Inaugural dissertation

for

obtaining the doctoral degree

of the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of the

Ruprecht-Karls-University

Heidelberg

Presented by
Stephen Jörg Krämer, M.Sc.
born in Schweinfurt, Germany

Oral examination:

Uncovering the mechanisms and information content of CpG-resolved DNA methylation programming during hematopoietic differentiation

Referees: Prof. Dr. Roland Eils
Prof. Dr. Christoph Plass

Acknowledgments

I am very much indebted to my thesis supervisor Prof. Dr. Matthias Schlesner, for valuable technical advice, personal guidance, and moral support whenever needed. He has created a collaborative and productive scientific working environment, and it was a pleasure to be a part of the team. I am very grateful for the support and mentorship of Prof. Dr. Roland Eils throughout my scientific life. Prof. Eils gave the foundational bioinformatics lectures in my study course, gave me the opportunity to participate in exiting iGEM competition teams, and generously supported me as a Ph.D. student. I am very thankful for the uncountable lively discussions with PD Dr. Daniel Lipka over the last years, which have made my work better and have made me a better scientist. It was a pleasure to dive deep into detailed epigenetic mechanisms and to ponder some of the big questions of stem cell biology in those discussions. I also extend my sincere appreciation to Dr. Michael Milsom, who always helped me see the biological side of my research when things started to look too much like an abstract statistical modeling task. Many other inspiring people have mentored me during my studies, and I would like to especially thank Dr. Holger Schäfer for teaching me how to survive in a wet lab and Prof. Dr. Franziska Matthäus and Prof. Dr. Thomas Höfer for giving me the mathematical tools to question some of the conventional intuitions encountered in the biological sciences.

Prof. Dr. Matthias Schlesner, PD Dr. Daniel Lipka, and Dr. Michael Milsom scientifically guided the main project of my thesis. I have pushed the project towards ambitious and risky hypotheses, where sometimes pursuing incremental research might have provided an easier path. I am very grateful to my supervisors for supporting me in this endeavor and actively and enthusiastically developing those ideas together with me.

I am very grateful to Prof. Dr. Christoph Plass for co-initiating the main project of my thesis, acting as the second referee of my thesis, and creating a division that has established excellent epigenetic research at the DKFZ and provided the foundation for projects like mine. I am grateful to Prof. Dr. Stefan Fröhling for providing generous support to my research projects. I would also like to thank Dr. Jan Korbel for agreeing to be a member of my thesis advisory committee and Dr. Pei-Chi Wei and Prof. Dr. Ingrid Lohmann for agreeing to be examiners on the thesis defense committee. Finally, I thank the organizers of the DKFZ Ph.D. Program, which has provided the framework for my thesis.

I would like to thank my colleagues in the Division of Theoretical Bioinformatics (led by Prof. Dr. Roland Eils), the Biomedical Informatics, Data Mining, and Data Analytics Lab (led by Prof. Dr. Matthias Schlesner), the Division of Applied Bioinformatics (led by Prof. Dr. Benedikt Brors), the Health Data Science Unit (led by Prof. Dr. Carl Herrmann), and the Computational Cancer Epigenomics Lab (led by Prof. Dr. Pavlo Lutsik): Dr. Charles Imbusch, Christian Heyer, Dr. Dr. Daniel Hübschmann, Dr. Dorett Odoni, Irmengard Sax, Dr. Ivo Buchhalter, Dr. Jeongbin Park, Dr. Kortine Kleinheinz, Dr. Lars Feuerbach, Matthias Bieg, Dr. Nagarajan Paramasivam, Dr. Naveed Ishaque, Dr. Qi Wang, and Dr. Zuguang Gu have brainstormed problems with me and helped me out with scripts and code snippets at many occasions. I especially want to acknowledge that Dr. Charles Imbusch and Dr. Qi Wang offered me generous help at the beginning of my thesis work and that Dr. Dr. Daniel Hübschmann gave me much invaluable advice over the years. I would like to thank the brilliant students whom I had the chance to supervise: Thore Bürgel, Eva Schitter, Alexander Mattausch, Jürgen Stachon, Theodor Pöschl, and Philipp Rentzsch.

I would also like to thank my experimental collaboration partners. In particular, I am indebted to the labs of PD Dr. Daniel Lipka, Dr. Mick Milsom, and Prof. Dr. Christoph Plass. I would like to extend special thanks to Dr. Sina Stäble, Dr. Maximilian Schönung, Dr. Mark Hartmann, Dr. Jens Langstein, Dr. Ruzhica Bogeska, Julia Knoch, and PD Dr. Dieter Weichenhan. Dr. Sina Stäble has, for several years, played a special role as my tandem experimental collaboration partner in multiple projects. I am further grateful to the labs of Prof. Dr. Frank Rosenbauer, Prof. Dr. Carsten Müller-Tidow, Prof. Dr. Maria Carolina Florian, Prof. Dr. Stefan Fröhling, and Prof. Dr. Carl Herrmann for exciting collaborations.

I am also very grateful to the many people and groups which offered crucial organizational, technical, and administrative support. I would like to extend my special thanks to the system administrators in Heidelberg and Augsburg: Frank Thommen, Martin Kaim, and Martin Lang. I thank the project managers who kept things running: Dr. Jan Eufinger, Dr. Julia Ritzerfeld, Dr. Matthias Binapfl, and Dr. Yoshimi Felbert. Many thanks to Corinna Sprengart, Manuela Schäfer, Tara Buschle, and Wiebke Hartung for their generous help in organizational matters. I thank the invaluable core facilities at the DKFZ: the ITCF team, the GPCF team, the scOpenLab team, and the ODFC team. I particularly thank Dr. Ivo Buchhalter, Dr. Lena Weise, Ingrid Scholz, Jules Kerssemakers, Dr. Philip Kensche, and Michael Heinold.

It has been said that (the pursuit of) scientific discovery plus time equals comedy. It was such great fun to discover this comedy with my close friends, who are some of the most brilliant scientists I have ever encountered: Daniel, Domi, Max, and Tim: thank you so much!

I am deeply grateful to my family. I especially want to thank all the Krämers, Trumps, and Bosslers for many recharging weekends. I am deeply indebted to my parents Jürgen and Lydia, for their unwavering support throughout my life. Finally, I am deeply grateful to my wife Feli for her love, support, and many interesting discussions about science and scientific life in live life science.

Abstract

DNA methylome remodeling is an essential molecular mechanism underlying all stages of hematopoietic differentiation. However, current datasets only cover a fraction of the genome and are often limited to specific hematopoietic cell types. A comprehensive, genome-wide atlas of the DNA methylation dynamics during hematopoietic differentiation is still missing. Preliminary evidence suggests that the single-cell landscape of the hematopoietic stem and progenitor cell (HSPC) compartment is characterized by a structured continuum of epigenetically-defined cell states. Significant advances in charting this epigenetic state manifold have recently been achieved for the chromatin accessibility and histone modification layers. However, despite its potential importance, the landscape of single-cell DNA methylome states in the HSPC compartment remains largely unexplored. This project aimed to comprehensively map the genome-wide DNA methylation dynamics during hematopoietic differentiation and leverage this atlas as a reference to analyze the single-cell DNA methylome landscape in the HSPC compartment and among mature hematopoietic cells. The functional importance and rich information content of differentially methylated regions (DMRs) are well-established. However, the DNA methylation layer inherently possesses the capability to encode information at CpG resolution. The role and extent of differentially methylated CpG (DMCpG) programming within DMR regions is largely unexplored. This project therefore aimed to evaluate the role and mechanisms of DMCpG programming during hematopoietic differentiation.

Using high-coverage tagmentation-based whole-genome bisulfite sequencing data for 25 hematopoietic populations, I have compiled a genome-wide, dual-layer DMR/DMCpG atlas, which maps, annotates, and integrates DMR and DMCpG programming during hematopoietic differentiation. Loss of stemness was associated with lineage-independent gain of DNA methylation, while lineage specification was accompanied by hierarchical DNA methylation dynamics, characterized by unidirectional loss of DNA methylation. Different DMCpGs within focal DMR intervals were often distinctly programmed and thus contained heterogeneous information content. In particular, most of the DMRs were seeded and progressively expanded through subsequent programming of specific DMCpGs at different stages of differentiation. Mature hematopoietic cells exhibited systematic seed DMCpG hypomethylation in DMRs associated with alternative cell fates. This seed hypomethylation likely represents epigenetic memory of alternative fate explorations in progenitor cells. Collectively, these findings suggest a hierarchical model of DNA methylation programming, in which information is encoded through DMR programming and through DMCpG programming within DMR regions. This model represents a significant extension of the commonly accepted paradigm of regional DNA methylation programming.

Using the dual-layer DMR/DMCpG atlas as a reference, single-cell methylome states for 312 HSPCs, as well as for a total of 136 mature B cells, T cells, CFU-Es, and monocytes, could

be dissected with high resolution. The HSPC compartment was characterized by a structured continuum of single-cell DNA methylome states. Multiple lines of evidence suggested that differentiation starts from apex HSCs possessing a lineage-naive DNA methylome state. Exit from the apex HSC state was initiated by balanced, multi-lineage DMR seeding. This early DMR programming was strictly restricted to specific DMR seeding regions, which often comprised only one or two DMCpGs. This contrasts with the conventional paradigm that functionally relevant DMRs always contain at least several DMCpGs. Further differentiation within the HSPC compartment was accompanied by continuous, gradually more lineage-specific accumulation of hypomethylation, leading to progressive DMR expansion.

The dual-layer DMR/DMCpG atlas provides an essential resource for studying the epigenetic regulation of the hematopoietic differentiation process and serves as a valuable reference for the analysis of single-cell bisulfite sequencing data. This work highlights the highly-resolved, progressive, and stable nature of DNA methylome remodeling during hematopoietic differentiation and reveals several aspects of the structure and information content of the DNA methylome layer which go beyond the currently accepted paradigms. It appears likely that the DNA methylome remodeling mechanisms active in other differentiation systems and related processes, such as tumor evolution, share the same principles of hierarchical DNA methylation programming with CpG resolution. However, in many systems, the information content of the DNA methylome may be convoluted by a combination of this programming mechanism and other programming mechanisms characterized by stochastic regional accumulation of DNA methylation alterations. The analysis strategies presented in this work provide a basis for the further development of computational methods capable of dissecting the rich but complex information content of the DNA methylome with high resolution.

Zusammenfassung

Epigenetische Programmierung mittels DNA-Methylierung ist ein wesentlicher molekularer Mechanismus, der allen Stadien der hämatopoetischen Differenzierung zugrunde liegt. Aktuelle Datensätze decken jedoch nur einen Bruchteil des Genoms ab und sind häufig auf bestimmte hämatopoetische Zelltypen beschränkt. Ein umfassender, genomweiter Atlas der DNA-Methylierungsdynamik während der hämatopoetischen Differenzierung fehlt. Vorläufige Erkenntnisse deuten darauf hin, dass die Einzelzelllandschaft der hämatopoetischen Stamm- und Vorläuferzellen (HSPCs) durch ein strukturiertes Kontinuum epigenetisch definierter Zellzustände gekennzeichnet ist. Bei der Kartierung dieses epigenetischen Zustandsraums mittels der Vermessung der Chromatinzugänglichkeitslandschaft und Histonmodifikationslandschaft von einzelnen HSPCs wurden kürzlich vielversprechende Fortschritte erzielt. Trotz ihrer potenziellen Bedeutung sind die Einzelzell-DNA-Methylomzustände von HSPCs dagegen noch weitgehend unerforscht. Dieses Projekt zielte darauf ab, die genomweite DNA-Methylierungsdynamik während der hämatopoetischen Differenzierung umfassend abzubilden und diesen Atlas als Referenz für die Analyse der Einzelzell-DNA-Methylomlandschaft in HSPCs und reifen hämatopoetischen Zellen zu nutzen. Die funktionelle Bedeutung und der reichhaltige Informationsgehalt von differenziell methylierten Regionen (DMRs) sind gut belegt. Allerdings besitzt DNA-Methylierung die Fähigkeit, Informationen mit CpG-Auflösung zu codieren. Das Ausmaß der Programmierung von differenziell methylierten CpGs (DMCpGs) innerhalb der DMR-Regionen ist weitgehend unerforscht. Dieses Projekt zielte daher darauf ab, die Rolle und Mechanismen der CpG-aufgelösten Programmierung während der hämatopoetischen Differenzierung systematisch zu untersuchen.

Unter Verwendung von Tagmentierungs-basierten Bisulfit-Sequenzierungsdaten für 25 hämatopoetische Populationen habe ich einen genomweiten, zweischichtigen DMR/DMCpG-Atlas zusammengestellt, der die DMR- und DMCpG-Programmierung während der hämatopoetischen Differenzierung kartiert, annotiert und integriert. Verlust des Stammzellcharakters war mit einem differenzierungslinienunabhängigen Gewinn an DNA-Methylierung verbunden, während die Spezifizierung von bestimmten Differenzierungslinien jeweils mit einer hierarchischen DNA-Methylierungsdynamik einherging, die durch einen unidirektionalen Verlust von DNA-Methylierung gekennzeichnet war. Verschiedene CpGs innerhalb fokaler DMR-Intervalle wurden oft unterschiedlich programmiert und trugen daher heterogene Informationsinhalte. Insbesondere wurden die meisten DMRs zunächst in kleinen Initiierungsregionen angelegt und dann durch anschließende Programmierung spezifischer benachbarter CpGs in verschiedenen Differenzierungsstadien schrittweise erweitert. Reife hämatopoetische Zellen zeigten eine systematische Hypomethylierung in den Initiierungsregionen von DMRs, die mit alternativen Zellschicksalen verbunden sind. Die Teilhypomethylierung dieser DMRs scheint epigenetisches Gedächtnis über die Exploration alternativer Zellschicksale in Vorläuferzellen darzustellen. Zusammengenommen legen diese Ergebnisse ein hierarchisches Modell der epigenetischen Programmierung mittels DNA-Methylierung nahe, bei dem Informationen durch DMR-Programmierung und durch CpG-Programmierung innerhalb von DMR-Regionen codiert werden. Dieses Modell stellt eine bedeutende Erweiterung des allgemein akzeptierten Paradigmas eines auf regionaler Ebene programmierten DNA-Methylierungs-Layers dar.

Unter Verwendung des zweischichtigen DMR/DMCpG-Atlas als Referenz konnten Einzelzell-DNA-Methylomzustände für 312 HSPCs sowie für insgesamt 136 reife B-Zellen, T-Zellen, CFU-Es und Monozyten mit hoher Auflösung untersucht werden. Die DNA-Methylierungslandschaft der HSPCs war durch ein strukturiertes Kontinuum von Einzelzell-DNA-Methylomzuständen gekennzeichnet. Mehrere Ergebnisse legten nahe, dass die Differenzierung bei Apex-HSCs beginnt, die einen differenzierungsliniennaiven DNA-Methylomzustand besitzen. Der Ausstieg aus dem Apex-HSC-Zustand wurde durch ausgewogenes Initiieren von verschiedenen DMRs eingeleitet, die mit mehreren Differenzierungslinien assoziiert waren. Diese frühe DMR-Programmierung war streng auf bestimmte DMR-Initiierungsregionen beschränkt, die oft nur ein oder zwei CpGs umfassten. Dies steht im Gegensatz zum gegenwärtig akzeptierten Paradigma, dass funktionsrelevante DMRs typischerweise drei oder mehr CpGs enthalten. Die weitere Differenzierung von HSPCs ging mit einer kontinuierlichen, allmählich stärker differenzierungslinienspezifischen Akkumulation der Hypomethylierung einher, was zu einer fortschreitenden DMR-Erweiterung führte.

Der zweischichtige DMR/DMCpG-Atlas stellt eine wesentliche Ressource für die Untersuchung der epigenetischen Regulation des hämatopoetischen Differenzierungsprozesses dar und dient als wertvolle Referenz für die Analyse von Einzelzell-Bisulfit-Sequenzierungsdaten. Diese Arbeit unterstreicht die hochaufgelöste, progressive und stabile Natur des DNA-Methylom-Umbaus während der hämatopoetischen Differenzierung und deckt mehrere Aspekte der Struktur und des Informationsgehalts des DNA-Methyloms auf, die über die derzeit akzeptierten Paradigmen hinausgehen. Es scheint wahrscheinlich, dass ähnliche Mechanismen der epigenetischen Programmierung mittels DNA-Methylierung auch in anderen Differenzierungssystemen und verwandten Prozessen, wie der Tumorentwicklung, aktiv sind. In vielen Systemen könnte das DNA-Methylom programmiert werden durch eine Kombination von systematischen, CpG-aufgelösten Mechanismen und von Mechanismen, die durch eine stochastische regionale Anhäufung von DNA-Methylierungsänderungen gekennzeichnet sind. Die in dieser Arbeit vorgestellten Analysestrategien bilden eine Grundlage für die Weiterentwicklung bioinformatischer Methoden, mit denen der reichhaltige, aber komplexe Informationsgehalt des DNA-Methyloms mit hoher Auflösung analysiert werden kann.

Contents

List of Figures **xvii**

List of Tables **xix**

List of Abbreviations **xxi**

1 Introduction **1**

1.1 Evolving models of the hematopoietic system 1

1.1.1 The classical model of hematopoiesis 1

1.1.2 Terminology: fate potential and fate of progenitor cells 4

1.1.3 The early split model of hematopoiesis: evidence of early lineage segregation 4

1.1.4 Heterogeneity of fate restriction states in the HSPC compartment . 5

1.1.5 Heterogeneity of transcriptional states in the HSPC compartment . 7

1.1.6 The continuum model of hematopoiesis 8

1.2 Epigenetic regulation of cell identity, cell function and cell state 9

1.3 Function and information content of DNA methylation 11

1.4 Analysis of DNA methylation 13

1.4.1 Profiling of DNA methylation 13

1.4.2 Computational analysis of DNA methylation data 15

1.5 Epigenetic dynamics during hematopoiesis 17

1.5.1 The continuous chromatin accessibility and histone modification landscapes of the hematopoietic system 17

1.5.2 DNA methylome remodeling during hematopoiesis 18

1.5.3 The DNA methylome state dynamics during hematopoiesis are still largely uncharted 20

1.5.4 The single-cell DNA methylation state landscape in the HSPC compartment is almost unexplored 21

1.6 Aim of the project 23

2 Results **25**

2.1 Extensive whole-methylome maps for 25 populations across the hematopoietic system . . . 25

2.1.1 Uniform alignments and bias-aware methylation calling . . . 25

2.1.2 Global methylome differences between hematopoietic populations reflect their differentiation hierarchy . . . 29

2.2 A DMR/DMCpG atlas of hematopoietic methylome remodeling at unprecedented scale . . . 32

2.2.1 Enhanced, integrated DMCpG and DMR calling with FDR control at the DMCpG level . . . 32

2.2.2 High-resolution detection and separation of adjacent focal DMRs with distinct programming patterns . . . 36

2.2.3 Leveraging genomic region classification for improved proximity-based DMR-to-gene annotations . . . 37

2.3 Characterization of lineage- and population-specific DMR programming modules . . . 40

2.3.1 Clustering analysis reveals population- and lineage-specific DMR programming modules . . . 40

2.3.2 DMR programming modules are associated with matching gene expression modules . . . 44

2.3.3 Co-regulation within CREs by DMR programming and enhancer establishment . . . 51

2.4 DMR seeding and expansion during hematopoietic differentiation . . . 53

2.4.1 DMRs expand progressively during hematopoietic differentiation . 53

2.4.2 A map of DMR expansion states across the hematopoietic system . 55

2.4.3 Widespread methylome programming in MPP populations . . . 58

2.4.4 Methylome programming often starts with small seed regions . . . 60

2.4.5 Mature cells exhibit hypomethylated seed regions in DMRs associated with alternative fates . . . 61

2.5 Hierarchical DNA methylation programming at the DMR and DMCpG level 62

2.5.1 Rationale: DMR expansion as the result of DMCpG-resolved programming within DMR intervals . . . 62

2.5.2 Characterization of lineage- and population-specific DMCpG programming modules . . . 62

2.5.3 A novel, hierarchical approach for the annotation of DNA methylation programming patterns . . . 65

2.5.4 Terminology: DMR and DMCpG programming . . . 65

2.5.5 DMCpG-resolved programming within DMR intervals is a ubiquitous mechanism . . . 66

2.5.6 Correlation between the extent of early DMCpG programming and the breadth of DMCpG programming across the mature cell types . 69
2.5.7 Successive programming of distinct DMCpGs within DMRs underlies DMR expansion 71
2.5.8 Typical mechanisms of DMCpG-resolved programming within DMRs 75
2.5.9 High compactness and clear separation of DMR and DMCpG clusters independent of DMR buildup or DMCpG location 77
2.5.10 Systematic quantification of DMCpG programming pattern co-occurrence within DMR regions 78
2.5.11 Epigenetic memory of early alternative fate exploration is maintained throughout differentiation in the form of partially expanded DMR states 81
2.5.12 Each DMCpG programming pattern is associated with specific transcription factors 83
2.6 Highly resolved, hierarchical DMR/DMCpG programming in single cells . 88
2.6.1 Engineering of a single-cell bisulfite sequencing analysis pipeline to generate high-quality methylome maps for 312 HSPCs 88
2.6.2 Dual-layer DMCpG sets: novel features for scBS-seq analysis with unprecedented resolution capabilities 89
2.6.3 Mapping the structured continuum of single-cell DNA methylome states in the HSPC compartment 91
2.6.4 Staggered activation of DMCpG programming modules underlies progressive DMR expansion in single cells 94
2.6.5 Highly-resolved characterization of cell type-specific DNA methylome states in mature hematopoietic cell types 96
2.6.6 Exploring partial DMR expansion and the origins of widespread epigenetic memory in single mature cells 99
2.7 Publications, manuscripts and open source software packages 101
2.7.1 Manuscripts in preparation 101
2.7.2 Publications of data analysis contributions in collaboration projects 101
2.7.3 Software packages for complex data visualizations in Python 101
2.7.4 Automated, multidimensional M-bias filtering with bistro 102
2.7.5 Other software packages 103

