminiscent of the epigenetic antagonism described by Glass *et al.*, who showed that *IDH1/DNMT3A* dual mutation resulted in a methylation landscape similar to normal CD34+ cells [113]. LMC6 was enriched in samples with co-occurring *NPM1/TET2* and *NPM1/IDH* mutations (Fig. 3.3), and was also closely associated with *MLL/KMT2A* rearrangements and monocytic M1 FAB classification (Fig. 3.4A).

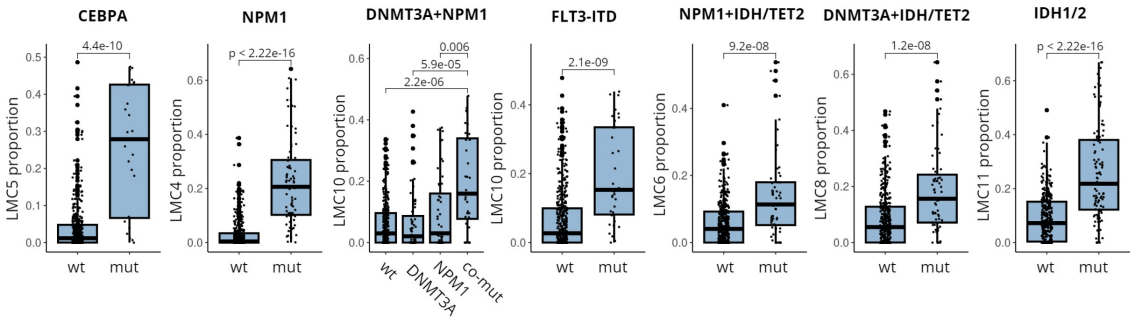

Figure 3.3: Methylation signatures associated with epigenetic regulator mutations. Boxplots showing association of selected LMCs with mutations and combinations of mutations in epigenetic regulators. Wilcoxon's p-values shown.

The remaining components, LMC2 and LMC3, were not associated with any specific epigenetic regulator mutations. To gain insight into these signatures, I considered their association to cytogenetic features, which might also disrupt epigenetic regulators or transcription factors. I found that LMC2 was increased in patients with inv(16), however many LMC2-high patients did not carry this alteration. I hypothesised that other events in these patients might converge on a similar epigenetic disruption. Since inv(16) results in disruption of the RUNX1 binding partner CBFβ, I investigated samples with other RUNX1/CBFβ-associated alterations. I found that LMC2 was the dominant signature in patients carrying a range of alterations convergent on the disruption of RUNX1 signaling (Fig. 3.4B), being especially high in a subset of ckAML carrying *RUNX1* mutations and/or deletions in chromosome 21q, in which the MDR has been mapped to the *RUNX1* locus [162]. LMC2 was also enriched in patients with mutations in *CBFβ*, and consistent with this, both RUNX1 and CBFβ transcription factor binding sites were highly enriched among LMC2-hypermethylated CpG sites (Appendix, Table 8.4). Finally, I found that LMC3 was enriched in patients with ckAML compared to those with normal karyotype or with fewer than three CNAs (Fig. 3.4C).

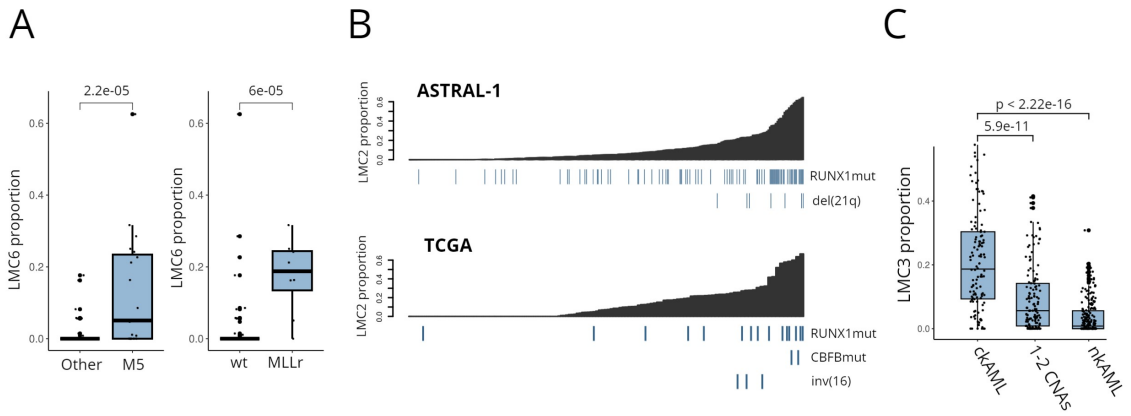

Figure 3.4: Epigenetic signatures linked to MLL rearrangement, RUNX1 disruption and ckAML. A. Boxplots showing association of LMC6 with M5 (monocytic AML) FAB classification (left) and KMT2A/MLL rearrangement (right). Analysis is based on estimated LMC proportions in the BEAT AML cohort. Wilcoxon's p-values shown. **B.** Bar plots showing association of LMC2 with RUNX1/CBFB disruption in ASTRAL-1 and TCGA cohorts. Patient samples are ordered by increasing LMC2 proportion. **C.** Boxplots comparing LMC3 proportion in relation to patterns of copy number variation. Patients are classified as normal, intermediate or complex karyotype according to the presence of CNAs as detected by Conumee [normal karyotype (nk)AML; no CNAs, intermediate karyotype; 1-2 CNAs, complex karyotype (ck)AML; $\geq$ 3 CNAs]. Wilcoxon's p-values shown.

3.1.4 DNA methylation subgroups defined by unique molecular, cytogenetic and clinical features

Next, to define methylation-based subgroups, I applied consensus k-means clustering to the matrix of LMC proportions [163]. For this I excluded the lymphoid-like and neutrophil-like components, LMC1 and LMC9, such that only leukemic-cell signatures – and not signatures of the microenvironment – would drive clustering of patients (Fig. 3.5). I selected k = 9 as the optimum number of clusters based on the Cumulative Distribution Function (CDF) [163] (Appendix, Fig. 8.2). Each of the resulting clusters was either dominated by one LMC, or two co-occurring LMCs, and could be linked to a defining molecular or cytogenetic feature; an LMC2-high cluster representing RUNX1 disruption, an LMC3-high cluster enriched for ckAML, an LMC4-high cluster enriched for *NPM1* mutations, an LMC5-high cluster enriched for *CEBPA* mutations, an LMC11-high cluster enriched for *IDH* mutations, an LMC7-high cluster enriched for LMMP/MMP-like AML, and the remaining clusters associated with combinations of *IDH, DNMT3A* and *NPM1* mutated samples (Fig. 3.5)

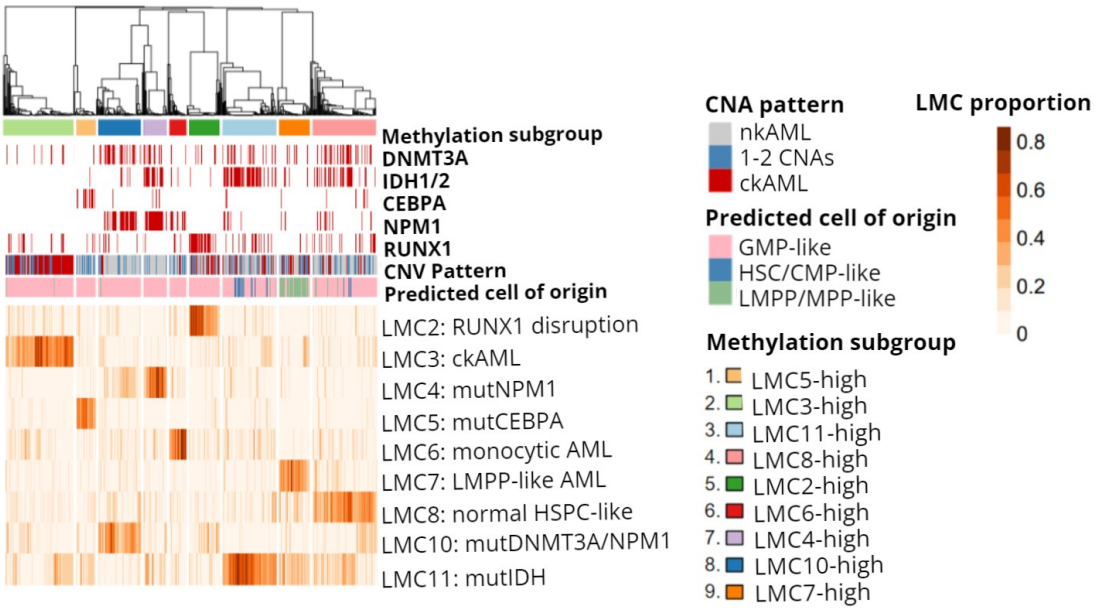

Figure 3.5: DNA methylation-based subgroups. Heatmap of LMC proportions used to define methylation-based subgroups by consensus k-means clustering. Lymphoid-like (LMC9) and neutrophil-like (LMC1) components, and low purity samples (lower tertile InfiniumPurify score) are excluded for clustering. Heatmap is annotated for mutations in *DNMT3A, IDH1/2, CEBPA, NPM1* and *RUNX1*, CNV patter. (complex karyotype, normal karyoptype, or otherwise) and predicted cell of origin (as defined in Fig. 3.2).

I investigated the clinical significance of each of these LMCs by performing survival analyses using data from the BEAT AML cohort. As expected, given its enrichment in ckAML, LMC3 was exclusively associated with poor clinical outcome [34]. Meanwhile LMC5 was associated with improved overall survival, consistent with the known association of *CEBPA* mutation with favorable prognosis [164]. LMC10 was also associated with favorable outcome, albeit with marginal statistical significance (Fig. 3.6).

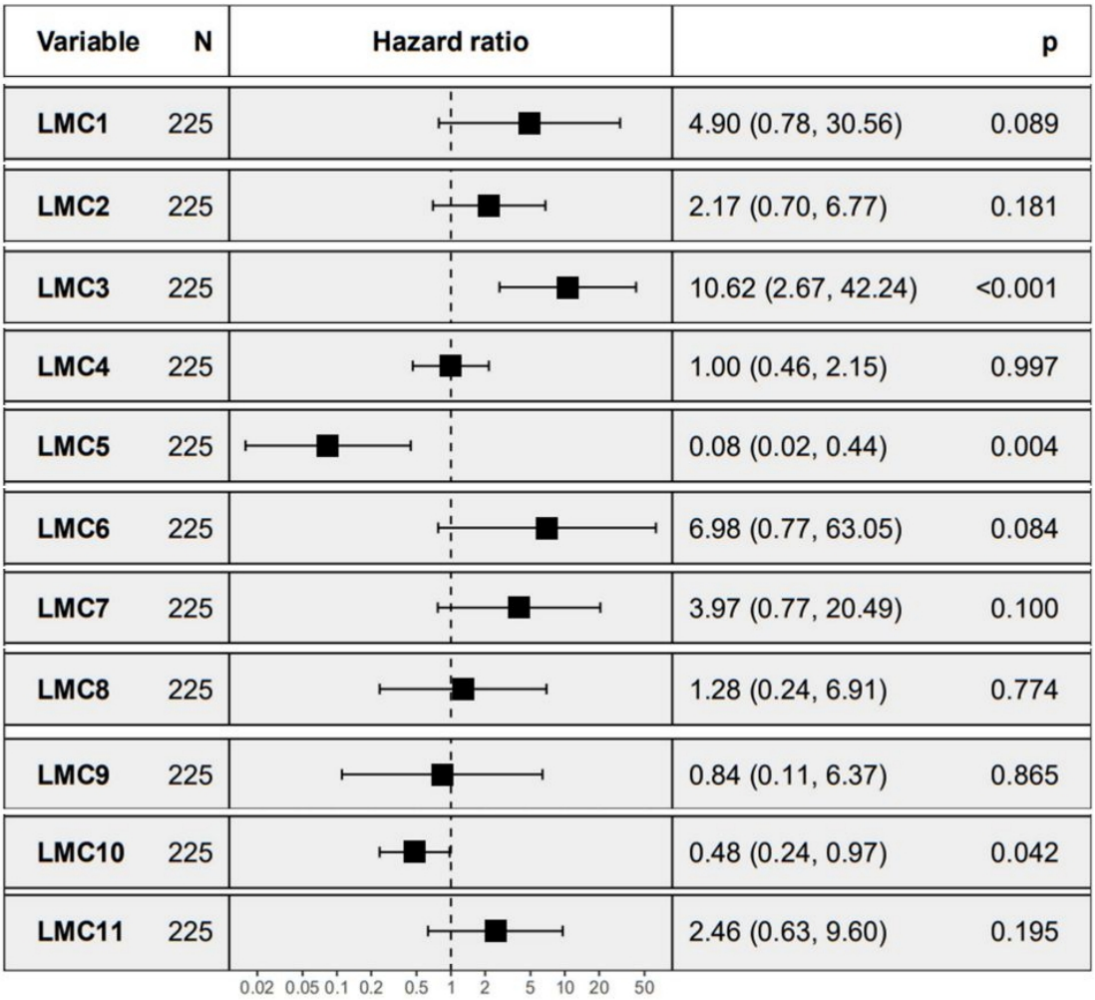

Figure 3.6: Survival analysis of LMCs. Forest plot showing the association of each latent methylation component (LMC) with overall survival. Survival analysis was performed using clinical data from the BEAT AML cohort. Cox proportional hazard ratio's and p-values are shown.

3.1.5 Del(5q) AML is an epigenetically distinct subgroup defined by a signature of DNA hypermethylation

In the previous section I noted that LMC3 was enriched in ckAML samples. However, the levels of LMC3 were also highly variable within this subgroup; not all ckAML samples carried LMC3, and some non-ckAML samples also carried the signature. This prompted me to investigate in more detail, whether any specific CNA patterns might distinguish LMC3-high from LMC3-low patients. I compared LMC3 proportions between ckAML patients with or without each individual chromosome/arm-level CNA. After multiple testing correction, I found only one significant association: LMC3 was strongly enriched in patients carrying a deletion in chromosome 5q (Fig. 3.7A-B). I validated this association in two independent datasets from TCGA and BEAT-OSU AML cohorts and found a similar pattern (Appendix, Fig. 8.3). Moreover, I showed that LMC3 was not specific to ckAML samples, but that patients with an isolated 5q deletion also carried significantly higher LMC3 proportion compared to non-ckAML samples lacking 5q deletions (Fig. 3.7B). Importantly, LMC3 showed no independent association with del(7q) or any other recurrent CNAs (Fig. 3.7A).

Since del(5q) AML has been shown to be enriched for LSCs [165], which may be epigenetically distinct from leukemic blasts [159], I considered whether LMC3 might reflect an LSC-specific epigenetic signature. I used a previously described 450K dataset from sorted AML LSCs (CD34+CD38+/CD38-) and blasts (CD34-) to compare the levels of LMC3 between LSC and blast counterparts of del(5q) samples[159]. I found that LMC3 was present in both del(5q) LSCs and blasts (Appendix, Fig. 8.4A). The signature did appear higher in del(5q) LSCs compared to blasts, however this could not be statistically determined, since the dataset contained only a single del(5q) patient.

Three del(5q) patients in the ASTRAL-1 dataset had unusually low LMC3 proportions, so I decided to investigate these as outliers: I found that one of these samples carried a co-occurring *IDH* mutation and therefore resembled other *IDH* mutated samples, being dominated by LMC11 ; the second was affected by whole genome duplication and therefore retained three copies of 5q, and the third carried a deletion spanning only the distal region of 5q and not the MDR. These outlying cases strengthened my hypothesis that the LMC3 methylation signature results from the loss of some entity from the 5q deleted region.

3.1.6 A hypermethylation signature enriched at developmental genes

To begin to disentangle the source of this methylation signature, I investigated the epigenetic differences between LMC3-high AML and other AML patients. I began by defining a set of CpG sites which are differentially methylated in LMC3-high AML; firstly I compared the methylomes of the LMC3-high AML cluster to all remaining AML clusters, and secondly to the methylomes of normal HSCs. This way, the analysis could leverage all 850,000 CpG sites covered by the assay, and I could consider methylation changes from a

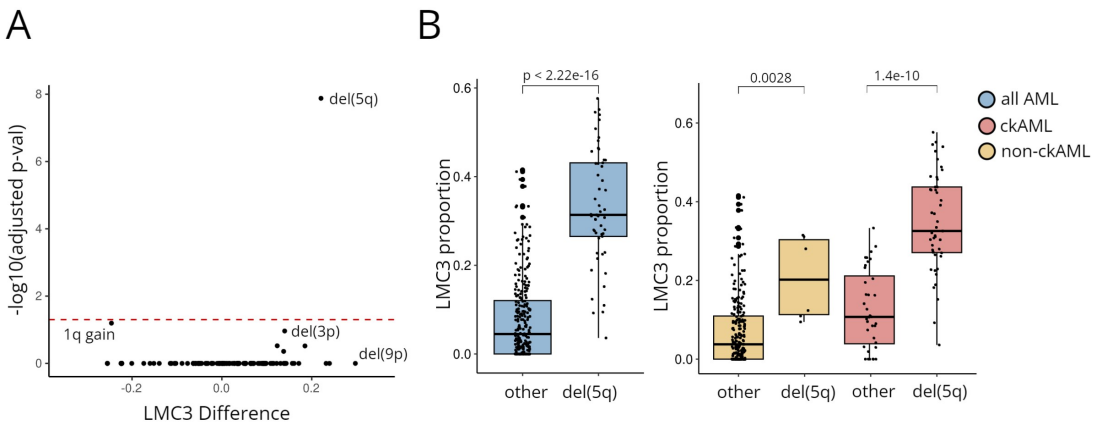

Figure 3.7: Linking the LMC3 methylation signature to del(5q) AML. A. Volcano plot summarising the comparison of LMC3 levels against each copy number gain and loss among ckAML patients. The x-axis depicts the difference in mean LMC3 proportion between patients with and without the specified CNA, and the y-axis shows the -log10 transformed, Bonferroni-corrected Wilcoxon p-value for each comparison. The dotted red line indicates an adjusted p-value of 0.05. **B.** Boxplots comparing LMC3 proportion in patients with and without del(5q), among all (left), all non-ckAML (middle) and all ckAML (right) patients from our cohort. Wilcoxon's p-values shown.

normal state as well as between malignant states. Both comparisons revealed a majority of hypermethylated CpG sites in LMC3-high AML (Fig. 3.8A). Taking the overlap from both comparisons, I defined a set of LMC3-specific hypermethylation (2020 CpG sites) and hypomethylation (426 CpG sites) events. I investigated the enrichments of TFBSs and gene ontologies for both of these sets. LMC3 hypomethylated regions showed few enrichments of gene ontology terms or TFBSs; therefore I focused my analysis of LMC3 as a signature of hypermethylation. LMC3 hypermethylated sites were distributed throughout CpG islands and shores as well as at CpG poor intragenic regions and gene bodies, but were not noticably enriched at any of these regions compared to other CpG sites on the EPIC array. Locus enrichment analysis of LMC3-specific hypermethylated CpG sites revealed an enrichment at sites of H3K27me3, binding sites for the PRC2 complex and for the H3K9/36 demethylase, KDM4A (Fig. 3.8B). This pattern of hypermethylation may be relevant in the context of LSCs, since methylation of polycomb targets in other cancers has been suggested to lock cells in an undifferentiated state [166]. Furthermore, gene ontology enrichment analysis revealed that these hypermethylated sites were enriched for various developmental gene sets (Appendix, Table 8.5). This included an enrichment of homeobox genes, which are believed to play an important role in AML biology (Fig. 3.8C) [167]. Such regions were largely unmethylated both in healthy HSCs and throughout the differentiated hematopoietic lineage (Appendix Fig. 8.4B).

To evaluate whether this hypermethylation translates to the level of gene expression, I performed gene set variation analysis (GSVA) [168], and investigated the pathway-level

changes in gene expression that correlated with LMC3. In line with the enrichment of developmental genes among del(5q) hypermethylated sites, numerous gene sets associated with developmental processes and stem cell differentiation were found among the top gene sets negatively correlating with LMC3 (Appendix, Table 8.6).

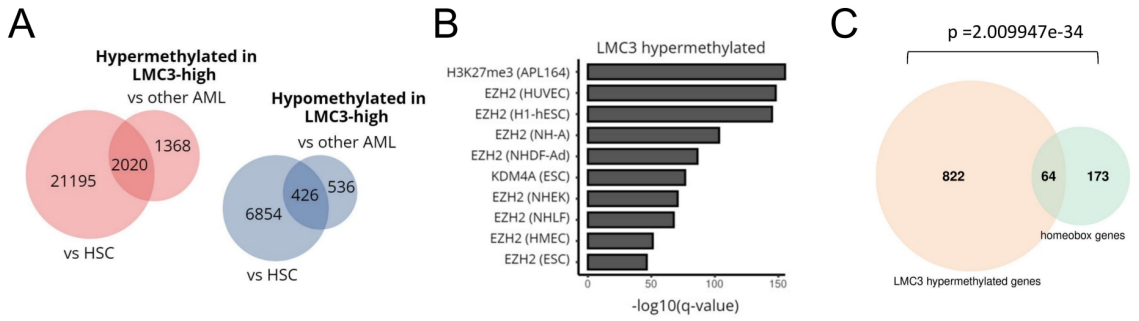

Figure 3.8: Locus and gene set enrichments of LMC3-hypermethylated CpG sites. A. Venn diagram showing the number of overlapping CpG sites detected as hypo/hypermethylated in the LMC3-high AML subgroup by comparison to other AML subgroups and normal HSCs. **B.** Bar plot showing top 10 locus enrichments among LMC3-high hypermethylated CpG sites, ranked by -log10 transformed p-values. **C.** Venn diagram showing the intersection of homeobox genes among del(5q) hypermethylated CpG sites. Hypergeometric p-value is shown indicating enrichment of the gene set.

3.1.7 LMC3 carries prognostic significance independently of del(5q)

Del(5q) AML, being part of the subgroup of so-called "typical ckAML", is known to be associated with dismal clinical outcome [33]. The above association of LMC3 with lower overall survival was therefore unsurprising. Yet, given that a substantial number of AML patients without del(5q) also carry high levels of LMC3, I decided to investigate the clinical significance of this methylation signature beyond the known prognostic indication of del(5q) itself. For this I took advantage of publicly available clinical and drug sensitivity data from the BEAT-OSU AML cohort, which contained a sufficiently large number of LMC3-high patients both bearing and lacking del(5q). I separated patients into two groups based on LMC3 proportion – LMC3-high (above mean levels of LMC3) and LMC3-low (below mean levels of LMC3) – and further separated the LMC3-high group into two groups based on the presence or absence of del(5q). Irrespective of del(5q) status, I found that LMC3-high patients had significantly lower overall survival (Fig. 3.9A) as well as overall reduced drug sensitivity (Fig. 3.9B).

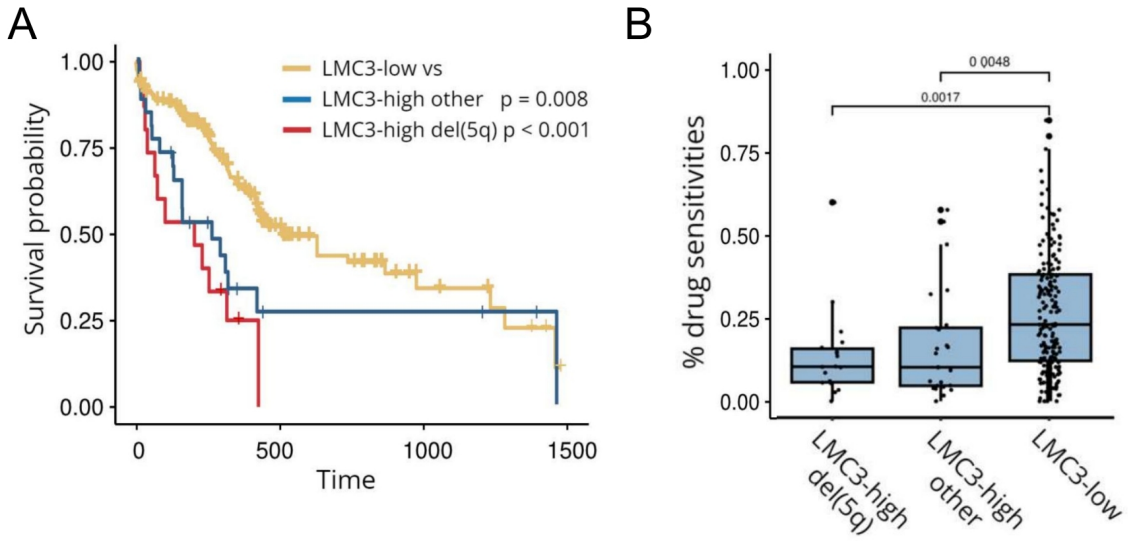

Figure 3.9: LMC3 is associated with poor prognosis and reduced drug sensitivity independently of del(5q) status. A. Kaplan Meier survival plot comparing overall survival of patients in the BEAT AML cohort. Patients are separated into LMC3-high and LMC3-low groups based on mean LMC3 proportion, and LMC3-high patients further stratified as del(5q) or otherwise. Pairwise log rank p-values shown. **B.** Boxplot comparing overall drug sensitivity based on *ex vivo* drug sensitivity screens on BEAT AML samples. Patients are separated as in (**A**). For each patient, the proportion of drug sensitivities is quantified after assigning binary sensitive/resistant calls for each sample to each drug. Samples within the top 20% of area under the curve (AUC) values for a given drug were considered sensitive.

3.1.8 LMC3 is present at low levels in a small subgroup of del(5q) MDS

Individuals with MDS often carry interstitial deletions in chromosome 5q which overlap with the MDR in AML. Although the majority of del(5q) MDS do not develop to aggressive disease, some such patients are at risk of progression to ckAML. DNA methylation patterns were previously found to separate del(5q) MDS into two subgroups, with the more hypermethylated subgroup demonstrating poorer overall survival [169]. This led me to consider whether a similar methylation signature might be present in a subset of del(5q) MDS. To test this I estimated LMC proportions in a del(5q) MDS 450K dataset [169]. As expected, MDS methylomes for the most part resembled those of untransformed HSPCs, being enriched for the HSPC-like LMC8, however LMC3 was also present at low levels in a small number of samples (Fig. 3.10A). Clustering of MDS methylomes revealed a clear separation of two del(5q) MDS subgroups; one of which harboured an LMC3-like methylation signature, though at lower levels than del(5q) AML (Fig. 3.10B).

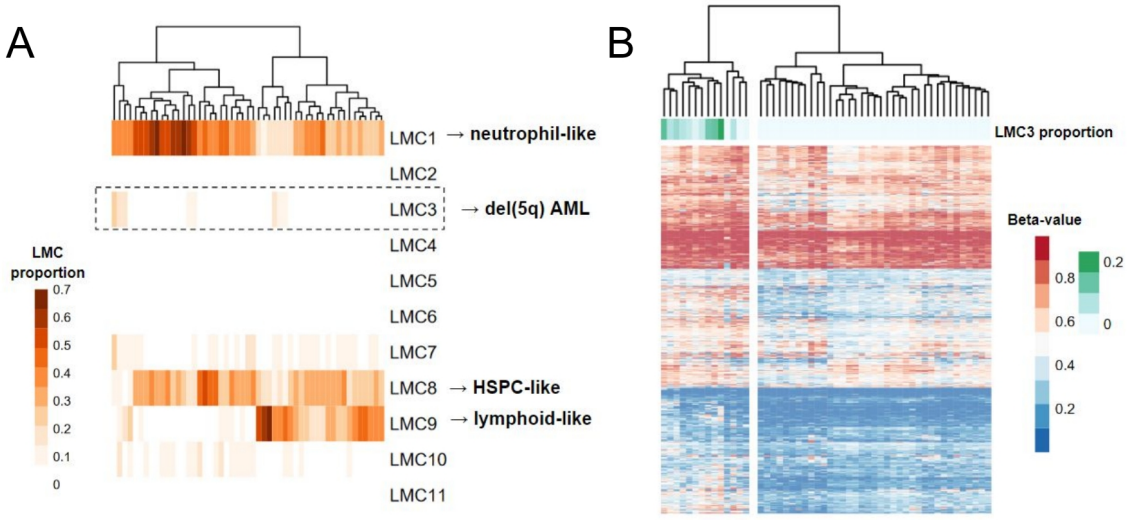

Figure 3.10: LMC3 is present at low levels in a small subgroup of del(5q) MDS. A. Heatmap of LMC proportions defined in ASTRAL-1 AML and estimated here in 450K data from del(5q) MDS samples [169]. **B.** Heatmap resulting from hierarchical clustering of the same MDS samples based on CpG sites which are hypermethylated in the LMC3-high AML subgroup. LMC3 proportion is annotated above.

3.2 Discussion

Improved epigenetic characterisation of AML in the elderly

AML is an epigenetically heterogeneous cancer type, and numerous attempts have been made to characterise AML based on DNA methylomes (see **Introduction** for details). These studies have highlighted the complex DNA methylation landscape of AML, wherein distinct epigenetic patterns have been linked to many of the most common mutational and cytogenetic events [112, 113, 114, 115, 32].

The results presented in this chapter represent the largest comprehensive DNA methylation-based characterisation of elderly AML patients to date. My approach differed from previous studies in two main ways; first, instead of subgrouping patients based on clustering of "bulk" methylation signals, I incorporated a methylome deconvolution method. The main advantage of this strategy is that it allowed me to identify and exclude methylation signatures derived from non-leukemic cell-types, and thus focus on cancer-specific signatures. Moreover, this approach offered the possibility to decipher intra-tumor heterogeneity among leukemic cells themselves, insofar as LMCs can be expected to represent different leukemic cell states. Secondly, my study focused on elderly AML patients, in which certain cytogenetic alterations are especially prevalent. What proved particularly important is the fact that our cohort included a much higher frequency of ckAML compared to previous analyses. CkAML is typically viewed as a genetic disease and has never been comprehensively characterised from an epigenetic perspective. Previous work by Giacopelli *et al.* revealed that ckAML is associated with a DNA methylation pattern characterised by hypermethylation at PRC2 targeted regions and features of inflammation [114], however, perhaps due to the relatively small number of ckAML samples in that study, the authors did not narrow down their observation to any particular subgroup within ckAML. Owing to the large number of ckAML patients in our cohort, it was possible to show that the methylation signature I identified was specifically associated with the 5q deletion, and not with ckAML more generally, or with any other recurrent CNAs in this subgroup.

Opening an epigenetic perspective on ckAML and del(5q)

It is quite surprising that despite the widespread disruption of epigenetic enzymes in almost all other AML subgroups – and the usual absence of such mutations in del(5q) AML – an epigenetic basis for del(5q) has never been postulated. Here I suggest that del(5q) AML represents an epigenetically distinct subgroup defined by a unique signature of DNA hypermethylation – a finding which could shed light on the mechanisms underlying leukemic progression in some of the most aggressive AML patients.

Although del(5q) has not been linked to epigenetic dysregulation before, it is interesting to note that some of the most highly-regarded candidate target genes of del(7q) – the second

most common CNA in ckAML, which often co-occurs with del(5q) – encode epigenetic regulators – most notably the H3K27me3 methyltransferase, *EZH2* [170], and the H3K4 methyltransferase *MLL3/KMT2C* [171]. Both of these genes are located within the MDR on chromosome 7 and are mutated in a subset of del(7q) AML patients. Given this background, and the fact that del(5q) and del(7q) co-occur in many ckAML patients, it is important to clarify that the methylation signature I identified was independently associated with del(5q) and not with del(7q). This leads to the enticing question which I will tackle in **Chapter 4** of this thesis: might there also be an epigenetic regulator targeted by the deletion on chromosome 5? If it is the case that both of the most recurrent CNAs in ckAML target epigenetic regulators, it may be time to reevaluate the idea that ckAML is driven by genetic aberrations, and to appreciate it rather as an epigenetically dysregulated subgroup.

In this chapter I also noted that a subset of individuals with del(5q) MDS harbour a similar methylation signature – though at lower levels – compared to del(5q) AML. Many cases of del(5q) MDS are not considered at high risk of progression to ckAML, but some – mostly those which acquire in addition a TP53 mutation – progress to ckAML with poor clinical outcomes [172]. Since I have also shown that this methylation signature is associated with worse overall survival in AML patients – even in patients which do not harbour 5q deletions – it would be interesting to investigate whether LMC3 may provide insight into the risk of progression of del(5q) MDS to AML.

"Epigenetic antagonism" in AML with co-occurring & conflicting mutations

Beyond the identification of a del(5q) methylation signature, a few further interesting observations from this analysis should be highlighted. One was the finding that co-occurring mutations in *DNMT3A/IDH* or *DNMT3A/TET2* resulted in a methylation signature resembling normal HSPCs. A similar phenomenon was previously described by Glass *et al.*, for *IDH* mutations, which they termed "epigenetic antagonism", but this was not previously shown for *TET2* mutations [113]. It would be interesting to investigate this finding at the single-cell level to clarify whether the observed normal-like methylome actually reflects the compounded effects of these two epigenetic aberrations within the same cells, or whether separate subpopulations of mutated cells exist which appear to antagonise when examined in data from mixed populations.

An epigenetic signature of RUNX1 disruption

My analysis also highlighted the epigenetic similarity of AML with aberrations affecting the RUNX1 signaling network. I identified a methylation signature common to AML with *RUNX1* mutations and del(21q), in which the minimally deleted region peaks at the *RUNX1* locus. Additionally, mutations and inversions [inv(16)] affecting the RUNX1 binding partner, CBFβ, resulted in the same signature. Individually, some of these events have been linked to DNA methylation changes in previous studies [173, 174], but were not

previously described to converge on a common epigenetic signature.

In summary, these findings highlight a need to explore epigenetic mechanisms of disease progression in ckAML, and may in particular open a new perspective towards understanding the del(5q) subgroup.

Chapter 4

Investigating the epigenetic underpinnings of del(5q) AML

4.1 Results

In the previous chapter I have shown that AML with del(5q) represents an epigenetically distinct subgroup, which is separable from other AML subgroups based on DNA methylation patterns. Del(5q) AML has not previously been characterised from an epigenetic perspective, and the mechanisms underlying this recurrent deletion remain largely elusive. I therefore hypothesised that LMC3 might offer some insight into the pathogenesis of this aggressive subgroup. In this chapter I will set out to understand what links the deletion on chromosome 5 to an aberrant epigenetic state, reflected in the DNA methylation landscape.

4.1.1 Mutations in epigenetic regulators are rare in del(5q) AML

To first exclude the possibility that LMC3 is caused by a mutation in an epigenetic regulator, I examined the frequency of mutations in a range of epigenetic regulators in del(5q) AML. Interestingly, I found that epigenetic regulator mutations were significantly less frequent in del(5q) patients compared to other AML cases (Fig. 4.1) – an association that was not apparent for ckAML in general. I concluded that LMC3 is not caused by a mutation in an epigenetic regulator, and considered instead whether the loss of an epigenetic regulator encoded from the deleted chromosome might explain the observed hypermethylation signature.

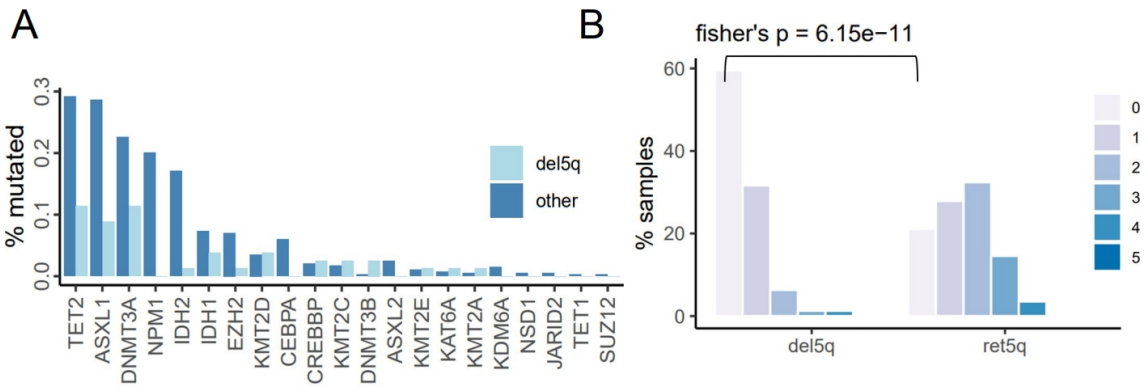

Figure 4.1: Mutations in epigenetic regulators are rare in del(5q) AML. A. Bar plot comparing the frequency of each individual epigenetic regulator mutation in del(5q) and other AML samples. **B.** Bar plot illustrates the frequency of epigenetic regulator mutations (genes listed in **A.**) in del(5q) AML and other AML samples. Fisher's exact test was used to compare the frequency of having at least one epigenetic regulator mutation.