3 Discussion **105**
3.1 A comprehensive atlas of DNA methylome remodeling during hematopoietic differentiation 105
3.1.1 Generation of genome-wide, high coverage DNA methylation maps for 25 hematopoietic populations 105

3.1.2 A novel, dual-layer atlas capturing hierarchical DMR and DMCpG programming, with rich annotations 106
3.2 Integrated DMCpG and DMR calling with FDR control at the DMCpG level 107
3.2.1 Technical considerations: FDR control and dispersion estimates in a multi-group setting with few replicates 107
3.2.2 An innovative procedure for integrated DMR and DMCpG calling and filtering 108
3.2.3 Advantages and limitations: a focal DMR model and a sensitivity tradeoff 109
3.2.4 High coverage data enable conservative, autosome-wide DMCpG detection while retaining high sensitivity 110
3.3 Hierarchical hypomethylation dynamics during hematopoiesis 110
3.3.1 Methylome remodeling during hematopoiesis predominantly involves unidirectional loss of methylation 110
3.3.2 Capturing hierarchical hypomethylation dynamics through DMR clustering analysis 111
3.3.3 A functional role for intermediate DMR methylation levels occurring during progressive methylation loss 113
3.3.4 Dissecting functionally distinct hypomethylated regions for improved cell type characterization 114
3.3.5 Limitations of enrichment-based DMR cluster characterizations . . 114
3.4 Specific roles for DNA hypermethylation during hematopoiesis 115
3.4.1 Loss of stemness is accompanied by lineage-independent gain of methylation 115
3.4.2 Lineage-specific gain of methylation occurs only for lymphoid populations 115
3.5 DMR expansion dynamics reveal mechanisms of hierarchical methylome programming 116
3.5.1 DMR expansion dynamics in differentiation systems could have high information content but are underexplored 116
3.5.2 Novel methods allow tracking of DMR expansion dynamics across bulk populations and single-cell clusters 117
3.5.3 Neglecting to align for asymmetric DMR expansion causes artificial DMR symmetry and obscures DMR expansion dynamics 119
3.5.4 Progressive DMR expansion is a common mechanism of DMR programming during hematopoietic differentiation 120
3.5.5 Progressive DMR expansion is the result of systematic, heterogeneous programming of individual DMCpGs within DMRs 120

3.5.6 DMR programming starts in small seed regions, demanding refined data analysis strategies 122
3.6 From a regional to a hierarchical model of DNA methylation programming 123
3.6.1 The classical model of regional DNA methylation programming . . 123
3.6.2 Findings challenging the classical model of regional DNA methylation programming 124
3.6.3 A novel, hierarchical model of DNA methylation programming . . . 125
3.6.4 A new paradigm for the analysis of DNA methylation data 127

4 Methods 131
4.1 Genome-wide DNA methylation profiling using T-WGBS for bulk populations and scBS-seq for single cells 131
4.1.1 Isolation and T-WGBS of 25 hematopoietic bulk populations 131
4.1.2 Alignment, methylation calling, and quality control for the T-WGBS samples 132
4.1.3 Isolation and scBS-seq of LSK cells, LSK CD150$^+$ cells, and cells of different mature cell types 133
4.1.4 Alignment, methylation calling, and quality control for the scBS-seq samples 133
4.2 Construction of a genome-wide dual-layer DMR/DMCpG atlas 134
4.2.1 Integrated DMCpG and DMR calling with FDR control at the DMCpG level 134
4.2.2 Calculation of DMCpG and DMR methylation levels 136
4.2.3 Annotation of genomic regions and potential target genes 136
4.2.4 Control of replicate homogeneity 136
4.3 Clustering analysis and annotation of DMR and DMCpG programming patterns 137
4.3.1 DMR and DMCpG clustering analysis 137
4.3.2 Annotation of marked and regulated populations for the loss of methylation clusters 138
4.3.3 Grouping and naming of DMR and DMCpG clusters based on the population-specificity of their regulatory profiles 139
4.3.4 Compilation of DMR cluster target gene sets 141
4.3.5 Compilation of hierarchical DMCpG sets 141
4.3.6 Single-cell clustering analysis 142
4.4 Profiling of a three-tier single-cell RNA-seq data set covering the hematopoietic system 142
4.4.1 Generation of 10x Genomics single-cell RNA-seq data for LSK, LK, and total bone marrow cells 142
4.4.2 Clustering and cell type annotation 143

4.4.3 Differential expression testing and computation of DMR cluster target gene set expression scores . . . 144
4.5 Enrichment analysis . . . 145
4.5.1 Clustering of hematopoietic enhancer regions . . . 145
4.5.2 DMR cluster annotation through gene set and region set enrichment analysis . . . 145
4.5.3 DMR subregion-resolved transcription factor binding motif enrichment analysis . . . 146
4.6 Quantification and visualization of DMR seeding and expansion during hematopoietic differentiation . . . 147
4.6.1 Classification of DMR expansion states . . . 147
4.6.2 Asymmetry-aware visualization of DMR profiles and DMR expansion 148
4.7 Data and code . . . 149
4.7.1 Code availability . . . 149
4.7.2 Data availability . . . 150
4.7.3 Programming languages and software packages . . . 150

5 Supplementary Materials **151**
5.1 Supplementary figures . . . 151
5.2 Supplementary tables . . . 162

6 Bibliography **171**
Publications . . . 175
Manuscripts . . . 177
Software packages . . . 179
References . . . 208

List of Figures

1 Classical model of hematopoiesis. 2

2 Early split model of hematopoiesis. 5

3 The continuum model of hematopoiesis. 9

4 Epigenetic regulation acts at different genomic scales. 11

5 Whole-genome DNA methylation data for 25 hematopoietic populations. . . 26

6 DNA methylation data with high genome-wide coverage for 25 hematopoietic populations. 28

7 Hematopoietic differentiation is accompanied by lineage-specific shifts of the CpG methylation level distribution. 30

8 Global methylome similarities between hematopoietic populations reflect their differentiation hierarchy. 31

9 Schematic of the integrated DMR and DMCpG calling strategy. 33

10 Comprehensive overview of the DMR and DMCpG detection and selection workflow. 35

11 Effective identification of focal, densely regulated DMR regions. 37

12 Gene and genomic region annotations for the hematopoietic DMRs, using the innovative gtfanno algorithm. 39

13 Clustering analysis reveals population- and lineage-specific DMR programming modules. 41

14 DMR programming modules are associated with matching gene expression modules. 48

15 Co-regulation within CREs by DMR programming and enhancer establishment. 52

16 DMR regions expand progressively during differentiation. 55

17 A DMR expansion state layer for the hematopoietic DMR/DMCpG atlas. . . 57

18 Seeding and expansion of DMRs in MPP populations occurs in all DMR clusters. 59

19 DMR regulation in MPP populations is restricted to small seed regions. . . 60

20 Mature populations exhibit a mixture of DMR expansion states across the hematopoietic DMRs. 61

21 Clustering analysis reveals population- and lineage-specific DMCpG programming modules. 64

22 Heterogeneous programming of DMCpGs within DMR intervals is a common mechanism behind hematopoietic DNA methylation remodeling. 67

23 Each DMR cluster is characterized by a distinctive set of frequently occurring DMCpG programming patterns. 67

24 Within each DMR cluster, DMRs exhibit a characteristic series of DMCpG programming patterns with increasingly narrowing population specificity. . 69

25 The extent of early DMCpG programming and the breadth of DMCpG programming across the mature cell types are correlated. 71

26 Exemplary case study: subsequent pan-hematopoietic, pan-myeloid, and neutrophil-specific DMCpG programming underlies the expansion of an *Elane* promoter DMR. 73

27 The progressive expansion of individual DMRs during hematopoietic differentiation can be effectively tracked based solely on the information about the DMCpG-PPs present within the DMRs. 75

28 Characteristic combinations of DMCpG clusters underlie seeding and expansion of the DMRs in the C3, D2, and M1 clusters. 77

29 Programming within DMRs typically involves a predominant DMCpG-PP shaping the population-specificity of the DMR, in combination with less frequent, less population-specific DMCpG-PPs which indicate preceding DMR seeding and expansion steps during differentiation. 80

30 Strong correlation between the extent of early DMCpG programming and the breadth of hypomethylation across the mature populations. 82

31 DMR subregions exhibiting different DMCpG programming patterns are enriched in distinct transcription factor binding motifs. 87

32 Histogram showing the autosomal CpG coverage for 312 HSPCs. 89

33 All hierarchical DMCpG sets exhibit sufficient coverage for robust methylation level estimation. 90

34 Structured heterogeneity of DNA methylome states in the HSPC compartment reveals a prominent role for multi-lineage priming upon exit from a lineage-naive apex HSC state. 92

35 PCA analysis suggests lineage-independent DNA methylation programming upon exit from the apex HSC state, followed by an increasingly biased accumulation of lineage-specific DNA methylation programming. 94

36 Staggered programming of DMCpGs belonging to different DMCpG clusters along differentiation trajectories. 95

37 Single mature cells exhibit completed DMRs in cell type-specific DMR clusters and partially expanded DMRs in DMR clusters associated with alternative fates. 98

S1 Boxplots showing the distribution of the autosomal CpG coverage (number of methylation calls per CpG dinucleotide) in each replicate. 151
S2 Methylation levels in the Ensembl Regulatory Build intervals are highly similar between replicates of the same population. 152
S3 DMR cluster compactness and separation is maintained when DMR methylation levels are considered directly. 153
S4 DMCpG cluster compactness and separation is maintained when DMCpG methylation levels are considered directly. 154
S5 Within most DMR clusters, the predominant majority of all contained DMCpGs is characterized by one or two characteristic DMCpG programming patterns which closely match the overall programming pattern of the DMR cluster. 155
S6 Programming within DMRs typically involves a predominant DMCpG–PP shaping the population-specificity of the DMR, in combination with less frequent, less population-specific DMCpG-PPs which indicate preceding DMR seeding and expansion steps during differentiation. 156
S7 Predominant agreement between DMR expansion state classifications based on DMCpG programming patterns or based on individual DMCpG methylation levels. 157
S8 DMCpGs from the same DMCpG cluster share highly related programming patterns independently of the DMR cluster membership of the DMRs they reside in. 158
S9 DMRs from the same DMR cluster share highly related programming patterns, independently of which DMCpG clusters the individual DMRs contain. . . 159
S10 Strong correlation betweeen the extent of early DMCpG programming and the breadth of hypomethylation across the mature populations within each DMCpG cluster. 160
S11 Programming with different DMCpG programming patterns is associated with the activity of distinct transcription factors. 161

List of Tables

1 Hematopoietic differentiation involves extensive methylome remodeling. . . 34

S1 Overview of the tagmentation-based whole-genome bisulfite sequencing (T-WGBS) dataset. 162

S2 References for the surface marker definitions of the 25 hematopoietic populations analyzed by tagmentation-based whole-genome bisulfite sequencing. . 164

S3 Surface marker definitions for the 25 hematopoietic populations analyzed by tagmentation-based whole-genome bisulfite sequencing. 165

S4 Alignment statistics for the uniformly processed tagmentation-based whole-genome bisulfite sequencing data. 166

S5 CpG methylation levels and CHH conversion rates for all T-WGBS replicates. 168

List of abbreviations

ATAC-seq	Assay for transposase-accessible chromatin sequencing
BH	Benjamini-Hochberg
BKY	Benjamini-Krieger-Yekutieli
bp	Base pair
cCRE	Candidate cis-regulatory element
cDC	Conventional dendritic cells
cDC1	Conventional type 1 dendritic cells
cDC2	Conventional type 2 dendritic cells
CDP	Common dendritic cell progenitor
CFU-E	Colony-forming unit-erythroid
CLP	Common lymphoid progenitor
cMoP	Common monocyte progenitor
CMP	Common myeloid progenitor
CRE	Cis-regulatory element
DC	Dendritic cells (comprising both conventional and plasmacytoid dendritic cells)
DKFZ	Deutsches Krebsforschungszentrum (German Cancer Research Center)
DMCpG	Differentially methylated CpG
DMR	Differentially methylated region
ESLAM	EPCR$^+$CD45$^+$CD48$^-$CD150$^+$
FACS	Fluorescence-activated cell sorting
FDR	False discovery rate
GMP	Granulocyte/macrophage progenitor
GOM	Gain of methylation
HSC	Hematopoietic stem cell
HSPC	Hematopoietic stem and progenitor cell
kb	Kilobase (equal to 1000 base pairs)
Lin-	Lineage-negative
LK	Lineage-negative c-Kit$^+$
LMPP	Lymphoid-primed multipotent progenitor

LOM	Loss of methylation
LSK	Lineage-negative Sca1$^+$cKit$^+$
mb	Megabase (equal to 1,000,000 base pairs)
MDP	Monocyte/dendritic cell progenitor
MDS	Multidimensional scaling
MEP	Megakaryocyte/erythrocyte progenitor
MHT	Multiple hypothesis testing
MkP	Megakaryocyte progenitor
MPP	Multipotent progenitor
NGS	Next generation sequencing
NK cell	Natural killer cell
OCRs	Open chromatin regions
ODCF	Omics IT and Data Management Core Facility
OTP	One Touch Pipeline
PBAT	Post-bisulfite adapter tagging sequencing
PCR	Polymerase chain reaction
pDC	Plasmacytoid dendritic cells
preMegE	Pre-megakaryocyte/erythrocyte progenitor
RRBS	Reduced representation bisulfite sequencing
s.d.	Standard deviation
scBS-seq	Single-cell bisulfite sequencing
scRNA-seq	Single-cell RNA sequencing
SNR	Signal-to-noise ratio
T-WGBS	Tagmentation-based whole-genome bisulfite sequencing
TAD	Topologically associating domain
TAD	Topologically associating domain
TBM	Total bone marrow
TF	Transcription factor
TFBM	Transcription factor binding motif
TFBS	Transcription factor binding site
TSS	Transcription start site
UMAP	Uniform Manifold Approximation and Projection
UTR	Untranslated region
WGBS	Whole-genome bisulfite sequencing

Chapter 1

Introduction

1.1 Evolving models of the hematopoietic system

Various types of mature blood cells execute essential tasks within the mammalian body. For example, erythrocytes (red blood cells) supply oxygen to tissues, myeloid immune cells (macrophages, neutrophils) and lymphocytes (T cells, B cells, natural killer cells) fight infections and thrombocytes support wound healing through blood clotting. Most mature hematopoietic cells have limited lifespans and do not have the ability to proliferate [1]. The blood system must therefore constantly be replenished in a process called hematopoiesis. In a healthy adult human, approximately 2.7×10^{13} new blood cells are produced daily in order to maintain steady state levels, most of them red blood cells [2]. The majority of the adult blood cells are derived from hematopoietic stem cells (HSCs), which reside at the apex of the hematopoietic system [3, 4].

1.1.1 The classical model of hematopoiesis

Early breakthroughs in the study of the structure of the hematopoietic system were made possible by technological advances enabling multi-colored fluorescence-activated cell sorting (FACS). This technology could be used to define and isolate cell populations characterized by specific combinations of surface markers, as reviewed in [5]. The differentiation potential of these populations could then be queried by in vitro assays and in vivo transplantation studies. Several seminal studies described immunophenotypic progenitor populations identified as hematopoietic stem cells, multipotent progenitor cells, or as oligopotent hematopoietic progenitor populations capable of producing specific subsets of the hematopoietic lineages [5]. The lineage potential of these immunophenotypic populations formed the basis for a hierarchical, tree-like model of the hematopoietic system often referred to as the classic model of hematopoiesis. The concept of modeling hematopoietic differentiation through a hierarchy of populations was first introduced by Kondo et al. [6] and Akashi et al. [7]. This initial model was then progressively refined over the next two decades, by introducing more granularly defined hematopoietic cell populations and generating refined analyses of their differentiation relationships. Figure 1 presents an updated version of the classical model of hematopoiesis, comprising 30 key immunophenotypic populations characterized within the framework of this classical paradigm.

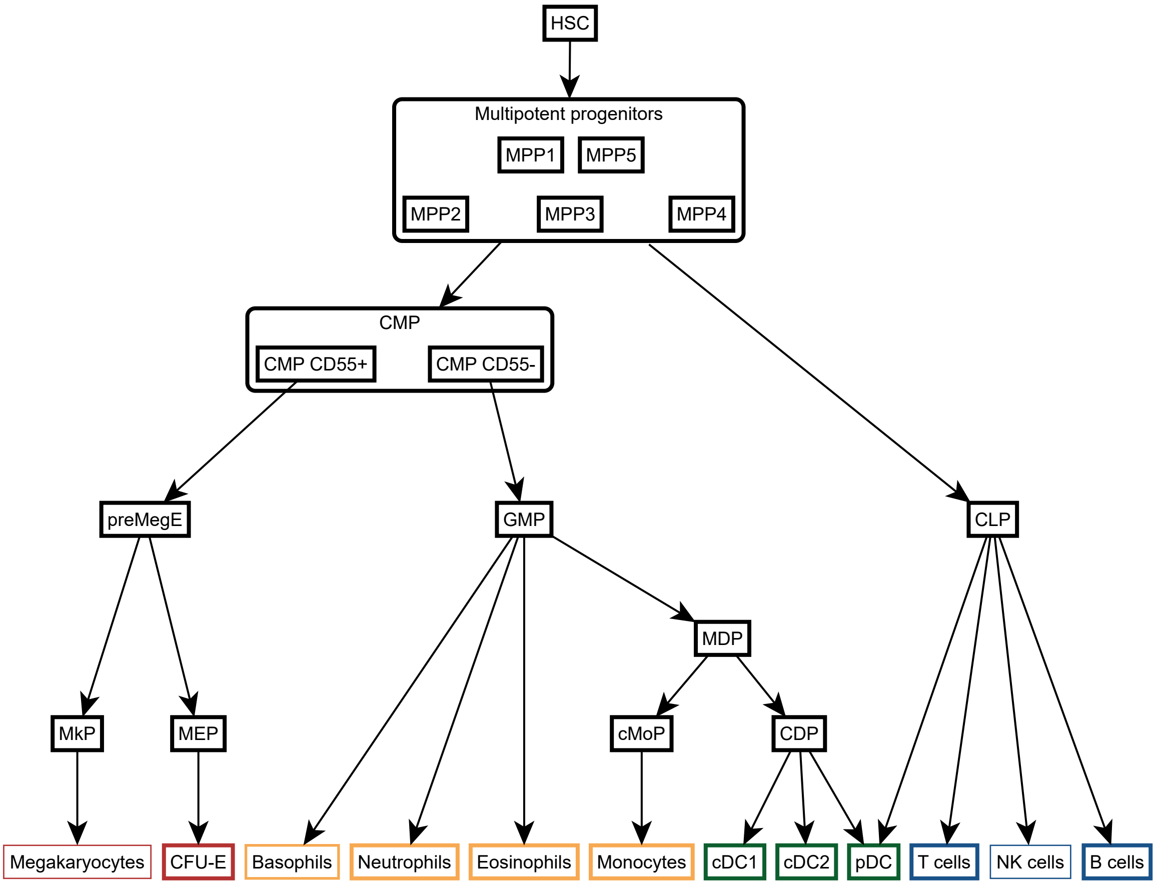

Figure 1: Classical model of hematopoiesis. The hematopoietic system is modelled as a hierarchical tree of discrete immunophenotypic populations [5, 8]. Within the framework of the model, these populations are assumed to capture relatively homogeneous groups of cells at discrete stages of hematopoietic differentiation. At the top of the differentation system are hematopoietic stem cells (HSCs). During differentiation, subsequent progenitor stages are passed in a step-wise manner, concomitant with increasing restriction of lineage production capacity. Various studies have found that the first lineage segregation occurs between the erythroid/myeloid and the lymphoid lineages [7, 9]. An alternative differentiation route involving an early split between the erythroid and lymphoid/myeloid lineages has also been suggested [9–11]. Thick frames around populations indicate that data for these populations are described in this thesis.
CDP, common dendritic cell progenitor; cDC, conventional dendritic cell; pDC, plasmacytoid dendritic cell; CFU-E, colony forming unit-erythroid; CLP, common lymphoid progenitor; cMoP, common monocyte progenitor; CMP, common myeloid progenitor; GMP, granulocyte/macrophage progenitors; MDP, monocyte-dendritic cell progenitor; MEP, megakaryocyte/erythrocyte progenitor; MPP, multipotent progenitor; MkP, megakaryocyte progenitor; preMegE, pre-megakaryocyte/erythroid progenitor.

At the apex of the classic model of hematopoiesis are hematopoietic stem cells (HSCs), from which the majority of adult blood and immune cells originate [8]. HSCs first progress through a series of multipotent progenitor (MPP) populations with decreasing self-renewal capability and increasingly biased lineage output [12–16]. Both the HSC population and these MPP populations are contained within the immunophenotypic LSK compartment [17, 18]. Taken together, these cells are often referred as hematopoietic stem and progenitor (HSPC) cells. The MPP1 and MPP5 populations have been reported to be the MPP populations which are most similar to the HSC population in terms of self-renewal capability and the potential for multi-lineage generation [12, 14, 15]. Due to the relatively low restriction of self-renewal capacity and multi-lineage fate potential in the MPP1 population, this population is alternatively referred to as the short-term HSC (ST-HSC) population. The MPP2, MPP3 and MPP4 populations have been broadly characterized as MPP populations with strongly restricted capacity for self-renewal and lineage biases towards the megakaryocyte/erythroid (MPP2), myeloid/erythroid (MPP3) and lymphoid (MPP4) lineages, respectively. [13, 14, 19, 20].