4.1.2 Narrowing down candidate genes in the minimally deleted region

A number of candidate TSGs have been localised to the minimally deleted region at 5q31.2, however there remains no clear mechanistic explanation that convincingly links the loss of any such gene to leukemic progression [41, 39, 40]. Finding a methylation signature enriched in del(5q) patients encouraged me to consider the question of del(5q) target genes from an alternative perspective. I decided to interrogate the minimally deleted 5q region and to examine the genes within this interval as potential drivers of epigenetic dysregulation.

I began by analyzing the overlap of significantly deleted segments using methylation based (Conumee) copy number profiles from 79 del(5q) cases from the ASTRAL-1 cohort [14]. This allowed me to define a MDR of approximately 1.5MB — an interval containing 20 genes, flanked by *MYOT* and *SIL1*. This definition was largely in agreement with previously defined MDRs (Fig. 4.2). To narrow down a list of likely candidates, I first correlated the expression of each gene within the region against the levels of LMC3. To remain cautious, I considered the 50 genes flanked by *Il9* and *UBE2D2* as an accepted and conservative definition of the MDR [43, 44]. Only three of these genes demonstrated a significant correlation which was maintained in independent datasets: the histone Lysine Demethylase 3B (*KDM3B*); the Eukaryotic Translation Termination Factor 1 (*ETF1*), and Catenin Alpha 1 (*CTNNA1*) (Fig. 4.2A). All three of these genes have been studied as candidate tumor suppressors in del(5q) AML before [52, 53, 44], with *CTNNA1* being the most acknowledged candidate given evidence for its promoter methylation as a possible inactivating "second hit" [44]. I, however, find multiple lines of evidence to argue that *KDM3B* is the most likely del(5q) target gene.

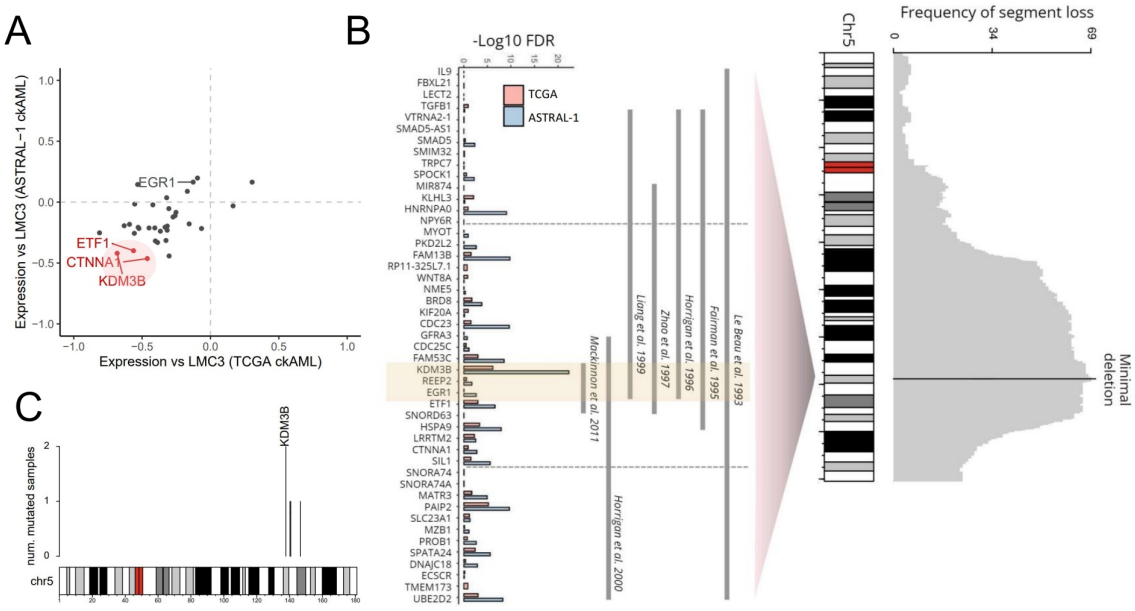

Figure 4.2: The 5q minimally deleted region peaks at the location of H3K9me1/2 demethylase KDM3B. A. Scatter plot summarising the correlations between LMC3 and gene expression of the genes within the minimally deleted 5q region (genes flanked by *IL9* and *UBE2D2* inclusively) among ckAML samples from the ASTRAL-1 (y-axis) and TCGA (x-axis) cohorts. Genes highlighted in red are those whose expression was significantly negatively correlated with LMC3 in both tested cohorts. *EGR1* – a renowned candidate for haploinsufficiency in del(5q) AML – is labelled for comparison. **B.** Density plot showing the frequency of deleted segments on chromosome 5 among del(5q) samples in the ASTRAL-1 cohort. The location of the MDR at 5q31.2 is shown. Minimally deleted intervals described in the literature are indicated in the zoomed region. Dotted gray lines indicate the boundaries of the minimally deleted region that could be defined in the ASTRAL-1 cohort. Bar plots show the -log10 adjusted p-value from differential expression analysis comparing del(5q) to other AML in the ASTRAL-1 (pink) and TCGA (blue) cohort, for all genes within the MDR. **C.** Plot of the mutational frequency of 5q genes in del(5q) samples from TCGA AML. *KDM3B* is labelled as the only gene on chromosome 5 which is affected by mutations in more than one del(5q) AML sample.

4.1.3 Evidence for *KDM3B* as the target of the 5q deletion

Firstly, it is important to note that *CTNNA1* and *ETF1* both lie outside of the minimally deleted intervals which Horrigan *et al.*, Zhao *et al.*, Liang *et al.*, and MacKinnon *et al.*, have proposed [42, 57, 40, 56] – *KDM3B* is therefore the only one of the three MDR genes correlating with LMC3 which is contained within all versions of the MDR described in the literature. One study by MacKinnon *et al.*, in fact narrowed down the MDR to an interval containing only four genes (*KDM3B, EGR1, ETF1* and *REEP2*), of which only *KDM3B* is downregulated in del(5q) AML [42] (Fig. 4.2B).

Next, I performed differential gene expression analyses comparing del(5q) to other AML patients in our own dataset of 40 ckAML samples, and in the TCGA dataset. Strikingly, I found that *KDM3B* was the most significantly downregulated of all genes within the MDR in both datasets (Fig. 4.3A-B). In fact, even beyond the minimally deleted region, no other gene on the entire chromosome arm showed as significant a reduction in expression that was conserved in both datasets.

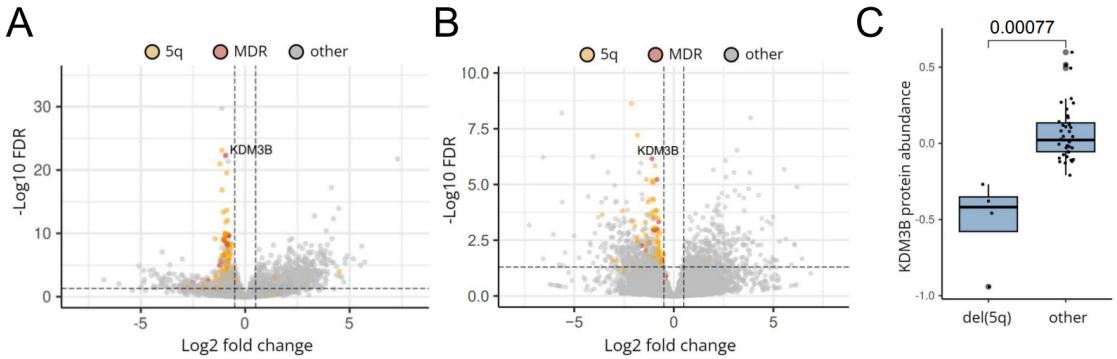

Figure 4.3: KDM3B gene and protein levels are reduced in del(5q) AML. A-B. Volcano plots resulting from differential expression analysis comparing del(5q) to other AML cases in the TCGA (**A**) and ASTRAL-1 ckAML (**B**) RNA-seq datasets. Genes located on chromosome 5q (yellow) and within the minimally deleted region (red) are highlighted. **C.** Boxplot comparing the protein abundance of KDM3B in del(5q) and other AML patient samples from the Kramer *et al.* dataset [175].

Thirdly, I examined the protein expression of MDR candidate genes using two proteomics datasets from AML patient samples, which both revealed a significant reduction in KDM3B protein expression in del(5q) patients (Fig. 4.3C & Appendix, Table 8.7). I further analyzed proteome data from the cancer cell line encyclopedia (CCLE) [176]. Here, KDM3B was the only MDR gene which exhibited a significant reduction in protein expression in del(5q) cell lines compared to other AML cell lines. While I observed a similar trend for ETF1 and CTNNA1, neither of these reached statistical significance (Appendix, Table 8.7).

Next, I inspected the mutation status of MDR genes in the TCGA and ASTRAL-1 cohorts. Here, I found that *KDM3B* was affected by somatic mutations in two del(5q) samples from the TCGA AML cohort, and one sample from the ASTRAL-1 cohort, while no mutations were detected in del(5q) samples for any other genes within the MDR, and no other gene on the whole chromosome arm was mutated in more than one del(5q) sample (Fig. 4.2C). In fact, considering all genes mutated in TCGA AML patients, *KDM3B* was among only seven genes which were mutated in more than one del(5q) sample, and was surpassed only by *TP53* in its mutational frequency in this subgroup. Moreover, no *KDM3B* mutations were detected in any sample without del(5q) in this dataset. This evidence that KDM3B can – rarely but recurrently – acquire a second hit, further supports its probable role in leukemogenesis.

Next, I performed a pan-cancer and pan-tissue analysis of *KDM3B* expression in TCGA and genotype-tissue expression portal (GTEx) data, where I observed higher expression of *KDM3B* in AML compared to every other tumor and normal tissue type. This suggests that *KDM3B* may be an important gene in this disease context, and that the disruption of KDM3B may have important biological consequences (Fig. 4.4).

Taken together with the fact that KDM3B is the only epigenetic regulator encoded from the minimally deleted region, I reasoned that the loss or haploinsufficiency of this gene might explain the observed methylation signature in del(5q) patients. Specifically, I hypothesised that depletion of KDM3B should prevent the demethylation of H3K9me1/2, resulting in increased H3K9me1/2, and consequently triggering *de novo* DNA methylation at its targeted regions.

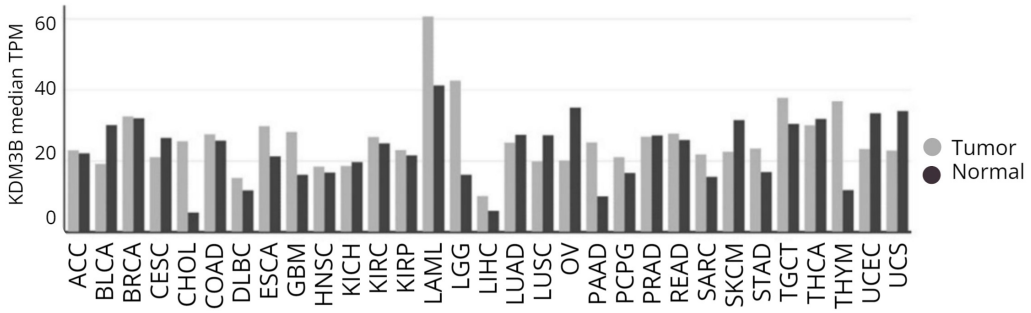

Figure 4.4: KDM3B expression is higher in AML than in other tumor and normal tissue types. Bar plot comparing KDM3B gene expression across different tumor and normal tissues from the TCGA and GTEx datasets. Analysis was performed in the GEPIA portal [177].

4.1.4 Mutual exclusivity of del(5q) and IDH mutations

KDM3B is a metabolically-regulated demethylase, whose activity depends on the availability of α-Ketoglutarate (α-KG), and therefore on the proper functioning of the IDH enzymes, which catalyze α-KG production [178]. It is therefore interesting to note that del(5q) occurs in a mutually exclusive pattern with mutations in *IDH1/2* (Fig. 4.5A), while these two subgroups share partially overlapping patterns of DNA hypermethylation, which can be seen upon clustering of AML bulk methylomes based on the most variable CpG sites (Fig. 4.5B). This observation may be comparable to what has been described for the mutual exclusivity of *TET2/IDH* mutations in AML, which reflects the metabolic dependency of TET enzymes on α-KG [12]. That is, since α-KG is depleted in *IDH* mutant tumors, the activity of KDM3B and other histone lysine demethylases should be therein impaired (Fig. 4.5C). In this way del(5q) and *IDH* mutation may converge on partially overlapping epigenetic disruptions.

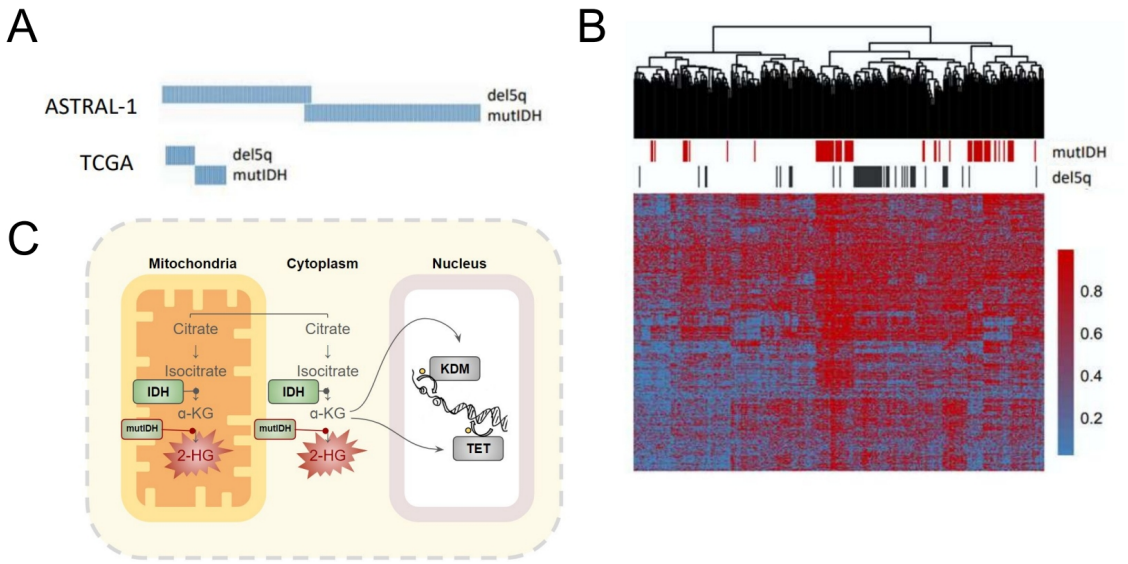

Figure 4.5: Del(5q) and *IDH* mutation are mutually exclusive events characterised by DNA hypermethylation. A. Mutual exclusivity of del(5q) and *IDH* (pooled *IDH1/2*) mutations in the TCGA and ASTRAL-1 AML cohorts. **B.** Hierarchical clustering of ASTRAL-1 AML samples based on the 5000 most variable CpG sites on the EPIC array, indicating co-clustering of del(5q) and *IDH*-mutated AML. **C.** Schematic describing the effects of IDH on epigenetic regulation including activity of TET DNA demethylases and histone lysine demethylases (KDMs).

4.1.5 Linking the del(5q) methylation signature to DNMT3B

Beyond the effects of KDM3B depletion, I speculated that the variability in LMC3 levels – within and outside of the del(5q) subgroup – might be explained by differences in the expression of other epigenetic regulators. To investigate this, I obtained a list of 1012 genes known to participate in epigenetic processes, and tested each gene's expression for a correlation with LMC3 (Fig. 4.6A). Considering the average correlation coefficients across three independent datasets from ASTRAL-1, BEAT and TCGA AML, I found that *KDM3B* exhibited the strongest negative correlation, while the *de novo* DNA methyltransferase, *DNMT3B*, showed one of the strongest positive correlations with LMC3 among epigenetic regulators (Fig. 4.6A). Features of del(5q) hypermethylation clearly align with the previously described activity of DNMT3B, which is known to predominantly methylate developmental genes and polycomb targeted sites [179]. Moreover, DNMT3B is known to have activity at sites of H3K9me1/2 and to interact with the H3K9me1/2 methyltransferase G9a [180, 181]. Overexpression of DNMT3B downstream of KDM3B depletion could thereby contribute to the del(5q) hypermethylation signature.

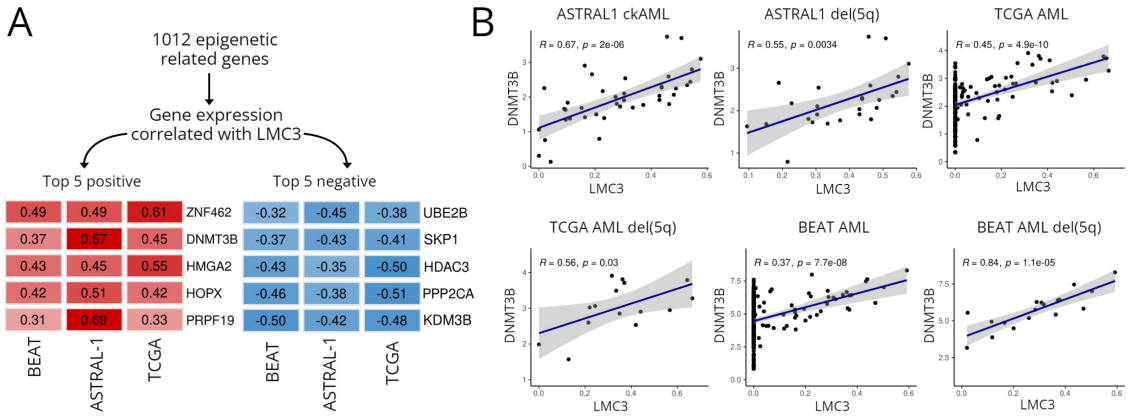

Figure 4.6: DNMT3B overexpression correlates with the del(5q) hypermethylation signature. A. Gene expression of all epigenetic-related genes was tested for correlation with LMC3 proportion in ASTRAL-1 (ckAML), TCGA and BEAT AML cohorts. Shown are the Pearson correlation coefficients for the top 5 positive and negative correlations, ranked by mean correlation across the three cohorts. **B.** Scatter plots correlating DNMT3B gene expression and LMC3 proportion in ASTRAL-1, TCGA and BEAT AML datasets, overall and among del(5q) samples. Pearson correlation coefficients and p-values are shown.

I assessed the expression patterns of *DNMT3B* pan-cancer using TCGA data, where I found it to be higher in AML compared to all other tumor types, except testicular germ cell tumors, which derive from embryonic origin (Fig. 4.7).

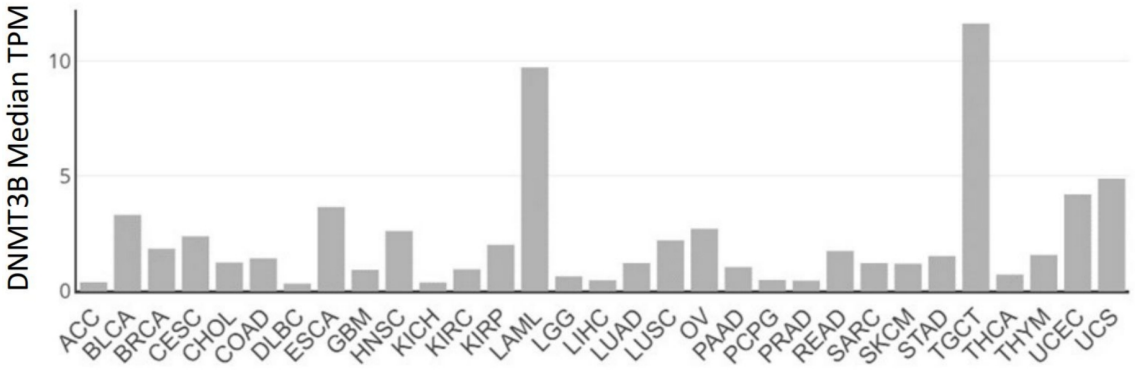

Figure 4.7: *DNMT3B* is overexpressed in AML compared to other non-embryonic tumor types. Bar plot comparing *DNMT3B* gene expression [median transcripts per million (TPM)] across all TCGA tumor types. LAML: acute myeloid leukemia; TCGT: testicular germ cell tumors.

There is emerging evidence to suggest that this methyltransferase may be an important gene in AML, where its overexpression has been linked to features of leukemic stemness and poor prognosis [182, 31]. *DNMT3B* is in fact the highest weighted of the 17 genes which contribute to the prognostic stemness signature, LSC17, as well as other previously established stemness signatures in pediatric AML [31, 183]. I noted a high correlation between LMC3 and the LSC17 score, which is maintained both overall and within the subgroup of complex karyotype patients (Fig. 4.8A). I confirmed this pattern in independent datasets. Importantly, since others have pointed out high LSC17 as a feature of "typical" ckAML [ckAML encompassing del(5q), del(7q) and/or del(17p)] [165] , I repeated this analysis on the subgroup of typical ckAML and found that LMC3/del(5q) correlated with LSC17 independently of the typical/atypical classification, i.e. del(5q) patients generally carried higher LSC17 scores compared to typical ckAMLs in which 5q is retained (Fig. 4.8A). I also reanalysed gene expression data from the LSC17 study, which revealed that *DNMT3B* is the top most significantly upregulated gene in LSCs by comparison to their leukemic blast counterparts (Fig. 4.8B). Taken together, these data suggest that DNMT3B might not only contribute to the del(5q) hypermethylation signature, but also may be linked to stemness features of the del(5q) phenotype.

Interestingly, while mutations in DNMT3B are rare in AML (3/480 patients), I found that they occur alongside 5q deletions in two out of three cases, which stands out in contrast to the relative paucity of epigenetic gene mutations in del(5q) AML (Fig. 4.1).

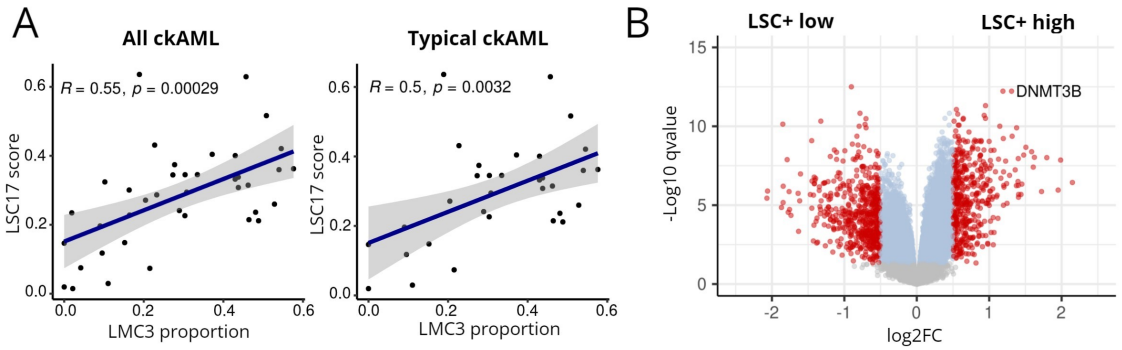

Figure 4.8: The del(5q) methylation signature correlates with features of leukemic stem cells. A. Scatter plots showing the correlation between LMC3 proportion and LSC17 score among ckAML (left) and among "typical" [containing del(5q), del(7q) and/or del(17p)] ckAML (right) from the ASTRAL-1 cohort. Pearson coefficients and p-values are indicated. **B.** Volcano plot resulting from differential expression analysis of LSC+ vs LSC- AML cells using gene expression data from the Ng *et al.* LSC17 study [31].

4.1.6 Overexpression of *DNMT3B* in del(5q) AML may be regulated by DNA methylation in LSCs

To consider possible mechanisms leading to *DNMT3B* overexpression, I examined its methylation status. In EPIC array data, l observed hypomethylation of a CpG site within the *DNMT3B* promoter in del(5q) patients, which correlated strongly with its gene expression, suggesting that *DNMT3B* itself might be activated through epigenetic mechanisms in this subgroup (Fig. 4.9A). Speculating that this event might be specific to LSCs, I examined its methylation status in del(5q) samples which were sorted to separate LSCs from leukemic blasts [159], and found that del(5q) LSCs were fully unmethylated at this region, while del(5q) blasts remained methylated (Fig. 4.9B). This suggests that the differences in methylation and expression of *DNMT3B* among AML samples may reflect the proportion of LSCs. However, since the associated LMC3 methylation signature appeared to be present in both LSCs and blasts (Appendix, Fig. 8.4), it may be that the signature is initiated by DNMT3B in LSCs, and propagated by maintenance methyltransferases to their blast progeny.

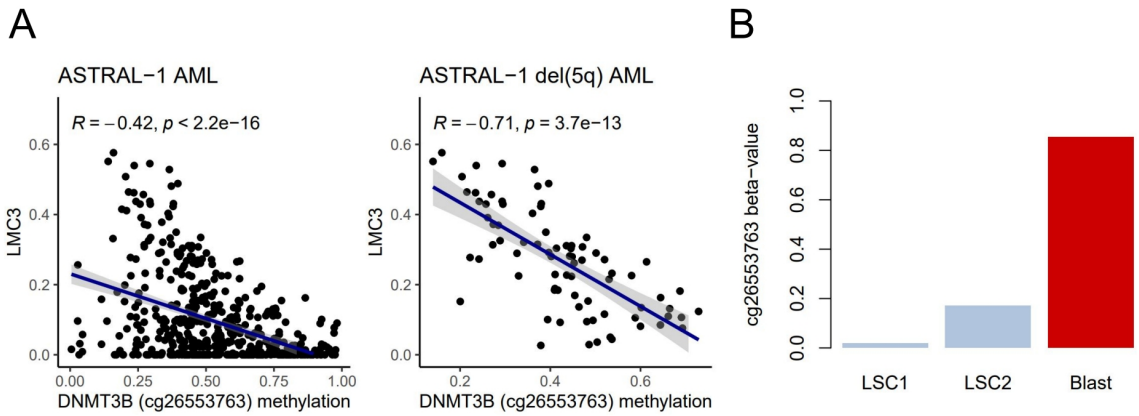

Figure 4.9: DNMT3B overexpression in AML may be regulated by methylation. A-B. Scatter plots correlating *DNMT3B* gene expression and promoter methylation at cg26553763 in all AML samples (**A**) and among del(5q) samples (**B**) from the ASTRAL-1 cohort. Pearson correlation coefficients and p-values shown. **C.** Bar plot comparing the methylation status of the DNMT3B promoter CpG site cg26553763 in LSCs and blasts in a del(5q) patient from the Jung *et al.* dataset [159].

4.1.7 Linking the del(5q) methylation signature to H3K9me1/2 methylation

Above I hypothesised that the del(5q) methylation signature might reflect increased levels of H3K9me1/2 along the genome, resulting from KDM3B deficiency. To support this link to H3K9me1/2, I next tested whether other regulators of H3K9me1/2 correlated with LMC3 levels. I found that LMC3 was positively correlated with the gene expression of two key H3K9me1/2 methyltransferases; *EHMT2/G9a* and *PRDM16* (Fig. 4.10A). Both of these enzymes have opposing activity to KDM3B, and both have been described to be overexpressed and linked to poor outcome and features of leukemic stemness in AML [184, 185]. I also quantified LMC3 proportions in methylation data from an AML patient overexpressing the H3K9me1 methyltransferase, *PRDM16* [186], and found levels of LMC3 comparable to that of del(5q) AML (Fig. 4.10B). Notably, *DNMT3B* and *EHMT2* expression also maintained a correlation with LMC3 independently of del(5q) status, i.e. del(5q) patients with higher expression of *DNMT3B/EHMT2* tend to have higher levels of LMC3 compared to del(5q) patients with lower expression of these genes (Fig. 4.10C).

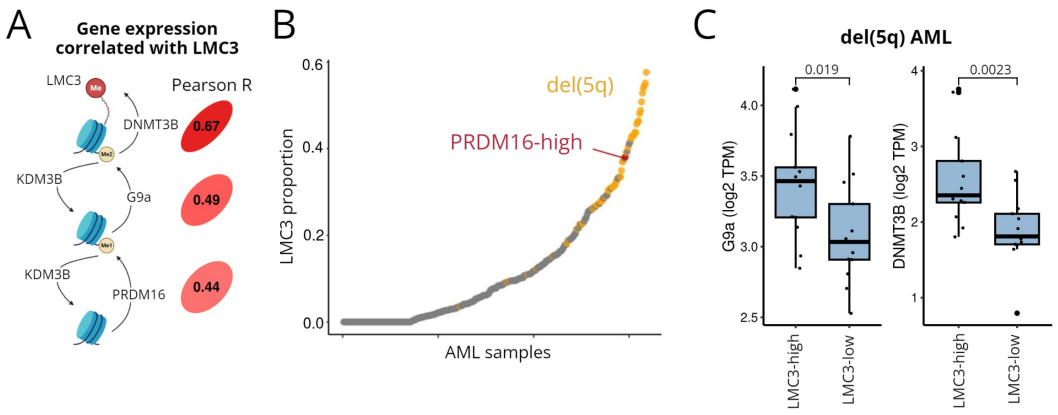

Figure 4.10: The del(5q) hypermethylation signature correlates with expression of H3K9me1/2 methyltransferases. A. Correlation plot depicting Pearson coefficients for the correlation of *DNMT3B, G9a* and *PRDM16* gene expression with LMC3 proportion in the ASTRAL-1 ckAML cohort. **B.** Dot plot comparing LMC3 levels in del(5q) AML (yellow) and in an AML sample overexpressing H3K9me1 methyltransferase, *PRDM16* (red). **C.** Boxplots comparing the gene expression (log2 transformed TPM) of *G9a* and *DNMT3B* in del(5q) samples separated by median LMC3 proportion. Wilcoxon's p-values shown.

Since H3K9me2 is associated with the formation of Lamina Associated Domain (LAD)s, I investigated the enrichment of LAD-associated genes among LMC3-hypermethylated and differentially expressed genes. I observed a significant enrichment of genes within LADs among LMC3-hypermethylated genes as well as among genes downregulated in del(5q) compared to other ckAML patients (Fig. 4.11).

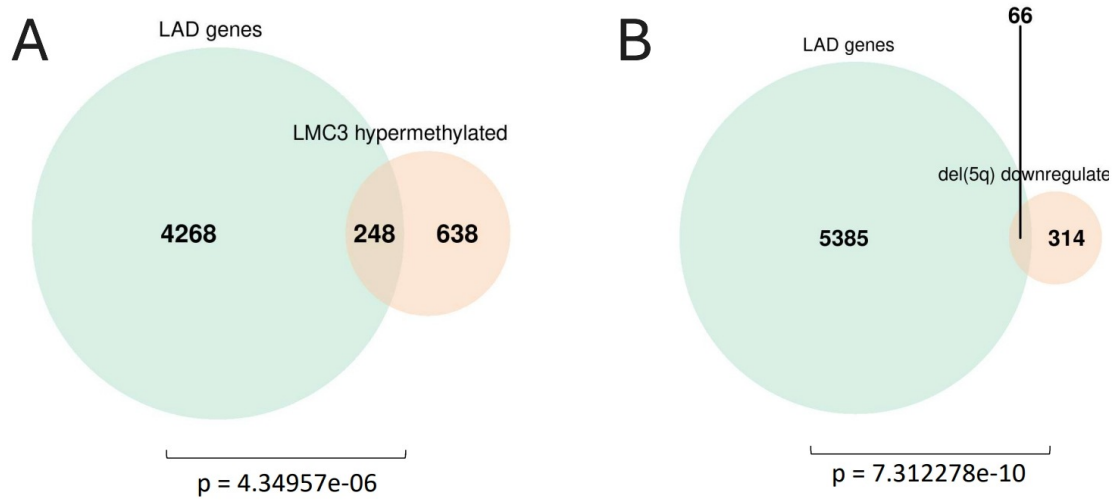

Figure 4.11: Genes within lamina-associated domains are dysregulated in del(5q) AML. A. Venn diagram showing the overlap of genes within lamina-associated domains (LAD genes) and genes with CpG sites hypermethylated in the LMC3-high AML subgroup (by comparison to all other AML subgroups and normal HSCs). Hypergeometric p-value is shown. **B.** Venn diagram showing the overlap of genes within lamina-associated domains (LAD genes) and genes downregulated in del(5q) AML (by comparison to other ckAML patients in the ASTRAL-1 cohort). Hypergeometric p-value is shown.

4.1.8 Epigenetic similarity of del(5q) and *MECOM*-overexpressing AML converge on overexpression of *DNMT3B*

AML with t(3;3)/inv(3) is a subgroup associated with overexpression of the *MECOM/EVI-1* oncogene [187]. This subgroup has been previously shown to exhibit a distinct hypermethylation signature [115], mediated by MECOM's interactions with DNMT3B [188]. Moreover, the H3K9me2 methyltransferase, G9a/EHMT2 is a known MECOM binding partner [189]. To support my hypothesised role for DNMT3B in bringing about the del(5q) hypermethylation signature, I investigated whether a similar methylation pattern might be detected in this subgroup. I found that AML overexpressing *MECOM/EVI-1*, or carrying the associated t(3;3)/inv(3) alterations (where gene expression was not available), did carry significantly higher levels of LMC3 compared to other patients without 5q deletions (Fig. 4.12A). In line with this, I found that *DNMT3B* was consistently upregulated in both del(5q) patients and MECOM-high/inv(3) AML (Fig. 4.12B), suggesting that the epigenetic similarity between these two subgroups converges on the overexpression of this *de novo* DNA methyltransferase.

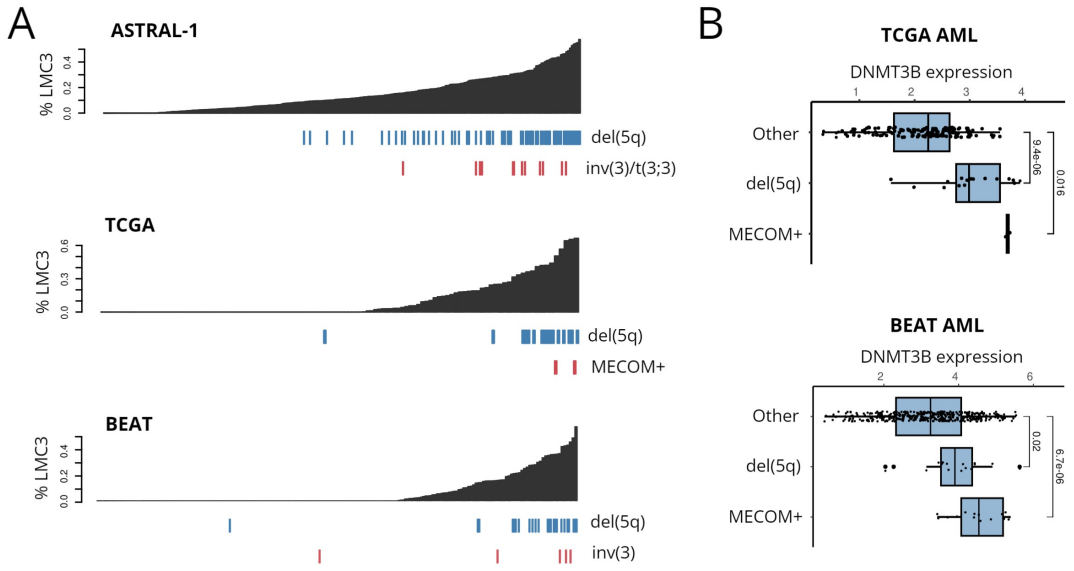

Figure 4.12: Epigenetic similarity of del(5q) and *MECOM/EVI-1*-overexpressing AML converge on overexpression of *DNMT3B*. A. Bar plots showing association of LMC3 with *MECOM* overexpression and the associated t(3;3)/inv(3) in ASTRAL-1, TCGA and BEAT AML cohorts. Samples are ordered by increasing LMC3 proportion. **B.** Boxplots comparing gene expression of *DNMT3B* [log transformed reads per kilobase million (RPKM)] in del(5q), *MECOM*-overexpressing AML and other AML samples from the TCGA and BEAT AML cohorts.

4.2 Discussion

Recurrent deletions in the genomes of cancer cells are typically viewed as suspected locations of TSGs. In del(5q) AML however, no gene has been found which fits the classical definition of a tumor suppressor, with evidence for recurrent mutation or otherwise inactivation of the second allele [190].

Here, building on results from **Chapter 3**, I decided to reevaluate the long-standing question of del(5q) TSGs, and to consider the possibility that del(5q) AML may be driven by previously overlooked epigenetic mechanisms. In this chapter, I have combined data on the copy number, transcriptional, protein and mutational patterns of genes within the MDR on chromosome 5. I have presented multiple lines of evidence suggesting that *KDM3B* - a H3K9me1/2 demethylase encoded from the MDR on chromosome 5q31.2 - is the most plausible del(5q) target gene. These findings lead me to hypothesise that haploinsufficiency of KDM3B disrupts the global epigenetic landscape of myeloid progenitor cells and thereby contributes to the development and progression of aggressive leukemias. Understanding how exactly these events unfold will require much further investigation, however I will consider this question in the following discussion, and in **Chapter 6** of this thesis I will examine one hypothesis in more detail.

KDM3B and (dys-)regulation of H3K9me1/2

KDM3B encodes a histone demethylase most known for its activity at lysines H3K9me1 and H3K9me2, and which has recently also been found to have activity at arginine H4R3me2s [191]. H3K9me2 is a repressive chromatin mark and a defining feature of heterochromatin, which together with H3K9me3 is associated with gene silencing and is found at transcriptionally inactive sections of the genome such as transposable elements and satellite repeats [192]. However, since KDM3B has activity at mono- and di-methylated lysines rather than trimethylated lysines, it may also play a role in organising facultative heterochromatin, for example in regulating LADs and the expression of tissue-specific genes [193, 194].

The role of H3K9me1 in the genome is more complex. As well as serving as a precursor for the heterochromatic marks H3K9me2/3, H3K9me1 is also enriched in the bodies of actively transcribed genes, as well as at active promoters and enhancers, and at chromatin boundary regions such as the edges of heterochromatin domains [195]. The repressive arginine modification, H4R3me2s, was also shown to be acted on by KDM3B in hematopoietic cells, where it supposedly controls the expression of numerous key hematopoietic regulators [191].

While these are the only three histone marks known to be directly influenced by KDM3B, it is important to recognise that many epigenetic marks are closely interconnected, such that KDM3B could indirectly influence the distribution of other histone modifications as

well as the methylation of DNA. For example, histone lysine methylation and acetylation are thought to be physically as well as functionally antagonistic and to occur in a mutually exclusive pattern, such that an increase in H3K9me may be sufficient to bring about a concomitant reduction in H3K9ac, which is associated with active transcription [196]. Moreover, repressive histone modifications are known to direct the establishment of *de novo* DNA methylation [78].

As is the case for many epigenetic regulators in cancer, both tumor-suppressing and tumor-promoting roles of KDM3B have been described across different cancer types. For example, overexpression of the gene was linked to poor recurrence-free survival in non-small cell lung cancer [197], and it was also suggested to promote the growth of liver cancer cells [198]. Meanwhile, KDM3B was shown to have an anti-proliferative effect in prostate cancer [199], and both tumor suppressor and oncogenic roles of KDM3B have been proposed in colorectal and breast cancer [200, 201, 202, 203]. Mutations in KDM3B were also described in Wilms tumor [204], and *de novo* and inherited mutations in KDM3B have been reported to result in a rare neurodevelopmental disorder referred to as Diets-Jongmans syndrome [205]. Interestingly, in a study of 14 individuals with Diets-Jongmans syndrome, two were found to develop cancer in childhood, one of which was a case of AML [205].

In the hematopoietic system, it is easy to imagine that genome-wide dysregulation of H3K9me1/2 could have dangerous implications. Patterns of H3K9me1/2 have been shown to regulate lineage commitment in HSCs [206], and regulators of H3K9me1/2 are known to be disrupted in other AML subgroups. For example, overexpression of the H3K9me1 methyltransferase *PRDM16* and the H3K9me2 methyltransferase *EHMT2/G9a* – events which may be analagous to depletion of KDM3B, and which I have shown correlate with a similar epigenetic signature as in del(5q) AML – have been previously linked to poor outcome and features of leukemic stemness [184, 185].

While KDM3B is not among the more widely studied epigenetic enzymes in AML, and has received relatively little attention in studies of del(5q) AML compared to other MDR genes [52], a few recent studies have highlighted its importance in hematopoietic development and malignancy. In a 2018 study, Li *et al.* showed that knockout of *KDM3B* resulted in downregulation of key hematopoietic regulatory genes, and defective hematopoiesis in mice, which displayed anemia and leukocytosis [191]. This study demonstrated that KDM3B is an essential epigenetic player during hematopoietic development. Evidence for tumor suppressor activity of KDM3B in del(5q) cell lines was suggested in a 2018 study by Xu *et al.* [52]. More recently, Gray *et al.* showed that depletion of KDM3B promotes amplification and rearrangements of the *MLL/KMT2A* locus in AML, through regulation of H3K9me1/2 and CTCF occupancy [207]. Given this background it is easy to imagine that KDM3B disruption, even to the extent of haploinsufficiency, could have profound phenotypic effects during myeloid development.

Although I identified a couple of instances where KDM3B appears to be biallelically in-

activated (with a deletion on one allele and a mutation on the other), it is clear that the vast majority of del(5q) patients lose only one copy of the gene. This raises the question – if loss of KDM3B is beneficial for disease progression, why are "double hits" not observed more often? I suggest that the heterozygous loss of KDM3B may represent a case of so-called "obligate haploinsufficiency" – a phenomenon proposed by Berger *et al.* wherein partial loss of the implicated gene results in a stronger tumorigenic outcome than complete loss [208]. This can happen due for example to the activation of compensatory mechanisms that would make up for the gene's complete absence, or if some level of expression of the gene is critical for cell survival. The *NPM1* gene for example, which is frequently mutated in AML, is typically inactivated on only a single allele, since complete loss of the protein is embryonically lethal and incompatible with cell growth. Considering that many of the common epigenetic gene mutations in AML – such as *DNMT3A, TET2* and *ASXL1* – are typically heterozygous events, it is tempting to speculate that "obligate haploinsufficiency" may be more common for epigenetic regulators than for other TSGs.

In the later part of this thesis, I will consider the hypothesis that the imbalance in H3K9me1/2 resulting from heterozygous loss of KDM3B could give rise to heterogeneity in epigenetic patterns from cell to cell. If del(5q) AML cells express lower than normal levels of KDM3B, the enzyme may only be able to cover a limited fraction of its binding sites. Assuming that the search for binding sites happens randomly, one might expect that the resulting patterns of H3K9me1/2 along the genome might differ from cell to cell. Such epigenomic heterogeneity could disrupt normal transcriptional patterns, and could potentially contribute to cancer progression and therapy resistance by increasing tumor plasticity. This hypothesis relies on the idea that KDM3B expression is partially reduced rather than completely abolished. However, as complete loss of KDM3B might result in compensatory changes in the expression of other histone lysine regulators, a shift in the balance of histone modifications, potentially leading to cell-to-cell heterogeneity, might also be possible in this scenario. I will investigate this hypothesis in **Chapter 6** of this thesis.

Notes on the haploinsufficiency of other chromosome 5 genes

In the **Introduction** of this thesis I have already outlined the prevailing hypotheses and pitfalls regarding the putative target gene(s) in del(5q) AML. In light of my findings here, a number of additional points should be discussed.

The data I have presented suggest that *KDM3B* is the most likely target of the 5q deletion, being the only gene in the MDR which is consistently downregulated at both the gene and protein expression levels in del(5q) patients, which is contained within all previously described versions of the MDR, and which is recurrently (albeit rarely) affected by somatic mutations in del(5q) patients. However it is not certain whether KDM3B acts alone or in combination with other haploinsufficient targets on chromosome 5. Since deletions on chromosome 5 typically span several MB and focal deletions are extremely rare, it

is difficult to rule out potential cooperative or combined effects. Perhaps most notable among putative del(5q) TSGs is *EGR1*, which lies adjacent to *KDM3B* at the peak of the MDR [49]. Although my own and previous analyses have highlighted the lack of consistent downregulation of this gene in del(5q) AML [61], contradicting the idea that loss of *EGR1* is sufficient to explain del(5q) leukemogenesis, it is still tempting to consider a possible involvement of EGR1, since it is recognised as a tumor suppressor in several other cancers [58].

Another of the most acknowledged putative target genes in del(5q) AML is *CTNNA1*, which encodes the alpha-catenin protein; an important regulator of cell adhesion and actin cytoskeletal organisation [44]. The main reason why *CTNNA1* has received more attention in this regard than other MDR genes is that it was shown to be affected by promoter hypermethylation in some del(5q) patients, providing an epigenetic route for silencing the remaining active allele [209]. In light of my findings, this idea deserves to be reappraised. In fact it was previously shown in the KG1α del(5q) cell line, that the hypermethylation of the *CTNNA1* promoter was accompanied by enrichment of inactivating histone marks including H3K9me2 [209]. This leads me to suspect that this hypermethylation, rather than representing a selective silencing mechanism targeting the *CTNNA1* gene as a "second hit", may rather be a bystander of the more widespread epigenetic changes which I believe are driven by the depletion of KDM3B.

One of the major roles of H3K9me2 is in the formation of LADs – regions of the genome which exist in close contact to the nuclear lamina, and which contain many tissue-specific genes, most of which are transcriptionally inactive or expressed at low levels [193]. I have shown that genes within LADs are significantly enriched among those differentially methylated/expressed in del(5q) AML. In this regard, it is interesting to note that one of the other 5q genes which has been previously proposed as a likely candidate is *LMNB1*, which encodes one of the two critical protein components of the nuclear lamina [210]. Although *LMNB1* lies outside of the 5q31.2 MDR, its locus would still be affected in the vast majority of del(5q) cases. Acquired Pelgar Huet Anomaly (PHA) or pseudo-PHA is a dysplastic change observed in myeloid malignancies, especially in high-risk MDS and AML with del(5q), which is characterised by hyposegmentation and clumping of the nuclear chromatin of myeloid cells [211, 210]. While the inherited form of PHA is caused by mutations in the Lamin B receptor, the acquired form was recently linked to the loss of *LMNB1* in del(5q) MDS/AML [210]. Considering their shared role in the regulation of H3K9me2/LADs, it is intriguing to theorise on the possible combined effect of KDM3B and LMNB1 disruption, and to consider whether loss of KDM3B may also contribute to the pseudo-PHA observed in AML cells.

Potential implications of KDM3B haploinsufficiency beyond epigenetic regulation

With this new perspective on del(5q) AML, it is worth returning to some of the unsolved questions about this intriguing subgroup: for example, why is it that del(5q) appears so often in the context of ckAML? Could the epigenetic dysregulation in del(5q) AML be directly linked in any way to the genomic instability and frequent *TP53* mutations which exemplify this subgroup, or are these events simply coincident? It is important to note that del(5q) is an early event in AML, which usually occurs before other CNAs and thus likely precedes the development of a complex karyotype [36, 37, 38, 172]. Therefore it is probably not simply that del(5q) is the CNA which is selected amid the large-scale genomic disarray, being most advantageous to the tumor, leading it to appear more often than other CNAs in ckAML. The fact that del(5q) seems to happen first rather suggests that a tumor with del(5q) is somehow more likely to acquire a *TP53* mutation and/or develop a complex karyotype than one without. How and why this occurs will be an important direction of future study.

Dysregulation of histone modifying enzymes in cancer is typically linked to altered transcriptional regulation. Nevertheless, it is also worth considering possible effects of KDM3B haploinsufficiency beyond the epigenetic regulation of histones. For example, as a post-translational modification, lysine methylation can be important in regulating the activity of non-histone proteins. In this regard, it is important to note that H3K9me2 methyltransferases like G9a have been shown to regulate the activity of the TP53 protein [212, 213]. It would be interesting to investigate whether KDM3B might also play a role in regulating the activity or stability of TP53, especially given the frequent coocurrence of TP53 mutations and del(5q) in AML. Among all CNAs in AML, del(5q) exhibits by far the most significant association with *TP53* mutations [214]. In other cancers, the cooccurrence of *TP53* mutations with other genomic events can indicate a synergy or functional interaction between the two events. It is intriguing to speculate whether – and in what way – the depletion of KDM3B might provide a greater selective advantage in the context of a *TP53* mutation.

Interestingly, KDM3B has also been previously linked to genomic instability in cancer. Saavedra *et al.* suggested that depletion of KDM3B can promote genomic instability by influencing histone protein metabolism and causing increased expression of histone proteins [215]. Increased histone protein production can contribute to chromatin over-compaction and replication stress. Moreover, many histone methyltransferases and demethylases, including demethylases of H3K9me2, have been shown to be recruited to sites of DNA damage, suggesting that the removal of H3K9me2 may play a role in the DNA damage response [216]. One possibility is that the removal of such repressive marks may be necessary to facilitate the loosening of chromatin that allows access of the DNA repair machinery at damaged sites. The involvement of KDMs in the DNA damage and repair pathways provides a tempting explanation for the frequent disruption of such enzymes in cancer,

and might provide a clue as to the link between del(5q) and genomic instability that would be worthwhile investigating in future studies in the context of ckAML.

Notably, Salzberg *et al.* have also identified the presence of large H3K9me2 blocks in AML cells, which they found to be enriched at sites of AML-specific mutations and chromosomal translocations [217]. Along similar lines, Gray *et al.* proposed that depletion of KDM3B and consequent accumulation of H3K9me2 can promote copy number amplifications and rearrangements of the *MLL/KMT2A* locus in AML – a mechanism that could conceivably extend to other genomic regions [207].

Linking del(5q) AML to other epigenetically dysregulated subgroups

To strengthen my argument for KDM3B as a target of del(5q) AML, I have also drawn attention to the epigenetic similarities between del(5q) AML and a number of other AML subgroups in which H3K9me1/2 regulation is disturbed. Firstly, I have commented on the mutual exclusivity and shared hypermethylation of del(5q) and *IDH*-mutated AML, and secondly on the epigenetic link between del(5q), *MECOM* and *PRDM16* overexpression. These may be indications of convergent evolution.

KDM3B belongs to the family of α-KG dependent dioxygenases. The activity of such enzymes is known to be inhibited in the presence of mutated *IDH*, owing to accumulation of the oncometabolite 2-HG, which is a competitive inhibitor of α-KG. By this mechanism, mutant *IDH* has been shown to promote an increase in histone lysine methylation, including the H3K9me2 mark [218, 178]. My observation that del(5q) and *IDH1/2* mutations occur in a mutually exclusive manner in AML patients, while exhibiting overlapping patterns of DNA hypermethylation, can therefore be taken as further evidence for KDM3B as a target of the deletion. Similarly, mutual exclusivity of mutations in *IDH* and *TET2* – another α-KG-dependent enzyme – is a well established phenomenon in AML [12]. My findings suggest that a similar metabolic dependency might underlie the mutual exclusivity of del(5q) and *IDH* mutations, and that the two events might converge on partially overlapping epigenetic pathways represented by reduced KDM3B activity. In this regard, it is interesting to note a recent study by Waarts *et al.* who showed using a clustered regularly interspaced short palindromic repeats (CRISPR) dependency screen that KDM3B is a selective dependency of *IDH/TET*-mutant HSCs [219].

Another AML subgroup for which a DNA hypermethylation signature has been previously described is AML overexpressing *MECOM/EVI-1*; a potent oncogene which can be activated through enhancer hijacking events and other genomic alterations on chromosome 3 [115]. In previous studies, MECOM was shown to drive DNA hypermethylation through its interactions with H3K9 methyltransferases as well as the DNA methyltransferase, DNMT3B [115]. Here, I have shown that AML overexpressing *MECOM* carry high levels of the same methylation signature as del(5q) patients, and that these patients concomitantly overexpress the *DNMT3B* gene to similar levels as del(5q). This is particularly

intriguing given that *MECOM*, in its longer isoform, encodes the H3K9me1 methyltransferase, PRDM3 – an enzyme with opposing activity to KDM3B. Several studies have reported on the biological correlates of AML overexpressing *MECOM* and its homolog, the H3K9me1 methyltransferase, *PRDM16* [186, 185]. Here, I have shown that both of these events result in a similar methylation pattern as del(5q) AML, lending weight to the assumption that this signature is related to the reduced activity of a H3K9me1/2 demethylase.

The overexpression of DNMT3B in AML is an intriguing phenomenon, as this *de novo* methyltransferase is predominantly active during embryogenesis and silenced in healthy adult tissues. Nevertheless, *DNMT3B* has been shown to be aberrantly expressed in some tumor types including colorectal and breast cancers, as well as myeloid and lymphoid leukemias [179, 220, 31, 221, 222]. Recently DNMT3B has emerged as a potentially important player in AML, where its expression is higher than other non-embryonic tumors, and where it has been linked to features of leukemic stemness and poor prognosis [182, 31]. In this chapter, I have identified a DNA methylation signature associated with DNMT3B in AML. I have shown that *DNMT3B* overexpression is a common feature of the del(5q) subgroup, and I have highlighted *DNMT3B* as the most strikingly upregulated gene in LSCs by comparison to non-self-renewing leukemic blasts. I hypothesise that DNMT3B is recruited to sites of H3K9me1/2 in KDM3B-depleted cells, resulting in a pattern of methylation in del(5q) AML resembling that of *MECOM/PRDM16*-overexpressing AML.

Looking forward, it will be important to understand how *DNMT3B* overexpression comes about in del(5q) patients, and if/how this is directly linked to the depletion of KDM3B. Interactions between DNMT3B and H3K9me1/2 have been extensively documented, and one could speculate that a rise in H3K9me1/2 might somehow trigger increased expression of the gene, though this would need to be experimentally tested. It may also be that *DNMT3B* expression marks a specific stage in HSPC development, or that it simply reflects an enrichment of LSCs – or a specific type of LSCs – in del(5q)/*MECOM/PRDM16*-overexpressing tumors.

Conclusion

In summary, I have suggested that haploinsufficiency of the H3K9me1/2 demethylase KDM3B, and overexpression of DNMT3B, may explain the epigenetic disruption in del(5q) AML. These findings shed new light on a highly aggressive and poorly understood AML subgroup and highlight a need for follow-up studies to disentangle the precise role of H3K9me1/2 in AML pathogenesis.

Chapter 5

EpiCHAOS: a metric for quantifying epigenomic heterogeneity in single cell data

In the previous chapter of this thesis, I proposed that KDM3B is a likely target for haploinsufficiency in del(5q) AML. I presented the hypothesis that haploinsufficiency of KDM3B could give rise to cell-to-cell heterogeneity in the repressive chromatin modifications, H3K9me1/2. With the ultimate aim of investigating this hypothesis, I was inspired to develop a computational strategy to quantify cell-to-cell epigenetic heterogeneity in single-cell data, which I have called epiCHAOS (epigenetic/Chromatin Heterogeneity Assessment of Single cells). In this chapter, I will present the development and validation of epiCHAOS, and demonstrate the functionality of the method through applications in a range of biological settings.

The main findings presented in this chapter also appear in a modified form in the results of Kelly *et al.* [223].

5.1 Results

5.1.1 Development and *in silico* validation of epiCHAOS

I designed epiCHAOS to assign a cell-to-cell epigenetic heterogeneity score at the level of cell clusters, or other user-defined groups of interest e.g. cell types, timepoints or treatment conditions. I initially focused on scATAC-seq data, since this is currently the most commonly used single-cell epigenomics modality. To compute epiCHAOS scores, data is first extracted from each single-cell group of interest in the form of a binarised matrix. This represents a peaks-by-cells or tiles-by-cells matrix in the case of scATAC-

seq (where 1's signify positions of open chromatin, and 0's represent closed chromatin or missing values), or any other regions-by-cells matrix in the case of other single-cell epigenomics data types. Depending on the biological question, cells may be stratified based on predefined clusters, cell types, treatment conditions, or any other annotated phenotype of interest. For each group or cluster, I then calculate the distances between all pairs of cells using a count-centered version of the Jaccard distance [224], which ensures that data with a greater or fewer number of 1's are not perceived as being more or less heterogeneous. Finally, I compute the mean of all pairwise distances per group or cluster as its epiCHAOS score. For simplicity of interpretation, I fit epiCHAOS scores to a range of 0-1, such that the lowest score in any given analysis will be equal to 0, and the highest will be equal to 1 (Fig.5.1A).

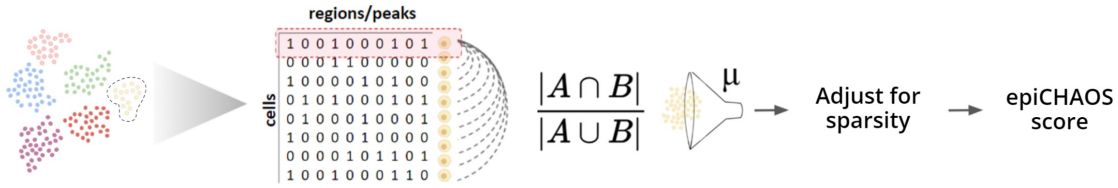

Figure 5.1: Calculation of epiCHAOS scores. Schematic describing epiCHAOS calculation. Using single-cell epigenomics data in binarised matrix form, epiCHAOS scores are assigned per cluster by computing the mean of all pairwise cell-to-cell distances using a chance-centered Jaccard index followed by regression-based adjustment for sparsity. μ = mean per cluster. Formula for the Jaccard index is shown.

To assess how epiCHAOS behaves in artificial situations of increasing heterogneity, I first generated *in silico* a series of 100 synthetic datasets mimicking the structure of binarised scATAC-seq matrices, in which I controlled the levels of heterogeneity, while fixing the total count. To do this I first generated a random binary matrix which would represent the first and expectedly most heterogeneous dataset in the series. Then, in each subsequent matrix, I incrementally introduced homogeneity by removing a defined number of 1's from selected n rows, and adding them to a different selected n rows. In this way, a constant number of 1's is maintained, while "cell-to-cell" (or column-to-column) heterogeneity gradually decreases. Using this system, I showed that the epiCHAOS score is highly correlated with the true controlled heterogeneity (Pearson R=0.99), reflecting the expected behaviour of the Jaccard Index (Fig. 5.2A).

Next, I verified that heterogeneity can be detected both in cases of increasing and decreasing counts, such that both heterogeneous gains or losses of chromatin accessibility would similarly be detected. To test this I simulated a further series of scATAC-like matrices in which I incrementally increased heterogeneity, while either increasing or decreasing genome-wide chromatin accessibility. Taking as baseline an scATAC-seq dataset from human monocytes, I incrementally perturbed the data with either addition or removal of 1's. Specifically, in the first series, I randomly selected 10, 20, 30, 40 and 50% of 1's, and replaced them by 0's, and in the second series, I randomly selected corresponding numbers

of 0's and replaced them by 1's. I showed that epiCHAOS correctly detects differences in heterogeneity both in cases where counts are added and removed (Fig. 5.2B).

Since single-cell epigenomics data are typically very sparse, it was important to confirm that epiCHAOS does not perceive differences in sparsity as differences in heterogeneity. I first investigated whether epiCHAOS correlated with the total number of 1's by generating a series of completely random datasets with varying total number of 1's. I found no correlation between epiCHAOS scores and total number of 1's between these synthetic datasets, which confirmed the expected behaviour of the count-centred Jaccard index (Fig. 5.2C).

In real single-cell datasets however, differences in coverage may be more complex since they are accompanied by differences in the number of missing values representing false negatives. Genome-wide differences in detected ATAC signals can appear due to technical reasons, i.e. reflecting differences in sequencing depth, but can also have biological sources. For instance, if a sample contains quiescent cells, where a large part of the chromatin is compacted, genome-wide ATAC-seq signals will be reduced. In cancer cells, deletions or gains of chromosomal material can result in a lower or higher number of fragments within the affected region, respectively. I used two different strategies to investigate how epiCHAOS would behave in biologically relevant datasets with varying sparsity. First, I implemented a previously published method, scReadSim [225], to simulate datasets with varying sequencing depth, using a scATAC-seq dataset from HSCs as a baseline [226]. Secondly, I called CNAs in scATAC-seq data from a liver cancer cell line [227], and investigated differences in epiCHAOS scores at regions of large CNAs in cells with and without the respective lesion. In both of these scenarios, I found that higher heterogeneity can be perceived in data with lower total number of counts or fragments, reflecting lower sequencing depth (Fig. 5.2D), or chromosomal deletions (Fig. 5.3). This necessitated an adjustment of the epiCHAOS score for sparsity. To do this, I implemented a linear regression-based adjustment for the total number of 1's across cell groups/clusters, which resulted in a sparsity-adjusted heterogeneity score. For applications to cancer samples, I used a similar approach but corrected separately for each chromosome. I showed that this adjusted score is no longer affected by differences in genome-wide chromatin accessibility or sequencing depth (Fig. 5.2E), or by the presence of large-scale deletions and gains in tumor samples (Fig. 5.3).

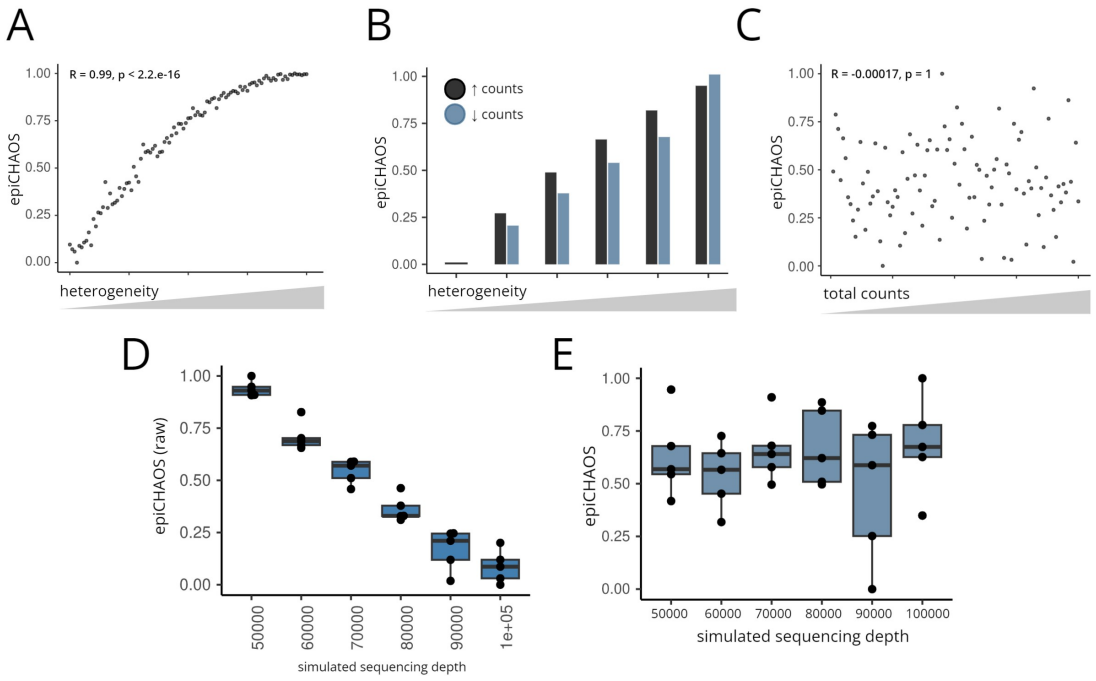

Figure 5.2: Validation of epiCHAOS in synthetic datasets. **A.** Scatter plot illustrating the correlation between epiCHAOS scores (epiCHAOS) and controlled heterogeneity across 100 synthetic datasets. Pearson correlation coefficient and p-value is shown. **B.** Bar plots illustrate increasing heterogeneity after perturbation of scATAC-seq data from sorted monocytes by either randomly adding or randomly removing 10-50 % of 1's. **C.** Scatterplot illustrating the absence of correlation between epiCHAOS scores (epiCHAOS) and total count across a series of 100 randomly generated datasets ordered by increasing total counts. Pearson correlation coefficient and p-value is shown. **D-E.** Boxplots comparing raw epiCHAOS scores before (**D**) and after (**E**) adjustment for sparsity across six simulated single-cell ATAC-seq datasets. Data were simulated using scReadSim with sequencing depth varying from 50,000 to 100,000 counts. ScATAC-seq data from hematopoietic stem cells subset from the Granja *et al.* 2019 dataset [226] were used as the baseline counts matrix.

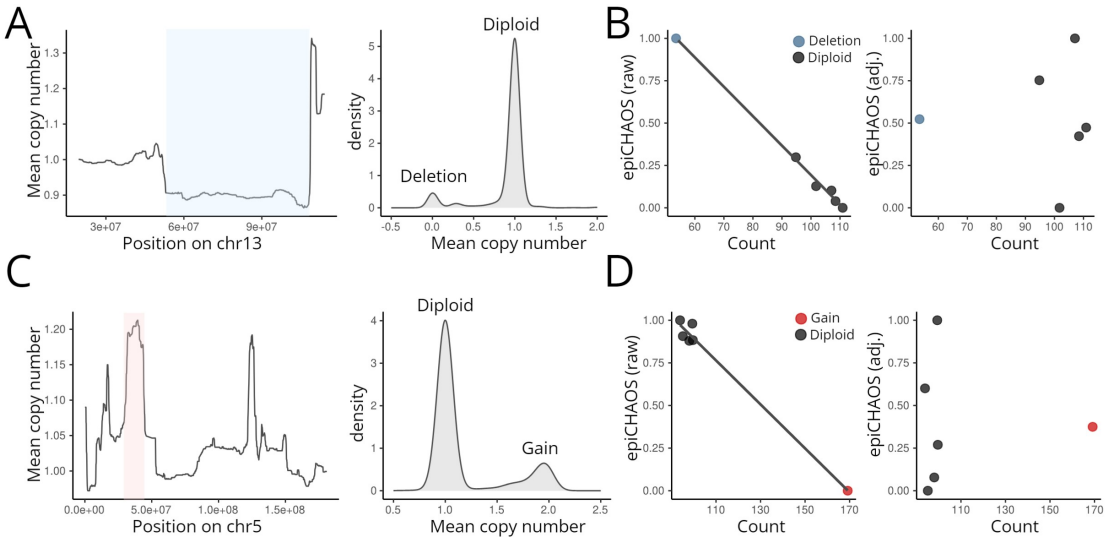

Figure 5.3: EpiCHAOS scores are not influenced by copy number. A. Selection of cells with subclonal copy number loss on chromosome 13 in the Hep-1 cell line. Line plot (left) shows the mean copy number among single cells on chromosome 13. Copy number was called using epiAneuFinder, where 1 = diploid, 0 = loss and 2 = gain. X-axis displays chromosome location. The highlighted region was selected as a region of subclonal deletion for which to test the effects of copy number alterations on epiCHAOS scores. Density plot (right) shows the distribution of cells with and without deletion in the selected region. **B.** After reducing the peaks-by-cells matrix for peaks within the deleted region (highlighted in **(A)**), one group of 100 deletion cells and five groups of 100 diploid cells were sampled for epiCHAOS calculation. Scatterplots show the relationship between epiCHAOS scores (epiCHAOS) and counts (average counts per cell in the group in the subsetted peaks matrix) before (left) and after (right) adjustment for total counts. EpiCHAOS scores were adjusted for counts by fitting a linear regression model of epiCHAOS scores against counts (average counts per cell in the group) and taking the residuals of the model as an adjusted score. Each point represents a group of either diploid or deleted cells on which epiCHAOS scores are computed. **C.** Selection of cells with subclonal copy number gain on chromosome 5 in the Hep-1 cell line. **C-D.** Plots show the same as in **(A-B)** in the situation of a copy number gain (increased counts).

As an additional *in silico* validation strategy, I created synthetic mixtures of various epigenetically distinct cell types. For this I utilised a previously published human hematopoietic dataset [226]. I synthesised mixtures of two to five cell types including all possible combinations of HSCs, monocytes, B-cells, CD8-T cells and plasmacytoid dendritic cells (pDCs) (Fig. 5.4A). To focus on regions which are differentially accessible between cell types, I subset the peaks-by-cells matrix to the 500 top marker peaks for each cell type, and then applied epiCHAOS to each individual cell type and mixture. As expected, I found that epiCHAOS scores were lowest in individual cell types, and increased in proportion to the number of cell types in the mixture, being highest when all five cell types were mixed (Fig. 5.4B).

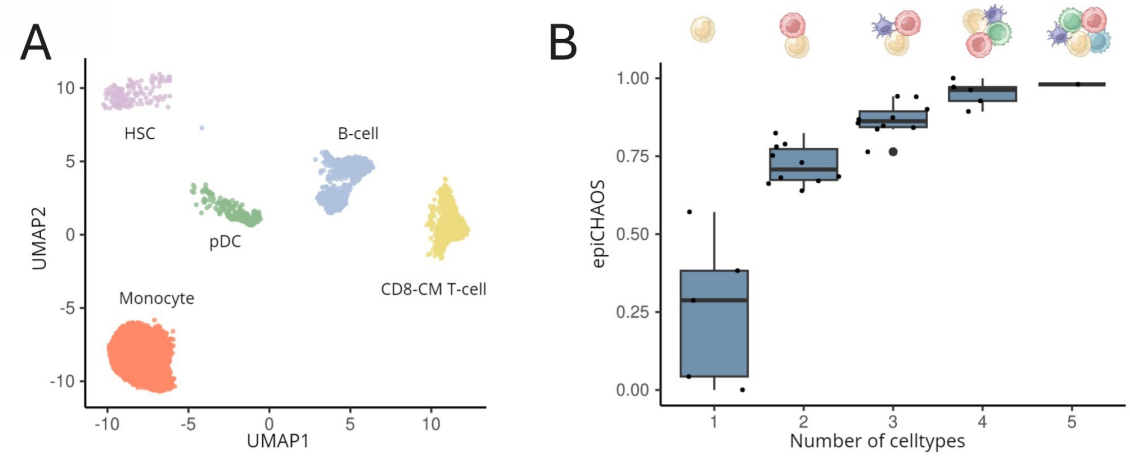

Figure 5.4: Validation of epiCHAOS in *in silico* cell-type mixtures. A. Uniform Manifold Approximation & Projection (UMAP) embedding illustrates scATAC profiles from five selected cell types of human bone marrow [226]. After selecting 500 top differentially accessible peaks for each cell type, in silico mixtures of two to five cell types in all possible combinations were created. **B.** Boxplots show the relationship between epiCHAOS scores (epiCHAOS) and number of cell types (x-axis) after *in silico* mixing.

5.1.2 EpiCHAOS scores are minimally influenced by technical noise and choice of clustering parameters

To ensure that epiCHAOS does not perceive "noisier" data as higher heterogeneity, I investigated the relationship between epiCHAOS scores and a range of established metrics of technical noise in scATAC-seq data; the fraction of reads in peaks (FRIP) score, transcription start site (TSS) enrichment score and nucleosome ratio. For this I used the same liver cancer cell line dataset (used for CNA analysis in the previous section) and separated cells into bins of increasing technical noise according to each specified metric. I found no correlation between epiCHAOS score and any of the above metrics of cell quality except in cases of extremely low TSS enrichment scores where epiCHAOS scores tended to be increased (Appendix, Fig. 8.5). In any case, such cells are routinely filtered out as part of standard quality control. I also confirmed that epiCHAOS was not influenced by the number of cells per cluster (Appendix, Fig. 8.5).

I also investigated how epiCHAOS behaves under different clustering parameters. For this, I selected as an example a scATAC-seq dataset from breast cancer [140]. I performed clustering in ArchR using a range of clustering resolutions from 0.1 to 0.9 and then investigated the resulting epiCHAOS scores. For each resolution, I tested the correlations between per-single-cell epiCHAOS scores and found that every comparison yielded a Pearson R greater than 0.9 (Appendix, Fig. 8.6). This suggests that overall, a similar pattern of epiCHAOS scores emerges in different cell groups regardless of clustering resolution.

5.1.3 EpiCHAOS reflects epigenetic heterogeneity associated with developmental plasticity

Having validated the performance of epiCHAOS in a range of synthetic datasets, I next moved on to real biological applications. First, in order to show that epiCHAOS generates results that make biological sense, I focused on a biological setting where there are prior expectations about epigenetic heterogeneity: developmental systems. Here, it is generally accepted that uncommitted, multipotent cell types, which need to retain their ability to differentiate down numerous different trajectories, should exhibit high cell-to-cell epigenetic heterogeneity, potentially with different cells being primed for different cell fates. By contrast, terminally differentiated cell types have already committed to a specific fate, and have no longer such a need to diversify. We would therefore expect such cells to be more fixed or stable at the epigenetic level, and to exhibit lower epigenetic heterogeneity [118, 228]. I decided to apply epiCHAOS to a range of developmental contexts to see how the results align with this accepted biological paradigm.