After progression through the HSPC compartment, the classical model of hematopoiesis proposes that progenitor cells further restrict their fate potential by stepwise transitions through different oligo- and bipotent progenitor populations. An initial breakthrough for the characterization of the hematopoietic differentiation trajectories after the HSPC compartment was the identification of the common myeloid progenitor (CMP) [7] and common lymphoid progenitor (CLP) populations [21]. The CMP population can differentiate into various myeloid and erythroid cell types. The CLP population gives rise to all lymphoid lineages, including B cells, T cells, and natural killer (NK) cells. The identification of the CMP and CLP populations has suggested an early lymphoid versus myeloid/erythroid lineage split during hematopoietic differentiation. Other findings have suggested an alternative differentiation route involving an early split between the erythroid lineage and the myeloid/lymphoid lineages [9–11]. The question of how the fate potential for the different hematopoieic lineages starts to segregate during early differentiation is still actively researched today. The concept of a progenitor cell population with oligopotent potential for the erythroid and myleoid lineages was further challenged by the finding that the CMP population can be functionally divided based on the CD55 surface marker [22]: The CMP CD55^{+} population demonstrates a propensity towards erythroid and megakaryocyte lineages, whereas the CMP CD55^{-} population predominantly gives rise to monocytes, dendritic cells and granulocytes.

Below the level of the CMP and CLP populations, additional oligo- and bipotent progenitor populations have been identified and incorporated into the hematopoietic differentiation model as further stepwise fate decisions. Cells originating from the CMP population can give rise to granulocyte/macrophage progenitor (GMP) cells [9, 23], which can differentiate into granulocytes, including neutrophils, eosinophils, and basophils. GMPs also generate monocytes and dendritic cells through the monocyte/dendritic cell progenitor (MDP) population [24]. MDPs can differentiate into the common monocyte progenitor (cMoP) population, which then gives rise to monocytes. Alternatively, MDPs can differentiate into the common dendritic cell progenitor (CDP) population, which gives rise to conventional type 1 and type 2 dendritic cells (cDC1 and cDC2) as well as to plasmacytoid dendritic cells (pDCs) [25–27].

Notably, pDCs can also develop through a lymphoid differentation route starting from lymphoid progenitors [28]. Cells from the CMP population can alternatively differentiate into a bipotent state with erythroid/megakaryocyte fate potential. This state is captured in the pre-megakaryocyte/erythrocyte progenitor (preMegE) population [23]. PreMegE cells further differentiate into megakaryocytes via the unipotent megakaryocyte progenitor (MkP) population or into erythroid cells through the colony-forming unit-erythroid (CFU-E) population [23]. The megakaryocyte/erythroid progenitor (MEP) population was initially introduced as a bipotent progenitor as indicated by its name. However, recent evidence suggests that the MEP population predominantly consists of unipotent erythroid progenitors [7, 29] .

1.1.2 Terminology: fate potential and fate of progenitor cells

Cell transplantation experiments are commonly employed to investigate the range of progeny that progenitors cells can produce. However, the behavior of a progenitor cell in a transplantation setting, which represents considerable stress for the transplanted cell and the host, differs significantly from the behavior of cells in an unperturbed setting, as reviewed in [8]. Adopting the terminology introduced by Haas et al. [8], this thesis distinguishes between two concepts: the fate potential and the fate of a progenitor cell. Fate potential refers to the capability of a progenitor cell to give rise to different mature blood cell types when compelled to expand in a transplantation setting, whereas the fate of a progenitor cell refers to its lineage output under homeostasis in an unperturbed in vivo environment.

1.1.3 The early split model of hematopoiesis: evidence of early lineage segregation

Technological advances have enabled researchers to explore the heterogeneity of fate potentials, fates, and molecular cell states of individual cells within the HSPC compartment and downstream progenitor populations. The first significant insights into the heterogeneity within these immunophenotypically defined progenitor populations emerged from the study of the CMP, MEP, and GMP populations. Various experimental strategies were applied to characterize cellular heterogeneity within these populations, including the scRNA-seq-based characterization of large number of individual cells [30, 31], in vitro single-cell lineage output assays [32], tracing of cell potential in transplantation-based experiments [33], and in situ tracing of cell fates under homeostasis using endogeneous barcodes [20]. Taken together, these studies have established that uni-lineage restriction largely occurs before the immunophenotypic CMP, MEP and GMP populations. Consequently, the CMP, MEP, and GMP populations are mainly composed of cells exhibiting uni-lineage fates, uni-lineage fate potentials, and uni-lineage-associated transcriptome states. The MEP gate was found to predominantly contain uni-lineage restricted erythrocyte progenitors [29] instead of bipotent megakaryocyte/erythroid progenitors as originally reported [7, 23]. Furthermore, single cell transcriptome-based clustering analysis of the cells in the CMP and GMP gates revealed distinct, uni-lineage-associated transcriptome states for neutrophil, basophil, eosinophil, monocyte, dendritic cell (DC), and megakaryocyte (MK) progenitor cells. Subsequent studies reported similar findings on uni-lineage associated transcriptome cell states for the human hematopoietic system [31]. Collectively, these findings led to the development of a revised, early split model of hematopoiesis in the adult bone marrow, as reviewed in [8]. This model proposes a hierarchical system with two primary tiers: a top tier encompassing multipotent stem and progenitor cells and a bottom tier consisting of uni-lineage restricted progenitors (Figure 2).

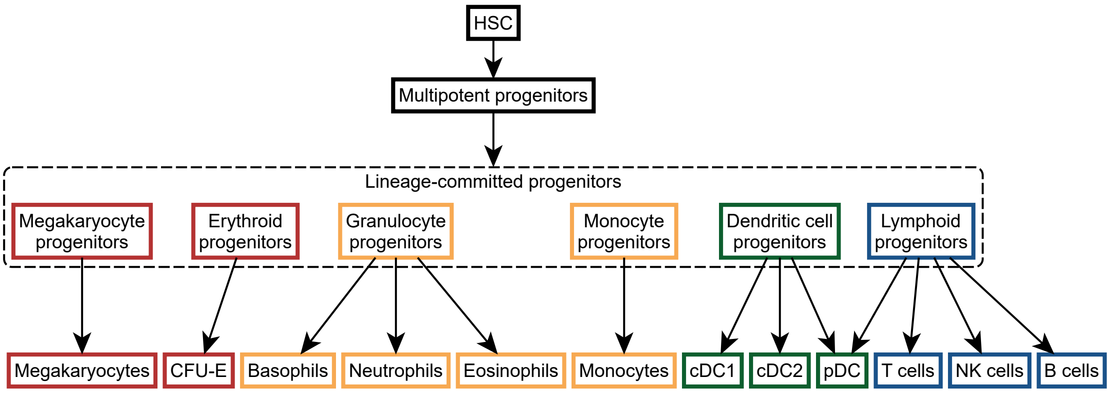

Figure 2: Early split model of hematopoiesis. The hierarchy of hematopoietic progenitors is modeled with two primary tiers: a top tier encompassing multipotent stem and progenitor cells and a bottom tier consisting of uni-lineage restricted progenitor cells. Cells in the second tier exhibit uni-lineage-restricted cell fates, as well as uni-lineage-associated transcriptome states. Figure adapted from schematic depictions of the early split model presented in [8] and [32].

1.1.4 Heterogeneity of fate restriction states in the HSPC compartment

Functional and immunophenotypic definition of the HSC and MPP cell identities

HSCs are functionally defined by two primary properties; first, by their ability for self-renewal, i.e., their ability to produce new HSC daughter cells through cell division. This ability is crucial for the life-long maintenance of hematopoiesis [1]. Second, by their ability to generate a wide range of mature blood cell types [8]. HSCs can differentiate into MPP cells, which retain a reduced capacity for self-renewal and multi-lineage cell production. The self-renewal potential and fate potential of a progenitor cell is typically assessed by measuring its ability to sustain multi-lineage engraftment in serial transplantation assays [34]. HSCs are capable of maintaining their repopulation potential for an extended period of time over more than one round of serial transplantation, while MPP cells lose their ability for multi-lineage engraftment over successive transplantations, or show restricted or biased fate potential and reduced repopulation capacity even after a single transplantation [14, 15]. The cell fates of HSC and MPP cells can also be tracked in situ in unperturbed hosts, typically by tracing endogenous barcodes, as reviewed in [35]. Importantly, the experimental profiles of individual HSPCs resulting from such experiments exhibit continuous spectra of self-renewal and multi-lineage production capabilities. The cutoff underlying the discretization of these continous spectra into a HSC population and progressively more restricted MPP populations is to a certain degree arbitrary [8]. As such, the experimental classification of HSCs based on functional read-outs can be conceptionalized as a label indicating cells in a state of maximal self-renewal and multilineage-production capability at the apex of the hematopoietic hierarchy, without a natural cutoff towards more differentiated states. Besides this functional definition, HSC and MPP cell populations can also be defined on the immunophenotypic level. Murine HSCs are often defined as LSK CD150$^+$ CD48$^-$ bone marrow cells. This surface marker defined population encompasses the CD34$^+$ long-term HSC population (LT-HSC) and the CD34$^-$ short-term HSC population (ST-HSC), which are differentiated by the substantially stronger long-term engraftment potential of LT-HSCs [12, 14]. However, surface marker defined populations must be treated with the caveat that they may contain heterogeneous cells, since the surface marker state of a progenitor cell is insufficient to completely infer its intracellular state and thus its cell fate and potential [8]. Furthermore, FACS gates represent discrete classification boundaries along a continuum of surface marker expression levels, analogously to the discretization of experimental read-outs of fate restriction and self-renewal outlined above.

Heterogeneity of fate restriction states in the immunophenotypic HSC population

The immunophenotypic HSC population comprises individual cells with considerably heterogeneous cell fate potentials and cell fates, as reviewed in [8]. In mice, only a small percentage of the immunophenotypic HSCs exhibit balanced multi-lineage fate potentials. Instead, the range of fate potentials observed within the HSC population is so extensive that almost every HSC appears to have a slightly different lineage output patterns concerning the types of cells produced and the dynamics of their production [8, 36–40]. Within the experimentally observed time windows, the observed fate potentials are often restricted to oligopotent or even unipotent lineage outputs. Furthermore, even stringently defined immunophenotypic HSCs exhibit a wide range of self-renewal potentials in transplantation experiments [36–40]. Recent studies have supplemented these findings by examining the cell fates of individual HSCs in vivo, by tracking clonal HSC lineage outputs through endogenous DNA and RNA barcodes in situ. As reviewed by Shang et al. [41], these studies confirmed that only a limited number of HSCs exhibit balanced multilineage fates [20, 42]. Many immunophenotypic HSCs exhibited broadly heterogeneous oligopotent and unipotent cell fates, in line with the findings on the HSC cell fate potentials from transplantation studies. Among the oligopotent cell fates, recurring patterns of fate restriction have emerged, in particular the existence of lymphy-myeloid-erythroid cell fates and myelo-erythroid cell fates. Furthermore, a considerable fraction of the immunophenotypic HSCs exhibit a unipotent megakaryocyte fate.

These cells often generate lineage-restricted MkP cells through a direct differentiation route, bypassing intermediate MPP cell states [20]. While these cells exhibit megakaryocyte cell fates, they typically exhibit multi-lineage fate potentials in transplantation experiments [20]. In summary, these discoveries challenge the traditional notion of a uniform population of HSCs with consistently high self-renewal potential and multipotency at the apex of the hematopoietic system, as formulated in the classical model of hematopoiesis. Rather, from the very top of the adult hematopoietic hierarchy in the bone marrow, there appears to be a spectrum of cells exhibiting a gradually decreasing capacity for self-renewal and multi-lineage output.

Heterogeneity of fate restriction states in the immunophenotypic MPP compartment

Similar to immunophenotypic HSCs, immunophenotypic MPP cells exhibit a high degree of heterogeneity with respect to their cell fate and cell fate potential [8]. Studies using single-cell in vitro assays [32], single-cell transplantation [33], and in situ lineage tracing [20] have revealed that MPP cells predominantly do not possess multilineage cell potential and cell fate. The MPP compartment is mainly characterized by a fraction of oligopotent cells in combination with progenitor cells with uni-lineage cell fates and cell fate potentials. In particular, the divergence between myeloid and erythroid lineages seems to develop within the HSPC compartment, upstream of the immunophenotypic CMP population which was originally proposed as the progenitor stage after which this lineage segregation occurs [7]. The exhibited cell fates and fate potentials form a complex mosaic of lineage outputs and reconstitution dynamics. Among the oligopotent MPP cells, recurring patterns of fate restriction have been observed, including cells with erythroid/myeloid, myeloid/lymphoid, and lymphoid/erythroid/myeloid lineage outputs, consistent with reports of immunophenotypically definable erythroid/myeloid and lymphoid/myeloid biased MPP subpopulations [11]. In addition to the direct MkP generation from HSCs, MPP cells with unilineage megakaryocyte fate have also been identified, representing a second differentiation route from the HSPC compartment towards megakaryocyte production.

1.1.5 Heterogeneity of transcriptional states in the HSPC compartment

Recent advances in single-cell transcriptomics have enabled researchers to densely sample a large number of cells across various differentiation stages within the hematopoietic system. Several studies have demonstrated that these cells can be used to construct a continuum of cell states. In this thesis, this continuum of cell states is referred to as the transcriptome state manifold, as proposed by [35]. This term references both the high-dimensional nature of the expression space in which cells are compared and the low-dimensional surface or graph representations commonly used for visualization of the state continuum [43, 44]. Initial studies examining the hematopoietic transcriptome state manifold have yielded differing assessments regarding the extent of transcriptional lineage-priming within HSPCs. While some studies observed no discernible patterns of lineage-associated gene expression [31], others have reported structured lineage-associated gene expression in HSPCs. For example, Giladi et al. [45] identified a hematopoietic core transcriptional signature for apex HSCs and demonstrated the anticorrelated expression of an erythroid gene module and a joint lymphoid/myeloid gene module within early progenitor cells [45]. These studies have been limited by the constraints of transcriptome state manifold analysis, which only permits population-level cell fate prediction and requires experimental validation to assess the fates and fate potentials of individual cells. Subsequent projects have combined transcripome state manifold analysis with lineage tracing using endogenous RNA barcodes [46, 47]. These studies identified multi-lineage progenitor cells, myeloid-erythroid progenitor cells, myeloid-lymphoid progenitor cells, and uni-lineage-restricted progenitor cells within the HSPC compartment. Although cells with differing fates are not strictly separated on the transcriptome state manifold, the average geodesic distance between oligopotent cells sharing similar cell fates is smaller than that between randomly drawn oligopotent cells, suggesting the presence of identifiable lineage-priming-associated transcriptome states in early progenitor cells. By leveraging this weak clustering of early progenitor cells with related fates, Pei et al. [47] identified transcriptional differences between lineage-primed HSPC cell states, proposing that early myeloid-erythroid restricted progenitor cells differ transcriptionally from other early progenitor cells, for example through reduced expression of lymphoid lineage marker genes. In conclusion, HSPCs can be situated within a structured continuum at the apex of a hematopoietic transcriptome state manifold. Though early progenitor cells with different cell fates may be partially differentiated by the expression of lineage-associated markers, the relationship between HSPC fate restriction and transcriptome state is insufficient for predicting HSPC fates based solely on their transcriptome state [35].

1.1.6 The continuum model of hematopoiesis

The extensive data on the heterogeneity of fate restriction states in the HSPC compartment, as well as a wealth of data establishing the existence of a continuous transcriptome state manifold underlying hematopoietic differentiation, have revealed a remarkable breadth of functional and molecular diversity between HSPCs. These findings have led to the development of the continuum model of hematopoiesis [8, 20]. In this model (Figure 3), the hematopoietic differentiation system is envisioned as a continuous landscape of cellular states, where the concept of cellular state may refer to the ability for self-renewal, the fate restriction state, or molecular states such as the transcriptome state. Differentiation within the continuum model does not involve step-wise passage through discrete intermediate progenitor stages. Instead, during differentiation, hematopoietic cells continuously acquire changes in their molecular states, accompanied by progressive fate restriction. Under this model, immunophenotypic progenitor populations are viewed as surface marker-guided samplings of specific subregions within the differentiation landscape, containing heterogeneous collections of cells that reflect the portions of the landscape being sampled. The sampled cell states contained in different immunophenotypic populations may overlap. Particularly strong overlap is expected for the MPP1-5 populations, which sample parts of the early HSPC differentiation landscape [20]. These populations contain heterogeneous cell fate restriction states, demonstrating the continuous nature of hematopoietic differentiation at early progenitor stages as well as the early emergence of lineage restriction along the differentiation continuum.

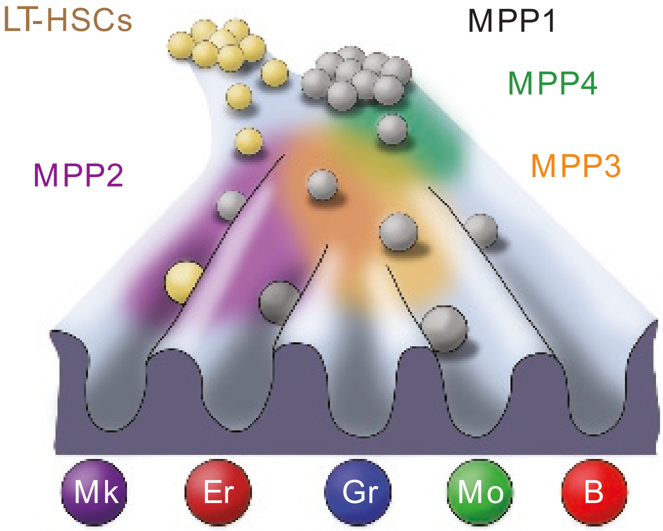

Figure 3: The continuum model of hematopoiesis. The hematopoietic differentiation system is envisioned as a continuous landscape of increasingly fate-restricted cell states, associated with an underlying continuum of molecular cell states, for example with regard to transcriptome states. Immunophenotypic progenitor populations are viewed as surface marker-guided samplings of specific subregions within the differentiation landscape, containing heterogeneous collections of cells that reflect the portions of the landscape being sampled. The MPP1-5 populations comprise overlapping subregions of the early hematopoietic differentiation continuum, and thus contain a continuum of cell fate restriction states, including cells with oligopotent fates as well as a considerable fraction of cells with uni-lineage fates. Illustration by Rodriguez-Fraticelli et al. [20] reproduced with permission.
Mk, megakaryocytes; Er, erythroid cells; Gr, granulocytes; Mo, monocytes; B, B cells.

1.2 Epigenetic regulation of cell identity, cell function and cell state

Epigenetic regulation occurs through multiple epigenetic layers

Multiple epigenetic layers in a cell regulate its function and transcriptomic and proteomic cell state without alteration of its DNA sequence. Epigenetic regulation often involves coordinated changes of stable epigenetic marks at multiple loci across the genome, which persist through cellular division [48]. These epigenetic layers include histone modifications, control of region-level chromatin accessibility, control of large-scale three-dimensional chromatin structures (e.g., through the formation or modulation of topologically associating domains, or TADs), RNA modifications, RNA factors (such as non-coding RNA), protein factors (such as transcription factors), and DNA base modifications. The most prevalent nucleotide base modification in mammalian cells is the covalent attachment of a methyl group to the C5 carbon of the cytosine nucleotide, resulting in 5-methylcytosine. In mammalian cells, cytosine methylation primarily occurs in a CpG sequence context. Throughout this text, the term “DNA methylation (DNAme)” refers to the generation of 5-methylcytosine in a CpG sequence context, unless explicitly stated otherwise. The process of cytosine methylation is catalyzed by three enzymes belonging to the DNA methyltransferase (DNMT) family, namely DNMT1, DNMT3A, and DNMT3B [49]. While DNMT3A and DNMT3B are responsible for the de novo establishment of DNAme, DNMT1 maintains DNAme during DNA replication. Methyl groups can also be removed through a process called active DNA demethylation [49]. This process is initiated by enzymes from the ten-eleven translocation (TET) family, including TET1, TET2, and TET3. During DNA demethylation by TET enzymes, 5-methylcytosine is oxidized, yielding 5-hydroxymethylcytosine (5hmC), which can act as an informative, albeit transient, mark by itself. The different epigenetic layers introduced here are intricately interconnected and act in concert to regulate cellular function and gene expression.