I selected scATAC-seq datasets from three different examples of developmental systems; human hematopoiesis (bone marrow mononuclear cells) [226], drosophila embryogenesis and mouse gastrulation [229]. Across hematopoietic cell types, epiCHAOS scores were highest in HSCs, followed by other hematopoietic progenitor cells (CLPs, CMPs, LMPPs, and early erythroid cells). As expected, I detected lower epiCHAOS scores in the more differentiated cells of the myeloid, erythroid and lymphoid lineages (Fig. 5.5A).

Similarly, in data from mouse gastrulation, epiCHAOS scores were highest in less differentiated cells, especially in the primitive streak and in primordial germ cells, and decreased throughout the formation of distinct meso-, endo- and ectodermal lineages (Fig. 5.5B).

In Drosophila embryogenesis, I also observed high epigenetic heterogeneity at the earliest multipotent stages including undifferentiated cells, blastoderm and germ cells (Fig. 5.5C), however I detected even higher epiCHAOS scores in neural cells, supposedly reflecting the extraordinary functional heterogeneity of the neural compartment [230].

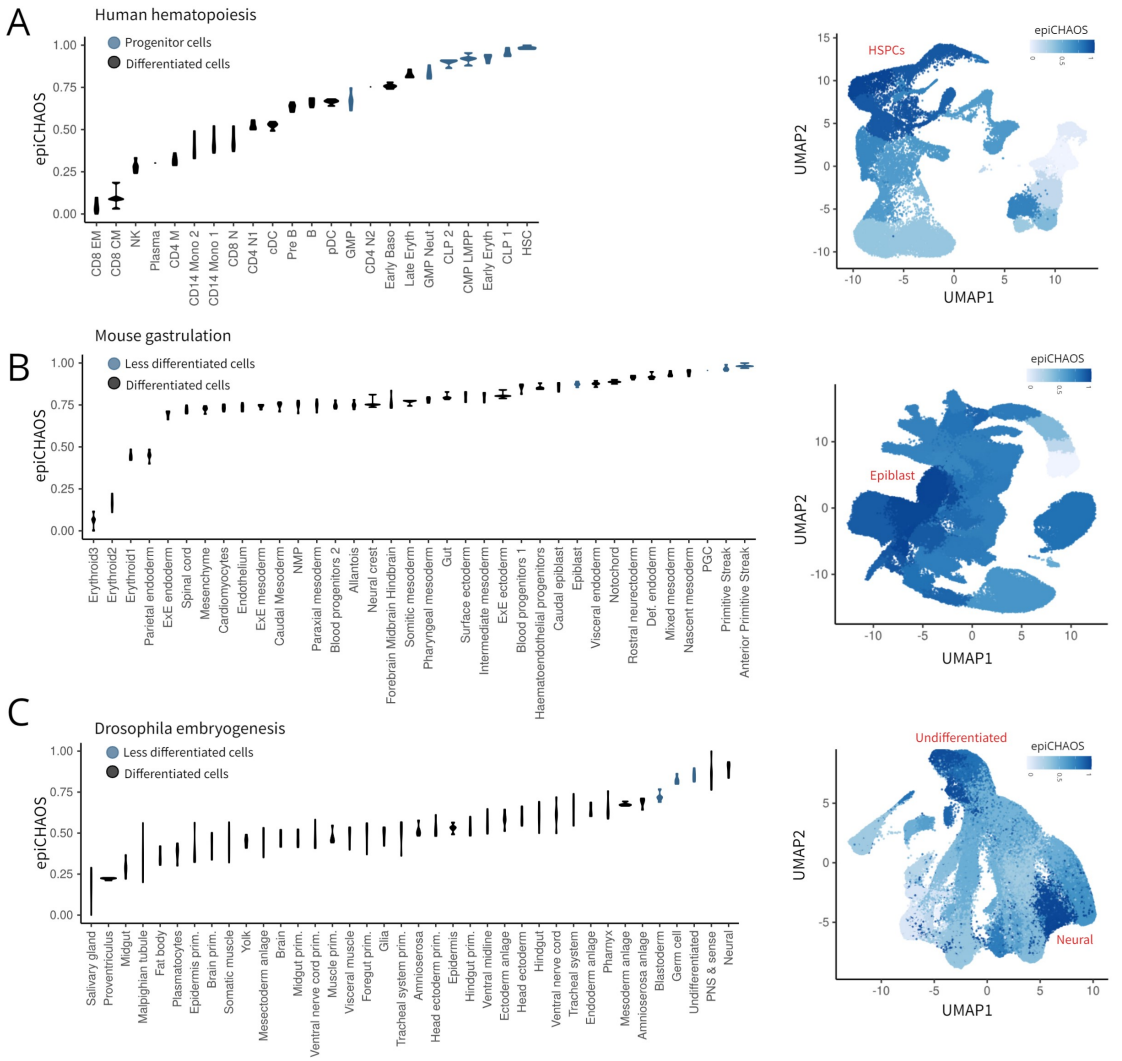

Figure 5.5: EpiCHAOS reflects epigenetic heterogeneity associated with developmental plasticity. Violin plots (left) showing epiCHAOS scores (epiCHAOS) computed in scATAC-seq data from (**A**) human hematopoiesis [226], (**B**) mouse gastrulation [229] and (**C**) drosophila embryogenesis [231]. EpiCHAOS scores were computed per-cell type as annotated in the original publications. Violins represent the scores computed in five random subsamples of 100 cells from each cell type, or once where fewer than 100 cells were available. Plots are ordered by epiCHAOS scores and progenitor cells and undifferentiated tissue types are highlighted in blue. UMAP embeddings (right) illustrating epiCHAOS scores in the same datasets as in violin plots. UMAPs are coloured by epiCHAOS scores computed per annotated cell/tissue type.

To investigate this outside of an embryonic context, I applied epiCHAOS to scATAC-seq data from the a pan-tissue atlas, where I found that epiCHAOS scores were generally higher in neural tissues compared to all other tissue types with the exception of placenta [232] (Fig. 5.6).

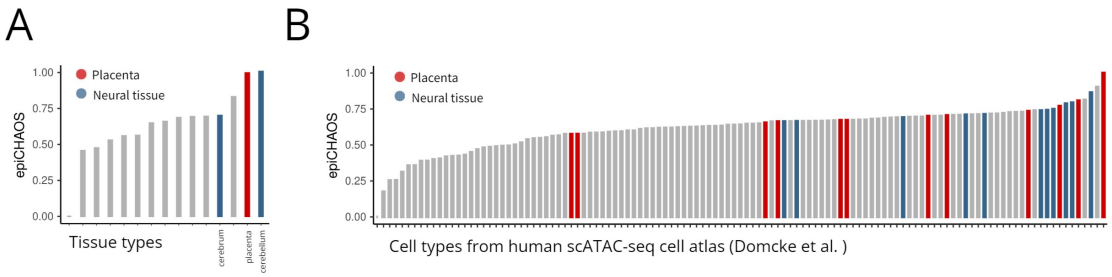

Figure 5.6: EpiCHAOS scores are increased in neural tissues and placenta. A-B Bar plots of epiCHAOS scores (epiCHAOS) in different tissues (**A**) and cell types (**B**) from the human scATAC-seq cell atlas [232]. Neural and placental tissues are highlighted.

To assess how epiCHAOS compared to previously established measurements of cellular plasticity, I correlated epiCHAOS scores with CytoTRACE – a scRNA-based metric designed to capture stemness/plasticity, which is based on the observation that the number of expressed genes per cell decreases during cellular differentiation [233]. In the hematopoietic system, epiCHAOS correlated moderately with CytoTRACE scores (Fig. 5.7), and in some cases the pattern of epiCHAOS scores across cell types was better representative of the differentiation trajectory than that of CytoTRACE. For example epiCHAOS detected higher heterogeneity in naive CD8+ and CD4+ T-cells compared to memory T cells, reflecting the expectation that naive T cells should have greater developmental potential. This was not detected by CytoTRACE, suggesting that epiCHAOS captures some features of cellular plasticity that might not be detectable at the transcriptional level. Similarly in gastrulation and embryogenesis datasets, epiCHAOS correlated with previously annotated metrics of developmental time (Fig. 5.7).

Next, I investigated whether this pattern of epigenetic heterogeneity was reflected by higher transcriptional heterogeneity in less differentiated cells (Fig. 5.8). For this I used a previously described metric of cell-to-cell transcriptional heterogeneity. While I generally observe higher transcriptional heterogeneity in less differentiated cells, overall correlations between epiCHAOS and transcriptional heterogeneity were moderate.

Collectively these data suggested that epiCHAOS can provide an accurate approximation of epigenetic heterogeneity indicative of developmental plasticity. This encouraged me to apply epiCHAOS to data from malignant settings, considering that it might also serve as an indicator of plasticity in cancer.

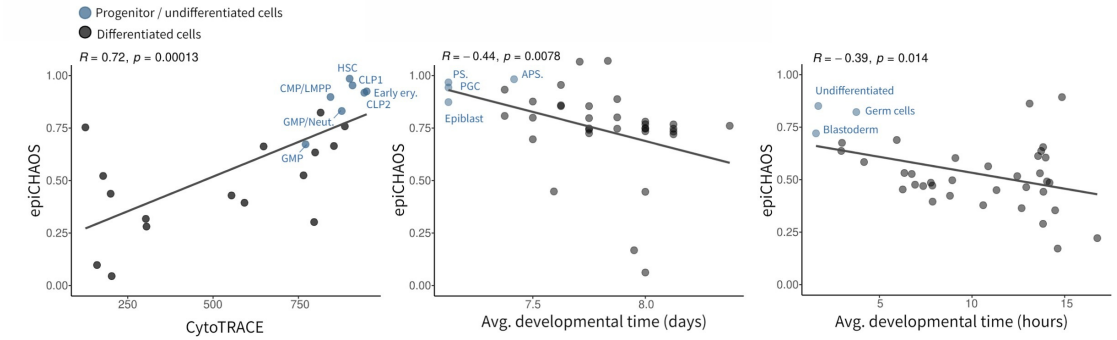

Figure 5.7: EpiCHAOS scores correlate with developmental time. Scatter plots correlating epiCHAOS scores (epiCHAOS) with developmental time as defined by (left) CytoTRACE score, averaged across cells (human hematopoietic system), (middle) developmental time in days at sample collection, averaged across cells (mouse gastrulation), or (right) predicted developmental time in hours (Drosophila embryogenesis). EpiCHAOS scores represent the average of pseudo-replicates shown in Fig. 5.5. Linear regression lines are displayed and Pearson correlation coefficients and p-values indicated.

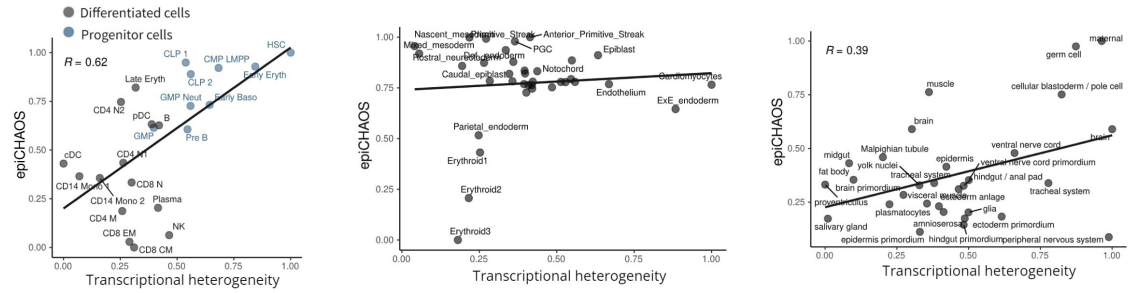

Figure 5.8: Correlation between epigenetic and transcriptional heterogeneity in developmental settings. Scatterplots show the correlation of epiCHAOS scores (epiCHAOS) with transcriptional heterogeneity scores in human hematopoiesis (left), mouse gastrulation (middle) and drosophila embryogenesis (right). Transcriptional heterogeneity is quantified using the associated scRNA-seq data (not data from the same cells) by taking the mean of pairwise euclidean distances between all cells within a group/cluster.

5.1.4 EpiCHAOS correlates with features of plasticity in malignant cells

To investigate epigenetic heterogeneity in malignancy, I selected two previously published scATAC-seq datasets from 16 breast [140] and 16 liver [234] cancer patients. To investigate whether epigenetic heterogeneity coincides with particular features of tumor cells, I calculated molecular signature scores for each malignant cluster by summarising the scATAC-seq gene score matrices across biological pathways. I applied epiCHAOS to each dataset, subset for malignant cell clusters, and correlated the resulting epiCHAOS scores against each molecular signature (Fig. 5.9A).

Epithelial-to-mesenchymal transition (EMT) – a process which is considered integral to breast cancer plasticity and metastasis – was among the top gene sets correlated with epiCHAOS scores across breast cancer cell clusters [235] (Fig. 5.9B). EMT-related signaling pathways such as TGF-beta and WNT signaling, were also correlated with epiCHAOS scores (Appendix, Fig. 8.7). Clusters with high epiCHAOS scores also displayed higher accessibility of genes within various previously described gene expression signatures of the more aggressive invasive and metaplastic breast cancer subtypes (Appendix, Fig. 8.7). Gene sets related to EMT were similarly correlated with epiCHAOS scores in liver cancer, and the average accessibility of previously described prognostic gene signatures of liver cancer was increased in clusters with high epiCHAOS scores (Fig. 5.9B, Appendix Fig. 8.7).

These recurring correlations between epiCHAOS and EMT scores hinted at a possible link between epigenetic heterogeneity and plasticity in cancer cells. To investigate this further I applied epiCHAOS to a scATAC-seq dataset from childhood ependymoma comprising multiple differentiated (astrocytes, ependymal cells, neural progenitor cells, and mesenchymal-like cells) as well as undifferentiated tumor cell types [236]. Fittingly, epiCHAOS scores were increased in the population of undifferentiated cells, which were previously shown to be enriched in more aggressive ependymomas [236, 237] (Fig. 5.9C-D). Among malignant cell types, the lowest epiCHAOS scores were detected in ependymal cells (Fig. 5.9C-D), which represent the latest stage of differentiation, and which are known to be associated with less aggressive disease [238].

In summary, these data indicate that the measure of cell-to-cell epigenetic heterogeneity provided by epiCHAOS can capture features of cancer cell plasticity and dedifferentiation, which correlate with aggressive tumor phenotypes.

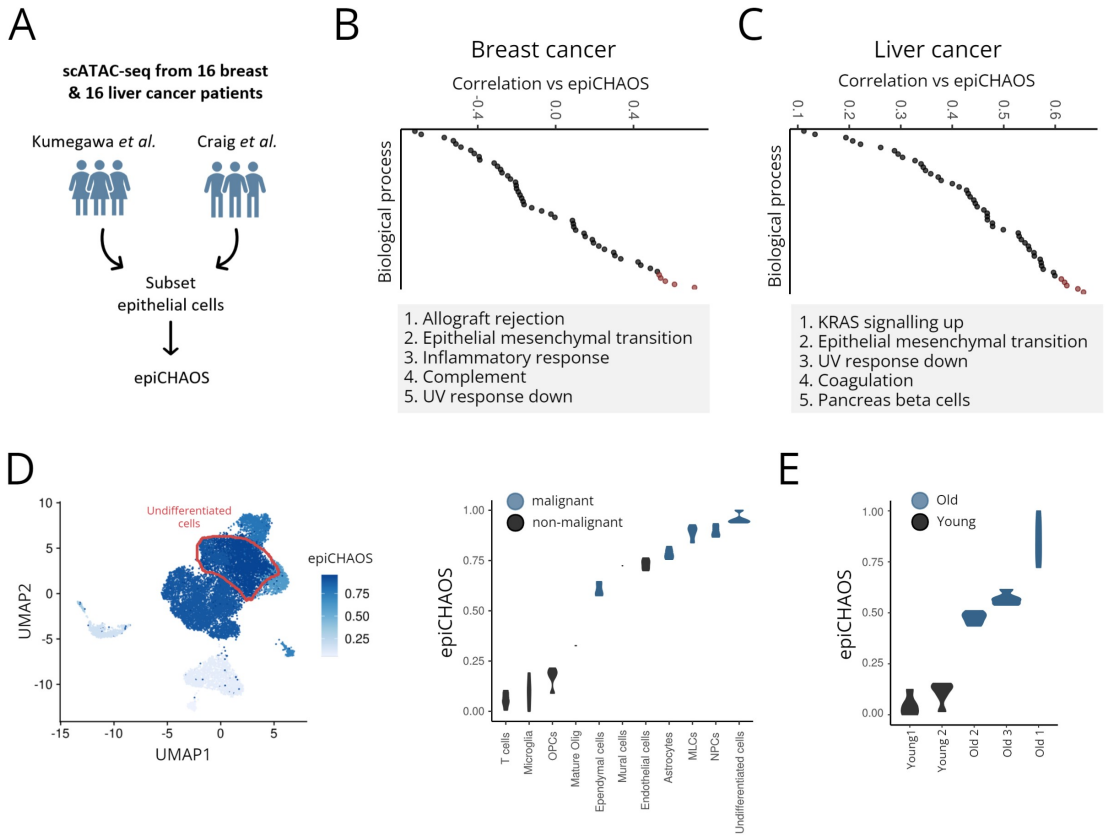

Figure 5.9: EpiCHAOS correlates with features of plasticity in malignant cells and epigenetic heterogeneity in aging. **A.** Schematic describing breast [140] and liver cancer [234] datasets used for epiCHAOS calculation. **B-C.** Dot plots illustrate ordered Pearson coefficients after correlation of per-cluster epiCHAOS scores against gene set scores for all Hallmark Gene Ontology biological processes in breast (**B**) and liver cancer (**C**). Top 5 correlations in each dataset are highlighted and labeled below. **D.** UMAP embedding (left) of scATAC-seq data from five primary and two metastatic childhood ependymoma tumors [236]. Cells are coloured by epiCHAOS scores (epiCHAOS) computed for each cell type. Undifferentiated cells are highlighted in red. Violin plot (right) ordered by epiCHAOS scores (epiCHAOS) for all malignant and non-malignant cell types annotated in ependymoma tumors. Malignant cell types are highlighted in blue. Violins represent the scores computed in five random subsamples of 100 cells from each cell type, or once where fewer than 100 cells were available. **E.** Violin plot ordered by epiCHAOS scores computed in scATAC-seq data from old (blue; n = 3) and young (black; n = 2) mouse HSCs [239]. Violins represent the scores computed in five random subsamples of 100 cells from each group.

5.1.5 EpiCHAOS reveals increased epigenetic heterogeneity associated with hematopoietic aging

Next I investigated whether epiCHAOS would detect differences in epigenetic heterogeneity associated with aging – a process which has been shown to correlate with accumulating stochastic variation in epigenetic marks [240]. For this I applied epiCHAOS to a scATAC-seq dataset from HSCs derived from old (3 mice, 24 months) and young (2 mice, 2 months) mouse bone marrow [239]. In agreement with previous studies, epiCHAOS detected increased epigenetic heterogeneity in aged HSCs compared to those from younger animals (Fig. 5.9E).

5.1.6 EpiCHAOS reveals elevated epigenetic heterogeneity at PRC2 targeted regions and promoters of developmental genes

This chapter has so far focused on comparisons of genome-wide epigenetic heterogeneity between different cell types or conditions. Next, I decided to investigate how different classes of genomic regions might differ in epigenetic heterogeneity within a single group of cells.

For this I focused on bone marrow HSCs from the previously described scATAC-seq dataset from human hematopoiesis [226]. To compare epigenetic heterogeneity across different genomic regions, I utilised annotated chromatin/transcription factor binding profiles from the encyclopedia of DNA elements (ENCODE) database, subsetting the peaks-by-cells matrix to regions overlappng with each of the included factors. Here, epiCHAOS detected especially high epigenetic heterogeneity at PRC2 targeted regions, as well as at binding sites for CCCTC-binding factor (CTCF) and Cohesin (Fig. 5.10A). Repeating this analysis for each of the hematopoietic cell types individually, I found that the pattern of epigenetic heterogeneity across regions was largely consistent across cell types, with the majority of between-cell-type Pearson correlations greater than 0.7. (Appendix, Fig. 8.8). To validate this finding, I compared epiCHAOS scores to the DNA methylation variation across CpG sites in HSCs (analysis of RRBS data performed by Martina Braun), which I summarised for each TFBS, as an independent measurement of per-region epigenetic heterogeneity. I detected a similar pattern of epigenetic heterogeneity at the level of DNA methylation, with PRC2 targets standing out as the most variably methylated regions, followed by binding sites for CTCF and cohesin (Fig. 5.10B). To discern whether this pattern translates to elevated transcriptional heterogeneity, I calculated transcriptional noise for each gene in HSCs using a previously described strategy based on the coefficient of variation [241]. Consistent with their higher epiCHAOS scores, and in line with previous reports [242, 117], I found that transcriptional noise was significantly increased at genes targeted by the PRC2 complex compared to non-PRC2 targets (Fig. 5.10C).

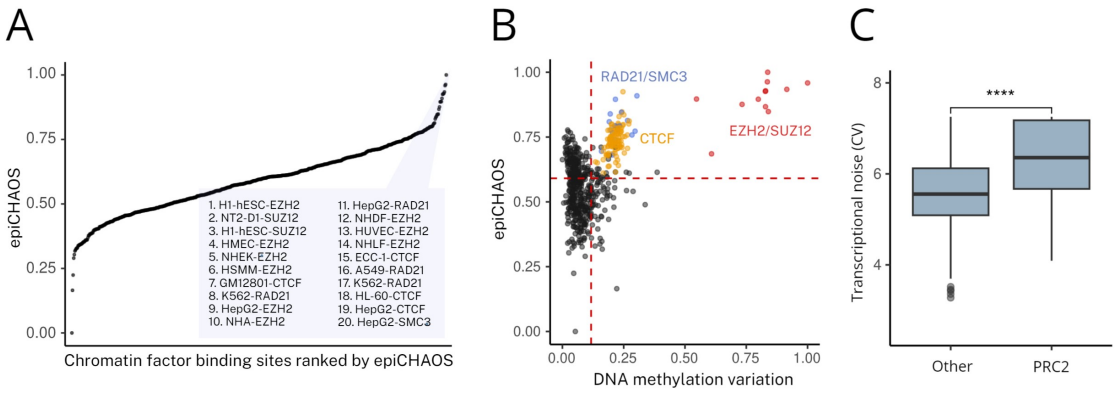

Figure 5.10: EpiCHAOS reveals increased heterogeneity at binding sites for PRC2, CTCF and cohesin. **A.** Ordered dot plot showing epiCHAOS scores (epiCHAOS) computed across region sets for each ENCODE chromatin factor binding site in HSCs, ordered by epiCHAOS scores. Top 20 region sets (cell type and binding site) are labeled. **B.** Scatter plot comparing epiCHAOS scores (epiCHAOS) with DNA methylation variation at each chromatin factor binding site as in (**A**). PRC2 targets (binding sites for EZH2/SUZ12, red), CTCF targets (orange) and Cohesin binding sites (binding sites for RAD21/SMC3, blue) are highlighted. X-axis represents the average of per-CpG methylation variances across 10 individuals at all CpG sites overlapping with the respective region set. Values are scaled to a 0-1 range. **C.** Boxplot comparing transcriptional noise, measured using the coefficient of variation (CV), between PRC2 target genes and other genes in HSCs using scRNA-sequencing (RNA-seq) data from Granja *et al.* [226].

Next, I compared epigenetic heterogeneity across sets of functionally related genes, calculating epiCHAOS scores at promoter-associated peaks subset for each gene ontology biological process. This analysis revealed elevated heterogeneity at genes associated with developmental processes including "cell fate commitment", "cell fate specification" and "somatic stem cell division" (Fig. 5.11A). This pattern also correlated quite well between different hematopoietic cell types, with the majority of between-cell type comparisons yielding a Pearson correlation coefficient above 0.7 (Appendix, Fig. 8.9). Nevertheless, some interesting distinctions also appeared between cell types, with certain gene sets showing increased heterogeneity in a cell-type specific manner; most notably, I found that the cell-to-cell accessibility of bivalent genes and genes related to cell fate specification appeared to be highly heterogeneous in HSCs, but relatively homogeneous in more differentiated cell types (Fig. 5.11B).

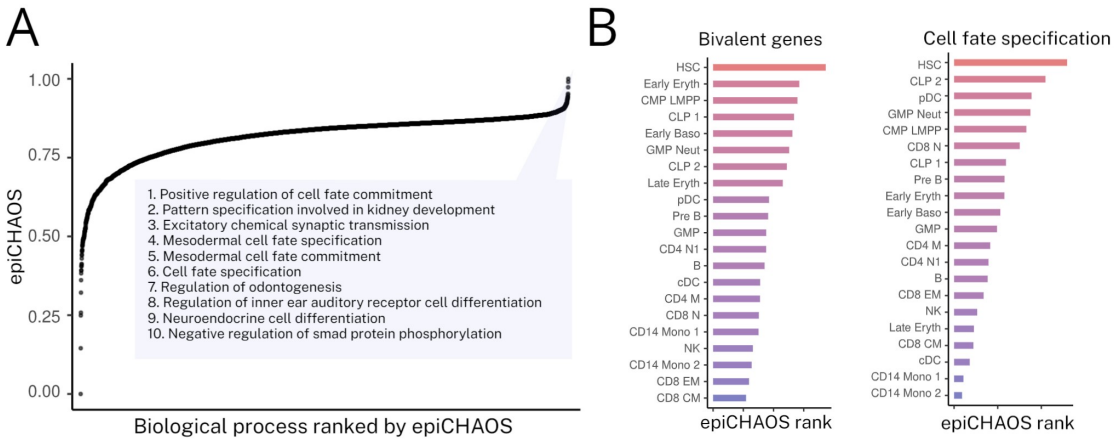

Figure 5.11: EpiCHAOS reveals increased heterogeneity of developmental genes. A. Ordered dot plot showing epiCHAOS scores (epiCHAOS) computed across promoters for each gene ontology biological process in HSCs, ordered by epiCHAOS scores. **B.** Bar plots comparing epiCHAOS ranks for the set of (left) bivalent genes and (right) genes within the "cell-fate specification" Gene Ontology term across different hematopoietic cell types. The higher the rank indicates that the selected gene set has higher epiCHAOS scores compared to other gene sets in that celltype. Ranks were -log(10) transformed for display.

These observations encouraged me to consider whether certain genomic regions might contribute a greater amount to the increased genome-wide heterogeneity I had observed in hematopoietic stem/progenitors compared to more differentiated cell types. This question required a means of assessing differential epiCHAOS scores between two cell groups or conditions. For this I used a permutation approach, comparing the difference in epiCHAOS scores between two groups of interest with that between pairs of 1000 randomly computed groups of cells sampled from the same pool of cells. This allowed the calculation of p-values to assess the significance of difference in epiCHAOS scores between two cell groups at each selected set of genomic regions. I performed this comparison for HSCs vs.

monocytes, HSCs vs. B-cells and HSCs vs. CD8T-cells, using the subset of ENCODE chromatin factor binding sites that were previously defined in the hematopoietic system. In all three comparisons I found that binding sites for the PRC2 complex component EZH2 displayed the greatest increase in epiCHAOS scores in the HSC group, hinting that epigenetic heterogeneity at these key genomic regions might be an important feature of HSC plasticity or developmental potential (Fig. 5.12).

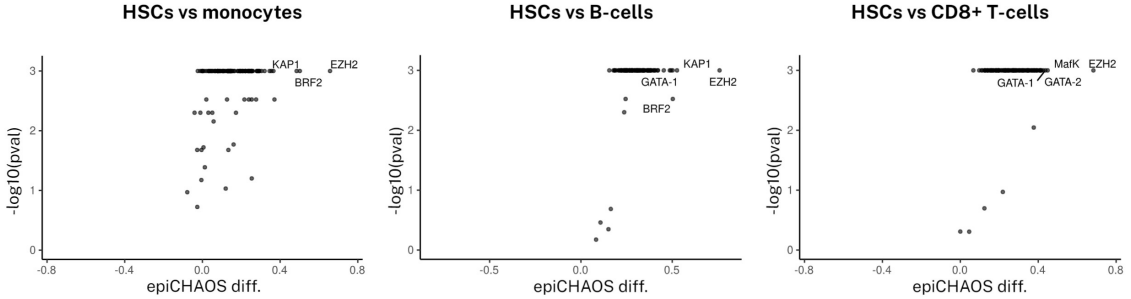

Figure 5.12: EpiCHAOS reveals increased heterogeneity at PRC2 target regions in HSCs compared to differentiated blood cells. Volcano plots illustrate differential heterogeneity between HSCs and monocytes (left), B-cells (middle) and CD8+ T-cells (right). Differential heterogeneity was tested for each ENCODE TFBS (binding sites measured in K562 cells). For each TFBS, the -log10(p-value) obtained by permutation test is displayed on the y-axis, and the difference in epiCHAOS scores between the two cell types is displayed on the x-axis, where a higher number indicates a higher heterogeneity in HSCs compared to the other tested cell type.

5.1.7 EpiCHAOS is applicable to single-cell epigenomics data from any modality

So far I have focused on demonstrating epiCHAOS' capabilities in scATAC-seq data, which is currently the most commonly used modality for single-cell epigenomics studies. Yet, technologies for studying DNA methylation and histone modifications at single-cell resolution are also emerging and are likely to soon come into wider use. To prove the utility of my metric for applications to diverse epigenomics data types, I tested epiCHAOS in three additional datasets from different single-cell modalities: (i) single-cell nucleosome, methylation and transcription sequencing (scNMT-seq) data from mouse gastrulation [243], (ii) single-cell targeted analysis of the methylome (scTAM-seq) [137] data from mouse hematopoiesis [244], and single-cell chromatin immunoprecipitation sequencing (scChIP-seq) data from breast cancer cells [138].

In the mouse gastrution dataset, I calculated DNA methylation-based epiCHAOS scores across cells at promoters, gene bodies and CpG islands, and found largely similar patterns as I previously observed at the level of chromatin accessibility, with increased heterogeneity in the epiblast compared to more differentiated germ layers (Fig. 5.13A). I also detected a moderate correlation between promoter-wide DNA methylation-based and ATAC-based

epiCHAOS scores from the same cells (Fig. 5.13B).

In scTAM-seq data from the hematopoietic system [137, 244], I found that DNA methylation-based heterogeneity was highest in more primitive hematopoietic progenitor cells such as HSCs and early MPPs and progressively decreased towards more differentiated progenitors such as GMPs and pre-B cells (Fig. 5.13C). This pattern was roughly in line with my previous observations at the level of chromatin accessibility (Fig. 5.5A).

Finally, I applied epiCHAOS to H3K27me3 scChIP-seq data from a previous study of breast cancer resistance to Capecitabine therapy [138]. Here, I detected increased epigenetic heterogeneity in resistant compared to sensitive cells – a finding that was reported, but not quantitatively investigated, in the original paper, and which supports the potential contribution of cell-to-cell epigenetic heterogeneity in the emergence of therapy resistance (Fig. 5.13D).

These data illustrate that epiCHAOS is not restricted to a single data type, but is broadly applicable for comparisons of cell-to-cell heterogeneity in any kind of single-cell epigenomics data.

5.2 Discussion

From inter- to intra-cluster assessment of epigenetic heterogeneity

Single-cell sequencing technologies have revolutionised the study of genetic, transcriptional and epigenetic heterogeneity in health and disease. At the epigenetic level, the main application of these techniques has been to highlight the presence of epigenetically distinct populations of cells – a layer that can be thought of as "macro-heterogeneity" or *inter*-cluster heterogeneity. In this thesis, my aim was rather to tackle the question of how cell-to-cell heterogeneity can be quantified *within* populations or groups of cells (i.e. "micro-heterogeneity" or *intra*-cluster heterogeneity) – such that it would be possible to determine whether a given group, condition or cluster of cells is more or less heterogeneous than another. Here, I have presented epiCHAOS; a quantitative metric of cell-to-cell epigenetic heterogeneity computed from single-cell epigenomics data. I have extensively validated the performance of epiCHAOS using a range of synthetic and real single-cell epigenomics datasets, and showcased its usage in a range of biological systems from development to malignancy and aging, to investigate both genome-wide and region-specific heterogeneity.

Epigenetic heterogeneity and plasticity in development, cancer, and aging

For biological validation of epiCHAOS, I focused on scenarios where biological expectations could be drawn from previous studies. For example, in developmental systems it is believed is that less differentiated and more developmentally plastic cell types should have higher heterogeneity compared to more differentiated and functionally specialised ones [228, 118].

Application of epiCHAOS to various datasets from developmental contexts revealed patterns of epigenetic heterogeneity in agreement with this biological paradigm. Subsequently, I showed that epigenetic heterogeneity is increased with age in the hematopoietic system, which reflects the expected stochastic decay of epigenetic information which has been established as a hallmark of aging [240, 245]. What has been described as "epigenetic drift" – the progressive noise in epigenetic marks, most notably DNA methylation, that occurs

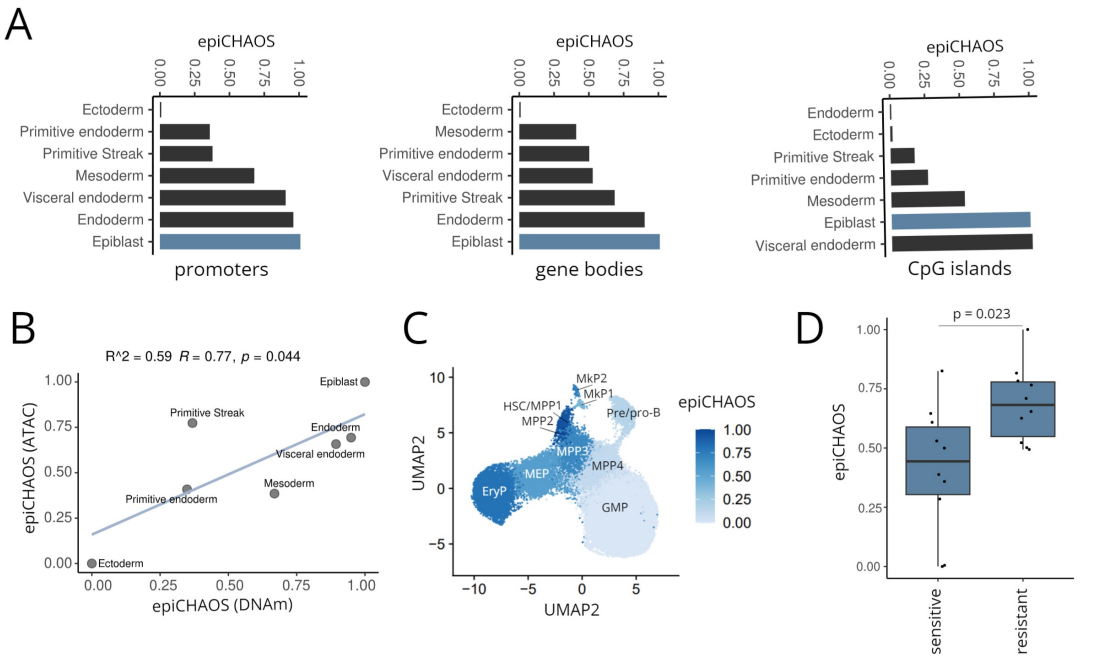

Figure 5.13: EpiCHAOS is applicable to multiple types of single-cell epigenomics data. A. Bar plots comparing epiCHAOS scores (epiCHAOS) across different lineages of mouse gastrulation using scNMT-seq DNA methylation data from Argelaguet *et al.* [243]. Methylation data summarised across promoters, gene-bodies or CpG islands were used for epiCHAOS computation. Epiblast is coloured in blue. **B.** Scatter plot comparing promoter-wide epiCHAOS scores across different gastrulation lineages using single-cell DNA methylation [epiCHAOS (DNAm)] and ATAC-seq [epiCHAOS (ATAC)] data from the same cells as in (**A**). Linear regression line is shown with Pearson correlation coefficient and p-value. **C.** UMAP embedding generated from scTAM-seq DNA methylation data from Scherer *et al.* [244]. Hematopoietic progenitor state and epiCHAOS scores are annotated. GMP: granulocyte monocyte progenitor, EryP: Erythroid progenitor, MEP: myeloid/erythroid progenitor, MPP: multipotent progenitor, HSC: hematopoietic stem cell, MkP: megakaryocyte progenitor. **D.** EpiCHAOS scores calculated using scChIP-seq data for H3K27me3 from Grosselin *et al.* [138]. Cells were obtained from a patient-derived xenograft (PDX) breast cancer model, separated based on sensitivity or resistance to Capecitabine. Ten subsamples of 100 cells each were taken per condition. Boxplot shows comparison of epiCHAOS scores (epiCHAOS) between sensitive and resistant cells. Wilcoxon's p-value is shown.

with aging – is thought to reflect a progressive decline in the fidelity of the replication-associated DNA methylation maintenance system. This is supported by the fact that different tissues exhibit signs of epigenetic aging to differing degrees, such that correlates with their proliferative history [246, 247]. Thus, studies of epigenetic aging have typically focused on DNA methylation, and the phenomenon has not been demonstrated before at the level of cell-to-cell chromatin accessibility.