Epigenetic regulation occurs at different genomic scales

Epigenetic regulation can act at varying levels of resolution, from large-scale genomic windows to nearly nucleotide-resolved programming (Figure 4). This regulation is centered around the establishment, maintenance, and control of the activity and interplay of cis-regulatory elements (CREs), including promoters, enhancers, silencers, and insulators. Such CREs have an essential role in the regulation of cell type-specific gene expression. On a per-CRE level of regulation, the establishment of individual CREs and modulation of their activity is controlled through concerted changes in DNAme, chromatin accessibility, histone modifications, and transcription factor (TF) binding, reviewed in [50]. Inactivated promoter regions are typically marked by DNA hypermethylation and repressive histone marks, such as H3K27me3, while active promoter and enhancer regions are generally marked by DNA hypomethylation, accessible chromatin states, and activating histone modifications, such as H3K27ac, and H3K4me3 (for active promoters) or H3K4me1 (for enhancers). The regulation of genes is often controlled by groups of several CREs concentrated in a window of around 100kb around the transcription start sites (TSS) of the genes, which act together to provide fine-tuned gene expression [51]. While the H3K27me3 repressive mark is associated with polycomb-repressed states involved in enhancer and promoter regulation, the H3K9me3 repressive mark is associated with more stably repressed heterochromatin states, often affecting larger genomic regions, reviewed in [52, 53]. Finally, the regulation of TADs represents a regulatory mechanism involving large-scale, 3D chromatin structure states, reviewed in [54]. TAD regions range from hundreds of kilobases to several megabases and may facilitate enhancer-promoter interactions within their domain. On the other hand, epigenetic information can also be encoded at very high resolution, even within individual CREs. In this regard, the DNAme layer stands out because of its ability to encode information on top of the DNA through covalent base modifications, theoretically allowing highly stable epigenetic programming with near base pair resolution (limited by the placement of CpG dinucleotides in mammalian cells). This ability can, for example, be used to control access to individual transcription factor binding sites (TFBSs), as detailed in the following section. In summary, epigenetic regulation is fundamentally characterized by the interplay of regulatory mechanisms acting at different levels of resolution, spanning from the megabase scale to the scale of a few base pairs. The role of DNAme programming may extend to the latter level of very high resolution.

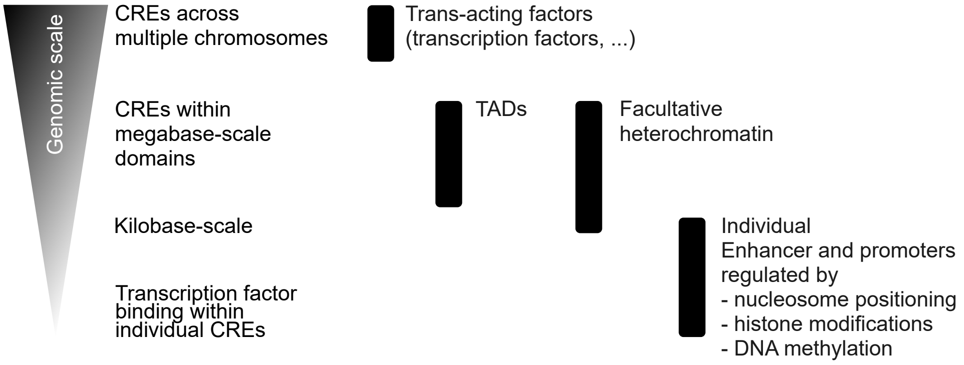
Genomic scale
CREs across multiple chromosomes
CREs within megabase-scale domains
Kilobase-scale
Transcription factor binding within individual CREs
Trans-acting factors (transcription factors, ...)
TADs
Facultative heterochromatin
Individual Enhancer and promoters regulated by
- nucleosome positioning
- histone modifications
- DNA methylation

Figure 4: Epigenetic regulation acts at different genomic scales. An important aspect of epigenetic programming is its ability to regulate the activity and interplay of cis-regulatory elements (CREs) at different genomic scales. Epigenetic programming resolution levels include the coordinated programming of various CREs across multiple chromosomes through transcription factors, the promotion of CRE interactions within large topologically associating domains (TADs) at the scale of several hundred kilobases to megabases, the control of individual CREs, and the programming within individual CREs to influence the binding of different transcription factors. As a DNA base modification, DNA methylation is theoretically uniquely poised for highly resolved programming within individual CREs through modulation of individual transcription factor binding sites.

1.3 Function and information content of DNA methylation

DNA methylation and transcription factor binding

Genome-wide studies have shown that a substantial number of TFs are sensitive to DNAme [55, 56]. DNAme can either inhibit or promote TF binding, depending on the specific TF and the modulated TFBS. For example, certain methyl-binding proteins, such as MeCP2, preferentially bind to methylated CpG sites and play a role in transcriptional repression [57]. While DNAme can regulate the access of TFs to their binding sites, TFs can also actively change the DNAme status at their binding sites. Pioneer TFs, such as FOXA1 [58], REST, and CTCF [59], can access their binding sites even in compacted chromatin with repressive marks. These pioneer TFs contribute to establishing a permissive chromatin environment for subsequent binding of other TFs, which may include the initiation of active DNA demethylation [49, 60]. Taken together, pioneering TFs can establish seed regions for CREs, which are later further activated and expanded by the recruitment of additional factors. The modulation of DNAme by pioneering TFs exemplifies highly resolved, progressive DNAme programming through seeding and subsequent extension of DNA hypomethylation within a larger locus capable of acting as a CRE.

DNA methylation and gene expression

The relationship between DNAme and gene expression is complex. DNAme can have both correlated or anticorrelated association with gene expression, depending on the genomic location and context [61]. Promoter methylation is generally associated with gene silencing. It effects gene silencing by inhibiting binding of activating TFs and recruiting inhibitory methyl-binding proteins [62]. Conversely, DNAme in gene bodies is often correlated with actively transcribed genes. This may support transcriptional fidelity by preventing spurious initiation from alternative promoters [63]. Hypomethylation of enhancer or promoter regions is generally associated with increased expression of the target genes. But many studies have only found weak correlation between gene-associated differentially methylated regions (DMRs) and gene expression, limited to a subset of the overall detected differentially expressed genes in the respectively studied systems. Moreover, hypomethylation in CREs is generally not sufficient to drive gene expression. Instead, hypomethylation of CREs is generally hypothesiszed to provide a permissive environment for gene expression, which requires additional epigenetic signals to activate transcription, such as histone modifications and the presence of specific TFs [49, 61].

DNA methylation and epigenetic memory

Within individual cells, DNAme plays an important role in maintaining stable information about the current cell identity. DNAme can furthermore encode information about past cellular states, for example in the context of differentiation or malignant transformation. For example, DNAme can act as a stable silencing mark for regions whose expression needs to be persistently suppressed in order to maintain cell identity, even through cell divisions and when challenged by perturbed environments. Particularly, DNAme stabilizes differentiation decisions by silencing lineage-inappropriate genes through promoter hypermethylation during differentiation [61]. The stable encoding of fate restriction during differentiation already occurs at the stem cell level. For example, DNAme plays an important role in maintaining stable cell fate potential profiles of individual HSC clones across serial transplantations [64] and physiologically during self-renewal divisions. Furthermore, the DNAme state of tumor cells can reflect their development history and this epigenetic memory can provide a useful marker for cancer diagnosis [65]. Additionally, DNAme can also act as a priming mechanism facilitating repeated responses to recurring environmental cues. For instance, murine immunophenotypic LT-HSC conserve epigenetic memory of infectious challenges, allowing increased transcriptional response to repeated challenges [66].

1.4 Analysis of DNA methylation

1.4.1 Profiling of DNA methylation

Because of the central role of DNAme in controlling cell function and cell state, profiling of DNAme is an important tool in biomedical research and clinical diagnosis. A variety of methods has been established, covering different use cases.

Methylation microarrays enable quantitative interrogation of methylation levels at cytosines selected to represent informative genomic loci. Modern microarrays cover large numbers of cytosines: for humans, the MethylationEPIC BeadChip v2 array covers over 935,000 CpG sites; for mouse, the Infinium Mouse Methylation BeadChip array interrogates over 285,000 CpG sites. The cost effectiveness of microarrays offers high-throughput capabilities, which are for example useful for clinical screenings.

Next generation sequencing (NGS)-based profiling methods allow for the comprehensive profiling of DNAme at single-base resolution and can be applied for genome-wide profiling or for profiling of specific, targeted regions. Many established NGS-based methods rely on bisulfite sequencing, where bisulfite treatment of genomic DNA is used to induce the deamination of unmethylated cytosines to uracils, while methylated cytosines are protected from the conversion. Consequently, bisulfite-treated DNA retains only methylated cytosines, while unmethylated cytosines are recognized as thymines in the subsequent PCR amplification steps. The altered DNA sequence is then profiled with NGS sequencing, and the cytosine methylation state is determined by tracking cytosine to thymine conversions. A challenge with bisulfite sequencing-based protocols is that bisulfite treatment can cause random strand breaks, leading to substantial degradation of genomic DNA. Different protocols for whole-genome bisulfite sequencing (WGBS) have been developed, with progressive success in reducing the required amounts of input DNA. The starting point for this process was the publication of the WGBS protocol of Lister et al. [67]. This initial WGBS protocol involved separate steps for DNA fragmentation, adapter ligation, bisulfite treatment, PCR-based library amplification and NGS sequencing. The protocol requires large amounts of DNA (between 200 ng and 5 µg of human DNA) to achieve genome-wide coverage, which makes it unsuitable for profiling rare cell populations such as stem cells [67–69].

Reduced representation bisulfite sequencing (RRBS) decreases the amount of required input DNA and the experimental costs by specifically targeting CpG-rich regions [70]. To achieve this, the DNA is digested with methylation-insensitive restriction enzymes, commonly with MspI. MspI introduces cuts after CpG sites in a CCGG sequence context. This leads to the generation of small fragments from CpG-rich genomic regions, which can be obtained by fragment size selection prior to sequencing. These fragments are then subjected to bisulfite-based sequencing. RRBS requires less input DNA (10-300 ng) but only covers a fraction of the genome [68]. Further advances yielded protocols achieving substantially reduced input DNA requirements while maintaining genome-wide DNAme profiling: prominent examples are the tagmentation-based whole-genome bisulfite sequencing (T-WGBS) [68] and post-bisulfite adapter tagging sequencing (PBAT) protocols [71]. T-WGBS uses a hyperactive Tn5 transposase for simultaneous DNA fragmentation and adapter addition, reducing the input DNA requirement for genome-wide coverage to around 20 ng. The PBAT protocol also eliminates the need for a dedicated fragmentation step, here the DNA is only fragmented by the bisulfite treatment. Sequencing adapters are added by random priming and no further amplification steps are required. This method requires around 100 ng of input DNA for genome-wide coverage. Variations of the PBAT protocol introducing additional DNA amplification steps have further reduced this input DNA requirement.

The subsequent development of single-cell bisulfite sequencing (scBS-seq) methods [72–75] represented a significant advancement in DNAme profiling, as it allowed researchers to interrogate the methylation states of individual cells. These protocols adapt bulk sequencing strategies for single-cell analysis and can be performed simultaneously with profiling of other omics layers, to provide a comprehensive view of the cellular state. For example, Clark et al. [74] and Hui et al. [75] proposed scBS-seq protocols that use a variation of the PBAT strategy. An example of a multi-omics single-cell profiling method is the single-cell nucleosome, methylation, and transcription sequencing (scNMT-seq) protocol [76]. This approach allows for parallel profiling of open chromatin regions, DNAme, and transcriptome states. The protocol adapts the Smart-seq2 bulk RNA-sequencing protocol [77] for transcriptome profiling, labels open chromatin with a GpC methyltransferase for open chromatin assessment and performs scBS-seq as outlined by Clark et al. [74].

The application of scBS-seq methods is accompanied by substantial technical challenges. The generated data are typically very sparse, meaning that only a fraction of the genome is covered in each individual cell. While some protocols offer up to 50% genome-wide coverage in theory, achieving this level of coverage with these protocols requires very high sequencing depth, which can be cost-prohibitive for studies aiming to profile hundreds of cells [74]. Consequently, to date the typical CpG coverage in larger studies ranges between 1% and 10% [75, 78, 79]. Furthermore, the regions covered in individual cells are stochastically sampled, unless targeted protocols are employed [80]. However, these targeted protocols severely limit the fraction of the genome that can be covered. Cost limitations also constrain the number of cells that can be profiled. Currently, a few hundred to a few thousand cells can be analyzed, depending on the desired genome coverage. This limits the ability to detect rare DNAme cell states. Finally, the sparse sampling of the methylome, combined with the sparse sampling of the queried cell populations, necessitates dedicated strategies for computational analysis. In summary, scBS-seq protocols offer valuable insights into the complex epigenetic landscape of DNAme in individual cells, but they also offer significant challenges such as data sparsity, cost constraints, and the need for specialized computational analysis.

1.4.2 Computational analysis of DNA methylation data

Differential methylation calling

A key challenge in the analysis of WGBS data is the statistical inference of loci exhibiting significant methylation changes between samples. Recent efforts for the statistical detection of DNAme differences have primarily been focused on the identification of differentially methylated regions (DMRs), commonly defined to encompass genomic loci with a size between 50 bp and several hundred bps, containing several (typically at least three) co-regulated CpGs. Early efforts for the statistical detection of DMRs predominantly applied a two-step procedure [81–85]. These methods first identified differentially methylated CpGs (DMCpGs) using different statistical tests. Next, these tools detected clusters of spatially adjacent DMCpGs, which were considered to represent coherently regulated DMRs. While these methods computed DMCpG locations as intermediate computation steps, their focus lay on the identification of DMRs, which were then used as atomic DNAme features for downstream analysis. The focus on DMR intervals is underscored by the fact that many of these methods assumed strong correlation of adjacent CpGs and consequently applied smoothing, binning, or other information-sharing techniques across neighboring CpGs.

However, these first-generation DMR detection methods faced statistical challenges and limitations. The high number of tests to perform, for example individual tests for 30 million CpGs in the human genome, together with the typically low percentage of truely differentially methylated CpGs, led to a considerable multiple hypothesis testing (MHT) problem [86]. High precision could then only be achieved through a significant tradeoff of sensitivity. To avoid this tradeoff, multiple first-generation DMR calling methods completely forgo MHT correction when default settings are applied [81–83]. Furthermore, when MHT is applied in the framework of these methods, it is only possible at the level of individual CpGs, and DMRs are subsequently identified using heuristic strategies without statistical inference [87, 88]. Second-generation tools have sought to address these limitations by employing segmentation algorithms to directly find regions with significantly different methylation levels between groups . One such tool, Metilene, uses change-point detection to identify DMRs [89]. Another tool, DMRseq [87], follows a two-stage procedure: First, it identifies candidate DMRs by segmenting the genome to locate candidate loci with apparent evidence of differential methylation. Second, it performs a hypothesis test for differential methylation for each candidate DMR while controlling the false discovery rate (FDR). In summary, the development of methods for the detection of differential methylation between samples has predominantly been focused on the identification of DMRs. Controlling the FDR of DMR detection is a crucial requirement for robust analysis of DNAme programming, but remains a field of active research.

Analysis of scBS-seq data

The computational analyis of scBS-seq data is challenging. Reasons include i) the sparse, stochastic single-cell DNAme profiles generated by current methods (typically 1-10% of CpGs covered per single cell); and ii) the limited number of cells which can be sampled with the currently available experiment methods (between hundreds and thousands of cells). Currently, a variety of data analysis approaches are being explored. The proposed methods differ in their definition of what should constitute the atomic feature of the cellular DNAme state vectors. Most methods use the methylation state of genomic intervals as basic DNAme programming units.

Several studies have directly computed DNAme state vectors based on the average methylation level of fixed genomic tiling windows [90–92], or based on the average methylation level in functionally defined genomic regions [73, 93, 94]. These vectors were then used in conventional clustering workflows based for example on hierarchical clustering or community detection clustering. A toolbox for performing such scBS-seq data analysis workflows was implemented in EpiScanpy [95]. Alternatively, the methylation states of individual CpGs can be used as the elements of the DNAme state vectors. While this approach has shown promising initial results for the resolution of heterogeneity in the hematopoietic stem cell compartment [75], it has so far not been systematically explored. One method building upon the findings of Hui et al. [75] is the scMelody clustering algorithm, which uses consensus clustering based on multiple distance metrics computed on CpG element-based DNAme state vectors [96]. Taken together, various methods separate the steps for the construction of DNAme state vectors, the calculation of pairwise cell-to-cell distances, and cell clustering. Cell clusters can then be used to apply conventional DMR calling strategies for bulk WGBS populations [75].

Alternative, model-based approaches use statistical inference and machine learning to perform one or more of these data analysis steps: i) imputation of the DNAme state vectors, in predefined genomic regions or genome wide; ii) testing for differential methylation or differential variability between regions; and iii) cell clustering [97–100]. The first method demonstrating the feasibility of genome-wide imputation of DNAme based on sparse DNAme data was the deepCpG algorithm. This method is based on deep neural networks and leverages the association between DNA sequence patterns and methylation states as well as the correlation of the DNAme states between neighbouring CpGs, within and across cells [97]. Many alternative imputation approaches have since been proposed, representing gradual improvements in imputation quality, but using the same principle of leveraging horizontal (across neighboring CpGs) and vertical (across similar cells) information sharing [101, 102]. Imputed DNAme state vectors can then be used for downstream data analysis. Furthermore, the scMET algorithm applies a hierarchical beta-binomial model starting from feature vectors based on predefined genomic regions, which allows for the detection of highly variable features as well as of features which are differentially methylated between conditions [99]. Other modeling approaches, again starting from region-based DNAme state vectors, are focused on infering optimal cell clusterings [100]. Taken together, the vast majority of the modeling-based DNAme analysis approaches are based on the assumption that adjacent CpGs share similar information content, such that information from individual CpGs can be shared within genomic regions, to alleviate problems conferred by the sparsity of scBS-seq data. Many methods leverage information sharing between cells in addition, which typically requires statistical inference or prior knowledge of cells with similar DNAme states.

1.5 Epigenetic dynamics during hematopoiesis

1.5.1 The continuous chromatin accessibility and histone modification landscapes of the hematopoietic system

Epigenetic regulation is crucial for controlling gene expression programs and cell fate decisions during hematopoietic differentiation. Assay for transposase-accessible chromatin sequencing (ATAC-seq) experiments have been instrumental in characterizing changes of the chromatin accessibility landscape during hematopoietic differentiation. Such studies have constructed large atlases of chromatin remodeling between immunophenotypic hematopoietic populations, demonstrating large-scale, lineage-specific remodeling of the chromatin accessibility landscape in mouse [51] and human [103]. These studies have suggested that the chromatin accessibility landscape may reflect cell identity better than the transcriptome landscape, highlighting the potential of epigenetic analysis of differentiation-related cell states. Lineage-specific changes in chromatin accessiblity activity are tightly associated with the activity of lineage specific TFs [104]. Chromatin accessibility has also been studied at the single cell level [105, 106]. Similarly to the construction of continuous hematopoietic transcriptome state manifolds, hematopoietic progenitor cells can be placed on a continuous chromatin accessibility state manifold. Progression along differentiation trajectories on this manifold is associated with gradually increasing accessibility of binding sites for lineage-specific TFs and gradual loss of accessibility for stemness-associated TFs.

Global histone modification programs controlling the establishment and activities of lineage-specific enhancer modules have also been mapped [107]. These enhancer modules could also be linked to a hierarchical network of TFs governing subsequent steps of hematopoietic differentiation. Single cell analyses of activating and repressing histone modification during hematopoieic differentiation have suggested that the landscape of histone modifications across hematopoietic progenitor also makes up a continuous state manifold [108]. Histone modifications appear to function in a hierarchical fashion, with H3K9me3-based heterochromatin state regulation being associated with global loss of stemness occurring across all differentiation trajectories. In contrast, modulation of polycomb-repressed chromatin states (via H3K27me3) and activation of lineage-specific enhancers and promoters (via H3K4me1 and H3K4me3) appear to occur in a lineage-specific fashion.

CREs associated with lineage-specific genes may gain activating histone marks [109] and accessible chromatin states [106] prior to the activation of gene expression in hematopoietic progenitor cells. Similar findings were observed for the TFBS of a subset of important hematopoietic TFs, which may become accessible in hematopoietic progenitor cells prior to activation of the TF target genes.

Taken together, extended characterizations of the chromatin accessibility and histone modification landscapes across immunophenotypic hematopoietic cell populations and single cells have captured atlases of large-scale epigenetic remodeling during hematopoieitic differentiation. Individual hematopoietic progenitor cells appear to reside on a continuous state manifold of chromatin accessibility or histone modification states. Along this state manifold, differentiating cells gradually lose epigenetic stemness signatures, while acquiring lineage-specific signatures. Several studies have suggested a time-lag between epigenetic priming of lineage specific genes and their gene expression. This suggests that the analysis of epigenetic cell states may provide a powerful approach for the dissection of early lineage priming, with perhaps greater resolution at early differentiation states than transcriptome-based analyses. The full potential of this approach remains to be evaluated in future studies.

1.5.2 DNA methylome remodeling during hematopoiesis

Genes involved in DNAme, such as *Tet2* and *Dnmt3a*, are frequently mutated in hematological malignancies and clonal hematopoiesis [110–113]. In healthy organisms, epigenetic regulation through the DNAme layer is critical for maintenance of HSC function and for healthy hematopoietic differentiation [114–117].