Moreover, in showing that epiCHAOS correlates with EMT signatures in cancer, and is increased in dedifferentiated tumor cell-types, I provide evidence to support the notion that epigenetic heterogeneity may be linked to cancer cell plasticity or stemness. However, since this analysis was limited to a small number of tumor types, it will be valuable to further explore these correlates in a broader pan-cancer context.

Differences in heterogeneity along the genome

As well as exploring the differences in genome-wide heterogeneity between cell types/states, I also applied epiCHAOS to investigate the preferential heterogeneity of specific genomic regions, where I found that in particular, there is elevated cell-to-cell heterogeneity at binding sites for polycomb complexes, CTCF and cohesin. The increased heterogeneity at polycomb targeted regions can be understood in the context of previous studies which have highlighted the increased transcriptional noise and DNA methylation-based heterogeneity of these loci. For example, Kar *et al.* have shown that PRC2 targets display higher cell-to-cell variation in gene expression, with a low transcriptional burst frequency giving rise to oscillatory patterns between transcriptionally "on" and "off" states over time [242]. Kumar *et al.* demonstrated that polycomb targeted genes are heterogeneously expressed within colonies of pluripotent stem cells compared to other gene categories [117]. This heterogeneity may be related to the fact that many PRC2 targets are associated with bivalent chromatin states – characterised by the presence of both active and repressive histone marks. Faure *et al.* showed that genes with bivalent chromatin exhibit higher transcriptional noise [248]. Similarly, Feinberg & Irizarry have demonstrated that DNA methylation stochasticity is increased at developmental genes – many of which are also targeted by PRC2 [249]. The novelty of my analyses is in demonstrating that these patterns of variability can be observed at the epigenetic level between single cells. Furthermore, my investigation of differential heterogeneity between hematopoietic cell types revealed that polycomb-targeted regions also appear to be the most preferentially heterogeneous in stem cells compared to more differentiated cells. Notably, PRC2 targeted regions also exhibit increased epigenetic variation in cancer by comparison to normal tissues [250]. It is interesting to speculate whether epigenetic heterogeneity at these specific genomic regions might play a role in tumor evolution and plasticity, and how this might be influenced by the frequent disruption of PRC2 complex components in cancer [251].

Beyond PRC2 binding sites, I also showed that epigenetic heterogeneity is increased at binding sites for the chromatin regulators, CTCF and cohesin. This observation also

seems to be consistent with recent genome-wide chromatin interaction studies in single cells, which revealed that CTCF/cohesin-mediated chromatin loops are highly variable between cells, compared to other regions of the genome [252].

Future perspectives

Understanding how epigenetic heterogeneity influences different stages of malignant evolution – from tumor initiation to progression, remission and relapse – and deciphering its role in the progression to metastasis and therapy response, could have important implications for our perception and treatment of cancer. In providing a computational strategy to explore these phenomena, epiCHAOS should be a valuable addition to the cancer research community, especially to those focusing on plasticity and stemness, and non-genetic mechanisms of therapy resistance. Going forward, it will be possible to investigate whether epigenetic heterogeneity is increased in pre-/cancerous tissues compared to their cells of origin, whether it changes with increasing disease severity, whether it increases the propensity for metastatic progression, therapeutic resistance (generally or to certain kinds of therapies), and whether it influences – or is influenced by – responses to other internal or external stresses. It will be possible to determine whether tumors become more homogeneous after treatment with the selection of certain subclones, or if subclones with higher heterogeneity are rather selected for. Future studies might also investigate if and how epigenetic heterogeneity is influenced by treatment with epigenetic-based therapies such as DNA hypomethylating agents, and whether mutations in epigenetic regulators in cancer cells can give rise to heterogeneity in a certain context. Furthermore it will be interesting to understand whether epigenetic heterogeneity arises in concert with, or independently of genomic heterogeneity, e.g. differences in mutational patterns between cells within a tumor. EpiCHAOS might also be complimented by additional metrics to evaluate the similarity across different clusters, such that both inter- and intra-cluster measurements are taken into consideration to better quantitatively appreciate the heterogeneity of complex tumors.

Beyond these questions, epiCHAOS should also yield novel biological insights outside of the cancer field, for instance in developmental biology, aging and immunity. It will be possible for example to investigate whether epigenetic heterogeneity is a general feature of aging or whether it is restricted to certain tissues such as blood, whether it is increased in the immune compartment during an immune or inflammatory response, and whether such heterogeneity is observed in other disease states in which plasticity programs are activated, such as wound healing and fibrosis.

While epiCHAOS is designed to be applicable to any kind of single cell data in the form of a binarised regions-by-cells matrix, an alternative metric is needed for calculations of transcriptional heterogeneity due to the continuous nature of these data. In this chapter I have utilised previously described strategies for computing heterogeneity in scRNA-seq data, based on pairwise euclidean distances [147]. In general I noted that the correlations

between epiCHAOS and transcriptional heterogeneity scores are not particularly strong, however it is unclear whether this observation is biologically meaningful, or only reflects the lack of a validated metric for quantifying transcriptional heterogeneity between cells. A potential problem with the existing strategy is that it detects higher heterogeneity in cells that express more genes. Similarly, the coefficient of variation – a metric which is widely used as a measurement of per-gene transcriptional noise, and which I used to measure this property of PRC2 target genes – is known to correlate with the level of gene expression, being generally increased for lowly expressed genes [253]. Establishing and validating a reliable metric for transcriptional heterogeneity will therefore be a worthwhile task of future studies, and the results presented in this chapter which rely on these existing metrics should be interpreted with caution.

Chapter 6

Investigating the epigenetic consequences of KDM3B haploinsufficiency in AML

6.1 Results

To investigate the epigenetic consequences of KDM3B haploinsufficiency in AML, we generated *KDM3B* heterozygous and homozygous deletions in the OCI-AML3 cell line using a CRISPR system (experimental work performed by Ashish Goyal) (Fig. 6.1). We confirmed by Western blot (performed by Fiona Brown-Burke) that the protein expression of KDM3B was reduced in the heterozygous deletion cells to approximately half of baseline levels, and was almost completely eradicated in the homozygous deletion cells. This was accompanied by a concomitant increase in H3K9me1 (Fig. 6.2A). My aim in this chapter was to assess whether the heterozygous deletion of *KDM3B* – which mimics the haploinsufficient expression levels observed in del(5q) AML patients – results in increased cell-to-cell epigenetic heterogeneity (see **Chapter 2: Discussion** for details on this hypothesis). For this we performed scATAC-seq, which would allow general comparisons of chromatin accessibility as well as heterogeneity between clones (experimental work performed by Fiona Brown-Burke, Oliver Mucke and Afzal Syed).

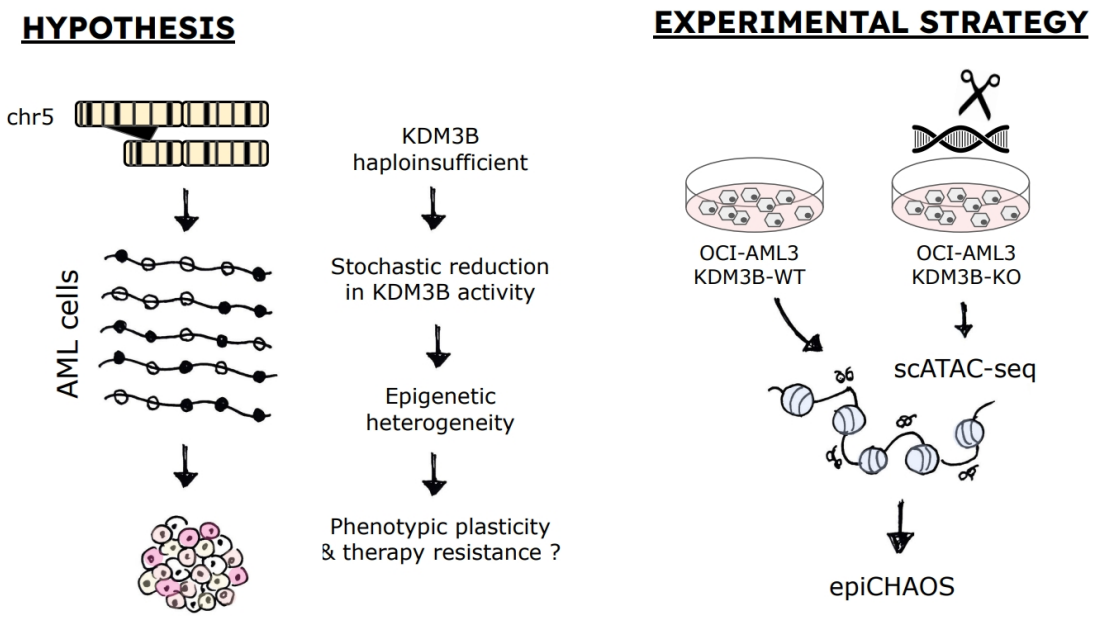

Figure 6.1: Hypothesis: the effects of KDM3B haploinsufficiency on epigenetic heterogeneity. Schematic describing the working hypothesis (left) and experimental strategy (right) used to investigate the effects of KDM3B haploinsufficiency in AML. I hypothesised that the heterozygous deletion of KDM3B would result in a haploinsufficient reduction of KDM3B protein levels and activity, such that a random subset of KDM3B targets may lose KDM3B activity in each cell. This may give rise to epigenetic heterogeneity between cells, and potentially contribute to phenotypic plasticity and therapy resistance. To investigate this, KDM3B-heterozygous deletion cells were generated from the OCI-AML3 cell line using CRISPR technology, and profiled by scATAC-seq. Using these data, epigenetic heterogeneity can be compared between samples by applying epiCHAOS.

6.1.1 Heterozygous deletion of *KDM3B* results in global chromatin compaction

Clustering of cells based on ATAC-seq signals did not reveal a clear separation of KDM3B-knockout (KO)/wild-type (WT) cells, but rather the majority of cells formed a gradient with differences in the proportion of KO/WT cells across clusters (Fig. 6.2B-C). Within this gradient, Cluster 5 was dominated by KO cells, while Cluster 3 was dominated by WT cells. One distinct cluster of cells also emerged – Cluster 1 – which was dominated by KO cells and separated from other WT and KO cells. This separation, and the fact that the cells in this cluster had mostly inaccessible chromatin, suggested that it likely represents a state of cell-cycle quiescence.

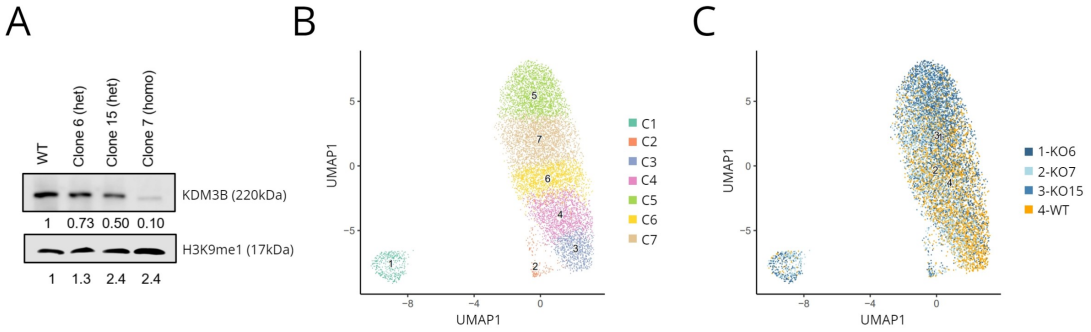

Figure 6.2: Heterozygous deletion of KDM3B in the OCI-AML3 cell line. A. Western blot (generated by Fiona Brown-Burke) showing KDM3B and H3K9me1 protein levels in KDM3B-heterozygous deletion (KO5, KO15), KDM3B-homozygous deletion (KO7) and WT OCI-AML3 cells. **B-C.** UMAP representations of the same cells profiled by scATAC-seq, coloured according to assigned cluster (**B**) and sample (**C**).

To understand the effects of KDM3B deletion I first inspected the differences in global chromatin accessibility between clones. I observed an overall compaction of chromatin in KDM3B-deletion cells, with the lowest accessibility in homozygous deletion cells, at regions associated with heterochromatin (Fig. 6.3). This pattern was observed genome-wide, and at heterochromatic regions as well as regions associated with H3K9me1/2, including H3K9me2-associated LADs (Fig. 6.3). Promoters and regions associated with more active chromatin also lost accessibility in KO cells but to a similar extent in homozygous and heterozygous deletion cells, in line with the idea that loss of one copy of KDM3B may be sufficient to affect transcriptional programs 6.3).

To better understand the differences in chromatin accessibility between states, I defined the set of differentially accessible peaks between the KO-dominated Cluster 5 and the remaining clusters (excluding the outlying Clusters 1 and 2 which likely reflect cell cycle effects). As expected, the majority of peaks showed lower accessibility in Cluster 5; only 53 peaks were preferentially accessible in Cluster 5, while 19,493 were preferentially accessible outside of Cluster 5. Nevertheless, I found a unique pattern of TF motif enrichment among

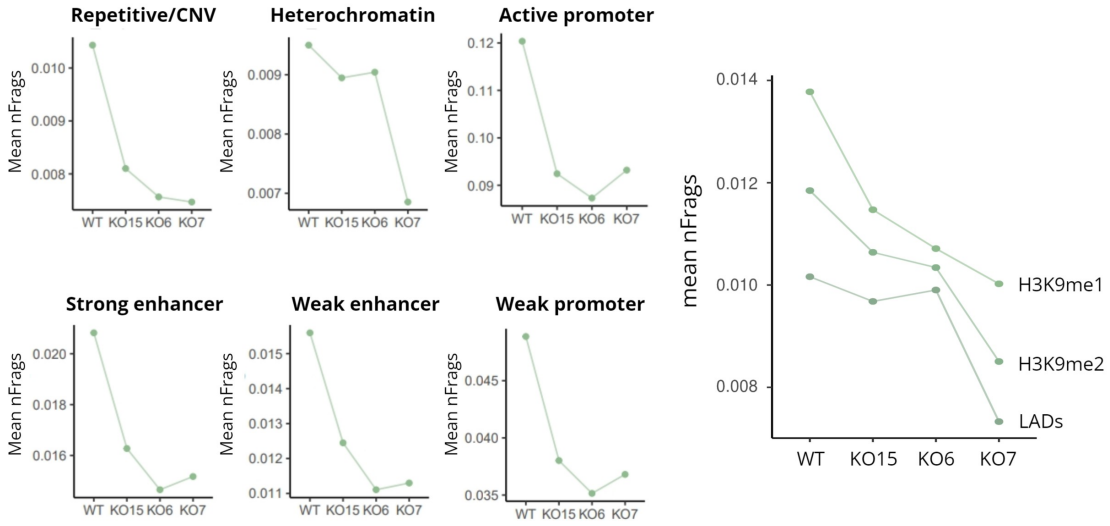

Figure 6.3: Heterozygous deletion of KDM3B results in global chromatin compaction. Line plots comparing the average chromatin accessibility at scATAC-seq regions overlapping with selected chromHMM states (left) and regions associated with H3K9me1/2 (right) in KDM3B-heterozygous deletion (KO5, KO15), KDM3B-homozygous deletion (KO7) and WT OCI-AML3 cells.

Cluster 5 marker peaks, including enrichment of CEBPB, GATA, SOX family members, and various homeobox TFs (Fig. 6.4A). Meanwhile the loss of chromatin accessibility in KO compared to WT cells appeared to be a genome-wide event, such that the top TF motif enrichments outside of Cluster 5 included ubiquitous factors such as YY1 and SP1 (Fig. 6.4B).

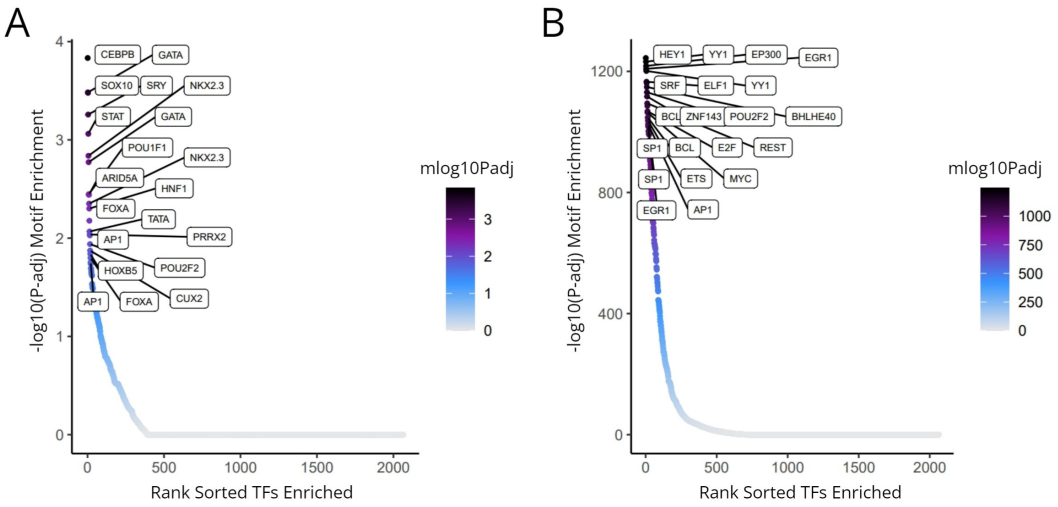

Figure 6.4: Transcription factor motif enrichments in KDM3B-KO cells. A-B. Transcription factor motif enrichments in (**A**) and outside of (**B**) Cluster 5 – the scATAC-seq cluster dominated by KDM3B-KO cells. Two outlying and potentially cell-cycle-related clusters (Clusters 1 and 2, see Fig. 6.2B) were excluded from the analysis. Motifs are ranked by enrichment score and coloured according to log10(adjusted p-value).

6.1.2 Heterozygous deletion of *KDM3B* results in cell-to-cell epigenetic heterogeneity at H3K9me1/2-associated regions

I next considered whether the heterozygous deletion of *KDM3B* could give rise to heterogeneity in epigenetic patterns from cell to cell. I applied epiCHAOS (see **Chapter 5** for details) to investigate this hypothesis. EpiCHAOS detected higher genome-wide epigenetic heterogeneity in both *KDM3B*-heterozygous clones compared to WT and homozygous deletion cells. This pattern was not evident at promoter-associated peaks, but rather at regions associated with H3K9me1 and H3K9me2, including LADs (Fig. 6.5).

In summary, these data suggest that the heterozygous loss of KDM3B in del(5q) AML can lead to accumulation of heterochromatic marks in a pattern that is heterogeneous from cell to cell.

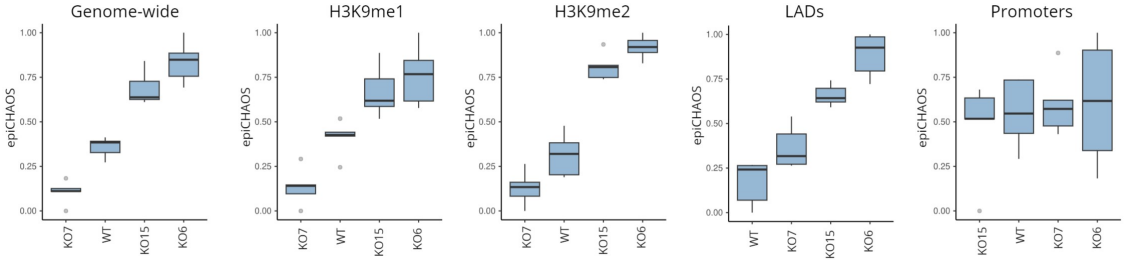

Figure 6.5: Heterozygous deletion of KDM3B results in cell-to-cell epigenetic heterogeneity. Boxplots comparing epiCHAOS scores (epiCHAOS) in KDM3B-heterozygous deletion (KO5, KO15), KDM3B-homozygous deletion (KO7) and WT OCI-AML3 cells which were profiled by scATAC-seq. The leftmost plot shows epiCHAOS scores computed across all peaks. In subsequent plots, peaks were subset for regions associated with H3K9me1/2, LADs and promoters, respectively. Boxes depict epiCHAOS scores computed in five random subsamples of 100 cells from each sample.

6.2 Discussion

In this chapter, I investigated the hypothesis that the heterozygous loss of the H3K9me1/2 demethylase KDM3B may give rise to cell-to-cell epigenetic heterogeneity in AML cells. The rationale behind this was that if the levels of the enzyme are not sufficient to occupy all of the regions which it normally targets, a random fraction of the genome may lose KDM3B activity in each cell. I hypothesised that the fraction of KDM3B-targeted regions would thereby be reduced in all cells, but that the specific distribution of targeted and non-targeted regions would differ from cell to cell, resulting in variable patterns of its associated histone marks. A similar phenomenon has been proposed in a 2019 study in breast cancer by Hinohara *et al.* [147]. There it was shown that overexpression of the H3K4 demethylases KDM5A/B results in transcriptional heterogeneity, and this was postulated as a mechanism of non-genetic resistance to endocrine therapies in breast cancer patients [147].

My data suggest that heterozygous deletion of the lysine demethylase, *KDM3B*, in the OCI-AML3 cell line results in increased genome-wide cell-to-cell epigenetic heterogeneity, which is detectable at the level of chromatin accessibility, and is especially apparent at regions associated with H3K9me1/2. Noting that these findings are preliminary and will require experimental repeats, they may provide a conceptual link between the heterozygous deletion of *KDM3B*, leukemic progression and therapy resistance in patients with del(5q) AML, that will need to be investigated in future studies.

Therapy resistance represents a major clinical hurdle in many cancer types including some of the more aggressive AML subgroups such as del(5q) AML [34]. Intra-tumor heterogeneity has been posited as a crucial driver of therapy resistance in cancer; however, historically, the primary focus in this domain has been on genetic heterogeneity – the existence of multiple mutationally distinct subpopulations of cells within a tumor [124]. Non-genetic heterogeneity has recently emerged as a compelling explanation for cancer plasticity and therapeutic tolerance in mutationally stable tumors [254]. As highlighted in a recent study by Nuno *et al.*, a significant portion of AML relapses do not involve changes in driver mutations, suggesting that epigenetic changes may be primarily responsible for therapy resistance and relapse in this disease [255].

Beyond epigenetic heterogeneity, I also observed a genome-wide reduction in chromatin accessibility in KDM3B depleted AML cells, likely reflecting a global accumulation of repressive histone marks. Such global chromatin compaction is itself an intriguing consequence of KDM3B haploinsufficiency, as it may have several downstream effects on genome stability, cell-cycle maintenance and DNA repair, that could also conceivably contribute to leukemic progression. For example, it has been shown that the strength of the DNA damage response can be impaired in the context of chromatin compaction [256], meaning that mutations and structural genomic alterations may be more likely to accumulate in cells with abnormally compact chromatin.

Interestingly, despite this overall compaction of chromatin, a small number of peaks gained accessibility in KDM3B-depleted cells. Among these, GATA transcription factor motifs were among the most enriched. GATA is a critical regulator of erythroid differentiation [257], and GATA binding sites were also found to be enriched in LSC-specific open chromatin compared to that of non-self-renewing blasts [258]. Since del(5q) AML is known to display an erythroid-biased differentiation state [259], and elevated expression of LSC markers [165], it would be interesting to further investigate whether these differences in chromatin accessibility could reflect a shift towards a more LSC-like or erythroid-like cell state.

Several outstanding questions remain regarding the role of KDM3B in AML biology, leaving significant scope for future studies. As a starting point, it will be important to assess the changes in the levels and patterns of H3K9me1/2 and other interconnected histone marks that result from KDM3B deletion, and it may be possible to assess epigenetic heterogeneity of these specific histone marks at the single cell level using techniques such as single-cell ChIP-seq. Furthermore, the consequences of this epigenetic heterogeneity have yet to be explored – most notably, whether this heterogeneity leads to increased resistance to AML therapies, as is observed in del(5q) patients. It will also be interesting to further explore the link between global chromatin compaction and genome instability in the context of del(5q) AML, where the depletion of KDM3B potentially precedes the acquisition of TP53 mutations and subsequent accumulation of CNAs.

Chapter 7

Conclusions & Future Perspectives

The work presented within this thesis lays important foundations for understanding the order and chaos of the epigenome in cancer cells.

First, my findings offer a novel perspective on the mechanisms driving one of the most aggressive forms of acute leukemia in adults; del(5q) AML. Through a comprehensive DNA methylation-based characterisation, I have shown that del(5q) AML is an epigenetically distinct subgroup, and I have provided compelling evidence that haploinsuffiency of KDM3B may be a critical event in this disease.

The role of KDM3B in AML remains relatively unchartered territory. Within the scope of this thesis, I have explored one hypothesis that might link haploinsufficiency of KDM3B to leukemic progression. My findings suggest that heterozygous deletion of KDM3B can promote epigenetic heterogeneity in chromatin accessibility at the single cell level – a phenomenon which could have important implications for leukemic evolution and therapy resistance. These data are, however, preliminary, and will require confirmation in a repeated experiment. Moreover, the downstream consequences of this epigenetic heterogeneity will deserve to be investigated in future studies. There is also much scope for investigation into how the heterozygous deletion of this gene may contribute to leukemogenesis in other ways. Future studies might be directed to teasing apart the potential influence of KDM3B depletion on genomic instability, TP53 mutation selection, and erythroid bias, which are typical features of del(5q) AML.

It is important to recognise that del(5q) is only one of the recurrent cancer CNAs whose target genes and underlying mechanisms have not yet been elucidated. This might raise speculation as to whether other recurrent CNAs in other cancer types might similarly target haploinsufficient epigenetic regulators that were overlooked in the search for classical oncogenes and tumor suppressors. A study by Zach *et al.* has shown for example that genomic regions of peak amplifications which recur in a pan-cancer context, but which do

not contain already known oncogenes, are strikingly enriched for genes involved in histone modification [260]. The findings and strategies presented here might therefore inspire a pan-cancer inquiry into haploinsufficient epigenetic regulators, and further efforts to understand the recurrent disruption of histone modifying enzymes in tumor cells.

The second major contribution of this doctoral thesis was in presenting a novel strategy for studying epigenetic heterogeneity at the single-cell level. With the recent expansion of single-cell epigenomics studies in cancer, and the growing interest in concepts of non-genetic cellular heterogeneity and plasticity in this domain, epiCHAOS could offer a valuable tool towards unravelling the role of epigenetic heterogeneity in tumor evolution.

Chapter 8

Materials and Methods

8.1 DNA methylation analyses

8.1.1 Preprocessing of EPIC array data from the ASTRAL-1 cohort

MethylationEPIC array data from AML 477 samples (after removal of three samples due to quality concerns) were obtained from patients from the ASTRAL-1 study [14, 154]. The cohort was enriched for elderly patients, with a median age of 77 (range 59-94 years). Clinical and molecular annotations (age, sex, karyotype and mutation status for a range of epigenetic regulators) for these samples were obtained from clinical collaborators prior to this study. EPIC array profiling and calling of CNAs to determine ckAML status was performed by Anna Riedel (DKFZ, Heidelberg) prior to this study [14].

Raw idat files were used to perform quality control, preprocessing and normalisation of the DNA methylation data using the RnBeads R package [261]. Filtering was applied to remove single nucleotide polymorphism (SNP)-overlapping probes, cross-reactive probes, sites outside of CpG context and CpGs mapping to sex chromosomes, as well as probes with detection p-value > 0.05 and sites covered by fewer than three beads. Beta-values were normalised and background subtraction carried out using the "scaling.internal" and "sesame.noobsb" methods, respectively.

8.1.2 Methylome deconvolution

Methylome deconvolution was performed using the MeDeCom R package [144] according to an established pipeline [156]. MeDeCom is a reference-free deconvolution method based on constrained non-negative matrix factorisation, which aims to decompose of a set of bulk methylomes to recover their constituent LMCs and the proportions of LMCs within each patient sample. For deconvolution, the 20,000 most variable CpG sites across patients were selected, in order to reduce computational demands while retaining the most informative

features in the dataset. Potential confounding factors including age, sex, batch, Sentrix ID and Sentrix position were adjusted for using ICA, implemented within the MeDeCom package. Deconvolution was performed using a range of K values (number of LMCs) from 2:15. K = 11 was selected as the optimum number of LMCs based on the cross-validation error. The regularisation parameter lambda = 0.01 was selected (after testing values of 0, 0.1, 0.01, 0.001 and 0.0001) in order to minimise the cross validation error.

8.1.3 Estimating sample purity

Sample purity was estimated from the methylation data using the InfiniumPurify method [262] InfiniumPurify infers tumor purity in a cancer-type-specific manner based on a kernel density estimation method, using a predefined set of CpG sites that are found to be differentially methylated between tumor (in this case, AML) and corresponding normal tissue. Samples with low estimated purity (lower tertile) were excluded from downstream analyses to avoid spurious interpretations due to the apparent abundance of lymphoid cells in some samples.

8.1.4 Estimation of LMC proportions in external datasets

To validate our interpretations of LMCs, the factorise regression function from the MeDeCom R package was used to derive an estimate of the original LMC proportions in two publicly available datasets comprising HumanMethylationEPIC/450K data from the TCGA and BEAT-OSU AML cohorts. Clinical and molecular annotations for TCGA-AML were retrieved from the Genomic Data Commons using the TCGAbiolinks package, and for BEAT AML were obtained from the supplementary tables of Tyner *et al.* [263].

8.1.5 Prediction of Leukemic Cell of Origin

A predicted cell of origin was assigned to each AML sample according to previously established methods [159]. HumanMethylation450K data from sorted hematopoietic progenitor cell states from healthy donors, including HSCs, MPPs, LMPPs, CMPs, GMPs and MEPs, was used as a reference. A set of 216 differentially methylated regions were identified by Jung *et a*l. from pairwise comparison between the differentiation states [159]. Hierarchical clustering of AML together with normal progenitor cell samples, on the methylation of CpGs sites within these regions, resulted in three groups of GMP-like (more differentiated), MEP/CMP-like and MPP/LMPP-like (less differentiated) AMLs. LMC proportions were compared between these three groups by Wilcoxon's test.

8.1.6 Biological interpretation of LMCs

To identify LMCs derived from non-leukemic cell types, LMC proportions were correlated against methylomes of normal hematopoietic cells using EPIC array data from Salas *et al.* [157] For further interpretation, sets of LMC-specific hypomethylated and hypermethylated CpG sites were defined for each LMC as those with a methylation beta value > 0.5

above or below the mean of the remaining LMCs. Gene ontology and transcription factor binding site enrichment analyses for the resulting CpGs were performed using the clusterProfiler [264] and genomic locus overlap enrichment analysis (LOLA) [265] R packages. To identify LMCs associated with mutations or cytogenetic subgroups, LMC proportions were compared between groups using Wilcoxon tests.

8.1.7 Defining LMC-based subgroups

To define LMC-based subgroups, consensus k-means clustering [163] was applied to the matrix of LMC proportions, excluding the lymphoid-like and neutrophil-like components, LMC1 and LMC9. Euclidean distance and complete inner linkage were used for clustering. K = 9 was selected as the optimum number of clusters based on inspection of the CDF and delta under the CDF curves.

8.1.8 Differential methylation analysis

A set of CpG sites which are differentially methylated in LMC3-high AML was defined by comparing the methylomes of the LMC3-high AML cluster (from consensus clustering) to (i) all remaining AML clusters, and (ii) normal HSCs. By this strategy the analysis was no longer restricted to those 20,000 CpG sites used for deconvolution, and could consider methylation changes from a normal state as well as between malignant states. Differential methylation between groups was computed at the level of CpG sites using RnBeads. CpG sites with absolute mean beta value difference > 0.2 and False Discovery Rate (FDR) adjusted p-value < 0.05 were considered differentially methylated. CpG sites which were consistently hypo/hypermethylated in both comparisons were taken forward for enrichment analyses. Among differentially methylated sites, enrichments of transcription factor binding sites and gene ontologies were calculated using the LOLA [265] and clusterProfiler [264] R packages, respectively. The list of homeobox genes was retrieved from Wilming *et al.* [266] and tested for enrichment among LMC3 hypermethylated CpGs using the hypergeometric test.

8.1.9 Comparison of DNA methylation in del(5q) LSCs and blasts

To compare LMC3 levels in del(5q) LSCs vs blasts, 450K methylation data from sorted LSCs (CD34+CD38+/CD38-) and blasts (CD34-) was downloaded from the (GSE63409) [159]. Conumee copy number profiles were generated to identify samples with 5q deletions. LMC proportions were estimated using Medecom's factor regression approach.

8.1.10 DNA methylation variation

For comparisons with epiCHAOS scores, DNA methylation variation at different genomic regions was computed in HSCs. Data from Adelman *et al.* [267] were used to calculate DNA methylation variability by computing variance per CpG site in HSC-enriched lineage-negative (Lin- CD34+ CD38-) samples across the eight male donors. A maximum quantile

threshold of 0.005 was established for missing values per site. Any sites that surpassed this threshold were removed (analysis performed by Martina Braun). For each ENCODE TFBS, the average of variances was calculated for all CpG sites overlapping with the respective regions.

8.2 Gene expression analysis

Differential gene expression analysis using RNA-seq datasets for the ASTRAL-1 ckAML and TCGA cohorts was performed using DEseq2 with default parameters [268]. Pan-cancer and pan-tissue gene expression analyses for *KDM3B* and *DNMT3B* were performed using TCGA and GTEx data from the gene expression profiling interactive analysis (GEPIA) portal [177]. To investigate pathway-level gene expression changes associated with the del(5q) methylation signature, GSVA was performed using the GSVA R package [168] and the resulting pathway-level expression scores were tested for correlation with LMC3 proportion. Stemness scores were computed for each patient using the weighted sum of expression of the 17 LSC17 genes as defined by Ng *et al.* [31].

To compare gene expression in LSC+ and LSC- blasts, microarray data were downloaded from Ng *et al.* [31] (GSE76009). Differential gene expression of microarray data was computed using the Limma R package [269]. The same dataset was used for comparisons of *KDM3B* gene expression between LSC and blast counterparts of del(5q) samples. Del(5q) status for the samples was obtained from the corresponding authors.

8.3 Copy number and mutational analyses

8.3.1 Mutual exclusivity of mutations & CNA patterns

Mutual exclusivity of del(5q) and *IDH* mutations was determined based on the Poisson-Binomial distribution as implemented in the Rediscover R package [270].

8.3.2 Investigation of the 5q minimally deleted region

Copy number profiles were generated from EPIC array data with the Conumee R package using 20 AML samples of known normal karyotype samples as "flat genome" controls. The minimally deleted region was defined from the ASTRAL-1 AML cohort using the Conumee segmentation results. In each sample, segments mapping to chromosome 5q were selected which had a p-value < 0.05 for loss at that region. The frequency of loss in 10KB bins was calculated and the minimal genomic interval having the highest frequency of loss was determined. This definition was compared to that of six studies of del(5q) AML or high-risk MDS identified from a literature search, which each proposed a variation on the minimal interval of interest, largely overlapping with our own definition. For all downstream analyses of MDR genes, the 50 genes flanked by *Il9* and *UBE2D2* were

considered as an accepted and conservative definition of the MDR [43, 44]. The expression of MDR genes was correlated against LMC3 proportions using Pearson correlation.

8.3.3 Assessment of del(5q) mutations

The mutation status of 5q MDR genes in the TCGA AML cohort was retrieved using the TCGAretriever package [271]. Single nucleotide variants were called in whole genome sequencing data from ASTRAL-1 patient samples using mutect2 [272] (analysis performed by Dr. Etienne Sollier, DKFZ, Heidelberg).

8.4 Protein expression analysis

Protein expression of MDR genes was examined using proteome data from the CCLE [176]. AML cell lines were selected using the DepMap portal [273] and identified as del(5q) or otherwise. Protein expression was compared between groups by Wilcoxon's test, excluding outliers from the analysis if their expression was > 3 standard deviation from the mean of all AML cell lines. Tandem mass tag (TMT) proteome data for 44 AML patients was retrieved from Kramer *et al.* [175]. Abundance values were calculated from reporter-ion intensity, log2 transformed and median centered at 0. A second mass-spectrometry-based proteome dataset of 177 AML samples including 3 del(5q) samples was obtained from Jajavelu *et al.* [274]. KDM3B protein expression was compared between del(5q) and other AML patient samples by Wilcoxon's tests.

8.5 Clinical data analyses

Clinical data from the BEAT AML cohort were used to investigate the prognostic relevance of LMCs. Hazard ratios and p-values were calculated using cox proportional hazard models to test the association of each LMC with overall survival. To further assess the prognostic value of LMC3, samples from this cohort were separated into LMC3-high and LMC3-low groups based on the mean LMC3 level, and the LMC3-high group was further subdivided into del(5q)-positive and del(5q)-negative patients. Conumee copy number profiles were generated and used to reliably define samples as del(5q) or otherwise. Pairwise Log rank test p-values were computed using the survival and survminer R packages.

Ex vivo drug sensitivity data for 122 small molecule inhibitors was obtained from the BEAT AML cohort [263]. To compare global drug sensitivities between patient subgroups, binary sensitive or resistant calls were assigned for each sample to each drug, considering samples with the highest 20% of AUC values for a given drug as sensitive. Then for each sample, the proportion of drug sensitivities was quantified, and these proportions were compared between LMC3-high [+/-del(5q)] and LMC3-low subgroups.

8.6 Single-cell epigenome and transcriptome analyses

8.6.1 Calculating epiCHAOS scores

EpiCHAOS heterogeneity scores were calculated for each given cluster or otherwise defined group of cells using a binarised matrix of scATAC or other single-cell epigenomics data. First, pairwise distances were calculated between all cells within the cluster using a chance-centered version of the Jaccard index. The Jaccard index is defined as the intersection between two sets (in this scenario, the number of matching 1's or matching 0's between a pair of cells/columns) divided by the size of their union (in this scenario the total number of measured loci/rows). Chance centering was introduced to control for differences in the relative number of ones and zeros [224]. Afterwards, the mean of all pairwise distances per cluster was computed. To remove any further effect of sparsity a linear regression model of the raw heterogeneity scores was fitted against the total number of detected accessible peaks (reflected by the total number of 1's), averaged across cells in the respective cluster, and the residuals of this model were taken as the adjusted epiCHAOS scores. Finally, the scores were transformed to an interval of 0-1, and subtracted from 1 to convert the similarity metric to a distance metric.

ScATAC-seq data from the Hep-1 liver cancer cell line [227] was used to test whether epiCHAOS scores correlate with measures of technical noise. Cells were stratified into bins (20 bins of 100 cells each) based on various quality control metrics: FRIP scores, TSS enrichment scores and nucleosome ratio, which were calculated using ArchR [275].

DNA CNAs were inferred in the Hep-1 cell line using epiAneuFinder [276] with a windowSize=100,000 and minFrags=20,000. To investigate the influence of CNAs on epiCHAOS scores, the most prominent examples of large subclonal copy number gains (gain on chromosome 5) and deletions (deletion on chromosome 13) were selected by visual inspection, and cells were stratified based on the presence or absence of each alteration. EpiCHAOS scores were calculated across peaks in the affected chromosome and compared between cells with diploid or alternative states. To correct for CNAs when applying epiCHAOS to cancer datasets, a per-chromosome count-corrected epiCHAOS score was derived, where epiCHAOS scores were calculated per chromosome, implementing a linear regression-based adjustment for the total number of 1's on that chromosome, and then the average of per-chromosome scores were taken as the global epiCHAOS scores.

Unless otherwise specified, epiCHAOS was calculated using the entire peaks-by-cells matrix. To allow a more robust comparison between groups, epiCHAOS scores were calculated on five random subsamples of 100 cells from each group/cluster, except in groups/-clusters which contained fewer than 100 cells. Since the scNMT-seq data contained fewer than 100 cells in most groups, epiCHAOS scores were calculated only once for each cell type. ENCODE TFBS regions from the LOLA core database [265] were used for comparisons of heterogeneity at different genomic regions, for which scATAC peaks matrices were

subsetted to obtain peaks overlapping with each genomic region. Similarly, for comparisons across gene sets, data were subsetted for peaks overlapping with promoters of each gene set using the gene ontology ´´biological process" (GO:BP) gene sets from MsigDB [277].

8.6.2 Generating synthetic datasets with controlled heterogeneity

To test the performance of epiCHAOS, synthetic datasets were generated *in silico* in a way that simulates the structure of binarised scATAC-seq peak matrices. First a series of 100 synthetic datasets with controlled heterogeneity was created, in which each dataset has an equal total count. To do this a random binary matrix was created, which would represent the first dataset in the series. In each subsequent dataset, homogeneity was incrementally introduced by removing a set number of 1's from selected n rows, and adding them to a different selected n rows, in such a way that a constant number of 1's is maintained, while heterogeneity decreases.

To test situations where the genome-wide chromatin accessibility is increasing or decreasing, binarised data from an example scATAC-seq dataset were perturbed to create datasets of increasing heterogeneity with either addition or removal of 1's. Specifically, 10, 20, 30, 40 and 50% of 1's were selected at random and replaced by 0's, and corresponding numbers of 0's were selected at random and replaced by 1's.

To test that epiCHAOS is not influenced by differences in sparsity, a series of 100 random binary datasets was generated with each dataset having equal dimensions and incrementally increasing total number of 1's. Their epiCHAOS scores were then computed and tested for a correlation with their total count.

As an additional validation approach semi-synthetic scATAC-seq datasets were created by mixing data from distinct cell types. Using scATAC-seq data from human bone marrow [226] five cell types were selected; HSCs, Monocytes, CD8-CM T cells, B-cells and plasmacytoid dendritic cell (DC)s. The top 500 differentially accessible peaks for each cell type were identified and used to create *in silico* mixtures of two to five cell types in all possible combinations.

8.6.3 Simulating scATAC-seq data with varying sequencing depth

The scReadSim package was used to simulate scATAC-seq data of varying sequencing depths [225]. A subset of scATAC-seq data from HSCs from the Granja *et al.* dataset was used as input [226]. Data were reduced to 10,000 randomly selected peaks for ease of processing. Simulated scATAC-seq matrices comprising each 500 cells were generated with sequencing depth ranging from 50,000 to 100,000 counts, in increments of 10,000. EpiCHAOS scores were calculated across matrices on five subsamples of 100 cells from each condition.

8.6.4 scATAC-seq data processing and analysis

Publicly available scATAC-seq datasets for human hematopoiesis [226], mouse gastrulation [229], drosophila embryogenesis [231], breast cancer [140], liver cancer [234], ependymoma [236], HSC aging [239] and liver cancer cell lines [227]were downloaded from the respective publications. For analyses in developmental datasets and in ependymoma, processed counts matrices were used as provided by the authors, where cell types were previously annotated. For breast and liver cancer datasets, fragments files were downloaded, processed and analyzed using ArchR [275]. Cells with TSS enrichment scores less than 4 or number of fragments higher than 1,000 were removed, and doublets were filtered out using default parameters. Iterative Latent Semantic Indexing (LSI) was performed followed by clustering using the Seurat method. Gene score matrices were generated using ArchR and used for subsetting cancer datasets for epithelial cells based on inspection of epithelial cellular adhesion molecule (EPCAM) scores. After reclustering epithelial cells, peak calling was performed using macs2 [278]. To assign gene set/pathway scores to each cluster, gene set annotations were obtained from MSigDB using the msigdbr R package [277]. Gene scores were first averaged across all cells within each cluster, and then the mean score of all genes within a given gene set was computed to assign gene set scores per cluster.

8.6.5 Differential heterogeneity analysis

Differential heterogeneity analyses were performed for each region using a permutation approach, whereby the difference in epiCHAOS scores between two selected cell types were compared with that between pairs of 1,000 randomly computed groups sampled from the same pool of cells. P-values were calculated as the quantile of the distribution of sampled permutations for which the difference in heterogeneity scores was greater than that between the two test groups.

8.6.6 scRNA-seq analyses

scRNAseq datasets for human hematopoiesis [226], mouse gastrulation [229] and drosophila embryogenesis [231]were downloaded from the respective publications and analyzed using Seurat. Cell-to-cell transcriptional heterogeneity was calculated by computing pairwise euclidean distances according to the methods of Hinohara *et al.* [147]. Developmental potential was calculated per cell using CytoTRACE [233] and assigned as a mean per celltype for downstream analyses. Transcriptional noise per gene was estimated using the coefficient of variation as previously described [241]. A list of PRC2-target genes used for comparison of transcriptional noise was obtained from Ben-Porath *et al.* [279].

8.6.7 scChIP-seq analysis

ScChIP-seq counts matrices representing 50kb non-overlapping bins of H3K27me3 from human breast cancer PDX cells that were sensitive or resistant to Capecitibine (HBCx-95 and HBCx-95-CapaR) were downloaded from GSE117309 and processed as described

by Grosselin *et al.* [138]. Cells having a total number of unique reads in the upper percentile were removed as outliers, and genomic regions not represented in at least 1% of all cells were filtered out. Data corresponding to non-standard chromosomes and the Y chromosome were excluded. Cells with a total number of unique reads less than 1,600 were removed. Counts matrices were binarised and cells from each condition were subsampled to select ten groups of 100 cells each for epiCHAOS calculation.

8.6.8 scNMT-seq and scTAM-seq analysis

ScNMT-seq DNA methylation and ATAC data from mouse gastrulation [243], summarised across promoters, gene bodies and CpG islands, were accessed using the SingleCellMultiModal R package. ScTAM-seq data from mouse hematopoiesis were obtained from Scherer *et al.* [244], downloaded from https://figshare.com/ndownloader/files/42479346, and analyzed using Seurat.

8.6.9 scATAC-seq data analysis from KDM3B deletion cells

ScATAC-seq data were processed using cellranger-atac and analyzed in R using the ArchR package [275]. Cells called as doublets or with nFrags < 1000 or TSS enrichment score < 4 were filtered out to remove low quality cells. Clusters were called using the Seurat method after Iterative LSI. Peaks were called using macs2 version 2.1.2.1. Differentially accessible peaks were identified between the KO dominated Cluster 5 and all other clusters, excluding the likely cell cycle-related Clusters 1 and 2. Motif enrichment analysis was performed in ArchR using the ENCODE motif set, with FDR $<= 0.05$ & Log2FC $>= 1$ selected as cutoffs to define marker peaks. Genomic annotations for LADs were obtained from Guelen *et al.* [280], ChIP-seq peaks for H3K9me2 (K562 cells) from Salzberg *et al.* [217], and H3K9me1 (CD133+ cells) from Cui *et al.* [281].

Contributions

8.6.10 Publications

First author manuscripts & preprints

1. **Katherine Kelly,** Michael Scherer, Martina Maria Braun, Pavlo Lutsik, Christoph Plass, *EpiCHAOS: a metric to quantify epigenomic heterogeneity in single-cell data,* Genome Biol 25, 305 (2024). doi: https://doi.org/10.1186/s13059-024-03446-w

2. **Katherine Kelly,** Linda Welte, Etienne Sollier, Anna Riedel, Fiona Brown-Burke, Michael Scherer, Harold N. Keer, Mohammad Azab, Ekaterina Jahn, Hartmut Döhner, Konstanze Döhner, Pavlo Lutsik, Christoph Plass, *Acute Myeloid Leukemia with deletion 5q is an epigenetically distinct subgroup defined by heterozygous loss of KDM3B,* bioRxiv, 2024.11.13.623380 doi: https://doi.org/10.1101/2024.11.13.623380

Additional manuscript contributions

1. Etienne Sollier, Anna Riedel, Umut H. Toprak, Justyna A. Wierzbinska, Dieter Weichenhan, Jan Philipp Schmid, Mariam Hakobyan, Aurore Touzart, Ekaterina Jahn, Binje Vick, Fiona Brown-Burke, **Katherine Kelly**, Simge Kelekci et al., *Pyjacker identifies enhancer hijacking events in acute myeloid leukemia including MNX1 activation via deletion 7q,* bioRxiv, 2024.09.11.611224 doi: https://doi.org/10.1101/2024.09.11.611224

2. Mulet-Lazaro, Roger, Stanley van Herk, Margit Nuetzel, Aniko Sijs-Szabo, Noelia Díaz, **Katherine Kelly**, Claudia Erpelinck-Verschueren et al, *Epigenetic alterations affecting hematopoietic regulatory networks as drivers of mixed myeloid/lymphoid leukemia,* Nature Communications 15, no. 1 (2024): 5693. doi: https://doi.org/10.1038/s41467-024-49811-y

3. Julian List, Etienne Sollier, Fiona Brown-Burke, **Katherine Kelly**, Dietmar Pfeifer, Valeria Shlyakhto, Kristina Maas-Bauer, Milena Pantic, Christoph Plass, Michael Lübbert, *Genocopy of EVI1-AML with paraneoplastic diabetes insipidus: PRDM16*

overexpression by t(1;2)(p36;p21) and enhancer hijacking, Br J Haematol. 2024; 00: 1–5. doi: http://doi.org/10.1111/bjh.19922

8.6.11 Poster presentations

- DNA methylation-based chatacterisation of Acute Myeloid Leukemia points towards KDM3B as a del(5q) haploinsufficiency candidate. *EHA2023. Frankfurt, Germany, 09.06.2023*

- DNA methylation-based chatacterisation of Acute Myeloid Leukemia points towards KDM3B as a del(5q) haploinsufficiency candidate. *VIB conference "Tumor Heterogeneity, Plasticity, Therapy" Leuven, Belgium, 04.10.2023*

- EpiCHAOS: a metric to quantify epigenomic heterogeneity in single-cell data. *ECCB2024. Turku, Finland, 18-19.09.2024*

Appendix

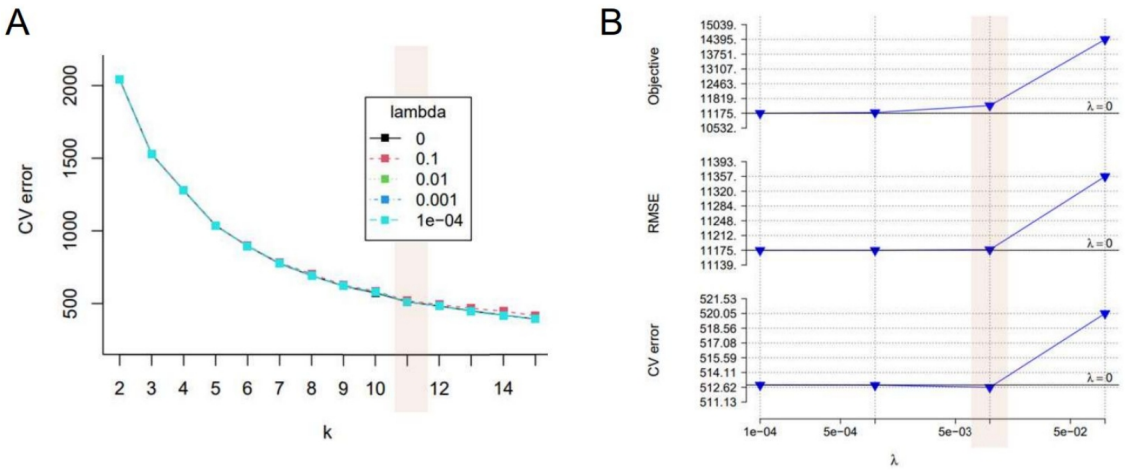

Figure 8.1: Selection of MeDeCom parameters. A. Plot of cross validation (CV) error used for K selection over 2-15 LMCs. The selected K value is highlighted. **B.** Selection of regularisation parameter Lambda (log scale) over a range of 0 to 0.1. The selected Lambda value is highlighted.

Table 8.1: Top 10 Gene Ontology enrichments among LMC1 hypomethylated CpG sites

ID	Description	qvalue
GO:0036230	granulocyte activation	7.89E-06
GO:0042119	neutrophil activation	7.89E-06
GO:0043312	neutrophil degranulation	2.10E-05
GO:0002446	neutrophil mediated immunity	2.10E-05
GO:0002283	neutrophil activation involved in immune response	2.10E-05
GO:0045785	positive regulation of cell adhesion	0.007284
GO:0051258	protein polymerization	0.008722
GO:0010506	regulation of autophagy	0.025752
GO:0051056	regulation of small GTPase mediated signal transduction	0.025752
GO:0016050	vesicle organization	0.025752

Table 8.2: Top 10 Gene Ontology enrichments among LMC9 hypomethylated CpG sites

ID	Description	qvalue
GO:0042110	T cell activation	4.67E-11
GO:1903131	mononuclear cell differentiation	7.71E-09
GO:0002768	immune response-regulating cell surface receptor signaling pathway	7.71E-09
GO:0002764	immune response-regulating signaling pathway	7.71E-09
GO:0030098	lymphocyte differentiation	7.88E-09
GO:0050851	antigen receptor-mediated signaling pathway	9.82E-09
GO:0002429	immune response-activating cell surface receptor signaling pathway	2.05E-08
GO:0002757	immune response-activating signal transduction	2.05E-08
GO:0050854	regulation of antigen receptor-mediated signaling pathway	1.03E-07
GO:1903039	positive regulation of leukocyte cell-cell adhesion	1.32E-07

Table 8.3: Top 20 LOLA enrichments among LMC10 hypomethylated CpG sites

collection	cellType	antibody	qValue
codex	Macrophage	CEBPB	0
encode	IMR90	CEBPB	6.57E-234
encode	MCF10A-Er-Src	c-Fos	5.72E-225
encode	MCF10A-Er-Src	c-Fos	9.14E-221
encode	MCF10A-Er-Src	c-Fos	1.06E-215
encode	MCF10A-Er-Src	c-Fos	2.26E-204
encode	MCF10A-Er-Src	STAT3	2.77E-198
codex	Monocyte	SPI1	8.04E-189
encode	HeLa-S3	CEBPB	3.88E-179
cistrome	HepG2 liver cells	CEBPa	3.49E-178
encode	A549	CEBPB	2.45E-177
encode	MCF10A-Er-Src	STAT3	1.22E-176
codex	Macrophage	SPI1	2.39E-163
encode	HepG2	CEBPB	8.36E-163
encode	MCF10A-Er-Src	STAT3	2.16E-162
encode	MCF10A-Er-Src	STAT3	1.16E-159
encode	HeLa-S3	p300(SC-584)	1.20E-129
encode	GM12891	PU.1	2.66E-122
encode	GM12878	RUNX3(SC-101553)	2.57E-118
cistrome	normal liver cells	CEBPa	1.74E-117

Table 8.4: Top 20 LOLA enrichments among LMC2 hypermethylated CpG sites

collection	cellType	antibody	qValue
codex	Acute Myeloid Leukemia	RUNX1	1.49E-48
codex	T-Cell	MYB	1.06E-26
codex	Hematopoietic Stem and Progenitor Cells	TCF7L2	1.10E-25
codex	Lymphoma cell	RUNX1	1.10E-25
codex	T-Cell	TAL1	1.50E-21
codex	Leukaemia cell	TCF3	3.15E-18
codex	Acute Myeloid Leukemia	RUNX1	7.59E-18
codex	T-Cell	MYB	3.01E-17
codex	B-cells	RUNX3	9.96E-17
encode	GM12878	RUNX3(SC-101553)	1.32E-16
codex	T-Cell	MYB	2.85E-13
codex	Leukaemia cell	RUNX1	7.96E-13
codex	Leukaemia cell	RUNX1	1.70E-12
codex	Leukaemia cell	CBFB	3.25E-12
codex	Hematopoietic Stem and Progenitor Cells	GATA2	9.49E-11
codex	T-Cell	MYB	1.22E-10
codex	T-cell acute lymphoblastic leukemia cells	NOTCH1	1.96E-10
codex	Acute Myeloid Leukemia	RUNX1	2.19E-09
codex	Hematopoietic Stem and Progenitor Cells	FLI1	3.37E-09
codex	T-Cell	MYB	3.29E-08

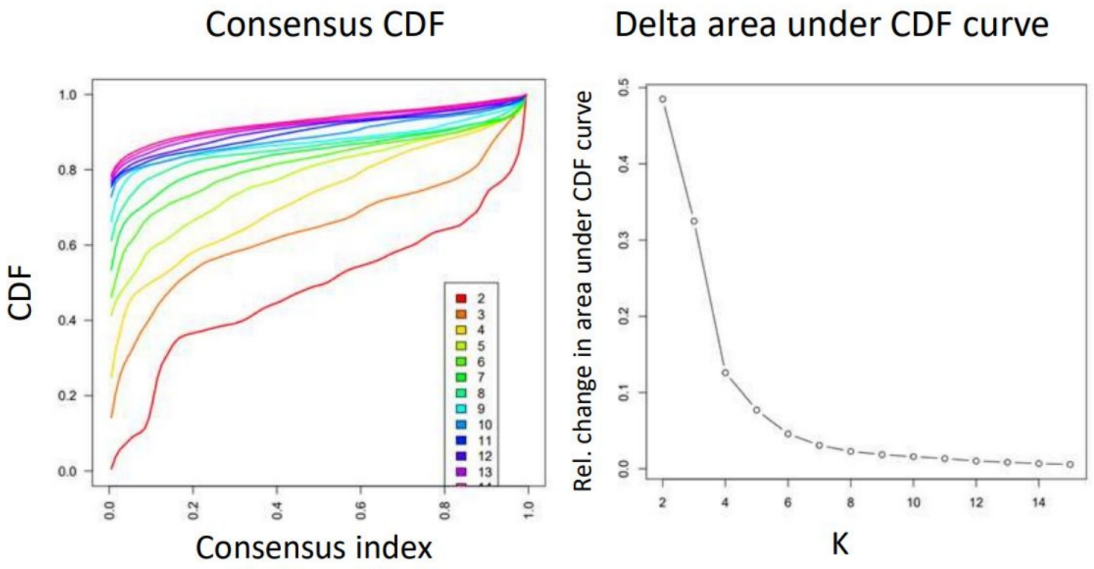

Figure 8.2: Selection of K for consensus clustering. Plots of the CDF and area under the CDF curve, used for K selection in consensus clustering.

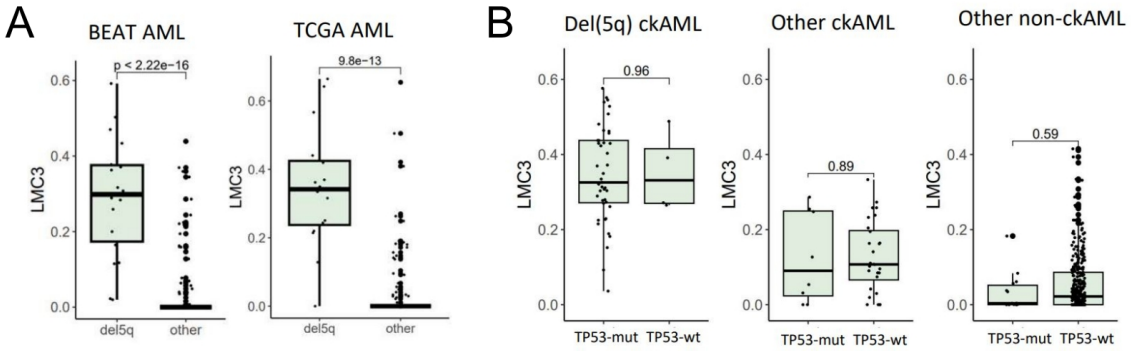

Figure 8.3: Del(5q) AML is defined by a unique DNA methylation signature. **A.** Boxplots comparing LMC3 proportion in del(5q) and 5q-retained AML samples from BEAT and TCGA AML cohorts. Wilcoxon's p-values shown. **B.** Boxplots comparing LMC3 proportion in TP53 mutated and wildtype AML, among del(5q) ckAML cases (left), non-del(5q) ckAML cases (middle), and non-del(5q), non-ckAML cases (right). Wilcoxon's p-values shown.

Table 8.5: Top 20 Gene Ontology enrichments among CpG sites hypermethylated in the LMC3-high AML subgroup

ID	Description	qvalue
GO:0007389	pattern specification process	5.68E-22
GO:0003002	regionalization	1.42E-21
GO:0048568	embryonic organ development	1.39E-17
GO:0048562	embryonic organ morphogenesis	9.62E-17
GO:0001501	skeletal system development	1.60E-16
GO:0045165	cell fate commitment	1.74E-14
GO:0048706	embryonic skeletal system development	3.93E-14
GO:0048705	skeletal system morphogenesis	4.86E-14
GO:0009952	anterior-posterior pattern specification	1.37E-13
GO:0048704	embryonic skeletal system morphogenesis	6.48E-12
GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules	3.22E-11
GO:0001708	cell fate specification	3.47E-11
GO:0021953	central nervous system neuron differentiation	3.47E-11
GO:0048736	appendage development	1.39E-10
GO:0060173	limb development	1.39E-10
GO:0035107	appendage morphogenesis	2.44E-10
GO:0035108	limb morphogenesis	2.44E-10
GO:0030900	forebrain development	5.23E-10
GO:0090596	sensory organ morphogenesis	2.18E-09
GO:0030326	embryonic limb morphogenesis	2.18E-09

Table 8.6: Top 20 Gene Ontologies negatively correlated with LMC3 obtained from GSVA analysis in ckAML samples

GOBP	Pearson R
Regulation of glycogen starch synthase activity	-0.608875967
Glomerulus vasculature morphogenesis	-0.574876515
Positive regulation of glycogen starch synthase activity	-0.539524829
Regulation of mesenchymal stem cell differentiation	-0.535494396
Embryonic appendage morphogenesis	-0.522072
Spinal cord oligodendrocyte cell differentiation	-0.519429794
Mesenchymal stem cell differentiation	-0.518130661
Digestive system development	-0.515509563
Embryonic hindlimb morphogenesis	-0.512169902
Nose development	-0.505983294
Synaptic assembly at neuromuscular junction	-0.503123939
Positive regulation of glycogen metabolic process	-0.502850725
Chloride transmembrane transport	-0.501355414
Hindlimb morphogenesis	-0.499991245
Epithelial cell proliferation involved in prostate gland development	-0.499539371
Male genitalia development	-0.491339939
Regulation of synaptic assembly at neuromuscular junction	-0.490940823
Regulation of skeletal muscle cell differentiation	-0.487533998
Regulation of muscle organ development	-0.480012261
Cell migration involved in kidney development	-0.469764292

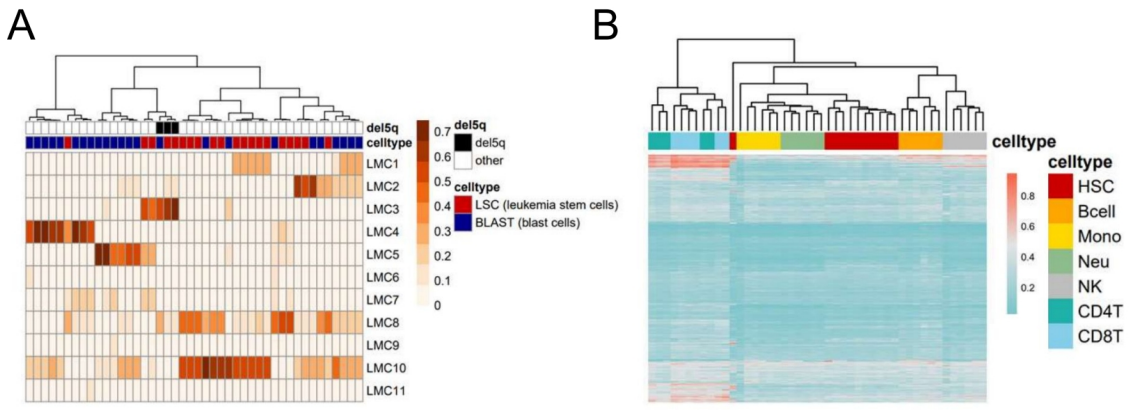

Figure 8.4: Comparisons of LMC3 in LSCs and blasts, and in the normal hematopoietic system. A. Heatmap showing estimated LMC proportions in sorted AML blasts and LSCs using 450K data from Jung *et al.* [159]. Samples deriving from a del(5q) patient are annotated. **B.** Heatmap showing the methylation levels of LMC3-hypermethylated CpG sites in the normal hematopoietic lineage.

Table 8.7: Comparison of protein expression of putative del(5q) target genes in del(5q) and other AML patient samples and cell lines from Kramer *et al.* 2022 [175], Jayavelu *et al.* 2022 [274] and CCLE [176] datasets. Data for EGR1 protein expression was not available in the Kramer *et al.* and Jayavelu *et al* datasets and is therefore not shown.

Dataset	Material	Protein	Wilcoxon p-value
CCLE	AML cell lines	KDM3B	0.011
Kramer et al.	AML patient samples	KDM3B	0.00077
Jayavelu et al.	AML patient samples	KDM3B	0.0055
Kramer et al.	AML patient samples	ETF1	0.0021
Jayavelu et al.	AML patient samples	ETF1	0.37
CCLE	AML cell lines	CTNNA1	0.5
Kramer et al.	AML patient samples	CTNNA1	0.87
Jayavelu et al.	AML patient samples	CTNNA1	0.87
CCLE	AML cell lines	EGR1	0.11

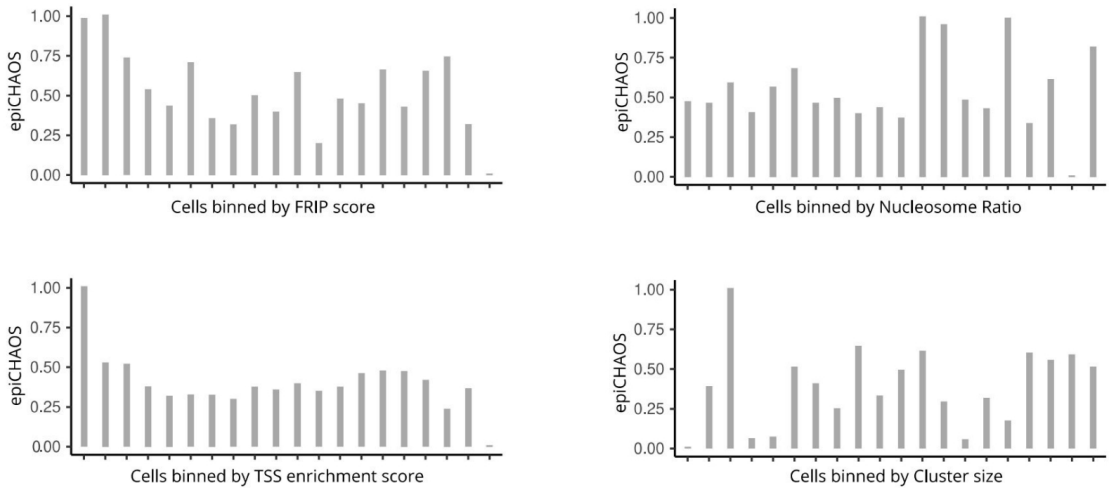

Figure 8.5: EpiCHAOS scores are not influenced by technical confounders. Bar plots of epiCHAOS scores (epiCHAOS) computed on 20 bins of 100 Hep-1 cells ordered by increasing (top left) FRIP scores, (bottom left) TSS enrichment scores, (top right) nucleosome ratio, and (bottom right) cluster size, where the number of cells in each group increases in increments of 25 cells, from 25 to 500 cells.

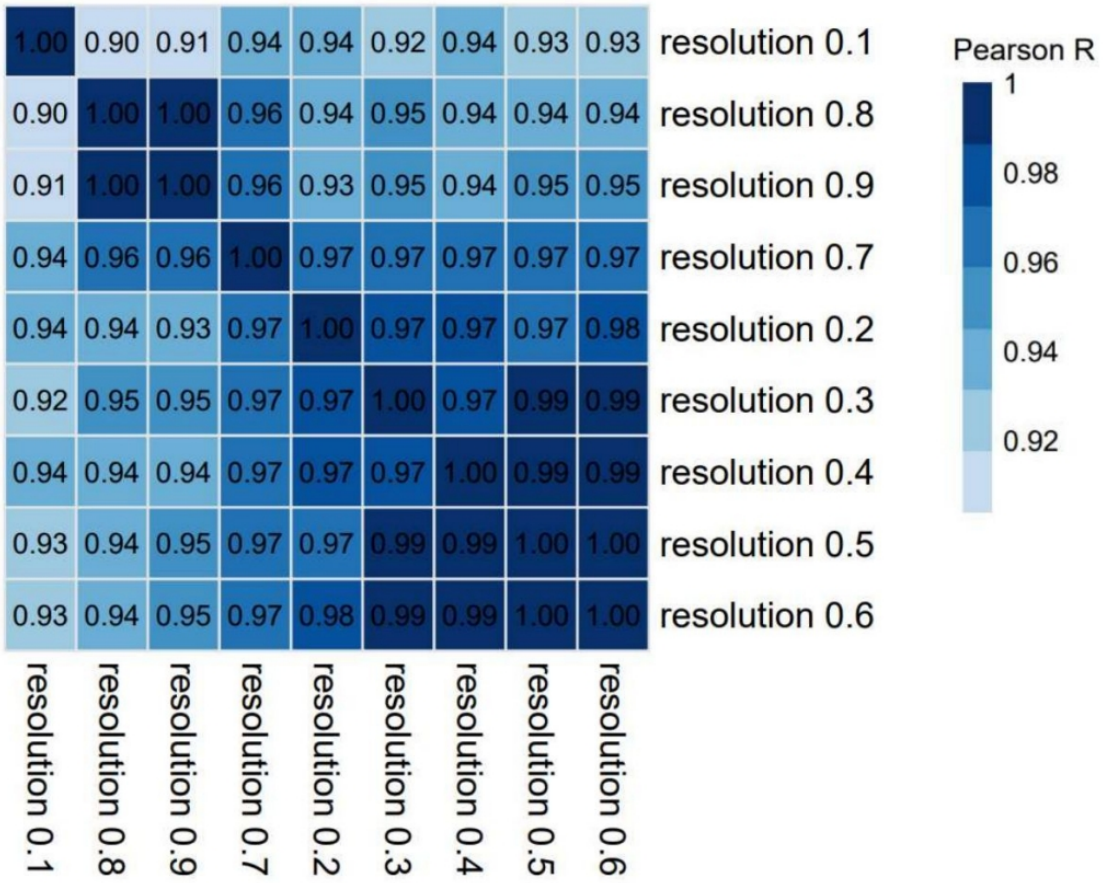

Figure 8.6: EpiCHAOS scores are minimally influenced by clustering parameters. Heatmap displays the Pearson correlation coefficients from per-single-cell correlation of epiCHAOS scores across different clustering resolutions from 0.1 to 0.9. Clusters were defined in scATAC-seq data from breast cancer epithelial cells from Kumegawa *et al.* [140] and epiCHAOS scores calculated for each cluster at each resolution.