Prompted by the relevance of DNAme for the control of hematopoiesis, Ji et al. [118] characterized multiple hematopoietic progenitor populations with a custom array platform covering 4.6 million CpGs across the genome. Their work identified widespread plasticity of DNAme during hematopoietic differentiation and resulted in the first large-scale atlas of hematopoietic DMRs, comprising several thousands regions. A subsequent study of human, female HSPCs, B cells and neutrophils leveraged WGBS to achieve a genome-wide characterization of hematopoietic methylome remodeling [119]. Both of these initial studies found that hematopoietic lineage specification involved both gain and loss of methylation. Hodges et al. [119] further reported that DMRs exhibit intermediate methylation states in HSPCs followed by bidirectional methylation programming, resulting in either loss of methylation compared to the HSPC level in neutrophils and gain of methylation in B cells, or vice versa. Other studies confirmed that human myeloid differentiation is associated with pronounced loss of DNAme associated with differentiation-related genes [120]. In contrast, Bock et al. assessed DNAme remodeling during hematopoietic populations across 19 murine, immunophenotypic populations through RRBS and found that DNAme changes were generally of "small, but informative", "modest" magnitude [121]. The authors further demonstrated that the TFBS for myeloid and lymphoid TFs are more strongly methylated in cells from the respectively opposing lineage. The study further concluded that DNAme and gene expression provide predominantly complementary information. Subsequently, an integrated, genome-wide analysis of DNAme and transcriptome data across the HSC and MPP1-4 populations identified a set of genes whose expression appeared to be partially regulated by DNAme programming [14, 122]. This study highlighted that substantial DNAme changes can be observed already between the early progenitor populations in the HSPC compartment. These changes comprise both gain and loss of DNAme across the MPP populations, which occur predominantly in a unidirectional, progressively increasing fashion across the MPP1, MPP2 and combined MPP3/4 populations. The important role of the DNAme state of HSPCs was further emphasized by the finding that clonally amplified HSC populations with distinct lineage potentials exhibit distinct DNAme profiles characterized by hypomethylation in lineage-associated CREs matching the HSC-specific lineage potentials [64]. A recent study, to which I contributed in parallel to the work on this thesis, demonstrated that distinct DNAme programming modules are involved in the specification of the pDC and cDC lineages [OWN1], indicating that the important role of DNAme programming extends from early differentiation steps to late lineage specification steps. The continued importance of DNAme programming in late stages of the hematopoietic differentiation process was also highlighted in a study of DNAme changes during B cell maturation [123]. Taken together, DNAme programming plays a crucial role during the hematopoietic differentiation process, from involvement in the maintenance and cell fate potential specification in HSPCs to a role in late lineage specification and cell type maturation processes.

Further technological advances allowed the interrogation of hematopoietic DNAme programming at the single-cell level, albeit with limited numbers of cells due to the cost restrictions of scBS-seq methods. A single-cell DNAme study of 122 human cells comprising cells from the immunophenotypic HSC, MPP, CMP, GMP, CLP, and an immature multi-lymphoid progenitor population (MLP0) demonstrated that the average DNAme levels of the binding sites of important hematopoietic TFs can be used to distinguish single cell DNAme states [93]. In this study, the immunophenotypic HSC population appeared as a relatively homogeneous population with high DNAme across most of the TFBS. The study further found initial evidence for heterogeneity of DNAme states within the MPP population. However, interpretation of the data is complicated by an apparent confounding effect of the cell donors, which was not explicitly addressed. The epigenetic variability between human individuals and the role of methylation quantitative trait loci (meQTLs) was more explicitly addressed in later studies [75]. Moreover, the focus on a limited set of regulatory regions underlying the study may obscure heterogeneity of HSPCs in other parts of the methylome. A subsequent study [75] attempted to further resolve the landscape of DNAme states in the murine HSPC compartment, through scWGBS analysis of 64 LSK cells and 84 EPCR+CD45+CD48−CD150+ (ESLAM) cells from the adult mouse bone marrow as well as within 121 human CD49f^{+} HSCs. The ESLAM cell population is highly purified for functionally defined HSCs with durabe repopulation activity (approximately 40% pure) [124, 125]. Using pairwise comparisons of single-CpG DNAme state vectors, the authors identified an epigenetic apex HSC state occuring in approximately 31% of the ESLAM cells and 5% of the LSK cells. Through the lense of the global CpG state vector comparison this apex HSC state appeared to reside at the top of a continuous landscape of epigenetic states in the HSPC compartment. To further investigate the functional role of DNAme programming in HSCs, Izzo et al. [126] performed multi-omics characterization experiments on individual Lin- and LT-HSCs with conditional *Tet2* or *Dnmt3a* knock-outs, including querying single cell methylome states with single-cell RRBS and single-cell ATAC-Seq in combination with a bisulfite conversion step. Disruption of DNAme programming through *Tet2* knock-out or *Dnmt3a* knock-out in single immunophenotypic LT-HSCs led to shifts towards the production of myeloid or erythroid lineages in in vitro assays, respectively. These shifts were associated with transcriptional priming signatures. The intensity of the transcriptional priming for erythroid or myeloid fates was correlated with the amount of DNAme loss in open chromatin regions (OCRs), indicating a direct relationship between the global disruption of the cellular DNAme landscapes and their transcriptome state. Moreover, the study found that *Tet2* knock-out led to an expansion of the frequency of cells with a transcriptionally-defined primitive HSC state. Taken together, the study emphasized the important role of DNAme programming in HSPCs, but did not offer further insights into the structure of the DNAme landscape in this compartment. The possibility to use single-cell DNAme analysis to differentiate epigenetic signatures towards the end of the hematopoietic differentiation system was also demonstrated, for example by applying targeted single cell methylome analysis to resolve epigenetic heterogeneity in the naive, non-switched and class-switched memory B cell populations [80].

1.5.3 The DNA methylome state dynamics during hematopoiesis are still largely uncharted

To fully understand the systems-level role of DNAme programming during hematopoietic differentiation, comprehensive, genome-wide studies of methylome remodeling during hematopoiesis are still required. Different consortia have undertaken considerable efforts on curating catalogues of CREs exhibiting differentiation-associated regulation during hematopoiesis. The Immunological Genome Project (ImmGen) [51] and the Validated Systematic Integration of Hematopoietic Epigenomes (VISION) [127] have compiled CRE atlases based on open chromatin and histone modification patterns, that do not take DNAme changes into account. The Encyclopedia of DNA Elements (ENCODE) [128] consortium has generated a large-scale CRE atlas across some mouse cell types, but did not systematically focus on broad coverage of the cell types within the hematopoietic system. The project gathered data across multiple organs and developmental stages during mouse ontogenesis, profiling DNAme, histone modifications, chromating accessibility, and CTCF binding. But the CRE atlas generated in the project is solely based on H3K4me3, H3K27ac, CTCF binding and chromatin accessibility, without consideration for DNAme. A recent study [129] has compiled an atlas of differentially methylated “blocks” (conceptionally related to DMRs) across 39 human cell types, including multiple blood cell types. However, this study provided limited resolution of the hematopoietic compartment, and was instead focused on investigating cell type-specific methylation differences between the major human cell types across multiple organs. Moreover, the atlas provided a limited amount of annotations for the identified informative methylation blocks. In summary, to my knowledge, no genome-wide CRE atlas comprising a systematic compilation of DMRs with a role in hematopoiesis exists to date.

The currently existing studies of hematopoietic DNAme remodeling have collected insufficient data for the generation of such an atlas, as they only provide limited genomic coverage through array-based assays [118, 120, 130], RRBS (which is primarily restricted to CpG-rich regions) [121], or sparse single-cell RRBS or scBS-seq assays (which only cover a small fraction of the genome in each individual cell) [75, 93]. Studies which did provide a high-coverage, genome-wide characterization of DNAme programming have been limited to few hematopoietic populations [14, 119, 122, OWN1, 123]. A recent study, with contributions from me in parallel to my work on this thesis, leveraged the high number of CpG sites contained in the recently created Infinium Mouse Methylation BeadChip array to generate a comprehensive catalogue of hematopoietic candidate CREs [OWN2]. While this work represented a major advance in the systematic study of genome-wide hematopoietic DNAme programming, it focused on providing a CRE atlas optimized for the cost-effective interrogation of hematopoietic samples through arrays. A fully genome-wide CRE atlas with highly resolved annotations of the DNAme information content is still crucially missing.

1.5.4 The single-cell DNA methylation state landscape in the HSPC compartment is almost unexplored

Although initial findings from studies of single-cell chromatin accessibility and histone modification states in early HSPCs suggest that epigenetic single-cell states may be highly informative of cell fate restriction states [105, 106, 108], our understanding of the heterogeneity of DNAme states within the HSPC compartment remains limited. To my knowledge, few systematic analyses of DNAme states in early hematopoietic progenitor cells have been attempted. Hui et al. [75] found that a primitive HSC state may be characterized by a specific DNAme signature and presented initial evidence for continuous changes in DNAme states within the HSPC compartment. However, the study also found that the structure of the DNAme landscape within the HSPC compartment is strongly influenced by the genomic regions considered, which leads to substantially different cell clustering within ESLAM and LSK populations. Farlik et al. [93], on the other hand, did not find evidence for heterogeneity in immunophenotypic HSC population. While their analysis of single immunophenotypic MPP cells provided initial evidence for epigenetic heterogeneity in the MPP compartment, this aspect was not explicitly discussed in the study. Due to the focus on the TFBS of selected TFs, the study's methodological capability for global DNAme state comparison was limited. In conclusion, while the exploration of single-cell DNAme states in the HSPC compartment represents a promising avenue of research, systematic mapping of these states has largely not been undertaken yet. This may be partially due to the technical challenges associated with analyzing sparse single-cell DNAme data, as methods for comparing and clustering sparse DNAme-based cell profiles are still in the early stages of development (section 1.4.2).

1.6 Aim of the project

This project was guided by three biologically motivated research questions that aimed to expand our understanding of the dynamic DNA methylome remodeling during hematopoietic differentiation. All of these biologically motivated subprojects required substantial software and method development efforts.

Generation of a comprehensive, deeply annotated atlas of the dynamic DNA methylome changes during hematopoietic differentiation. To the best of my knowledge, this project presents the first systematic attempt to generate a broad atlas of the sites of methylome remodeling in mouse or human. The first requirement for the generation of such a resource was the statistically robust, genome-wide identification of the DMR regions arising during hematopoiesis. Next, detailed annotations of the methylome state within these DMR regions across populations selected to broadly cover the hematopoietic system were necessary, including separate annotations of the information content of each CpG within the DMRs. Finally, the atlas had to be complemented with detailed annotations of the association of hematopoietic methylome dynamics with changes on other omics layers, including the interplay with gene expression, transcription factor binding, and enhancer establishment.

Evaluation of the role of DMCpG programming during differentiation. The functional importance and rich information content of DMR programming are well-established, along with the methods for analyzing DMR programming. On the other hand, DNA methylation inherently possesses the capability to encode information at near-nucleotide level resolution. This raises the possibility that a more nuanced understanding of DNA methylome dynamics may require considering a hierarchical system of DNAme programming, occurring both at the level of DMRs and at the level of individual DMCpG sites within these DMRs. A prime example of the highly resolved programming abilities of the DNA methylation mark is its direct and multifaceted role in the epigenetic programming at individual transcription factor binding sites. DNA methylation can regulate access of transcription factors to individual binding sites, while the DNA methylation status at transcription factor binding sites can, in turn, be influenced by the binding of specific pioneering transcription factors. To the best of my knowledge, the concept of viewing methylome programming as a hierarchical system of DMR and DMCpG programming represents a novel paradigm that has not been systematically explored thus far. This project aimed to apply and evaluate this new paradigm of DNAme data analysis in the context of murine hematopoiesis. The murine hematopoietic system is a widespread model system for the study of differentiation processes. As such, it is ideally suited for the evaluation of new paradigms of epigenetic programming during differentiation.

Exploration of the structure of the DNA methylome state landscape in single HSPCs. Technical limitations have held back attempts to resolve cell-to-cell heterogeneity of DNAme states in early hematopoietic progenitor cells. In this project, I hypothesized that a comprehensively annotated, highly resolved atlas of hematopoietic DNAme programming (the first goal of this project), enriched by a thorough investigation of how information is encoded into the DNA methylome (the second goal of this project) could provide a rich resource for the analysis of single-cell DNAme states during early hematopoietic differentiation. The exploration of the early hematopoietic DNAme state landscape was intended to further our understanding of several important aspects of DNAme programming in HSPCs, including questions such as: Is there a distinct epigenetic DNAme state signature for primitive HSCs? Can this signature be associated with the regulation of certain gene modules, interplay with certain transcription factors, or the expression of certain surface markers? How heterogeneous and structured is the landscape of DNAme states in the HSPC compartment? What is the information content of these potentially heterogeneous DNAme states? Can they be informative about early fate restriction during hematopoiesis?

Software and methods development. To address these biological research questions, substantial development effort with regards to data analysis methods and software engineering was required. A novel procedure for integrated DMR and DMCpG calling with robust statistical properties and inference capabilities in a multi-group setting was needed as a basis for the generation of a highly resolved, genome-wide atlas of DNAme programming with annotations of both DMR and DMCpG sites. Several software solutions were needed to obtain optimal annotations for the atlas, including innovative concepts for DMR-to-gene annotations and software for the visualization of complex relationships within the data. Additionally, new software for the analysis of single-cell methylome data had to be developed, including a start-to-end workflow for the alignment and methylation calling of scBS-seq data.

Chapter 2

Results

2.1 Extensive whole-methylome maps for 25 populations across the hematopoietic system

2.1.1 Uniform alignments and bias-aware methylation calling

Methylome-wide DNAme data were generated or collected from previous studies for 25 hematopoietic cell populations (Figure 5), chosen to cover i) the hematopoietic stem and progenitor cell (HSPC) compartment with the hematopoietic stem cell (HSC) population and the multipotent progenitor populations MPP1, MPP2, MPP3, MPP4, MPP5; ii) the megakaryocyte-erythroid lineage with the common myeloid progenitor (CMP) subpopulation CMP CD55$^+$ and the megakaryocyte/erythrocyte progenitor (MEP), pre-megakaryocyte/erythroid progenitor (preMegE), megakaryocyte progenitor (MkP) and colony forming unit-erythroid (CFU-E) populations; iii) the myeloid lineage with the granulocyte/macrophage progenitor (GMP), common monocyte progenitor (cMoP), monocyte, neutrophil, and eosinophil populations; iv) the dendritic cell lineage with the CMP subpopulation CMP CD55$^-$ and the monocyte-dendritic cell progenitor (MDP), common dendritic cell progenitor (CDP), conventional type 1 dendritic cell (cDC1), conventional type 2 dendritic cell (cDC2), and the plasmacytoid dendritic cell (pDC) populations; and v) the lymphoid lineage with the common lymphoid progenitor (CLP), B cell and T cell populations.

The sequencing data for nine of these 25 populations were generated in previous studies. This includes the HSC, MPP1 and MPP2 populations [14, 122] and the MDP, CDP, cMoP, cDC1, cDC2 and pDC populations [OWN1]. At least three replicates were available for all of these populations, except for the cMoP and MDP populations (two replicates). For the remaining populations, new sequencing data were generated for this study, in at least three

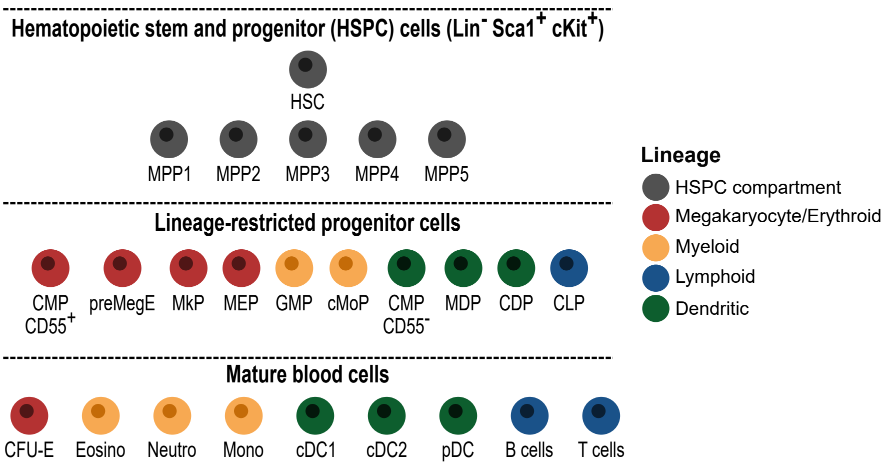

Figure 5: Whole-genome DNA methylation data for 25 hematopoietic populations. For this thesis, tagmentation-based whole-genome bisulfite sequencing (T-WGBS) data for 25 immunophenotypically defined hematopoietic cell populations were analyzed. The populations included in the analysis were chosen to provide high resolution within the hematopoietic stem and progenitor cell (HSPC) compartment and across the megakaryocyte/erythroid, myeloid, lymphoid and dendritic cell lineages.
CDP, common dendritic cell progenitor; cDC, conventional dendritic cell; pDC, plasmacytoid dendritic cell; CFU-E, colony forming unit-erythroid; CLP, common lymphoid progenitor; cMoP, common monocyte progenitor; CMP, common myeloid progenitor; Eosino, eosinophil; GMP, granulocyte/macrophage progenitors; HSC, hematopoietic stem cell; MDP, monocyte-dendritic cell progenitor; MEP, megakaryocyte/erythrocyte progenitor; MPP, multipotent progenitor; Meg, megakaryocyte; MkP, megakaryocyte progenitor; Mono, monocyte; Neutro, neutrophil; preMegE, pre-megakaryocyte/erythroid progenitor.

replicates per population. Moreover, an additional replicate for the HSC population was generated, complementing the three HSC population replicates published in earlier studies. The experimental sample generation was performed by collaborators (Methods, section 4.1.1), using the same experimental protocol as in the previous studies. Briefly, cells were isolated by fluorescence-activated cell sorting (FACS) from the bone marrow or spleen of C57BL/6J mice (aged between 8 and 12 weeks). Genome-wide DNAme was measured by tagmentation-based whole-genome bisulfite sequencing (T-WGBS) [68]. An overview of all T-WGBS samples used in this study, including their provenance, data accessibility, and number of replicates is provided in Table S1. The surface marker definitions used for the 25 populations are documented in Tables S2 and S3. The experimental generation of all samples is detailed in the doctoral thesis of Sina Stäble [131], who performed a large part of this experimental work as part of her doctoral project. Her thesis describes the surface marker definitions chosen for the different hematopoietic populations in detail and presents representative sorting schemes for all populations.

Alignments and methylation calling were performed with a uniform workflow for all samples. Read alignment was performed by the Omics IT and Data Management Core Facility (ODCF) at the German Cancer Research Center. Alignments were carried out using an updated version of the T-WGBS alignment workflow described by Wang et al. [68]. This workflow was implemented by Matthias Bieg as a Roddy Workflow as part of the automated One Touch Pipeline alignment framework [132] maintained by the ODCF. The resulting alignments yielded 2.0 ± 0.7 × 10^9(mean ± s.d.) properly paired and deduplicated reads per population. Table S4 provides a detailed overview of library size and alignment quality control parameters on the replicate and population-level.

The T-WGBS workflow introduces gap repair nucleotides during the tagmentation reaction [68]. Additional read positions can be affected by M-bias to an extent varying between samples [68, 81]. Methylation calling was performed with the bistro software package, which provides automatic detection and filtering of methylation calls affected by either problem. The bistro software package was developed by me [SOFT1] and its capability of automatic M-bias removal has been successfully applied in several projects [OWN1, OWN3]. The bistro methylation calling algorithm was parametrized to always remove the T-WGBS gap repair nucleotides and to automatically identify and remove methylation calls affected by M-bias. This automatic filtering strategy specifically removes only those methylation calls at a given read position which are likely affected by M-bias, and thus retains as much coverage as possible at the read position.

I merged the methylation calling information per CpG motif, assuming that CpG motif methylation is generally symmetric. For population-level analyses, I further merged the CpG methylation calls across all replicates. The mean CpG coverage (the number of methylation calls obtained for a CpG dinucleotide) was comparable for all populations except the MPP3, MPP4 and MPP5 populations (Figure 6A). To allow for the highly resolved analysis of heterogeneity within the MPP3, MPP4 and MPP5 populations, these populations were sequenced with deeper coverage. The resulting average CpG coverages were 67 (MPP3), 78 (MPP4) and 62 (MPP5). The CpG coverage of the remaining populations ranged from 19 (cMoP) to 48 (B cells), with an average of 35 ± 10 (mean ± s.d.). Three replicates of the HSC population have been previously published [14, 122]. Reprocessing of these replicates yielded a CpG coverage of 34. For this study, an additional HSC replicate was generated, to increase the coverage for this important reference population. By merging the published and newly generated sequencing data, the CpG coverage for the HSC population could be increased to 44. The CpG coverage distributions for all populations are described in Table S1. Similar CpG coverage for the replicates within a population can be important for certain statistical tests, such as differentially methylated region (DMR) calling. I therefore verified that the replicates within each population exhibited comparable coverage levels (Figure S1 and Table S1).

To quantify how extensively the methylome was covered in our dataset, I computed for each population the proportion of CpGs exceeding various CpG coverage thresholds (Figure 6B). Across all populations, 96% to 97% of the autosomal CpGs were covered at least once. On average, across the populations, 93.07 ± 1.88% (mean ± s.d.) of the autosomal CpGs had a minimum of ten methylation calls. Thus, all populations exhibited coverage sufficient for

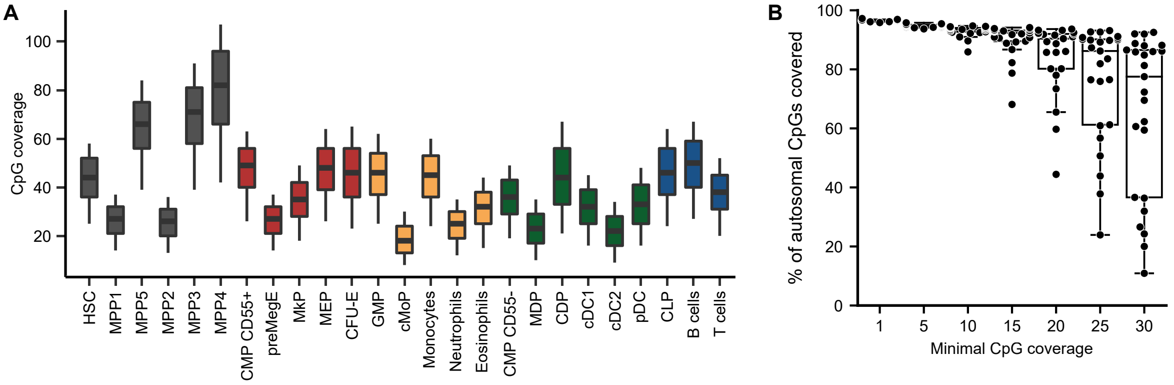

Figure 6: DNA methylation data with high genome-wide coverage for 25 hematopoietic populations. T-WGBS was performed on three or more replicates for all populations, except for the cMoP and MDP populations (two replicates). A uniform alignment and methylation calling pipeline was applied to the T-WGBS data of each replicate. Methylation calls from all replicates within a population were combined for population-level analyses. Coverage statistics per replicate are shown in Figure S1. Alignments were performed by the Omics IT and Data Management Core Facility (ODCF) at the German Cancer Research Center, using an updated version of the T-WGBS alignment workflow described Wang et al. [68]. Methylation calling was performed using the bistro software package [SOFT1], which offers automatic detection and filtering of methylation calls affected by gap repair nucleotides or M-bias.
(A) Box plots showing the distribution of the autosomal CpG coverage (number of methylation calls per CpG dinucleotide) in each population. Whiskers represent the 10th and 90th percentiles.
(B) Percentage of CpGs whose coverage exceeds different CpG coverage thresholds in the hematopoietic populations. Individual percentages for each population are displayed as dots. Box plots summarize the distribution of the percentages across the populations, whiskers extend to the furthest observation within 1.5 times the interquartile range.

statistical analyses across the vast majority of the autosomal CpGs. Furthermore, across the populations, on average 84.58 ± 12.24% (mean ± s.d.) of the CpGs surpassed twenty methylation calls, with over 86% of the autosomal CpGs exhibiting at least 20 methylation calls in 18 of the 25 populations. Thus, the majority of the populations broadly exhibited high autosomal CpG coverage, enabling highly resolved analyses across the methylome.

In summary, by integrating and uniformly processing previously published datasets and a large set of new experimental data from collaborators, I have compiled a resource of whole-genome DNA methylation data that substantially expands upon existing resources [14, 75, 93, 118, 121, 122]. This dataset encompasses a broader range of hematopoietic populations than previous studies, offering unprecedented resolution within the HSPC compartment in combination with extensive coverage of all major hematopoietic lineages except the Megakaryocyte lineage. While megakaryocyte progenitor populations, such as the MkP population, are available, a more mature Megakaryocyte population is missing. Moreover, this dataset significantly improves the scope and density of methylome coverage for those populations whose methylome was partially characterized in earlier studies, using array-based assays [118], RRBS sequencing [121] (primarily restricted to CpG-rich regions), or low-input and single cell protocols with sparse methylome coverage [75, 93].

2.1.2 Global methylome differences between hematopoietic populations reflect their differentiation hierarchy

To verify that the CpG methylation levels observed across the biological replicates for each population were highly homogeneous, I first asserted that replicates within each population had highly similar global CpG methylation levels (Table S5). Next, I focused the replicate methylome comparison on regions where methylation changes are likely to be biologically meaningful. For this purpose, I used the genome-wide set of candidate murine cis-regulatory elements identified in the Ensembl Regulatory Build [133]. I performed unsupervised hierarchical clustering based on the average methylation levels in these regions (Figure S2). Replicates clustered by population and then by hematopoietic branch, except for the replicates from the MEP and CFU-E, and the CMP CD55^{+}, MkP and preMegE populations, which could not be separated through this global methylome clustering. Additionally, few individual replicates with comparatively low coverage were attached as outer leaves to the branches of their closest related populations in the dendrogram (replicates MPP2-1, MPP5-4, PreMegE-3, CMP CD55^{+}-2, cDC2-2; see also the replicate CpG coverage overview in Table S1). Together, these data indicated high homogeneity between the replicates in our dataset, but also pointed out that highly related hematopoietic populations were in part difficult to distinguish based on global methylome comparisons alone.

Hematopoietic differentiation is accompanied by significant changes of the CpG methylation level distribution, which is generally shifted towards lower methylation levels in more differentiated cell types compared to the HSC population [93]. To investigate global methylation level shifts between the hematopoietic populations in our dataset, I compared their autosomal CpG methylation level distributions (Figure 7). The highest methylation levels were observed in HSCs (median: 94%). Methylation levels generally decreased towards more differentiated populations, with considerably differences between the hematopoietic lineages. The lowest methylation levels were observed in the CFU-E population (median: 82%). By comparison, the median CpG methylation levels for monocytes, cDC2 cells and T cells were 88%, 90% and 91% respectively. The CpG methylation level distributions were bimodal in all populations. Most CpGs showed high methylation levels, while a small subset of the CpGs were almost completely unmethylated. While most CpGs in the HSC population exhibited very high or low methylation levels, a considerable number of CpGs showed intermediate methylation levels. A total of 1,043,389 CpGs (5% of all autosomal CpGs) showed methylation levels between 30% and 70% in HSCs. This indicated the presence of considerable fractions of both methylated and unmethylated alleles for these CpGs in the HSC population. The presence of CpGs with such heterogeneous methylation states increased in more differentiated populations. The percentage of CpGs with methylation levels between 30% and 70% was 8%, 10%, 12% and 17% for B cells, neutrophils, eosinophils and CFU-Es respectively. Taken together, hematopoietic differentiation was accompanied by large, lineage-specific shifts of the CpG methylation level distribution, dominated by loss of methylation. A significant number of CpGs showed epigenetic heterogeneity already within the HSC population. The percentage of CpGs with intermediate methylation levels increased towards more differentiated populations.

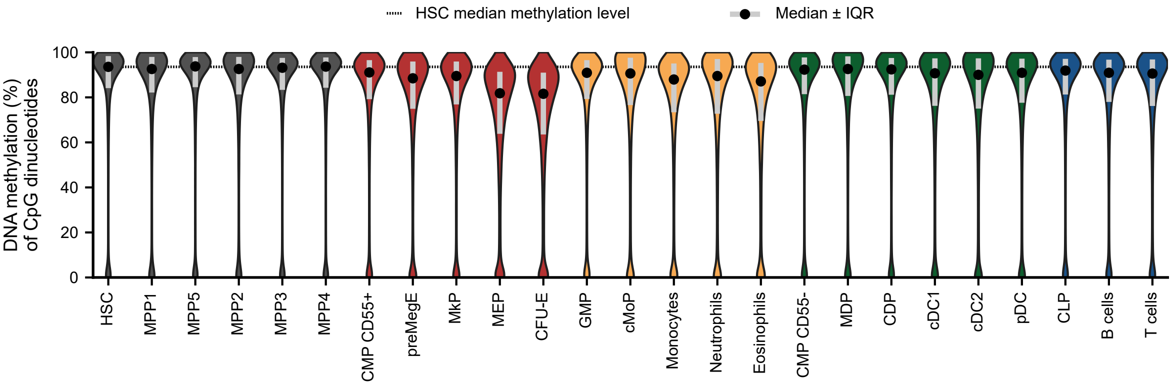

Figure 7: Hematopoietic differentiation is accompanied by lineage-specific shifts of the CpG methylation level distribution. Violin plots show the distribution of autosomal CpG methylation levels for each population. Median methylation levels are denoted by dots within the violins, the gray box represents the interquartile range (IQR). The horizontal dotted line indicates the median autosomal CpG methylation level for the HSC population.

The relationships between immunophenotypically defined populations have traditionally been modeled using a differentiation hierarchy. Previous studies have indicated that the pairwise methylome similarities between distinct hematopoietic populations may be closely aligned with the expectations derived from this conventional hierarchy. For example, Farlik et al. [93] used a custom classifier-based distance metric to translate population methylome similarities into a graph structure that mirrored the classic hematopoietic hierarchy. I therefore investigated whether methylome similarities between the populations considered in this study similarly conformed to these expectations. For this purpose, I computed the average methylation levels in 500 bp tiles across the genome for each replicate, keeping information for all tiles with at least 20 methylation calls. I then performed dimension reduction of these vectors using multidimensional scaling (MDS). In the MDS projection, the distances between the populations were in line with the expectations drawn from the classically assumed hierarchy (Figure 8). The resolution of this approach was quite high. For example, among the MPP populations, the populations with the closest proximity to the erythroid, myeloid and lymphoid lineage were MPP2, MPP3 and MPP4 respectively, in line with recently reported lineage biases in the progeny of these populations [14, 134]. This analysis demonstrated that MDS based on global methylome data was sufficient to recapitulate meaningful relationships between the hematopoietic populations in our study. The observed population similarities were in line with expectations drawn from classically assumed differentiation hierarchies as well as from recently proposed relationships between these populations.

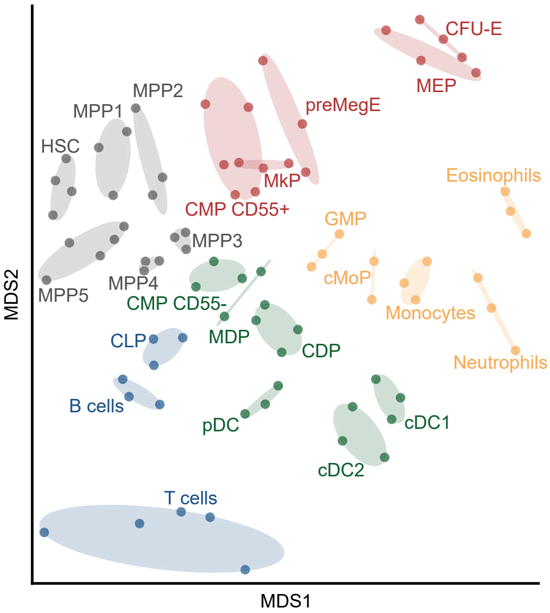

Figure 8: Global methylome similarities between hematopoietic populations reflect their differentiation hierarchy. The average methylation levels of 500 bp tiles across all autosomes were computed for each replicate and then used for dimension reduction by MDS to visualize the global methylome similarities between the populations.

2.2 A DMR/DMCpG atlas of hematopoietic methylome remodeling at unprecedented scale

2.2.1 Enhanced, integrated DMCpG and DMR calling with FDR control at the DMCpG level

In this study, I have generated high coverage, genome-wide DNA methylome maps for a wide array of hematopoietic progenitor and mature populations. This enabled the generation of a comprehensive atlas of DMRs arising during hematopoietic differentiation from early progenitor cells towards all lineages. I aimed to identify DMRs that possess robust statistical properties and demonstrate biologically relevant methylation shifts at a high signal-to-noise ratio. For this purpose, I developed and applied a multistep procedure for identifying and filtering DMRs, with FDR control on the level of individual differentially methylated CpGs (DMCpGs). DMR calling was performed for all autosomes. Sex chromosomes were not considered because DNAme analysis on these chromosomes requires special considerations, for example regarding the role of DNAme in dosage compensation for X-linked genes [135]. The schematic in Figure 9 summarizes the key steps of the DMR calling procedure. Briefly, candidate hematopoietic DMRs were identified by performing pairwise DMR calling between the HSC population and all other populations using the DSS DMR detection algorithm [82] and then merging these DMR intervals. Hematopoietic DMCpGs were identified (FDR $\leq$ 1%) and filtered based on a minimal significant methylation level shift of 20% compared to the HSC population in at least one population. Next, candidate DMR regions were filtered to only include those containing at least two DMCpGs and showing a methylation level shift of at least 30% compared to the HSC population in at least one population. These DMRs were trimmed to end with DMCpGs, but could contain up to 50% of not differentially methylated CpGs. The analysis resulted in a dual-layer DMR/DMCpG atlas of methylome remodeling during hematopoietic differentiation.

The developed DMR calling procedure, its rationale, as well as key intermediate and final results, are described in detail below and summarized in Figure 10. The initial set of candidate hematopoietic DMR regions comprised 143,177 DMRs, with a minimal DMR size of 3 CpGs. These DMR regions represent heuristically determined candidate DMR regions, since the DSS DMR detection algorithm does not perform any multiple testing correction on the DMCpG or the DMR level.

To filter for hematopoietic DMR regions with robust statistical properties and biologically relevant methylation levels shifts, I first further characterized the structure of these DMRs. For this purpose, pairwise DMCpG calling was performed between the HSC population and all other populations across all autosomal CpGs (using the statistical DMCpG test implemented in

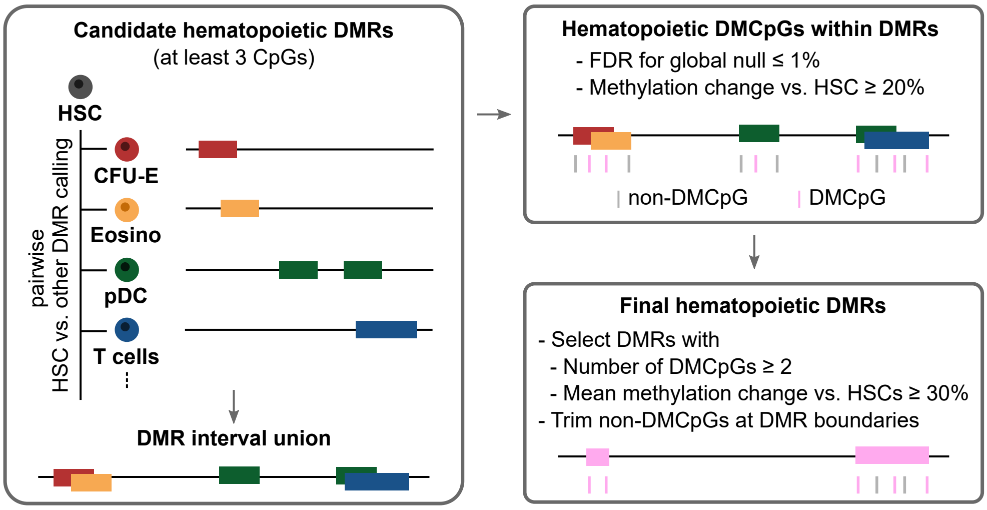
Candidate hematopoietic DMRs
(at least 3 CpGs)
HSC
HSC vs. other DMR calling
pairwise
CFU-E
Eosino
pDC
T cells
DMR interval union
Hematopoietic DMCpGs within DMRs
- FDR for global null ≤ 1%
- Methylation change vs. HSC ≥ 20%
non-DMCpG
DMCpG
Final hematopoietic DMRs
- Select DMRs with
- Number of DMCpGs ≥ 2
- Mean methylation change vs. HSCs ≥ 30%
- Trim non-DMCpGs at DMR boundaries

Figure 9: Schematic of the integrated DMR and DMCpG calling strategy. Candidate hematopoietic DMRs are identified by performing pairwise DMR calling between the HSC population and all other populations using the DSS DMR detection algorithm [82] and then merging these DMR intervals. Individual pairwise DMRs must contain at least 3 CpGs and at least 50% regulated CpGs as identified by the algorithm. Autosomal hematopoietic DMCpGs are identified using the statistical DMCpG test from the DSS package. The two-stage step-up method of Benjamini, Krieger and Yekutieli (BKY-method) is used to control the FDR at the DMCpG level (FDR $\leq$ 1%). These DMCpG sites are filtered for those located within candidate DMR regions and exhibiting a significant methylation level shift of at least 20% relative to the HSC population in at least one population. Next, DMR regions are filtered to only include those containing at least two DMCpGs and showing an average methylation level shift of at least 30% compared to the HSC population in at least one population. Lastly, the selected DMRs are trimmed to end with DMCpGs, but can contain up to 50% of not differentially methylated CpGs. A step-by-step workflow chart presenting the results of the individual testing and filtering steps is provided in Figure 10. DMCpG, differentially methylated CpG; DMR, differentially methylated region; DSS, dispersion shrinkage for sequencing data, Bioconductor package for differential analysis of sequencing data.

DSS, which also underlies its heuristic DMR detection algorithm). To identify hematopoietic DMCpGs, the global null hypothesis that a CpG was not differentially methylated in any of these pairwise comparisons was tested. The Bonferroni test was used to deal with the complex dependency structure of the individual p-values conservatively. The FDR for the global null hypothesis tests was controlled at the level of 1%, using the two-stage step-up method of Benjamini, Krieger and Yekutieli [136]. This analysis revealed 1,136,816 autosomal DMCpGs (5.6% of all autosomal CpGs). Of these, 633,560 DMCpGs were located within the candidate hematopoietic DMR regions, representing 55.7% of all autosomal DMCpGs. This indicates that the candidate DMR regions cover the majority of the observed autosomal DMCpGs. However, focusing on these DMRs disregards a significant portion of methylome programming that occurs in regions not meeting the DMR definition. Within the hematopoietic DMRs, 81.4% of all CpGs were DMCpGs, suggesting that while a strong majority of the CpGs within the candidate DMR regions appears to be differentially methylated, excluding the remaining CpGs within each DMR from analysis may be beneficial. Most of the DMCpGs sites (97.37%) exhibited strong methylation level shifts of at least 20% compared to the HSC population in at least one significantly differentially methylated population (Methods, section 4.2.1) and broad coverage across the dataset (at least 8 methylation calls in each population). Further analysis was restricted to this set of high signal-to-noise ratio (high-SNR) DMCpGs.

Using this map of high-SNR DMCpGs within the candidate DMR regions, high-SNR DMR regions were identified according to the following criteria: i) DMRs had to possess at least two DMCpGs; ii) the DMR coverage across the DMCpGs had to be at least 30 in each population; and iii) the average methylation level shift over all DMCpGs had to be at least 30% in at least one pairwise comparison between the HSC population and another significantly differentially methylated population (Methods, section 4.2.1). For further analyses, such as calculating the average DMR methylation level, only the DMCpGs were used to prevent the dilution of methylome programming signals by CpGs with insignificant or low methylation level shifts within the DMR. Therefore, the DMR boundaries were trimmed at the outermost DMCpGs. The final high-SNR, trimmed DMRs therefore contain at least two DMCpGs or greater.

Taken together, I have compiled a dual-layer DMR/DMCpG atlas comprising 122,613 hematopoietic DMRs exhibiting strong methylation level changes during hematopoiesis. The atlas features a CpG-level map that distinguishes strongly differentially methylated DMCpGs from potentially unregulated CpGs within the DMR regions. The DMRs contained 594,071 DMCpGs (84.27% of all 704,973 CpGs located within these DMRs). The key properties of the atlas are summarized in Table 1. The atlas was complemented with an autosome-wide map of DMCpGs outside the DMR regions. More than half of all detected DMCpGs were located within the DMR regions, suggesting that the DMR atlas covers a large part of the biologically relevant methylome changes accompanying hematopoietic differentiation. To generate this dual-layer atlas, I have developed a testing and filtering procedure for the identification of DMRs with robust statistical properties and a high SNR. The procedure is based on the DMCpG test and DMR calling algorithm provided by the DSS package [82]. It integrates the heuristic identification of candidate DMR regions implemented in DSS with FDR-controlled analysis of differential methylation at the CpG level.

Table 1: Hematopoietic differentiation involves extensive methylome remodeling. This table summarizes the number of features included in the dual-layer hematopoietic DMR/DMCpG atlas, along with data on the occurrence of DMCpGs within DMR regions.