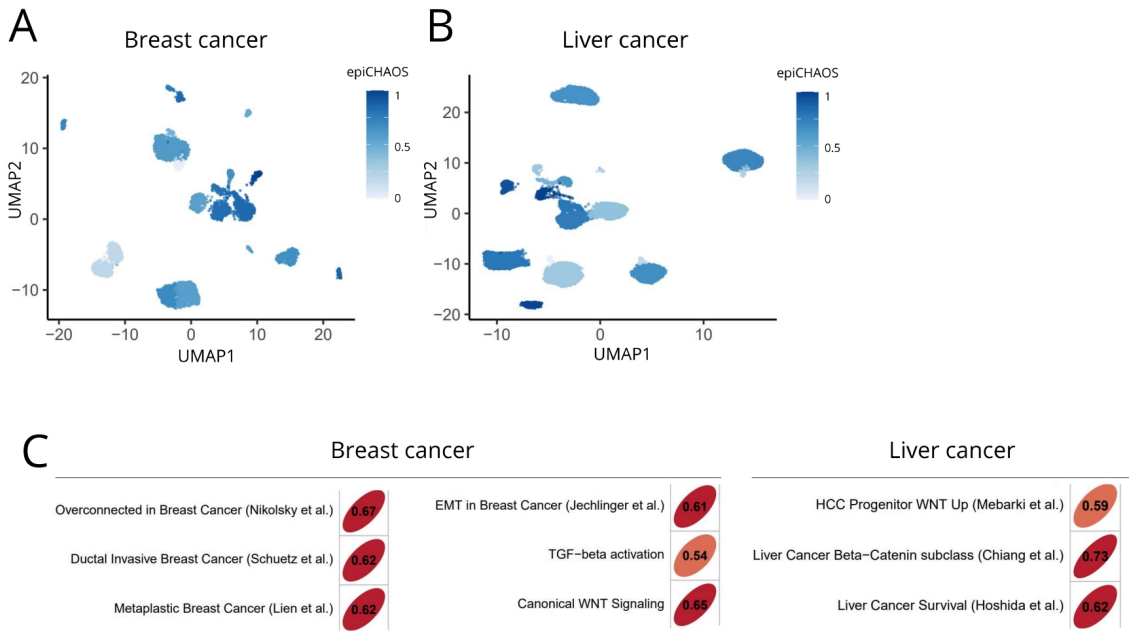

Figure 8.7: Correlating epiCHAOS scores to cancer relevant pathways in breast and liver cancer. A-B. UMAP representations of scATAC-seq clusters from (**A**) breast and (**B**) liver cancer samples after subsetting epithelial cells. Clusters are coloured by epiCHAOS scores (epiCHAOS). **C.** Correlation plots of epiCHAOS scores with selected gene sets from MSigDB. Pearson correlation coefficients are shown.

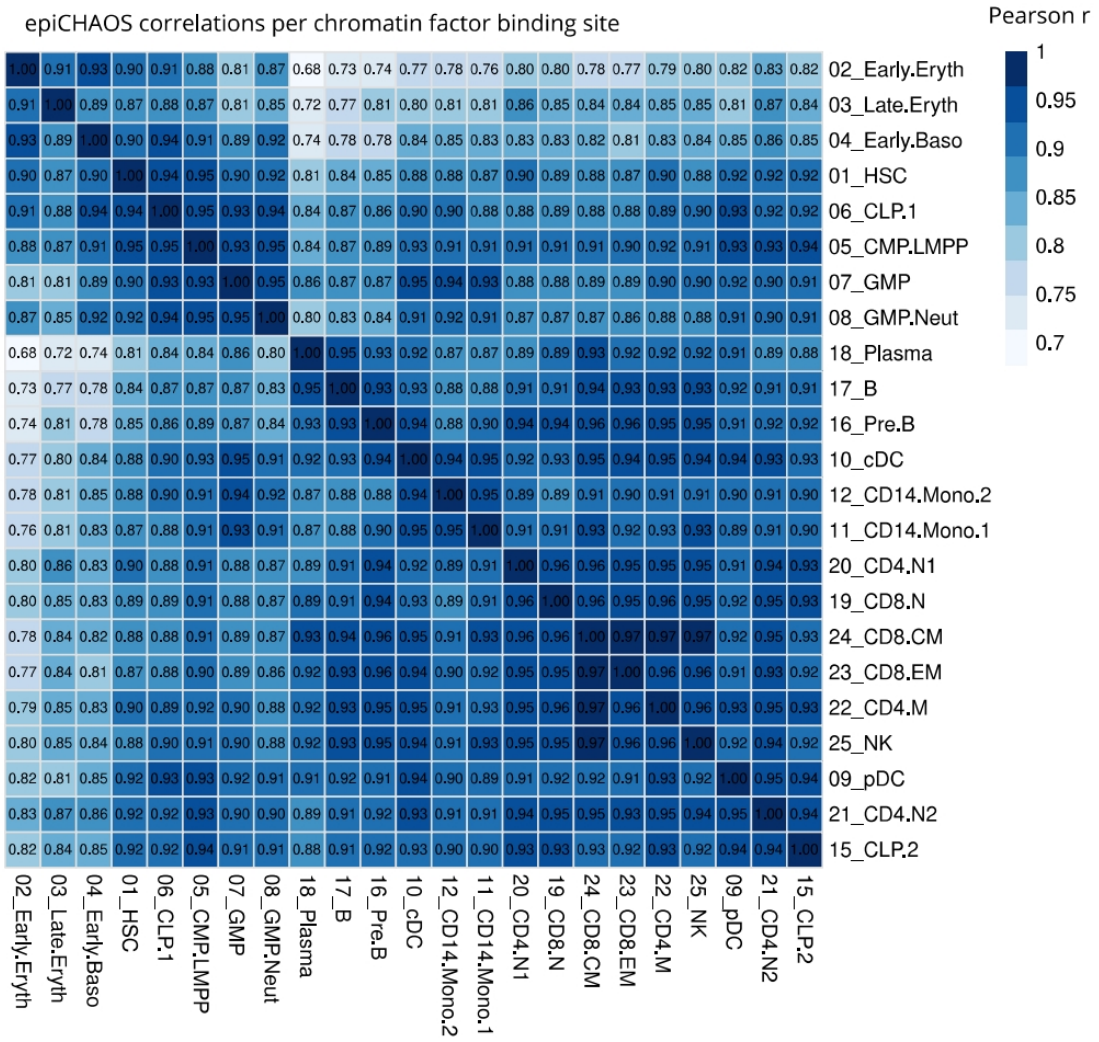

Figure 8.8: Patterns of heterogeneity at different genomic regions are correlated across hematopoietic cell types. Correlation heatmap depicting the similarity in per-region epiCHAOS scores between cell types from human bone marrow [226]. Correlations were computed across all chromatin factor binding sites from the ENCODE TFBS database. Pearson's correlation coefficients are shown.

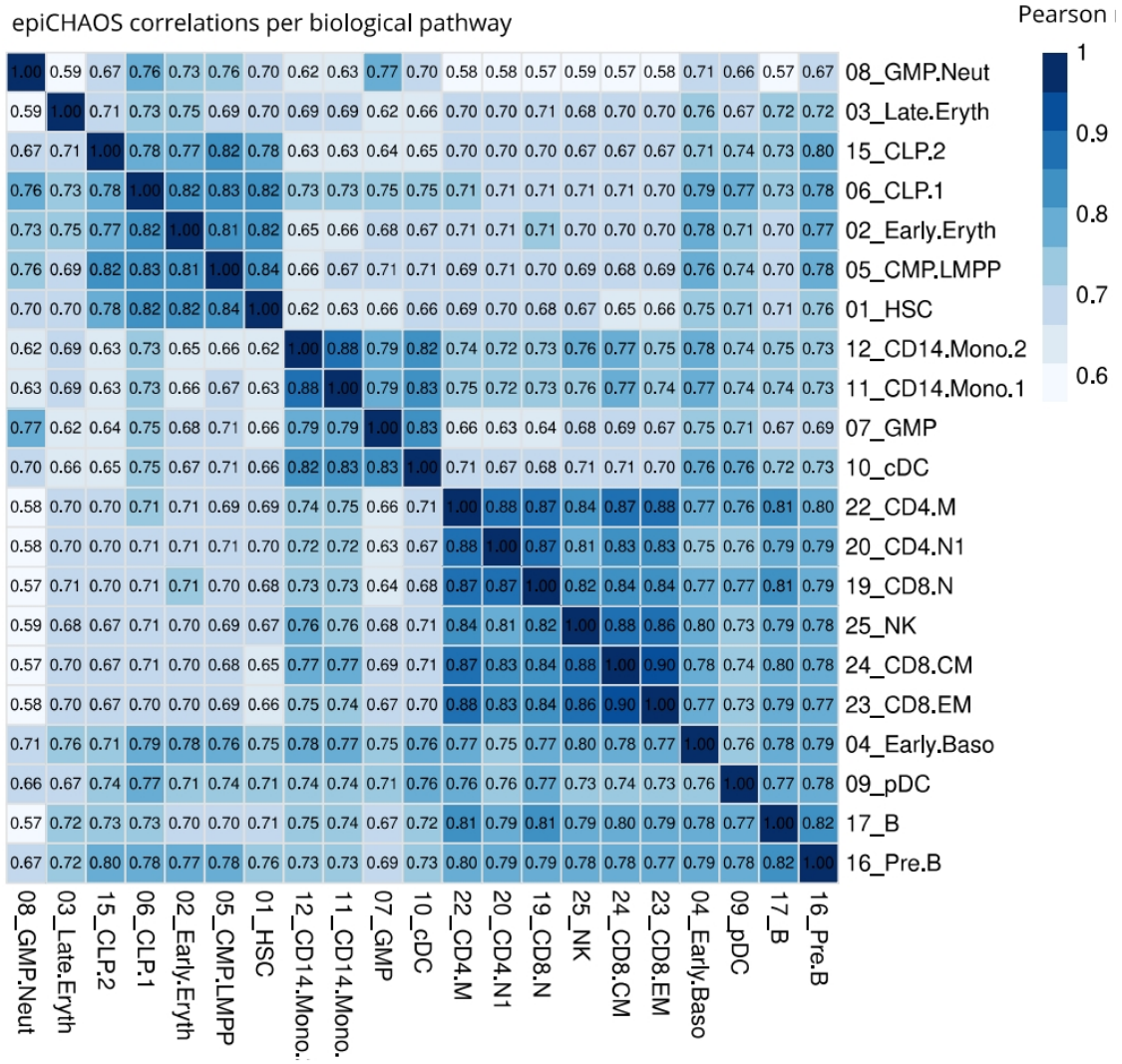

Figure 8.9: Patterns of heterogeneity at different gene sets are correlated across hematopoietic cell types. Correlation heatmap depicting the similarity in per-gene-set epiCHAOS scores between cell types from human bone marrow [226]. Correlations were computed across all gene ontology biological processes (GO:BP). Pearson's correlation coefficients are shown.

Acronyms

2-HG	2-Hydroxyglutarate.
ACT	Antibody-guided Chromatin Tagmentation.
AML	Acute myeloid leukemia.
ATAC-seq	assay for transposase-accessible chromatin with sequencing.
AUC	area under the curve.
CCLE	cancer cell line encyclopedia.
CDF	Cumulative Distribution Function.
ChIP-seq	Chromatin Immunoprecipitation with sequencing.
CLP	common lymphoid progenitor.
CMP	common myeloid progenitor.
CNA	copy number alteration.
CRISPR	clustered regularly interspaced short palindromic repeats.
CTCF	CCCTC-binding factor.
CUT&RUN	Cleavage Under Targets & Release Using Nuclease.
DC	dendritic cell.
DNA	deoxyribonucleic acid.
DNMT	DNA methyltransferase.
EMT	Epithelial-to-mesenchymal transition.
ENCODE	encyclopedia of DNA elements.
EPCAM	epithelial cellular adhesion molecule.
ERRBS	Enhanced Reduced Representation Bisulfite Sequencing.
FAB	French American British.
FDR	False Discovery Rate.
FRIP	fraction of reads in peaks.
GEPIA	gene expression profiling interactive analysis.
GMP	granuloyte/monocyte progenitor.
GSVA	gene set variation analysis.
GTEx	genotype-tissue expression portal.
HAT	histone acetyltransferase.

HDAC histone deacetylase.
HELP HpaII tiny fragment enrichment by ligation-mediated polymerase chain reaction.
HSC hematopoietic stem cell.
HSPC hematopoietic stem or progenitor cell.

ICA independent component analysis.

KDM lysine demethylase.
KMT lysine methyltransferase.
KO knockout.

LAD Lamina Associated Domain.
LMC latent methylation component.
LMPP lymphoid-primed multipotent progenitor.
LOLA genomic locus overlap enrichment analysis.
LSC leukemic stem cell.
LSI Latent Semantic Indexing.

MB megabase.
MDR Minimally Deleted Region.
MDS myelodysplastic syndrome.
MEP megakaryocyte/erythroid progenitor.
MPP multipotent progenitor.

NK natural killer.

PDX patient-derived xenograft.
PRC polycomb repressive complex.

RNA ribonucleic acid.
RNA-seq RNA-sequencing.
RPKM reads per kilobase million.

SAM S-adenosylmethionine.
SNP single nucleotide polymorphism.

TCGA the cancer genome atlas.
TF transcription factor.
TFBS transcription factor binding site.
TPM transcripts per million.
TSG tumor suppressor gene.
TSS transcription start site.

UMAP Uniform Manifold Approximation & Projection.

WT wild-type.

α-KG α-Ketoglutarate.

List of Figures

1.1 Healthy and malignant hematopoiesis . . . 2
1.2 Epigenetic modifications of the DNA and histones . . . 10

3.1 Linking LMCs to hematopoietic cell types . . . 25
3.2 Linking LMCs to hematopoietic progenitor cell states . . . 26
3.3 Methylation signatures associated with epigenetic regulator mutations . . . 27
3.4 Epigenetic signatures linked to MLL rearrangement, RUNX1 disruption and ckAML . . . 28
3.5 DNA methylation-based subgroups . . . 29
3.6 Survival analysis of LMCs . . . 30
3.7 Linking the LMC3 methylation signature to del(5q) AML . . . 32
3.8 Locus and gene set enrichments of LMC3-hypermethylated CpG sites . . . 33
3.9 LMC3 is associated with poor prognosis and reduced drug sensitivity independently of del(5q) status . . . 34
3.10 LMC3 is present at low levels in a small subgroup of del(5q) MDS. . . . 35

4.1 Mutations in epigenetic regulators are rare in del(5q) AML . . . 40
4.2 The 5q minimally deleted region peaks at the location of H3K9me1/2 demethylase KDM3B . . . 42
4.3 KDM3B gene and protein levels are reduced in del(5q) AML . . . 43
4.4 KDM3B expression is higher in AML than in other tumor and normal tissue types . . . 45
4.5 Del(5q) and *IDH* mutation are mutually exclusive events characterised by DNA hypermethylation . . . 46
4.6 DNMT3B overexpression correlates with the del(5q) hypermethylation signature . . . 47
4.7 DNMT3B is overexpressed in AML compared to other non-embryonic tumor types . . . 48
4.8 The del(5q) methylation signature correlates with features of leukemic stem cells . . . 49
4.9 DNMT3B overexpression in AML may be regulated by methylation. . . . 50
4.10 The del(5q) hypermethylation signature correlates with expression of H3K9me1/2 methyltransferases . . . 51
4.11 Genes within lamina-associated domains are dysregulated in del(5q) AML . 52

4.12 Epigenetic similarity of del(5q) and *MECOM/EVI-1*-overexpressing AML converge on overexpression of *DNMT3B* 53

5.1 Calculation of epiCHAOS scores 62
5.2 Validation of epiCHAOS in synthetic datasets 64
5.3 EpiCHAOS scores are not influenced by copy number 65
5.4 Validation of epiCHAOS in *in silico* cell-type mixtures 66
5.5 EpiCHAOS reflects epigenetic heterogeneity associated with developmental plasticity 69
5.6 EpiCHAOS scores are increased in neural tissues and placenta 70
5.7 EpiCHAOS scores correlate with developmental time 71
5.8 Correlation between epigenetic and transciptional heterogeneity in developmental settings 71
5.9 EpiCHAOS correlates with features of plasticity in malignant cells and epigenetic heterogeneity in aging 73
5.10 EpiCHAOS reveals increased heterogeneity at binding sites for PRC2, CTCF and cohesin 75
5.11 EpiCHAOS reveals increased heterogeneity of developmental genes 76
5.12 EpiCHAOS reveals increased heterogeneity at PRC2 target regions in HSCs compared to differentiated blood cells 77
5.13 EpiCHAOS is applicable to multiple types of single-cell epigenomics data . 79

6.1 Hypothesis: the effects of KDM3B haploinsufficiency on epigenetic heterogeneity 84
6.2 Heterozygous deletion of KDM3B in the OCI-AML3 cell line 85
6.3 Heterozygous deletion of KDM3B results in global chromatin compaction . 86
6.4 Transcription factor motif enrichments in KDM3B-KO cells 87
6.5 Heterozygous deletion of KDM3B results in cell-to-cell epigenetic heterogeneity 88

8.1 Selection of MeDeCom parameters 105
8.2 Selection of K for consensus clustering. 107
8.3 Del(5q) AML is defined by a unique DNA methylation signature. 108
8.4 Comparisons of LMC3 in LSCs and blasts, and in the normal hematopoietic system. 109
8.5 EpiCHAOS scores are not influenced by technical confounders 110
8.6 EpiCHAOS scores are minimally influenced by clustering parameters 111
8.7 Correlating epiCHAOS scores to cancer relevant pathways in breast and liver cancer 112
8.8 Patterns of heterogeneity at different genomic regions are correlated across hematopoietic cell types 113
8.9 Patterns of heterogeneity at different gene sets are correlated across hematopoietic cell types 114

List of Tables

8.1 LMC1 Gene Ontology enrichments 105
8.2 LMC9 Gene Ontology enrichments 106
8.3 LMC10 TFBS enrichments 106
8.4 LMC2 TFBS enrichments 107
8.5 LMC3 Gene Ontology enrichments 108
8.6 LMC3 Gene Set Variation Analysis 109
8.7 Differential protein expression of MDR genes 110

Bibliography

[1] Rory M Shallis et al. "Epidemiology of acute myeloid leukemia: Recent progress and enduring challenges". In: *Blood reviews* 36 (2019), pp. 70–87.

[2] Lars Velten et al. "Human haematopoietic stem cell lineage commitment is a continuous process". In: *Nature cell biology* 19.4 (2017), pp. 271–281.

[3] Nicolas Goardon et al. "Coexistence of LMPP-like and GMP-like leukemia stem cells in acute myeloid leukemia". In: *Cancer cell* 19.1 (2011), pp. 138–152.

[4] Hartmut Döhner et al. "Diagnosis and management of AML in adults: 2022 recommendations from an international expert panel on behalf of the ELN". In: *Blood, The Journal of the American Society of Hematology* 140.12 (2022), pp. 1345–1377.

[5] Zachary R Chalmers et al. "Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden". In: *Genome medicine* 9 (2017), pp. 1–14.

[6] Lars Bullinger, Konstanze Döhner, and Hartmut Döhner. "Genomics of acute myeloid leukemia diagnosis and pathways". In: *Journal of clinical oncology* 35.9 (2017), pp. 934–946.

[7] Timothy A Graubert and Elaine R Mardis. "Genomics of acute myeloid leukemia". In: *The Cancer Journal* 17.6 (2011), pp. 487–491.

[8] Etienne Sollier et al. "Pyjacker identifies enhancer hijacking events in acute myeloid leukemia including MNX1 activation via deletion 7q". In: *bioRxiv* (2024), pp. 2024–09.

[9] SMNR Ziemin-van der Poel et al. "Identification of a gene, MLL, that spans the breakpoint in 11q23 translocations associated with human leukemias." In: *Proceedings of the National Academy of Sciences* 88.23 (1991), pp. 10735–10739.

[10] Timothy J Ley et al. "DNMT3A mutations in acute myeloid leukemia". In: *New England Journal of Medicine* 363.25 (2010), pp. 2424–2433.

[11] Myunggon Ko et al. "Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2". In: *Nature* 468.7325 (2010), pp. 839–843.

[12] Maria E Figueroa et al. "Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation". en. In: *Cancer Cell* 18.6 (Dec. 2010), p. 553.

[13] Elli Papaemmanuil et al. "Genomic classification and prognosis in acute myeloid leukemia". In: *New England Journal of Medicine* 374.23 (2016), pp. 2209–2221.

[14] Ekaterina Jahn et al. "Clinical impact of the genomic landscape and leukemogenic trajectories in non-intensively treated elderly acute myeloid leukemia patients". In: *Leukemia* 37.11 (2023), pp. 2187–2196.

[15] N Hosono et al. “Recurrent genetic defects on chromosome 7q in myeloid neoplasms”. In: *Leukemia* 28.6 (2014), pp. 1348–1351.

[16] Cancer Genome Atlas Research Network et al. “Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia”. en. In: *N. Engl. J. Med.* 368.22 (May 2013), pp. 2059–2074.

[17] Sheng Li, Christopher E Mason, and Ari Melnick. “Genetic and epigenetic heterogeneity in acute myeloid leukemia”. en. In: *Curr. Opin. Genet. Dev.* 36 (Feb. 2016), pp. 100–106.

[18] Altekruse Sf. “SEER cancer statistics review, 1975-2007”. In: *http://seer. cancer. gov/csr/1975_2007/results_merged/sect_13_leukemia. pdf* (2009).

[19] Frederick R Appelbaum et al. “Age and acute myeloid leukemia”. en. In: *Blood* 107.9 (May 2006), p. 3481.

[20] Seishi Ogawa. “Genetics of MDS”. en. In: *Blood* 133.10 (Mar. 2019), pp. 1049–1059.

[21] Ina Radtke et al. “Genomic analysis reveals few genetic alterations in pediatric acute myeloid leukemia”. In: *Proceedings of the National Academy of Sciences* 106.31 (2009), pp. 12944–12949.

[22] Hervé Dombret and Claude Gardin. “An update of current treatments for adult acute myeloid leukemia”. In: *Blood, The Journal of the American Society of Hematology* 127.1 (2016), pp. 53–61.

[23] Courtney D DiNardo et al. “Azacitidine and Venetoclax in Previously Untreated Acute Myeloid Leukemia”. en. In: *N. Engl. J. Med.* (Aug. 2020).

[24] Charles A Schiffer and Richard M Stone. “Morphologic Classification and Clinical and Laboratory Correlates”. In: *Holland-Frei Cancer Medicine. 6th edition.* BC Decker, 2003.

[25] Hartmut Döhner et al. “Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel”. In: *Blood, The Journal of the American Society of Hematology* 129.4 (2017), pp. 424–447.

[26] Brenton Thomas Tan et al. “The cancer stem cell hypothesis: a work in progress”. In: *Laboratory investigation* 86.12 (2006), pp. 1203–1207.

[27] Tsvee Lapidot et al. “A cell initiating human acute myeloid leukaemia after transplantation into SCID mice”. In: *nature* 367.6464 (1994), pp. 645–648.

[28] Liran I Shlush et al. “Tracing the origins of relapse in acute myeloid leukaemia to stem cells”. In: *Nature* 547.7661 (2017), pp. 104–108.

[29] Fumihiko Ishikawa et al. “Chemotherapy-resistant human AML stem cells home to and engraft within the bone-marrow endosteal region”. In: *Nature biotechnology* 25.11 (2007), pp. 1315–1321.

[30] François M Vallette et al. “Dormant, quiescent, tolerant and persister cells: Four synonyms for the same target in cancer”. In: *Biochemical pharmacology* 162 (2019), pp. 169–176.

[31] Stanley W K Ng et al. “A 17-gene stemness score for rapid determination of risk in acute leukaemia”. en. In: *Nature* 540.7633 (Dec. 2016), pp. 433–437.

[32] Michael A Koldobskiy et al. “A Dysregulated DNA Methylation Landscape Linked to Gene Expression in MLL-Rearranged AML”. en. In: *Epigenetics* 15.8 (Aug. 2020), pp. 841–858.

[33] Krzysztof Mrózek et al. “Complex karyotype in de novo acute myeloid leukemia: typical and atypical subtypes differ molecularly and clinically”. en. In: *Leukemia* 33.7 (July 2019), pp. 1620–1634.

[34] Krzysztof Mrózek. “Cytogenetic, molecular genetic, and clinical characteristics of acute myeloid leukemia with a complex karyotype”. In: *Seminars in oncology*. Vol. 35. 4. Elsevier. 2008, pp. 365–377.

[35] Frank G Rücker et al. “TP53 alterations in acute myeloid leukemia with complex karyotype correlate with specific copy number alterations, monosomal karyotype, and dismal outcome”. en. In: *Blood* 119.9 (Mar. 2012), pp. 2114–2121.

[36] Aino-Maija Leppä et al. “Single-cell multiomics analysis reveals dynamic clonal evolution and targetable phenotypes in acute myeloid leukemia with complex karyotype”. In: *Nature genetics* (2024), pp. 1–14.

[37] J F Spinella et al. “DELE1 haploinsufficiency causes resistance to mitochondrial stress-induced apoptosis in monosomy 5/del(5q) AML”. In: *Leukemia* 38.3 (Mar. 2024).

[38] M Mallo et al. “Impact of adjunct cytogenetic abnormalities for prognostic stratification in patients with myelodysplastic syndrome and deletion 5q”. In: *Leukemia* 25.1 (2011), pp. 110–120.

[39] S K Horrigan et al. “Polymerase chain reaction-based diagnosis of del (5q) in acute myeloid leukemia and myelodysplastic syndrome identifies a minimal deletion interval”. en. In: *Blood* 88.7 (Oct. 1996), pp. 2665–2670.

[40] S K Horrigan et al. “Delineation of a minimal interval and identification of 9 candidates for a tumor suppressor gene in malignant myeloid disorders on 5q31”. en. In: *Blood* 95.7 (Apr. 2000), pp. 2372–2377.

[41] Z Hu et al. “A novel nuclear protein, 5qNCA (LOC51780) is a candidate for the myeloid leukemia tumor suppressor gene on chromosome 5 band q31”. en. In: *Oncogene* 20.47 (Oct. 2001), pp. 6946–6954.

[42] Ruth N MacKinnon et al. “A cryptic deletion in 5q31.2 provides further evidence for a minimally deleted region in myelodysplastic syndromes”. en. In: *Cancer Genet.* 204.4 (Apr. 2011), pp. 187–194.

[43] M M Le Beau et al. “Cytogenetic and molecular delineation of the smallest commonly deleted region of chromosome 5 in malignant myeloid diseases”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 90.12 (June 1993), pp. 5484–5488.

[44] Ting Xi Liu et al. “Chromosome 5q deletion and epigenetic suppression of the gene encoding alpha-catenin (CTNNA1) in myeloid cell transformation”. en. In: *Nat. Med.* 13.1 (Jan. 2007), pp. 78–83.

[45] Rebekka K Schneider et al. “Rps14 haploinsufficiency causes a block in erythroid differentiation mediated by S100A8 and S100A9”. en. In: *Nat. Med.* 22.3 (Mar. 2016), pp. 288–297.

[46] Benjamin L Ebert et al. “Identification of RPS14 as a 5q- syndrome gene by RNA interference screen”. en. In: *Nature* 451.7176 (Jan. 2008), pp. 335–339.

[47] Manuela Santarosa and Alan Ashworth. “Haploinsufficiency for tumour suppressor genes: when you don’t need to go all the way”. In: *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1654.2 (2004), pp. 105–122.

[48] Alfred G Knudson Jr. “Mutation and cancer: statistical study of retinoblastoma”. In: *Proceedings of the National Academy of Sciences* 68.4 (1971), pp. 820–823.

[49] John M Joslin et al. “Haploinsufficiency of EGR1, a candidate gene in the del(5q), leads to the development of myeloid disorders”. en. In: *Blood* 110.2 (July 2007), pp. 719–726.

[50] Angela Stoddart et al. “EGR1 Haploinsufficiency Confers a Fitness Advantage to Hematopoietic Stem Cells Following Chemotherapy”. en. In: *Exp. Hematol.* 115 (Nov. 2022), pp. 54–67.

[51] Rebekka K Schneider et al. “Role of casein kinase 1A1 in the biology and targeted therapy of del (5q) MDS”. In: *Cancer cell* 26.4 (2014), pp. 509–520.

[52] Xin Xu et al. “KDM3B shows tumor-suppressive activity and transcriptionally regulates HOXA1 through retinoic acid response elements in acute myeloid leukemia”. en. In: *Leuk. Lymphoma* 59.1 (Jan. 2018), pp. 204–213.

[53] Christèle Dubourg et al. “Evaluation of ETF1/eRF1, mapping to 5q31, as a candidate myeloid tumor suppressor gene”. en. In: *Cancer Genet. Cytogenet.* 134.1 (Apr. 2002), pp. 33–37.

[54] Natallia Mikhalkevich and Michael W Becker. *Alpha-Catenin Is Dispensable for Normal Hematopoietic Stem Cell Function.* 2009.

[55] Jeffrey Fairman et al. “Physical mapping of the minimal region of loss in 5q-chromosome.” In: *Proceedings of the National Academy of Sciences* 92.16 (1995), pp. 7406–7410.

[56] N Zhao et al. “Molecular delineation of the smallest commonly deleted region of chromosome 5 in malignant myeloid diseases to 1-1.5 Mb and preparation of a PAC-based physical map”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 94.13 (June 1997), pp. 6948–6953.

[57] J C Liang et al. “Spectral karyotypic study of the HL-60 cell line: detection of complex rearrangements involving chromosomes 5, 7, and 16 and delineation of critical region of deletion on 5q31.1”. en. In: *Cancer Genet. Cytogenet.* 113.2 (Sept. 1999), pp. 105–109.

[58] Anja Krones-Herzig et al. “Early growth response 1 acts as a tumor suppressor in vivo and in vitro via regulation of p53”. In: *Cancer research* 65.12 (2005), pp. 5133–5143.

[59] J Yu et al. “A network of p73, p53 and Egr1 is required for efficient apoptosis in tumor cells”. In: *Cell Death & Differentiation* 14.3 (2007), pp. 436–446.

[60] Veronique Baron et al. “The transcription factor Egr1 is a direct regulator of multiple tumor suppressors including TGFβ1, PTEN, p53, and fibronectin”. In: *Cancer gene therapy* 13.2 (2006), pp. 115–124.

[61] Timothy A Graubert et al. “Integrated genomic analysis implicates haploinsufficiency of multiple chromosome 5q31. 2 genes in de novo myelodysplastic syndromes pathogenesis”. In: *PloS one* 4.2 (2009), e4583.

[62] Vera Adema et al. “Pathophysiologic and clinical implications of molecular profiles resultant from deletion 5q”. In: *EBioMedicine* 80 (2022).

[63] Ursula SA Stalmann et al. “Genetic barcoding systematically compares genes in del (5q) MDS and reveals a central role for CSNK1A1 in clonal expansion”. In: *Blood Advances* 6.6 (2022), pp. 1780–1796.

[64] Stijn NR Fuchs et al. “Collaborative effect of Csnk1a1 haploinsufficiency and mutant p53 in Myc induction can promote leukemic transformation”. In: *Blood Advances* 8.3 (2024), pp. 766–779.

[65] Joshua D Tompkins et al. "Epigenetic stability, adaptability, and reversibility in human embryonic stem cells". In: *Proceedings of the National Academy of Sciences* 109.31 (2012), pp. 12544–12549.
[66] Sangita Pal and Jessica K Tyler. "Epigenetics and aging". In: *Science advances* 2.7 (2016), e1600584.
[67] Gerda Egger et al. "Epigenetics in human disease and prospects for epigenetic therapy". In: *Nature* 429.6990 (2004), pp. 457–463.
[68] Aaron D Goldberg, C David Allis, and Emily Bernstein. "Epigenetics: a landscape takes shape". In: *Cell* 128.4 (2007), pp. 635–638.
[69] Nehmé Saksouk, Elisabeth Simboeck, and Jérôme Déjardin. "Constitutive heterochromatin formation and transcription in mammals". In: *Epigenetics & chromatin* 8 (2015), pp. 1–17.
[70] Aimée M Deaton and Adrian Bird. "CpG islands and the regulation of transcription". In: *Genes & development* 25.10 (2011), pp. 1010–1022.
[71] Peter A Jones. "Functions of DNA methylation: islands, start sites, gene bodies and beyond". In: *Nature reviews genetics* 13.7 (2012), pp. 484–492.
[72] Masaki Okano et al. "DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development". In: *Cell* 99.3 (1999), pp. 247–257.
[73] Timothy H Bestor. "The DNA methyltransferases of mammals". In: *Human molecular genetics* 9.16 (2000), pp. 2395–2402.
[74] Xiaoji Wu and Yi Zhang. "TET-mediated active DNA demethylation: mechanism, function and beyond". In: *Nature Reviews Genetics* 18.9 (2017), pp. 517–534.
[75] Netanel Loyfer et al. "A DNA methylation atlas of normal human cell types". In: *Nature* 613.7943 (2023), pp. 355–364.
[76] Cizhong Jiang and B Franklin Pugh. "Nucleosome positioning and gene regulation: advances through genomics". In: *Nature Reviews Genetics* 10.3 (2009), pp. 161–172.
[77] Andrew J Bannister and Tony Kouzarides. "Regulation of chromatin by histone modifications". In: *Cell research* 21.3 (2011), pp. 381–395.
[78] Howard Cedar and Yehudit Bergman. "Linking DNA methylation and histone modification: patterns and paradigms". In: *Nature Reviews Genetics* 10.5 (2009), pp. 295–304.
[79] B Lehnertz et al. "Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin". In: *Curr. Biol.* 13.14 (July 2003).
[80] Emmanuelle Viré et al. "The Polycomb group protein EZH2 directly controls DNA methylation". en. In: *Nature* 439.7078 (Dec. 2005), pp. 871–874.
[81] Emily L Putiri and Keith D Robertson. "Epigenetic mechanisms and genome stability". en. In: *Clin. Epigenetics* 2.2 (Dec. 2010), pp. 299–314.
[82] Jeffrey A Yoder, Colum P Walsh, and Timothy H Bestor. "Cytosine methylation and the ecology of intragenomic parasites". In: *Trends in genetics* 13.8 (1997), pp. 335–340.
[83] Panagiotis Karakaidos, Dimitris Karagiannis, and Theodoros Rampias. "Resolving DNA damage: epigenetic regulation of DNA repair". In: *Molecules* 25.11 (2020), p. 2496.

[84] Haico Van Attikum and Susan M Gasser. "The histone code at DNA breaks: a guide to repair?" In: *Nature Reviews Molecular Cell Biology* 6.10 (2005), pp. 757–765.

[85] Sergei Chuikov et al. "Regulation of p53 activity through lysine methylation". In: *Nature* 432.7015 (2004), pp. 353–360.

[86] Douglas Hanahan and Robert A Weinberg. "The hallmarks of cancer". In: *cell* 100.1 (2000), pp. 57–70.

[87] Douglas Hanahan. "Hallmarks of cancer: new dimensions". In: *Cancer discovery* 12.1 (2022), pp. 31–46.

[88] Manel Esteller et al. "The epigenetic hallmarks of cancer". In: *Cancer Discovery* 14.10 (2024), pp. 1783–1809.

[89] M Esteller et al. "Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors". In: *J. Natl. Cancer Inst.* 92.7 (Apr. 2000).

[90] J G Herman et al. "Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 91.21 (Oct. 1994), p. 9700.

[91] Jean-Pierre Issa. "CpG island methylator phenotype in cancer". In: *Nature Reviews Cancer* 4.12 (2004), pp. 988–993.

[92] Daniel J Weisenberger et al. "CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer". In: *Nature genetics* 38.7 (2006), pp. 787–793.

[93] Mikhail F Denissenko et al. "Cytosine methylation determines hot spots of DNA damage in the human P 53 gene". In: *Proceedings of the National Academy of Sciences* 94.8 (1997), pp. 3893–3898.

[94] Celestia Fang et al. "Cancer-specific CTCF binding facilitates oncogenic transcriptional dysregulation". In: *Genome biology* 21 (2020), pp. 1–30.

[95] Kyung Lock Kim et al. "Dissection of a CTCF topological boundary uncovers principles of enhancer-oncogene regulation". In: *Molecular Cell* 84.7 (2024), pp. 1365–1376.

[96] Melanie Ehrlich. "DNA hypomethylation in cancer cells". In: *Epigenomics* 1.2 (2009), pp. 239–259.

[97] Wanding Zhou et al. "DNA methylation loss in late-replicating domains is linked to mitotic cell division". In: *Nature genetics* 50.4 (2018), pp. 591–602.

[98] Nicolle Besselink et al. "The genome-wide mutational consequences of DNA hypomethylation". In: *Scientific Reports* 13.1 (2023), p. 6874.

[99] G Howard et al. "Activation and transposition of endogenous retroviral elements in hypomethylation induced tumors in mice". In: *Oncogene* 27.3 (2008), pp. 404–408.

[100] Csaba Bödör et al. "EZH2 mutations are frequent and represent an early event in follicular lymphoma". In: *Blood, The Journal of the American Society of Hematology* 122.18 (2013), pp. 3165–3168.

[101] Sebastian Stasik et al. "EZH2 mutations and impact on clinical outcome: an analysis in 1,604 patients with newly diagnosed acute myeloid leukemia". In: *Haematologica* 105.5 (2020), e228.

[102] Kimberly H Kim and Charles WM Roberts. "Targeting EZH2 in cancer". In: *Nature medicine* 22.2 (2016), pp. 128–134.