# autosomal DMCpGs	1,136,816
% of autosomal CpGs	5.6%
# autosomal DMRs	122,613
Within autosomal DMRs	
# CpGs	704,937
# DMCpGs	594,071
% of autosomal DMCpGs located within DMRs	52.3%
% of all CpGs within DMRs which are DMCpGs	84.3%

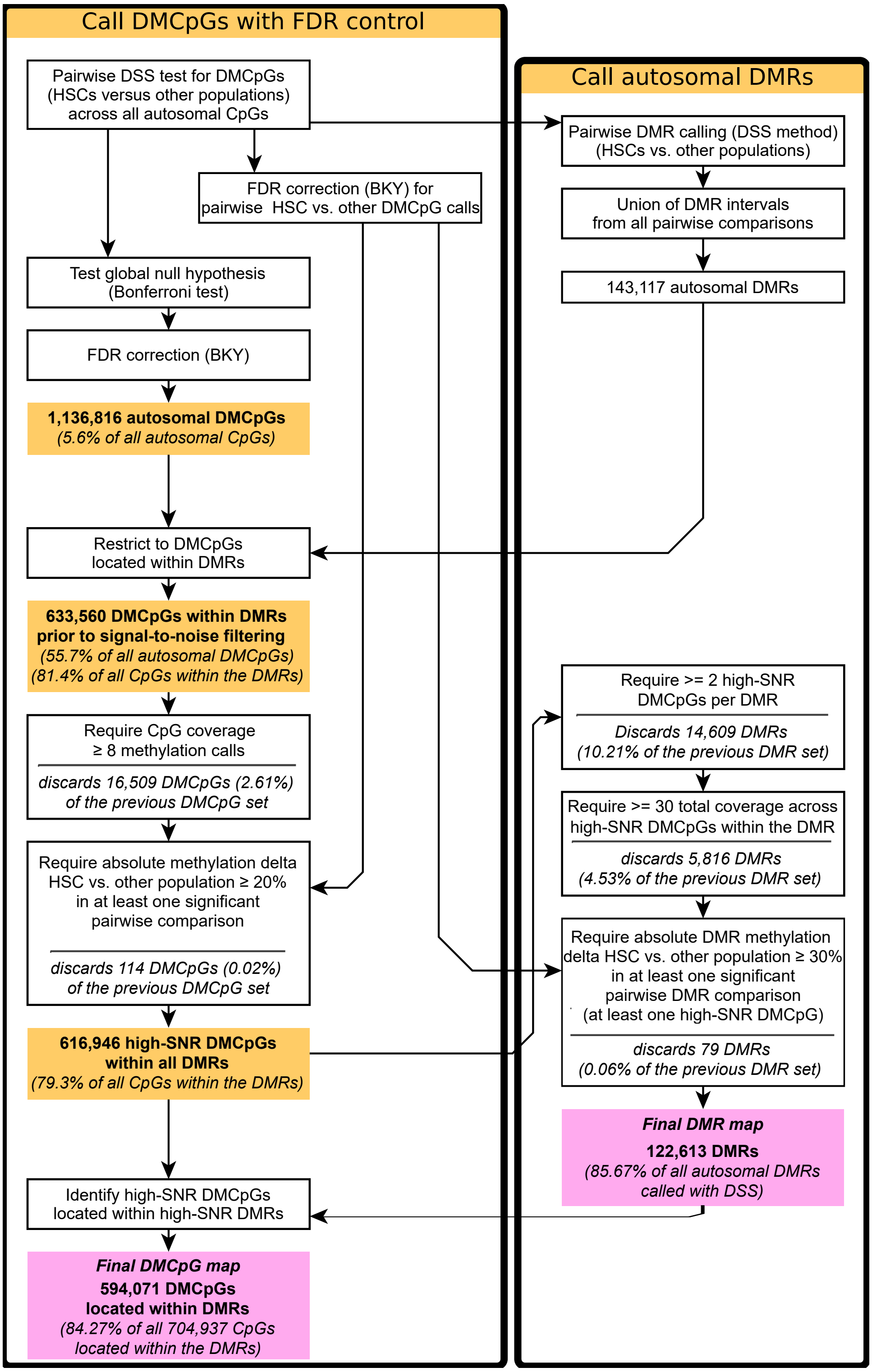
Call DMCpGs with FDR control
Pairwise DSS test for DMCpGs (HSCs versus other populations) across all autosomal CpGs
FDR correction (BKY) for pairwise HSC vs. other DMCpG calls
Test global null hypothesis (Bonferroni test)
FDR correction (BKY)
1,136,816 autosomal DMCpGs
(5.6% of all autosomal CpGs)
Restrict to DMCpGs located within DMRs
633,560 DMCpGs within DMRs prior to signal-to-noise filtering
(55.7% of all autosomal DMCpGs)
(81.4% of all CpGs within the DMRs)
Require CpG coverage ≥ 8 methylation calls
discards 16,509 DMCpGs (2.61%) of the previous DMCpG set
Require absolute methylation delta HSC vs. other population ≥ 20% in at least one significant pairwise comparison
discards 114 DMCpGs (0.02%) of the previous DMCpG set
616,946 high-SNR DMCpGs within all DMRs
(79.3% of all CpGs within the DMRs)
Identify high-SNR DMCpGs located within high-SNR DMRs
Final DMCpG map
594,071 DMCpGs located within DMRs
(84.27% of all 704,937 CpGs located within the DMRs)
Call autosomal DMRs
Pairwise DMR calling (DSS method) (HSCs vs. other populations)
Union of DMR intervals from all pairwise comparisons
143,117 autosomal DMRs
Require >= 2 high-SNR DMCpGs per DMR
Discards 14,609 DMRs
(10.21% of the previous DMR set)
Require >= 30 total coverage across high-SNR DMCpGs within the DMR
discards 5,816 DMRs
(4.53% of the previous DMR set)
Require absolute DMR methylation delta HSC vs. other population ≥ 30% in at least one significant pairwise DMR comparison (at least one high-SNR DMCpG)
discards 79 DMRs
(0.06% of the previous DMR set)
Final DMR map
122,613 DMRs
(85.67% of all autosomal DMRs called with DSS)

Figure 10: Comprehensive overview of the DMR and DMCpG detection and selection workflow. An introductory schematic of the DMR and DMCpG calling strategy is provided in Figure 9. This chart details the initial number of detected DMRs and DMCpGs and the impact of the various filtering steps underlying the feature selection for the dual-layer hematopoietic DMR/DMCpG atlas. BKY, two-stage step-up method of Benjamini, Krieger and Yekutieli; DSS, dispersion shrinkage for sequencing data, Bioconductor package for differential analysis of sequencing data; High-SNR, high signal-to-noise ratio, refers to DMCpGs and DMRs exhibiting high methylation level shifts relative to the HSC population and having high coverage across all populations.

2.2.2 High-resolution detection and separation of adjacent focal DMRs with distinct programming patterns

The applied DMR calling procedure effectively identified highly focal loci of methylome programming. One example is the detection of three focal DMRs around the transcription start site (TSS) of the *Cd74* gene, which is highly expressed in dendritic cells, B cells, and macrophages [137, 138] (Figure 11A). These DMRs were separated by small genomic intervals lacking biologically significant differential methylation. In these intervals, hypomethylation was observed across all populations, including in the HSC population. While all three DMRs exhibited strong hypomethylation in the cDC1 and cDC2 population, they exhibited distinct regulatory patterns across the other populations. One DMR upstream of the TSS showed partial hypomethylation on one side of the DMR in the myeloid lineage. A second DMR, overlapping the TSS, exhibited partial loss of methylation in the pDC, MDP, and CDP populations. The third DMR in the first intron did not exhibit strong hypomethylation in other populations besides the cDC1 and cDC2 populations. This suggests that the applied DMR calling procedure effectively separates adjacent focal regions of methylome programming that may have different regulatory patterns despite their close genomic proximity. Another example of this high-resolution analysis is the identification of two DMRs near the TSS of the *Esam* gene (Figure 11B), which is a marker for hematopoietic stem cells in humans and mice [139]. A large DMR in the first intron exhibited strong hypomethylation in the HSC and MPP1 populations, while a smaller, directly adjacent DMR in the first exon was primarily characterized by gain of methylation in T cells. The locus plots in Figure 11A and Figure 11B were generated using the methlevels [SOFT2] and codaplot [SOFT3] Python packages developed by me. Taken together, these representative examples demonstrate that the applied DMR calling procedure was able to discern focal DMR regions which were closely adjacent but separated by unregulated or lowly regulated CpGs. The examples highlight that such DMRs may exhibit distinct regulatory patterns despite their close spatial proximity. This suggests that treatment of these focal DMRs as separate features may be highly beneficial for correctly capturing the methylome remodeling within regulatory elements.

The focal nature of the identified DMRs was reflected in their small interval sizes. The median DMR size in the DMR atlas was 183 bp, and the 90th percentile of the DMR size was 578 bp (Figure 11C). The median count of DMCpGs contained within the DMRs was 4 DMCpGs, and the 90th percentile of this count was 8 DMCpGs (Figure 11D). Because the identified DMRs were highly focal, the vast majority of the CpGs within these DMR intervals exhibited strong biological signals. In 76% of all DMRs, all contained CpGs were identified as DMCpGs (Figure 11E). For the remaining DMRs, only a small percentage of the contained CpGs were not identified as DMCpGs. Consequently, the average methylation levels of these focal DMRs are likely to provide an unambiguous measure of interval-level methylome remodeling patterns, in contrast to strategies yielding broader DMR intervals [87, 89], where unregulated CpGs might distort the mean. To improve the DMR methylation level signal even further, all CpGs which were not identified as DMCpGs were excluded from the calculation of the average DMR methylation levels and other DMR-based analyses. This further reduced dilution of the DMR programming signals. Taken together, the detection of focal DMR regions, coupled with the identification of the strongly regulated DMCpGs within these regions, allowed the compilation of DMR features that captured methylome programming with high resolution and a strong signal-to-noise ratio.

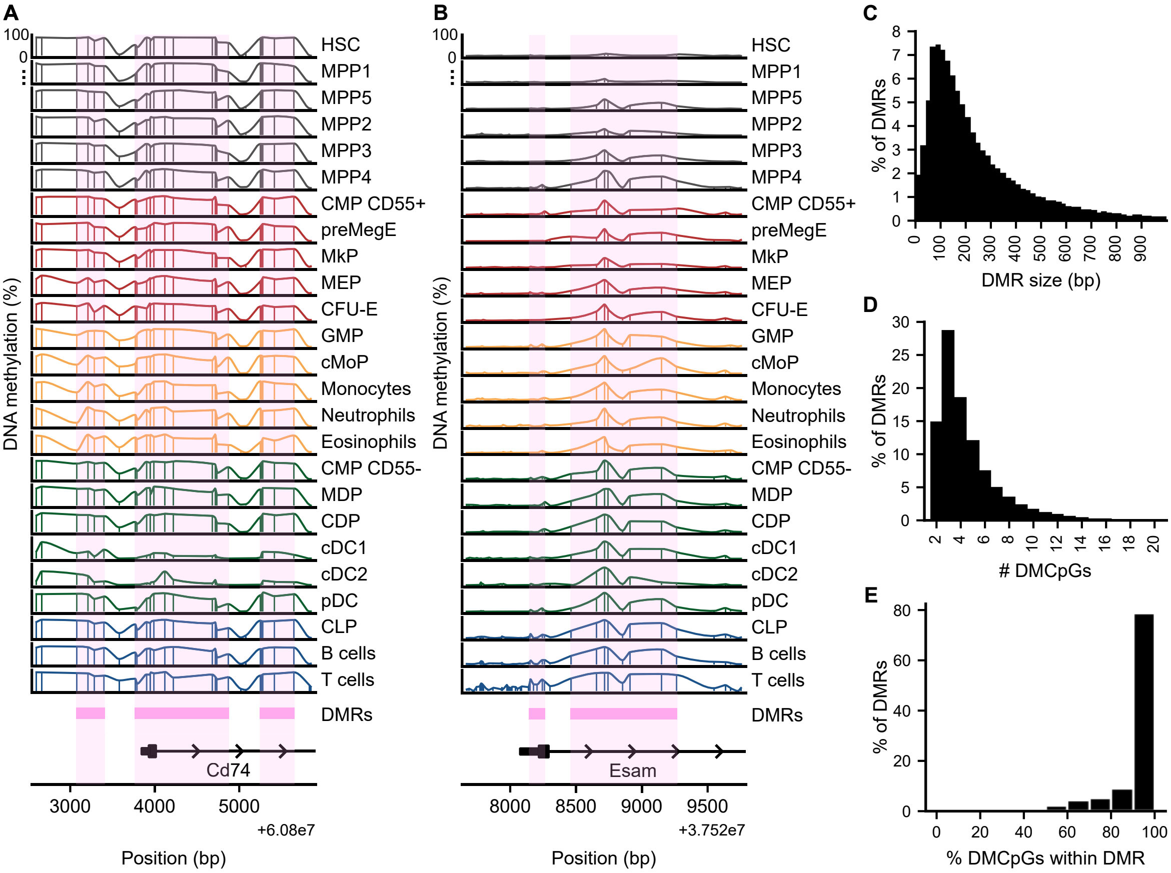

Figure 11: Effective identification of focal, densely regulated DMR regions.
(A,B) Representative locus plots depicting CpG methylation levels for all populations, as well as the detected DMR regions around the principal transcription start sites of *Cd74* (A) and *Esam* (B).
(C, D) Histograms showing the distribution of the DMR size, measured in base pairs (C) and by the number of DMCpGs within each DMR (D).
(E) For each DMR, the percentage of CpGs classified as DMCpGs was calculated. The distribution of this percentage is shown with a histogram.

2.2.3 Leveraging genomic region classification for improved proximity-based DMR-to-gene annotations

The DMRs were annotated against protein-coding genes to identify potential DMR target genes and to classify the genomic regions in which the DMRs resided. Only transcripts with strong experimental support were considered for annotation. The distances between the individual DMRs and their closest TSSs formed a distribution centered close to zero, with a strong decay towards larger distances (Figure 12A). Approximately one-third (35%) of all DMRs were located within ±15 kb of a TSS, and about two-thirds (66.9%) were located within ±50 kb of a TSS. This distribution closely resembles the feature-to-TSS distance distributions observed for various epigenetic features associated with cis-regulatory functions, such as open chromatin regions [51], enhancers [140], and TF binding sites [141]. This suggests that a significant fraction of the DMRs captured in this study are associated with gene-regulatory functions.

In a recent benchmark study [142], proximity-based methods for annotating cis-regulatory elements (CREs) with potential target genes have achieved competitive performance compared to signal correlation-based or machine learning-based annotation approaches. Proximity-based target gene annotation methods are especially attractive for the prediction of CRE-to-gene associations across cell types because they have less cell type-related bias than more complex methods [142]. Therefore, I have developed gtfanno, a novel, proximity-based algorithm to perform DMR-to-gene annotation, which exploits genomic region class residence information in addition to the identification of proximal TSSs. This algorithm was made available as a Python package [SOFT4] and has been successfully applied in various projects [OWN1–OWN3]. The different genomic region classes considered in the algorithm are illustrated in Figure 12B. Briefly, gene annotation with gtfanno was performed by first querying, for each DMR, residence in the following genomic location classes in this order of precedence: i) promoter (5000 bp upstream to 1000 bp downstream of a TSS); ii) 5'-untranslated region (UTR) or 3'-UTR; iii) intron or exon; iv) candidate cis-regulatory element (cCRE), within ±50 kb of a TSS but outside the promoter or gene body regions; and v) intergenic, if no annotation based on the preceding genomic location classes was possible. If a DMR was associated with multiple genes, all potential target genes for which the DMR was located in the genomic region class with the highest precedence were kept, and all others were discarded. For example, a DMR may have resided in the promoter region of two genes and the intron of a third gene. In this case, both promoter-based gene annotations would have been retained, but the intron-based gene annotation would have been discarded. Multiple genes were allowed to be annotated as potential target genes of a single DMR because the DMR-to-gene annotations are subject to a considerable degree of uncertainty. For example, when a DMR is located in the promoter region of two different genes, proximity-based statistics such as the distance of the DMR to the TSSs of the two genes alone are insufficient to determine with certainty which gene is more likely to be regulated by the DMR. Taken together, I have developed and implemented a novel proximity-based DMR annotation algorithm. This algorithm was applied to provide annotations for the DMRs in the hematopoietic DMR/DMCpG atlas. The annotation approach underlying the algorithm has been validated through successful use across several projects [OWN1–OWN3].

In total, 97074 (79%) of all DMRs could be associated with one or more potential target genes (Figure 12C). A considerable number of DMRs (10417 DMRs, 8.5% of all DMRs) were located in the promoter regions of 6360 individual genes. The most frequent genomic position of the DMRs was within introns (55244 DMRs, 45% of all DMRs). Such intronic DMRs were found for 9046 genes (Figure 12D). Following the global DMR-to-TSS distance distribution (Figure 12A), about one-third (30%) of these intronic DMRs were located very closely (within 15 kb) of the TSS of the gene they resided in, and about two-thirds (63%) were located within 50 kb of this TSS. Taken together, these annotations enrich the hematopoietic DMR/DMCpG atlas with a single genomic region annotation for each DMR as well as an annotation of one or more potential target genes for all non-intergenic DMRs (79% of all DMRs).

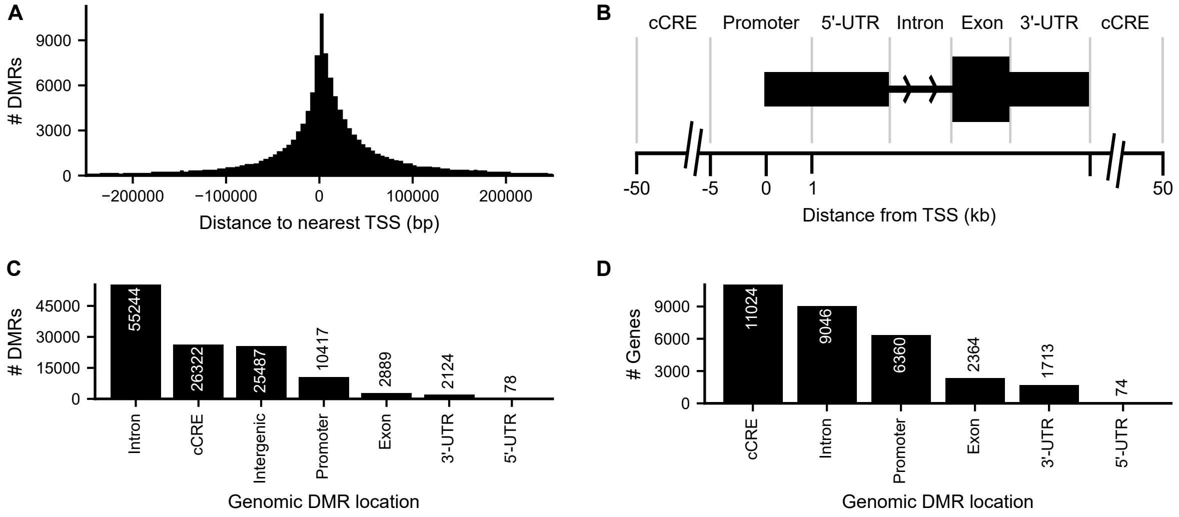

Figure 12: Gene and genomic region annotations for the hematopoietic DMRs, using the innovative gtfanno algorithm. The gtfanno annotation algorithm was implemented as a Python package [SOFT4].
(A) Histogram showing the distribution of the distances between each DMR and its nearest transcription start site (TSS). Negative values indicate that the DMR is upstream of the TSS, while positive values indicate downstream locations.
(B) Overview of the genomic region classes employed for gene and genomic region annotations with gtfanno. DMRs were assigned to a single class, with precedence given in the following order if multiple classifications were possible: Promoter, 5'-untranslated region (5'-UTR) or 3'-UTR, Intron or Exon, and cis-regulatory element (CRE).
(C) Number of DMRs associated with the different genomic region classes.
(D) Number of genes associated with DMRs from each genomic region class.

2.3 Characterization of lineage- and population-specific DMR programming modules

2.3.1 Clustering analysis reveals population- and lineage-specific DMR programming modules

DMR clustering analysis was performed with the aim of identifying sets of co-regulated DMRs with a well-defined relationship to distinct hematopoietic cell types. These DMR clusters were intended to serve as reference region sets for single-cell analysis of DNAme changes during hematopoietic differentiation. In particular, these reference region sets were intended to aid with the investigation of methylome programming in early progenitor cells. To ensure that the DMR cluster analysis was not biased by the bulk progenitor populations in our dataset, clustering was based on only the data from the mature populations (the CFU-E, monocyte, eosinophil, neutrophil, B cell, T Cell, cDC1, cDC2 and pDC populations). These mature populations comprise relatively homogeneous cells at functionally defined, mature endpoints of differentiation in the hematopoietic system. Thus, differing methylation levels between these populations are unambigously associated with functionally different hematopoietic cell types. In contrast, methylation level differences between potentially heterogeneous immunophenotypic progenitor populations may be caused by convoluted shifts in cell type composition. Consequently, functional annotation of DNAme differences between such populations is challenging.

DMR clustering was performed with the unsupervised Leiden community detection algorithm [143], using the correlation distance. The methylation level of each DMR was computed as the mean CpG methylation level across all DMCpGs within the DMR. This resulted in the identification of 28 DMR clusters with distinct regulatory patterns (Figure 13A). Cluster sizes varied widely (Figure 13B), ranging from 819 DMRs (C5 cluster) to 10385 DMRs (E4 cluster), with an average cluster size of 4377 ± 2233 DMRs (mean ± s.d.).