[103] Michael S Lawrence et al. “Discovery and saturation analysis of cancer genes across 21 tumour types”. In: *Nature* 505.7484 (2014), pp. 495–501.

[104] Yuan Cheng et al. “Targeting epigenetic regulators for cancer therapy: mechanisms and advances in clinical trials”. In: *Signal transduction and targeted therapy* 4.1 (2019), p. 62.

[105] Hervé Dombret et al. “International phase 3 study of azacitidine vs conventional care regimens in older patients with newly diagnosed AML with¿ 30% blasts”. In: *Blood, The Journal of the American Society of Hematology* 126.3 (2015), pp. 291–299.

[106] Rachael Straining and William Eighmy. “Tazemetostat: EZH2 inhibitor”. In: *Journal of the advanced practitioner in oncology* 13.2 (2022), p. 158.

[107] Michael Weller et al. “MGMT promoter methylation in malignant gliomas: ready for personalized medicine?” In: *Nature Reviews Neurology* 6.1 (2010), pp. 39–51.

[108] Huiyan Luo et al. “Liquid biopsy of methylation biomarkers in cell-free DNA”. In: *Trends in molecular medicine* 27.5 (2021), pp. 482–500.

[109] Wenhua Liang et al. “Non-invasive diagnosis of early-stage lung cancer using high-throughput targeted DNA methylation sequencing of circulating tumor DNA (ctDNA)”. In: *Theranostics* 9.7 (2019), p. 2056.

[110] David Capper et al. “DNA methylation-based classification of central nervous system tumours”. en. In: *Nature* 555.7697 (Mar. 2018), p. 469.

[111] Christian Koelsche et al. “Sarcoma classification by DNA methylation profiling”. In: *Nature communications* 12.1 (2021), p. 498.

[112] Maria E Figueroa et al. “DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia”. en. In: *Cancer Cell* 17.1 (Jan. 2010), pp. 13–27.

[113] Jacob L Glass et al. “Epigenetic Identity in AML Depends on Disruption of Nonpromoter Regulatory Elements and Is Affected by Antagonistic Effects of Mutations in Epigenetic Modifiers”. en. In: *Cancer Discov.* 7.8 (Aug. 2017), pp. 868–883.

[114] Brian Giacopelli et al. “DNA methylation epitypes highlight underlying developmental and disease pathways in acute myeloid leukemia”. en. In: *Genome Res.* 31.5 (May 2021), pp. 747–761.

[115] Sanne Lugthart et al. “Aberrant DNA hypermethylation signature in acute myeloid leukemia directed by EVI1”. en. In: *Blood* 117.1 (Jan. 2011), pp. 234–241.

[116] Albert Gough et al. “Biologically Relevant Heterogeneity: Metrics and Practical Insights”. en. In: *SLAS Discov* 22.3 (Mar. 2017), pp. 213–237.

[117] Roshan M Kumar et al. “Deconstructing transcriptional heterogeneity in pluripotent stem cells”. en. In: *Nature* 516.7529 (Dec. 2014), pp. 56–61.

[118] Elisabet Pujadas and Andrew P Feinberg. “Regulated noise in the epigenetic landscape of development and disease”. en. In: *Cell* 148.6 (Mar. 2012), pp. 1123–1131.

[119] Sarah E Jackson and John D Chester. “Personalised cancer medicine”. In: *International journal of cancer* 137.2 (2015), pp. 262–266.

[120] Corbin E Meacham and Sean J Morrison. “Tumour heterogeneity and cancer cell plasticity”. en. In: *Nature* 501.7467 (Sept. 2013), pp. 328–337.

[121] Ibiayi Dagogo-Jack and Alice T Shaw. “Tumour heterogeneity and resistance to cancer therapies”. en. In: *Nat. Rev. Clin. Oncol.* 15.2 (Feb. 2018), pp. 81–94.

[122] Douglas Hanahan. "Hallmarks of Cancer: New Dimensions". en. In: *Cancer Discov.* 12.1 (Jan. 2022), pp. 31–46.

[123] W K Hofmann, A Trumpp, and C Müller-Tidow. "Therapy resistance mechanisms in hematological malignancies". In: *International journal of cancer* 152.3 (Feb. 2023).

[124] Andriy Marusyk, Michalina Janiszewska, and Kornelia Polyak. "Intratumor heterogeneity: the Rosetta stone of therapy resistance". en. In: *Cancer Cell* 37.4 (Apr. 2020), p. 471.

[125] Arthur W Lambert, Diwakar R Pattabiraman, and Robert A Weinberg. "Emerging Biological Principles of Metastasis". en. In: *Cell* 168.4 (Feb. 2017), pp. 670–691.

[126] Anna Chapman et al. "Heterogeneous tumor subpopulations cooperate to drive invasion". en. In: *Cell Rep.* 8.3 (Aug. 2014), pp. 688–695.

[127] Jeff H Tsai and Jing Yang. "Epithelial–mesenchymal plasticity in carcinoma metastasis". In: *Genes & development* 27.20 (2013), pp. 2192–2206.

[128] Ning Li et al. "Whole genome DNA methylation analysis based on high throughput sequencing technology". In: *Methods* 52.3 (2010), pp. 203–212.

[129] Alexander Meissner et al. "Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis". In: *Nucleic acids research* 33.18 (2005), pp. 5868–5877.

[130] Marina Bibikova et al. "High density DNA methylation array with single CpG site resolution". In: *Genomics* 98.4 (2011), pp. 288–295.

[131] Ruth Pidsley et al. "Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling". In: *Genome biology* 17 (2016), pp. 1–17.

[132] Peter J Park. "ChIP–seq: advantages and challenges of a maturing technology". In: *Nature reviews genetics* 10.10 (2009), pp. 669–680.

[133] Peter J Skene, Jorja G Henikoff, and Steven Henikoff. "Targeted in situ genome-wide profiling with high efficiency for low cell numbers". In: *Nature protocols* 13.5 (2018), pp. 1006–1019.

[134] Benjamin Carter et al. "Mapping histone modifications in low cell number and single cells using antibody-guided chromatin tagmentation (ACT-seq)". In: *Nature communications* 10.1 (2019), p. 3747.

[135] Jason D Buenrostro et al. "ATAC-seq: a method for assaying chromatin accessibility genome-wide". In: *Current protocols in molecular biology* 109.1 (2015), pp. 21–29.

[136] Jason D Buenrostro et al. "Single-cell chromatin accessibility reveals principles of regulatory variation". In: *Nature* 523.7561 (2015), pp. 486–490.

[137] Agostina Bianchi et al. "scTAM-seq enables targeted high-confidence analysis of DNA methylation in single cells". en. In: *Genome Biol.* 23.1 (Oct. 2022), p. 229.

[138] Kevin Grosselin et al. "High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer". en. In: *Nat. Genet.* 51.6 (June 2019), pp. 1060–1066.

[139] Mario L Suvà and Itay Tirosh. "Single-cell RNA sequencing in cancer: lessons learned and emerging challenges". In: *Molecular cell* 75.1 (2019), pp. 7–12.

[140] Kohei Kumegawa et al. "GRHL2 motif is associated with intratumor heterogeneity of cis-regulatory elements in luminal breast cancer". en. In: *NPJ Breast Cancer* 8.1 (June 2022), p. 70.

[141] Paloma Cejas et al. “Subtype heterogeneity and epigenetic convergence in neuroendocrine prostate cancer”. en. In: *Nat. Commun.* 12.1 (Oct. 2021), p. 5775.

[142] Qionghua Zhu et al. “Single cell multi-omics reveal intra-cell-line heterogeneity across human cancer cell lines”. en. In: *Nat. Commun.* 14.1 (Dec. 2023), p. 8170.

[143] Paul Guilhamon et al. “Single-cell chromatin accessibility profiling of glioblastoma identifies an invasive cancer stem cell population associated with lower survival”. en. In: *Elife* 10 (Jan. 2021).

[144] Pavlo Lutsik et al. “MeDeCom: discovery and quantification of latent components of heterogeneous methylomes”. en. In: *Genome Biol.* 18.1 (Mar. 2017), p. 55.

[145] Fengying Wu et al. “Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer”. en. In: *Nat. Commun.* 12.1 (May 2021), p. 2540.

[146] Lichun Ma et al. “Tumor Cell Biodiversity Drives Microenvironmental Reprogramming in Liver Cancer”. en. In: *Cancer Cell* 36.4 (Oct. 2019), 418–430.e6.

[147] Kunihiko Hinohara et al. “KDM5 Histone Demethylase Activity Links Cellular Transcriptomic Heterogeneity to Therapeutic Resistance”. en. In: *Cancer Cell* 35.2 (Feb. 2019), pp. 330–332.

[148] Andrew E Teschendorff and Tariq Enver. “Single-cell entropy for accurate estimation of differentiation potency from a cell’s transcriptome”. en. In: *Nat. Commun.* 8 (June 2017), p. 15599.

[149] Dan A Landau et al. “Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia”. en. In: *Cancer Cell* 26.6 (Dec. 2014), pp. 813–825.

[150] Shicheng Guo et al. “Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA”. en. In: *Nat. Genet.* 49.4 (Apr. 2017), pp. 635–642.

[151] Gilad Landan et al. “Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues”. en. In: *Nat. Genet.* 44.11 (Nov. 2012), pp. 1207–1214.

[152] Hehuang Xie et al. “Genome-wide quantitative assessment of variation in DNA methylation patterns”. en. In: *Nucleic Acids Res.* 39.10 (May 2011), pp. 4099–4108.

[153] Michael Scherer et al. “Quantitative comparison of within-sample heterogeneity scores for DNA methylation data”. en. In: *Nucleic Acids Res.* 48.8 (May 2020), e46.

[154] Gail J Roboz et al. “Guadecitabine vs TC in relapsed/refractory AML after intensive chemotherapy: a randomized phase 3 ASTRAL-2 trial”. In: *Blood Advances* 8.8 (2024), pp. 2020–2029.

[155] Katherine Kelly et al. “Acute Myeloid Leukemia with deletion 5q is an epigenetically distinct subgroup defined by heterozygous loss of KDM3B”. In: *bioRxiv* (2024), pp. 2024–11.

[156] Michael Scherer et al. “Reference-free deconvolution, visualization and interpretation of complex DNA methylation data using DecompPipeline, MeDeCom and FactorViz”. en. In: *Nat. Protoc.* 15.10 (Oct. 2020), pp. 3240–3263.

[157] Lucas A Salas et al. “An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray”. en. In: *Genome Biol.* 19.1 (May 2018), p. 64.

[158] Yufang Qin et al. “InfiniumPurify: An R package for estimating and accounting for tumor purity in cancer methylation research”. en. In: *Genes Dis* 5.1 (Mar. 2018), pp. 43–45.

[159] Namyoung Jung et al. “An LSC epigenetic signature is largely mutation independent and implicates the HOXA cluster in AML pathogenesis”. en. In: *Nat. Commun.* 6 (Oct. 2015), p. 8489.

[160] David A Russler-Germain et al. “The R882H DNMT3A mutation associated with AML dominantly inhibits wild-type DNMT3A by blocking its ability to form active tetramers”. en. In: *Cancer Cell* 25.4 (Apr. 2014), pp. 442–454.

[161] Elisabeth R Wilson et al. “Focal disruption of DNA methylation dynamics at enhancers in IDH-mutant AML cells”. en. In: *Leukemia* 36.4 (Apr. 2022), pp. 935–945.

[162] Fernando P G Silva et al. “Identification of RUNX1/AML1 as a classical tumor suppressor gene”. en. In: *Oncogene* 22.4 (Jan. 2003), pp. 538–547.

[163] Matthew D Wilkerson and D Neil Hayes. “ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking”. en. In: *Bioinformatics* 26.12 (June 2010), pp. 1572–1573.

[164] Claude Preudhomme et al. “Favorable prognostic significance of CEBPA mutations in patients with de novo acute myeloid leukemia: a study from the Acute Leukemia French Association (ALFA)”. en. In: *Blood* 100.8 (Oct. 2002), pp. 2717–2723.

[165] Marius Bill et al. “Mutations associated with a 17-gene leukemia stem cell score and the score’s prognostic relevance in the context of the European LeukemiaNet classification of acute myeloid leukemia”. en. In: *Haematologica* 105.3 (Mar. 2020), pp. 721–729.

[166] Martin Sauvageau and Guy Sauvageau. “Polycomb group proteins: multi-faceted regulators of somatic stem cells and cancer”. en. In: *Cell Stem Cell* 7.3 (Sept. 2010), pp. 299–313.

[167] Raed A Alharbi et al. “The role of HOX genes in normal hematopoiesis and acute leukemia”. In: *Leukemia* 27.5 (2013), pp. 1000–1008.

[168] Sonja Hänzelmann, Robert Castelo, and Justin Guinney. “GSVA: gene set variation analysis for microarray and RNA-seq data”. en. In: *BMC Bioinformatics* 14 (Jan. 2013), p. 7.

[169] A Hecht et al. “Genome-wide DNA methylation analysis pre- and post-lenalidomide treatment in patients with myelodysplastic syndrome with isolated deletion (5q)”. In: *Ann. Hematol.* 100.6 (June 2021).

[170] Andres Jerez et al. “Loss of heterozygosity in 7q myeloid disorders: clinical associations and genomic pathogenesis”. en. In: *Blood* 119.25 (June 2012), pp. 6109–6117.

[171] C Chen et al. “MLL3 is a haploinsufficient 7q tumor suppressor in acute myeloid leukemia”. In: *Cancer Cell* 25.5 (May 2014).

[172] Martin Jädersten et al. “TP53 mutations in low-risk myelodysplastic syndromes with del (5q) predict disease progression”. In: *Journal of clinical oncology* 29.15 (2011), pp. 1971–1979.

[173] NSD Larmonie et al. “MN1 overexpression is driven by loss of DNMT3B methylation activity in inv (16) pediatric AML”. In: *Oncogene* 37.1 (2018), pp. 107–115.

[174] Ekaterina I Romanova et al. “RUNX1/CEBPA mutation in acute myeloid leukemia promotes hypermethylation and indicates for demethylation therapy”. In: *International Journal of Molecular Sciences* 23.19 (2022), p. 11413.

[175] Michael H Kramer et al. “Proteomic and phosphoproteomic landscapes of acute myeloid leukemia”. en. In: *Blood* 140.13 (Sept. 2022), pp. 1533–1548.

[176] David P Nusinow et al. “Quantitative Proteomics of the Cancer Cell Line Encyclopedia”. en. In: *Cell* 180.2 (Jan. 2020), 387–402.e16.

[177] Zefang Tang et al. “GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses”. en. In: *Nucleic Acids Res.* 45.W1 (July 2017), W98–W102.

[178] Chao Lu et al. “IDH mutation impairs histone demethylation and results in a block to cell differentiation”. en. In: *Nature* 483.7390 (Feb. 2012), pp. 474–478.

[179] Bilian Jin et al. “DNMT1 and DNMT3B modulate distinct polycomb-mediated histone modifications in colon cancer”. en. In: *Cancer Res.* 69.18 (Sept. 2009), pp. 7412–7421.

[180] Silvina Epsztejn-Litman et al. “De novo DNA methylation promoted by G9a prevents reprogramming of embryonically silenced genes”. en. In: *Nat. Struct. Mol. Biol.* 15.11 (Nov. 2008), pp. 1176–1183.

[181] Daniela Meilinger et al. “Np95 interacts with de novo DNA methyltransferases, Dnmt3a and Dnmt3b, and mediates epigenetic silencing of the viral CMV promoter in embryonic stem cells”. en. In: *EMBO Rep.* 10.11 (Nov. 2009), pp. 1259–1264.

[182] Sandrine Hayette et al. “High DNA methyltransferase DNMT3B levels: a poor prognostic marker in acute myeloid leukemia”. en. In: *PLoS One* 7.12 (Dec. 2012), e51527.

[183] Abdelrahman H Elsayed et al. “A six-gene leukemic stem cell score identifies high risk pediatric acute myeloid leukemia”. en. In: *Leukemia* 34.3 (Mar. 2020), pp. 735–745.

[184] Bernhard Lehnertz et al. “The methyltransferase G9a regulates HoxA9-dependent transcription in AML”. en. In: *Genes Dev.* 28.4 (Feb. 2014), pp. 317–327.

[185] Genki Yamato et al. “Genome-wide DNA methylation analysis in pediatric acute myeloid leukemia”. en. In: *Blood Adv* 6.11 (June 2022), pp. 3207–3219.

[186] Julian List et al. “Genocopy of EVI1-AML with Paraneoplastic Diabetes Insipidus: PRDM16 Overexpression By t (1; 2) and Enhancer Hijacking”. In: *Blood* 144 (2024), p. 6136.

[187] Stefan Gröschel et al. “A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia”. In: *Cell* 157.2 (2014), pp. 369–381.

[188] Vitalyi Senyuk et al. “The oncoprotein EVI1 and the DNA methyltransferase Dnmt3 co-operate in binding and de novo methylation of target DNA”. en. In: *PLoS One* 6.6 (June 2011), e20793.

[189] D Spensberger and R Delwel. “A novel interaction between the proto-oncogene Evi1 and histone methyltransferases, SUV39H1 and G9a”. In: *FEBS Lett.* 582.18 (Aug. 2008).

[190] Sangeetha Venugopal, John Mascarenhas, and David P Steensma. "Loss of 5q in myeloid malignancies–A gain in understanding of biological and clinical consequences". In: *Blood reviews* 46 (2021), p. 100735.

[191] Sihui Li et al. "JMJD1B Demethylates H4R3me2s and H3K9me2 to Facilitate Gene Expression for Development of Hematopoietic Stem and Progenitor Cells". en. In: *Cell Rep.* 23.2 (Apr. 2018), pp. 389–403.

[192] Jan Padeken, Stephen P Methot, and Susan M Gasser. "Establishment of H3K9-methylated heterochromatin and its functions in tissue differentiation and maintenance". In: *Nature Reviews Molecular Cell Biology* 23.9 (2022), pp. 623–640.

[193] Bas van Steensel and Andrew S Belmont. "Lamina-associated domains: links with chromosome architecture, heterochromatin and gene repression". en. In: *Cell* 169.5 (May 2017), p. 780.

[194] Patrick Trojer and Danny Reinberg. "Facultative heterochromatin: is there a distinctive molecular signature?" In: *Molecular cell* 28.1 (2007), pp. 1–13.

[195] Artem Barski et al. "High-resolution profiling of histone methylations in the human genome". In: *Cell* 129.4 (2007), pp. 823–837.

[196] Jun-ichi Nakayama et al. "Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly". In: *Science* 292.5514 (2001), pp. 110–113.

[197] Maithili P Dalvi et al. "Taxane-platin-resistant lung cancers co-develop hypersensitivity to JumonjiC demethylase inhibitors". In: *Cell reports* 19.8 (2017), pp. 1669–1684.

[198] Mi-Jin An et al. "Histone demethylase KDM3B regulates the transcriptional network of cell-cycle genes in hepatocarcinoma HepG2 cells". In: *Biochemical and biophysical research communications* 508.2 (2019), pp. 576–582.

[199] Hilal Saraç et al. "Systematic characterization of chromatin modifying enzymes identifies KDM3B as a critical regulator in castration resistant prostate cancer". In: *Oncogene* 39.10 (2020), pp. 2187–2201.

[200] Jiong Li et al. "KDM3 epigenetically controls tumorigenic potentials of human colorectal cancer stem cells through Wnt/β-catenin signalling". In: *Nature communications* 8.1 (2017), p. 15146.

[201] Aixia Hu et al. "KDM3B-ETF1 fusion gene downregulates LMO2 via the WNT/β-catenin signaling pathway, promoting metastasis of invasive ductal carcinoma". In: *Cancer Gene Therapy* 29.2 (2022), pp. 215–224.

[202] Elisa Paolicchi et al. "Histone lysine demethylases in breast cancer". In: *Critical reviews in oncology/hematology* 86.2 (2013), pp. 97–103.

[203] Yongxia Liu et al. "An epigenetic role for PRL-3 as a regulator of H3K9 methylation in colorectal cancer". In: *Gut* 62.4 (2013), pp. 571–581.

[204] Shazia Mahamdallie et al. "Identification of new Wilms tumour predisposition genes: an exome sequencing study". In: *The Lancet Child & Adolescent Health* 3.5 (2019), pp. 322–331.

[205] Illja J Diets et al. "De novo and inherited pathogenic variants in KDM3B cause intellectual disability, short stature, and facial dysmorphism". In: *The American Journal of Human Genetics* 104.4 (2019), pp. 758–766.

[206] Xiaoji Chen et al. "G9a/GLP-dependent histone H3K9me2 patterning during human hematopoietic stem cell lineage commitment". en. In: *Genes Dev.* 26.22 (Nov. 2012), pp. 2499–2511.

[207] Gray ZH et al. “Epigenetic balance ensures mechanistic control of MLL amplification and rearrangement”. In: *Cell* 186.21 (Oct. 2023), 4528–4545.e18.

[208] Alice H Berger, Alfred G Knudson, and Pier Paolo Pandolfi. “A continuum model for tumour suppression”. In: *Nature* 476.7359 (2011), pp. 163–169.

[209] Ying Ye et al. “Progressive chromatin repression and promoter methylation of CTNNA1 associated with advanced myeloid malignancies”. In: *Cancer research* 69.21 (2009), pp. 8482–8490.

[210] Andreea Reilly et al. “Lamin B1 deletion in myeloid neoplasms causes nuclear anomaly and altered hematopoietic stem cell function”. en. In: *Cell Stem Cell* 29.4 (Apr. 2022), p. 577.

[211] John M Cunningham et al. “Historical perspective and clinical implications of the Pelger-Huet cell”. In: *American journal of hematology* 84.2 (2009), pp. 116–119.

[212] Sergei Chuikov et al. “Regulation of p53 activity through lysine methylation”. en. In: *Nature* 432.7015 (Nov. 2004), pp. 353–360.

[213] Jing Huang et al. “G9a and Glp methylate lysine 373 in the tumor suppressor p53”. en. In: *J. Biol. Chem.* 285.13 (Mar. 2010), pp. 9636–9641.

[214] Etienne Sollier. “Enhancer hijacking in acute myeloid leukemia with complex karyotype”. Expected completion. Doctoral dissertation. Heidelberg University, 2024.

[215] Francisco Saavedra et al. “JMJD1B, a novel player in histone H3 and H4 processing to ensure genome stability”. en. In: *Epigenetics Chromatin* 13 (2020).

[216] Fade Gong and Kyle M Miller. “Histone methylation and the DNA damage response”. In: *Mutation Research/Reviews in Mutation Research* 780 (2019), pp. 37–47.

[217] Anna C Salzberg et al. “Genome-wide mapping of histone H3K9me2 in acute myeloid leukemia reveals large chromosomal domains associated with massive gene silencing and sites of genome instability”. In: *PloS one* 12.3 (2017), e0173723.

[218] Wei Xu et al. “Oncometabolite 2-Hydroxyglutarate Is a Competitive Inhibitor of α-Ketoglutarate-Dependent Dioxygenases”. en. In: *Cancer Cell* 19.1 (Jan. 2011), p. 17.

[219] M R Waarts et al. “CRISPR Dependency Screens in Primary Hematopoietic Stem Cells Identify KDM3B as a Genotype Specific Vulnerability in IDH2- and TET2-Mutant Cells”. In: *Cancer Discov.* (May 2024).

[220] Katsuhiko Nosho et al. “DNMT3B expression might contribute to CpG island methylator phenotype in colorectal cancer”. In: *Clinical cancer research* 15.11 (2009), pp. 3663–3671.

[221] Candace J Poole et al. “DNMT3B overexpression contributes to aberrant DNA methylation and MYC-driven tumor maintenance in T-ALL and Burkitt’s lymphoma”. In: *Oncotarget* 8.44 (2017), p. 76898.

[222] J Devon Roll et al. “DNMT3b overexpression contributes to a hypermethylator phenotype in human breast cancer cell lines”. In: *Molecular cancer* 7 (2008), pp. 1–14.

[223] Katherine Kelly et al. “EpiCHAOS: a metric to quantify epigenomic heterogeneity in single-cell data”. In: *bioRxiv* (2024), pp. 2024–04.

[224] Neo Christopher Chung et al. “Jaccard/Tanimoto similarity test and estimation methods for biological presence-absence data”. en. In: *BMC Bioinformatics* 20.Suppl 15 (Dec. 2019), p. 644.

[225] Guanao Yan, Dongyuan Song, and Jingyi Jessica Li. "scReadSim: a single-cell RNA-seq and ATAC-seq read simulator". en. In: *Nat. Commun.* 14.1 (Nov. 2023), p. 7482.
[226] Jeffrey M Granja et al. "Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia". en. In: *Nat. Biotechnol.* 37.12 (Dec. 2019), pp. 1458–1465.
[227] Shanshan Wang et al. "Single-cell multiomics reveals heterogeneous cell states linked to metastatic potential in liver cancer cell lines". en. In: *iScience* 25.3 (Mar. 2022), p. 103857.
[228] Hisham Mohammed et al. "Single-Cell Landscape of Transcriptional Heterogeneity and Cell Fate Decisions during Mouse Early Gastrulation". en. In: *Cell Rep.* 20.5 (Aug. 2017), pp. 1215–1228.
[229] Ricard Argelaguet et al. "Decoding gene regulation in the mouse embryo using single-cell multi-omics". en. June 2022.
[230] Zizhen Yao et al. "A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain". en. In: *Nature* 624.7991 (Dec. 2023), pp. 317–332.
[231] Diego Calderon et al. "The continuum of embryonic development at single-cell resolution". en. In: *Science* 377.6606 (Aug. 2022), eabn5800.
[232] Silvia Domcke et al. "A human cell atlas of fetal chromatin accessibility". en. In: *Science* 370.6518 (Nov. 2020).
[233] Gunsagar S Gulati et al. "Single-cell transcriptional diversity is a hallmark of developmental potential". en. In: *Science* 367.6476 (Jan. 2020), pp. 405–411.
[234] Amanda J Craig et al. "Genome-wide profiling of transcription factor activity in primary liver cancer using single-cell ATAC sequencing". en. In: *Cell Rep.* 42.11 (Nov. 2023), p. 113446.
[235] Fabiana Lüönd et al. "Distinct contributions of partial and full EMT to breast cancer malignancy". en. In: *Dev. Cell* 56.23 (Dec. 2021), 3203–3221.e11.
[236] Rachael G Aubin et al. "Pro-inflammatory cytokines mediate the epithelial-to-mesenchymal-like transition of pediatric posterior fossa ependymoma". en. In: *Nat. Commun.* 13.1 (July 2022), p. 3936.
[237] Austin E Gillen et al. "Single-Cell RNA Sequencing of Childhood Ependymoma Reveals Neoplastic Cell Subpopulations That Impact Molecular Classification and Etiology". en. In: *Cell Rep.* 32.6 (Aug. 2020), p. 108023.
[238] Johannes Gojo et al. "Single-Cell RNA-Seq Reveals Cellular Hierarchies and Impaired Developmental Trajectories in Pediatric Ependymoma". en. In: *Cancer Cell* 38.1 (July 2020), 44–59.e9.
[239] Y Meng et al. "Epigenetic programming defines haematopoietic stem cell fate restriction". In: *Nat. Cell Biol.* 25.6 (June 2023).
[240] David H Meyer and Björn Schumacher. "Aging clocks based on accumulating stochastic variation". In: *Nature Aging* (2024), pp. 1–15.
[241] Dominic Grün, Lennart Kester, and Alexander van Oudenaarden. "Validation of noise models for single-cell transcriptomics". en. In: *Nat. Methods* 11.6 (June 2014), pp. 637–640.
[242] Gozde Kar et al. "Flipping between Polycomb repressed and active transcriptional states introduces noise in gene expression". en. In: *Nat. Commun.* 8.1 (June 2017), p. 36.

[243] Ricard Argelaguet et al. “Multi-omics profiling of mouse gastrulation at single-cell resolution”. en. In: *Nature* 576.7787 (Dec. 2019), pp. 487–491.

[244] Michael Scherer et al. “Somatic epimutations enable single-cell lineage tracing in native hematopoiesis across the murine and human lifespan”. en. Apr. 2024.

[245] Andrei E Tarkhov et al. “Nature of epigenetic aging from a single-cell perspective”. en. In: *Nature Aging* 4.6 (May 2024), pp. 854–870.

[246] Shinji Maegawa et al. “Widespread and tissue specific age-related DNA methylation changes in mice”. In: *Genome research* 20.3 (2010), pp. 332–340.

[247] Jean-Pierre Issa et al. “Aging and epigenetic drift: a vicious cycle”. In: *The Journal of clinical investigation* 124.1 (2014), pp. 24–29.

[248] Andre J Faure, Jörn M Schmiedel, and Ben Lehner. “Systematic Analysis of the Determinants of Gene Expression Noise in Embryonic Stem Cells”. en. In: *Cell Syst* 5.5 (Nov. 2017), 471–484.e4.

[249] Andrew P Feinberg and Rafael A Irizarry. “Evolution in health and medicine Sackler colloquium: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 107 Suppl 1.Suppl 1 (Jan. 2010), pp. 1757–1764.

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

[251] Victoria Parreno, Anne-Marie Martinez, and Giacomo Cavalli. “Mechanisms of Polycomb group protein function in cancer”. en. In: *Cell Res.* 32.3 (Jan. 2022), pp. 231–253.

[252] Tim J Stevens et al. “3D structures of individual mammalian genomes studied by single-cell Hi-C”. In: *Nature* 544.7648 (2017), pp. 59–64.

[253] Jacob Stewart-Ornstein, Jonathan S Weissman, and Hana El-Samad. “Cellular noise regulons underlie fluctuations in Saccharomyces cerevisiae”. In: *Molecular cell* 45.4 (2012), pp. 483–493.

[254] Tali Mazor et al. “Intratumoral heterogeneity of the epigenome”. In: *Cancer cell* 29.4 (2016), pp. 440–451.

[255] Kevin Nuno et al. “Convergent epigenetic evolution drives relapse in acute myeloid leukemia”. en. In: *Elife* 13 (Apr. 2024).

[256] Matilde Murga et al. “Global chromatin compaction limits the strength of the DNA damage response”. en. In: *J. Cell Biol.* 178.7 (Sept. 2007), p. 1101.

[257] John J Welch et al. “Global regulation of erythroid gene expression by transcription factor GATA-1”. In: *Blood* 104.10 (2004), pp. 3136–3147.

[258] Sophie G Kellaway et al. “Leukemic stem cells activate lineage inappropriate signalling pathways to promote their growth”. In: *Nature Communications* 15.1 (2024), p. 1359.

[259] Alba Rodriguez-Meira et al. “Single-cell multi-omics identifies chronic inflammation as a driver of TP53-mutant leukemic evolution”. In: *Nature genetics* 55.9 (2023), pp. 1531–1541.

[260] Travis I Zack et al. “Pan-cancer patterns of somatic copy number alteration”. In: *Nature genetics* 45.10 (2013), pp. 1134–1140.

[261] Yassen Assenov et al. “Comprehensive analysis of DNA methylation data with RnBeads”. en. In: *Nat. Methods* 11.11 (Nov. 2014), pp. 1138–1140.

[262] Xiaoqi Zheng et al. "Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies". en. In: *Genome Biol.* 18.1 (Jan. 2017), p. 17.

[263] Jeffrey W Tyner et al. "Functional genomic landscape of acute myeloid leukaemia". en. In: *Nature* 562.7728 (Oct. 2018), pp. 526–531.

[264] Guangchuang Yu et al. "clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters". en. In: *OMICS* 16.5 (May 2012), p. 284.

[265] N C Sheffield and C Bock. "LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor". In: *Bioinformatics* 32.4 (Feb. 2016).

[266] Laurens G Wilming, Veronika Boychenko, and Jennifer L Harrow. "Comprehensive comparative homeobox gene annotation in human and mouse". en. In: *Database* 2015 (Sept. 2015).

[267] Emmalee R Adelman et al. "Aging Human Hematopoietic Stem Cells Manifest Profound Epigenetic Reprogramming of Enhancers That May Predispose to Leukemia". en. In: *Cancer Discov.* 9.8 (Aug. 2019), pp. 1080–1101.

[268] Michael I Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". en. In: *Genome Biol.* 15.12 (Dec. 2014), pp. 1–21.

[269] Matthew E Ritchie et al. "limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic acids research* 43.7 (2015), e47–e47.

[270] Juan A Ferrer-Bonsoms, Laura Jareno, and Angel Rubio. "Rediscover: an R package to identify mutually exclusive mutations". en. In: *Bioinformatics* 38.3 (Jan. 2022), pp. 844–845.

[271] Damiano Fantini. *TCGAretriever: Retrieve and Download Data from The Cancer Genome Atlas Project*. R package version 1.9.1. 2024. DOI: 10.32614/CRAN.package.TCGAretriever. URL: https://cran.r-project.org/web/packages/TCGAretriever/index.html.

[272] David Benjamin et al. "Calling Somatic SNVs and Indels with Mutect2". en. Dec. 2019.

[273] Aviad Tsherniak et al. "Defining a cancer dependency map". In: *Cell* 170.3 (2017), pp. 564–576.

[274] Ashok Kumar Jayavelu et al. "The proteogenomic subtypes of acute myeloid leukemia". en. In: *Cancer Cell* 40.3 (Mar. 2022), 301–317.e12.

[275] Jeffrey M Granja et al. "ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis". en. In: *Nat. Genet.* 53.3 (Mar. 2021), pp. 403–411.

[276] Akshaya Ramakrishnan et al. "epiAneufinder identifies copy number alterations from single-cell ATAC-seq data". en. In: *Nat. Commun.* 14.1 (Sept. 2023), p. 5846.

[277] Arthur Liberzon et al. "The Molecular Signatures Database (MSigDB) hallmark gene set collection". en. In: *Cell systems* 1.6 (Dec. 2015), p. 417.

[278] Yong Zhang et al. "Model-based analysis of ChIP-Seq (MACS)". en. In: *Genome Biol.* 9.9 (Sept. 2008), R137.

[279] Ittai Ben-Porath et al. "An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors". en. In: *Nat. Genet.* 40.5 (May 2008), pp. 499–507.

[280] Lars Guelen et al. “Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions”. In: *Nature* 453.7197 (2008), pp. 948–951.

[281] Kairong Cui et al. “Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation”. In: *Cell stem cell* 4.1 (2009), pp. 80–93.

Acknowledgements

It has been a great joy and privilege to complete my doctoral studies in the Division of Cancer Epigenomics at the DKFZ, and within the University of Heidelberg. I would first and foremost like to thank my supervisor Christoph Plass for this enriching opportunity and for his enduring support, encouragement and enthusiasm throughout this project. Likewise, I would like to thank my co-supervisor Pavlo Lutsik for his invaluable scientific guidance and kindness over the past years.

I feel extremely fortunate to have had the chance to spend these years in relative freedom, exploring such fascinating concepts, and it has been a pleasure to do so alongside such bright scientists and personalities as I have found in the B370 group.

I would like extend a special thanks to the past and present members of the AML group - Simge, Etienne, Anna and Elena, and especially Fiona, Michael and Ashish for their support and valuable scientific advice throughout this journey.

I would also like to thank the members of my Thesis Advisory Committee - Karsten Rippe and Sebastian Waszak, for providing constructive feedback and insightful discussions.

I am particularly grateful for those who I have shared the bioinformatics office with during the past years - especially Yunhee, Etienne and Nan - whose unique company has resulted in catastrophic amounts of laughter, the inheritence of personal sound effects, and many enjoyable scientific and linguistic discussions.

I am also indebted to the mentors, professors and classmates who guided me through earlier academic chapters at University College Dublin, Lund University, and King's College London. In particular, I want to thank Anita, Johan, and Madeleine for their wisdom, warmth, and excitement, and for instilling in me both the desire and the abilities to embark on this path.

Beyond that, I would like to acknowledge some special people in Heidelberg and elsewhere, who have contributed significantly, albeit intermittently, to my wellbeing throughout this journey. I am especially grateful for Smita, Isobel, Olivia, Moeed, the Vetters, and Ole - with whose help I find myself rounding up this chapter in a state of "happy chaos" rather than chaos proper.

Lastly, I am grateful to my family - my brother Matthew for maintaining my sense of home, and above all, my parents and Louise, who have largely made me the person I am, and who have given me a reason to pursue cancer research, however much they might disapprove. It is still my hope to one day write a “real book”, in which I may be allowed to stretch words further than this. In the meantime, this thesis will serve as a nice placeholder.