Two of the DMR clusters (H1 and H2) showed a distinct regulatory pattern compared to all other clusters (Figure 13A). The average DMR methylation level in these clusters was lowest in the HSC population, indicating that methylation was primarily gained in the DMRs of these clusters. The observation that the maximal hypomethylation in these clusters occured in the HSC population is also denoted with the “H” prefix of the cluster names H1 and H2. The remaining 26 DMR clusters (Figure 13A) exhibited high DMR methylation levels in the HSC population followed by loss of methylation in the downstream populations. In total, 11 166 DMRs (9.1% of all DMRs) were part of the gain of methylation clusters, while 111 395 DMRs (90.9% of all DMRs) were part of loss of methylation clusters. Specifically, the H1 DMR cluster comprised DMRs in which substantial gain of methylation was already

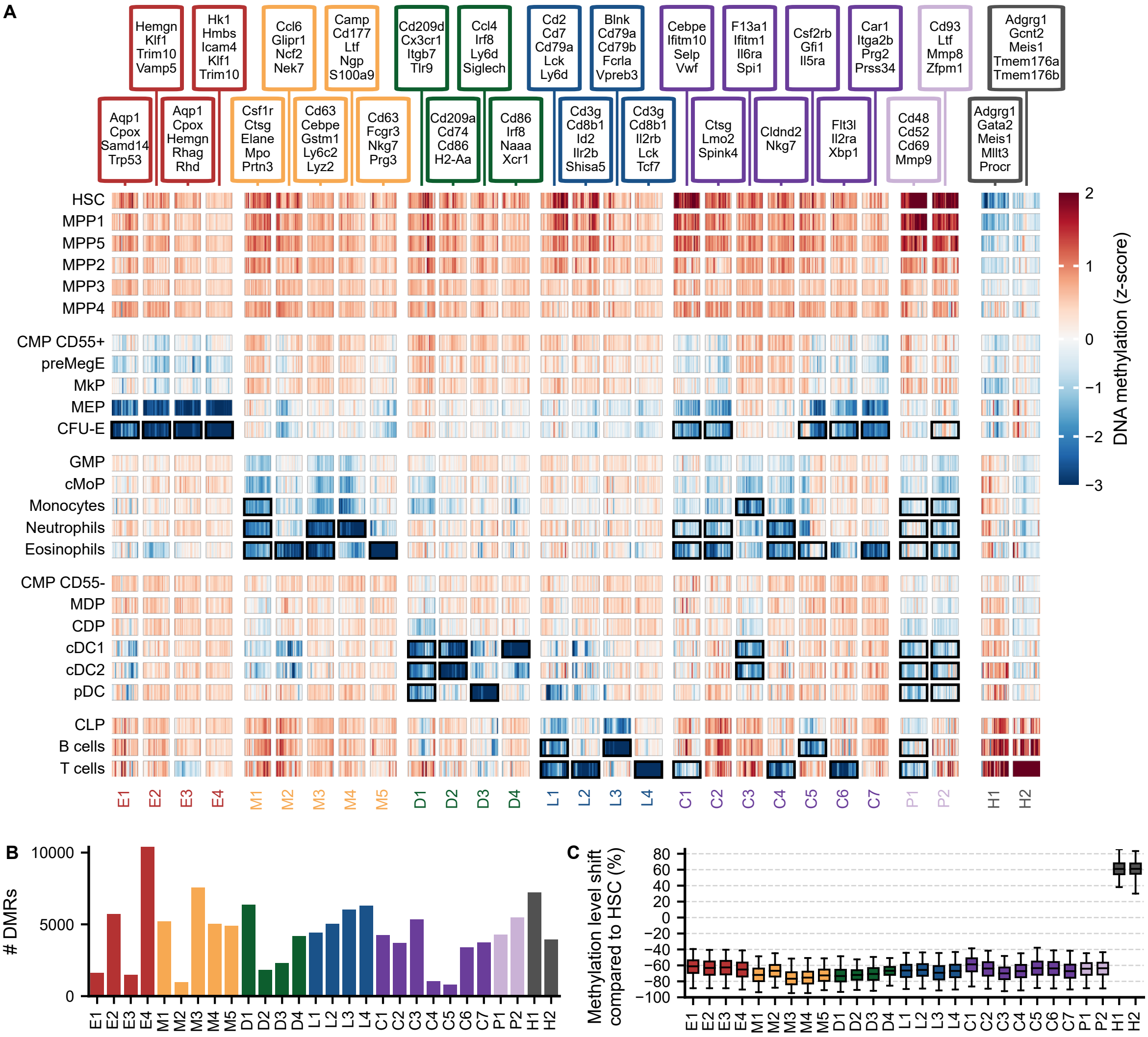

Figure 13: Clustering analysis reveals population- and lineage-specific DMR programming modules. (A) Heatmap showing z-score transformed DMR methylation levels of 200 randomly selected DMRs for each of the 28 identified DMR clusters. Clustering was performed on the DMR methylation levels of the mature populations and the HSC population, using Leiden clustering with the correlation distance. The H1 and H2 DMR clusters were characterized by low methylation levels in the HSC population, and subsequent gain of methylation in downstream populations. All other clusters were characterized by loss of methylation compared to the HSC population. For each of these clusters, black rectangles indicate the population with the lowest average DMR methylation level in the cluster as well as all mature populations with an average DMR methylation level within 15% of that value (referred to as the marked populations for that cluster). Clusters were grouped according to the lineage-specifity of their marked populations: within the cluster names, the prefixes E, M, D and L indicate erythroid, myeloid, lymphoid or dendritic cell lineage-specific clusters respectively; the prefix C (for cross-lineage) indicates DMR clusters marking populations across multiple lineages; the prefix P indicates pan-hematopoietic DMR clusters, marking 7 or more hematopoietic populations across three lineages. The DMR clusters were ordered by increasing population-specificity of their regulatory pattern within each group of DMR clusters, as indicated by the ordinal number within the cluster name.
(B) DMR cluster sizes, ranging from 819 DMRs (C6) to 10385 DMRs (E4).
(C) Distribution of the DMR methylation level shifts compared to the HSC population for each DMR cluster. The methylation shift for each DMR was computed as the DMR methylation level difference between the HSC population and the population with the most different methylation level in this DMR. Whiskers indicate the first and 99th percentile of the DMR methylation level shifts for each DMR cluster.

observed in the MPP populations and high methylation levels were maintained throughout the downstream hematopoietic populations. On the other hand, the H2 DMR cluster consisted of DMRs where methylation levels remained comparable to the HSC population across the MPP populations as well as across the populations of the erythroid, myeloid, and dendritic cell lineages. Gain of methylation was primarily observed in the lymphoid lineage, culminating in maximal gain of methylation in T cells. Taken together, DNA methylome remodeling during hematopoiesis appeared to primarily involve loss of methylation compared to the HSC population as a reference state, as previously described [14, 93, 122]. However, this study revealed two distinct DMR clusters characterized by gain of methylation compared to the HSC population. These clusters appear to represent an early, pan-hematopoietic DMR programming module associated with loss of stemness (H1 cluster) and a later, lymphoid-specific DMR programming module (H2 cluster).

The 26 loss of methylation DMR clusters captured methylome remodeling occuring at different levels of specificity, ranging from population-specific across lineage-specific to pan-hematopoietic DMR programming modules. To characterize these DMR clusters, I determined for each DMR cluster the set of the most hypomethylated populations, referred to as the "marked" populations for that cluster (Figure 13A). The set of marked populations for each DMR cluster was defined to include i) the population with the lowest average DMR methylation level in this cluster, and ii) all other populations with an average DMR methylation level in the cluster within 15% of this lowest average DMR methylation level. Seventeen DMR clusters showed lineage-specific regulatory patterns, with all of their marked populations belonging to same lineage. Specifically, I observed four erythroid-specific clusters (E1-E4), five myeloid-specific clusters (M1-M5), four dendritic cell-specific clusters (D1-D4) and four lymphoid-specific clusters (L1-L4). Among these lineage-specific clusters, one or more clusters specifically marking a single hematopoietic population were found for most mature hematopoietic populations, namely for the CFU-E (E1-E4 clusters), eosinophil (M2 and M4 clusters), neutrophil (M5 cluster), cDC1 (D4 cluster), pDC (D3 cluster), B cell (L3 cluster) and T cell (L2 and L4 clusters) populations. Within the group of lineage-specific DMR clusters, these highly population-specific DMR clusters were complemented by clusters exhibiting strong regulation over multiple or all populations from a specific lineage, such as the D1, M1, and L1 clusters which specifically marked all populations from the dendritic cell lineage, myeloid lineage, and lymphoid lineage respectively. Next, seven DMR clusters marked populations across multiple lineages, and therefore their names were prefixed with "C", for "cross-lineage" (clusters C1-C7). For example, the C7 cluster marks CFU-Es and eosinophils and the C3 clusters marks cDC1, cDC2 and monocytes. Finally, two DMR clusters (P1 and P2) exhibited pan-hematopoietic methylation loss with more than six marked populations across three lineages. In summary, the 28 DMR clusters captured DMR programming modules with distinct regulatory patterns. Two DMR clusters (H1 and H2) were characterized by gain of methylation and 26 DMR clusters were characterized by loss of methylation. The regulatory patterns exhibited by these DMR clusters suggest that coordinated DNAme programming in the hematopoietic system may occur at varying levels of specificity, ranging from population- and lineage-specific to general, pan-hematopoietic programming.

As demonstrated above, the regulatory pattern of each DMR cluster could be primarily characterized by markedly strong methylation level changes (compared to the HSC population) in a specific set of populations (the marked populations). However, all DMR clusters also exhibited smaller, progressively staggered methylation changes across multiple other populations in addition to their marked populations (Figure 13A). Such progressive methylation level changes often occured across progenitor and sibling populations of the populations marked in a DMR cluster. They could however also occur in populations apparently outside of the lineages marked by a DMR cluster. Some DMR clusters showed such partial methylation level changes only in a few populations. For example, the D4 DMR cluster marked the cDC1 population and showed progressive loss of methylation mainly across i) the CDP population followed by the cDC2 population, a progenitor and a sibling population to cDC1 within the dendritic lineage; and ii) the cMoP population followed by the monocyte population, which have been reported to be developmentally related to dendritic cells through a shared progenitor state captured in the MDP population [144]. Other DMR clusters showed progressive loss of methylation across a broader part of the hematopoietic system. For example, the D1 DMR cluster marked all mature dendritic populations and showed progressive methylation level changes among others across the dendritic progenitor populations CMP CD55^{-}, MDP and CDP, as well as outside of the dendritic lineage across the GMP, cMoP and monocyte populations. In summary, the regulatory patterns of the DMR clusters were characterized by a combination of distinctly strong methylation changes (compared to the HSC population) in a specific set of marked populations and smaller, progressively increasing methylation level changes across other populations.

To provide an initial, high-level characterization of the specificity of the regulatory pattern of each DMR cluster, I computed a specificity score for each cluster. The specificity score for a given DMR cluster was computed by first calculating the average DMR methylation levels for all mature populations in this cluster, and then calculating the overall mean of these values, weighted such that each lineage (erythroid, myeloid, lymphoid, dendritic cells) contributed equally to the mean. This score summarized how broadly methylation level changes were observed across the mature hematopoietic system. Within each group of DMR clusters (i.e., within the gain of methylation/H cluster group, each of the lineage-specific (E, M, D, and L) cluster groups, the multilineage/C cluster group, and the pan-hematopoietic/P cluster group), a range of specificity scores was observed, highlighting that the individual DMR clusters represent DMR programming modules of varying levels of population- or lineage-specificity. The relative specificity of the regulatory patterns was also denoted in the cluster names: the order of the DMR clusters within each group reflects the order of their specificity scores, either in ascendingly sorted order for the loss of methylation clusters, or in descendingly sorted order for the gain of methylation clusters. The cluster names thus reflect the observed specificity of the methylation loss or gain observed in the DMR cluster. For example, the M1 and M5 DMR clusters both exhibit a myeloid lineage-specific regulatory pattern, but the M1 DMR cluster shows broad occurrence of hypomethylation across the mature hematopoietic system in addition to marking all mature myleoid populations (the most unspecific regulatory pattern of the myeloid lineage clusters), while the M5 DMR cluster has the most population-specific regulatory pattern of all myeloid clusters (marking only neutrophils, with partial methylation loss observed only in a few populations). In summary, the DMR cluster nomenclature encodes basic properties of the regulatory patterns characterizing the individual DMR clusters: i) the lineages where markedly strong methylation changes occur in each DMR cluster are encoded with a prefix letter, and ii) the ordinal number after the prefix letter encodes how broadly methylation changes are observed across the entire mature hematopoietic system in that DMR cluster.

The scale invariance of the correlation distance was not of concern for this dataset because all DMRs exhibited considerable methylation level shifts during hematopoiesis (Figure 13C). All individual DMRs showed a methylation level shift compared to the HSC population of at least 30%. The median methylation level shift was 61% for both the H1 and H2 gain of methylation clusters and it ranged between -77% and -59% for the loss of methylation clusters. Furthermore, the strong DMR cluster compactness and separation observed for z-score transformed DMR methylation levels (Figure 13A) were largely retained when comparing the DMR methylation levels directly (Figure S3).

2.3.2 DMR programming modules are associated with matching gene expression modules

As detailed above, the DMR clusters were characterized by a combination of distinctly strong hypomethylation in a specific set of marked populations and smaller, progressively staggered hypomethylation in additional populations. I next investigated the relationship between the regulatory patterns of the DMR clusters and the expression of their potential target genes.

For this purpose, I first computed a comprehensive catalogue of hematopoietic cell type marker genes based on an in house single-cell RNA sequencing (scRNA-seq) dataset generated and provided by the Lipka lab at the German Cancer Research Center (Methods, section 4.4.1). Single-cell RNA-seq was performed by these collaborators using the 10X Genomics platform. To broadly cover the hematopoietic system, a total of 8495 cells was collected by FACS in three tiers: i) 1070 Lin$^-$Sca-1$^+$c-Kit$^+$ (LSK) cells, ii) 3441 Lin$^-$c-Kit$^+$ (LK) cells, and iii) 3984 total bone marrow cells (Figure 14A). Single-cell RNA-seq alignments were performed by the

Omics IT and Data Management Core Facility (ODCF) at the German Cancer Research Center using Cell Ranger.

I acknowledge that several other doctoral students have done independent work on the same scRNA-seq data set. I have discussed some aspects of the analysis of these data with several of these colleagues, including Maximilian Schönung, Sina Stäble, Mariam Hakobyan, and Abdelrahman Mahmoud. I have carried out an earlier version of the data analysis in cooperation with Sina Stäble. Sina Stäble has shown parts of this collaborative work in her thesis [131], in combination with independent work performed by her and other collaboration partners. Later, Maximilian Schönung carried out an independent, similar clustering analysis as the one presented in this thesis and provided it to me for reference. Nevertheless, the analysis in this thesis stands out as an independent and original analysis, with distinct goals, scope, and methodological complexity compared to other analyses of the data. The analysis presented in this thesis was conceptionalized and coded by me, and differs in various key aspects from the parallel efforts of my colleagues. My analysis uses a different analysis framework (Python/ScanPy instead of R). I use different computational strategies for example with regard to expression normalization and clustering (which I have provided in part to Sina Stäble for reproduction in her thesis). My analysis also differs in the focus on the identification of clean clusters of rare cell populations in the data set, such as the eosinophil population, which was of particular interest for this project.

To compile a comprehensive hematopoietic cell type marker catalogue based on the scRNA-seq dataset, I first conducted a full clustering and cell type annotation workflow optimized for i) the simultaneous detection of single cell clusters of varying sizes and compactness and ii) gene expression normalization across cell types with varying transcriptome compositions (Methods, section 4.4). Briefly, I used the standard scRNA-seq clustering workflow implemented in scanpy [145] with targeted custom modifications. Gene expression normalization was performed using the sctransform algorithm [146]. This approach is based on predicting gene expression levels in individual cells using a regularized negative binomial regression model where the cellular sequencing depth is utilized as the independent variable. The Pearson residuals from this model have been shown to represent normalized expression values with favorable properties [146, 147]. A challenge for the normalization of scRNA-seq data covering a range of cell types from progenitor to mature cells is that some of these cell types exhibit transcriptomes dominated by individual, markedly strongly expressed genes. This can skew standard gene expression normalization approaches [148, 149]. To address this challenge, I have exchanged the standard independent variable used for sctransform normalization (total sequencing depth) with an adjusted sequencing depth. This adjusted sequencing depth was computed while excluding all genes which in at least one cell possessed more than 5% of all the counts observed within that cell. This is likely to provide an improved size factor, following ideas initially proposed by Weinreb et al. [149] and recently promoted by an influential review of best practices in scRNA-seq data analysis [148]. Clustering was performed with the PARC algorithm [150] for community detection. This algorithm extends the standard Leiden clustering algorithm with several preprocessing steps pruning the k-nearest neighbor graph. This has been shown to improve the clustering in the presence of strongly differing cluster sizes and between-cluster similarities [150]. The clustering analysis identified 18 distinct cell clusters (Figure 14B). The clusters were annotated with cell type labels using literature-based cell type markers [29, 45, 151, 152]. The cell clusters captured i) early stem and progenitor cell types (identified as HSC, ST-HSC, MPP, and LMPP cell clusters); ii) committed progenitor cell types (identified as Early GMP, GMP, neutrophil progenitor, monocyte progenitor and erythroid progenitor cell clusters); and iii) differentiated cell types at the endpoints of the erythroid, megakaryocytic, myeloid, lymphoid and dendritic cell lineages. Taken together, 18 single-cell clusters corresponding to distinct hematopoietic cell types covering the hematopoietic system from HSPCs to the fully differentiated endpoints of the major hematopoietic lineages were identified.

Next, for each single-cell cluster, I collected the top 50 expression marker genes based on a Wilcoxon rank-sum test for differential expression. Briefly, enrichment of highly expressed genes within each single-cell cluster was tested against the background of all other clusters and multiple testing correction was performed with the Benjamini-Hochberg (BH) method (FDR $< 0.01\%$). Only enrichments with a $\log_2$ fold change > 1.25 and only genes which were expressed in more than 25 cells were considered. The union set of these marker genes comprised 589 genes with strong differential expression and cell type association during hematopoiesis. Of these expression marker genes, 520 genes (88%) had at least one associated DMR according to the DMR-to-gene annotation presented in section 2.2.3. For almost all of the hematopoietic marker genes with one or more annotated DMRs, at least one DMR was located very close to the TSS. For 73% of these marker genes (380 genes), a DMR was found in the promoter region, and for 95% of these marker genes (492 genes), a DMR was found within ±15 kb of the TSS. Taken together, I have generated a comprehensive catalogue of hematopoietic cell type markers. A large majority of these hematopoietic marker genes appeared associated with TSS-proximal DMRs. It appears likely that these DMRs can exert a regulatory influence on the expression of their associated genes.

This catalogue of hematopoietic cell type markers contained many genes with well-established roles during hematopoietic differentiation. Many of these genes had promoter DMRs annotated to them from DMR clusters whose regulatory pattern matched the known role of the gene. Many examples of such genes are detailed in Figure 13A and a few are highlighted in the following.

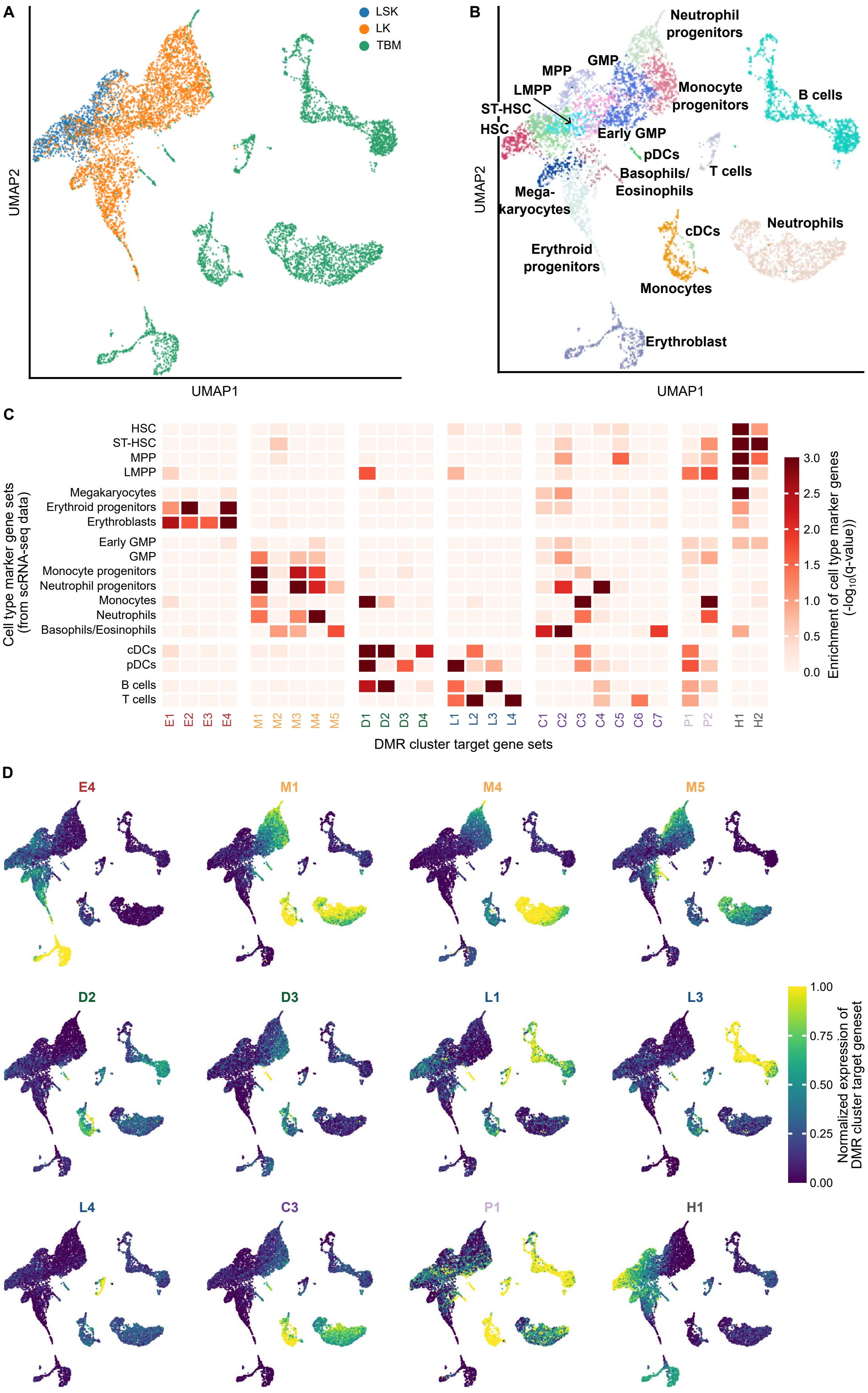
A
LSK
LK
TBM
UMAP2
UMAP1
B
Neutrophil progenitors
GMP
MPP
LMPP
ST-HSC
HSC
Monocyte progenitors
B cells
Early GMP
pDCs
Basophils/ Eosinophils
T cells
Mega-karyocytes
cDCs
Neutrophils
Erythroid progenitors
Monocytes
Erythroblast
UMAP2
UMAP1
C
Cell type marker gene sets (from scRNA-seq data)
HSC
ST-HSC
MPP
LMPP
Megakaryocytes
Erythroid progenitors
Erythroblasts
Early GMP
GMP
Monocyte progenitors
Neutrophil progenitors
Monocytes
Neutrophils
Basophils/Eosinophils
cDCs
pDCs
B cells
T cells
E1 E2 E3 E4
M1 M2 M3 M4 M5
D1 D2 D3 D4
L1 L2 L3 L4
C1 C2 C3 C4 C5 C6 C7
P1 P2
H1 H2
DMR cluster target gene sets
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Enrichment of cell type marker genes (-log10(q-value))
D
E4
M1
M4
M5
D2
D3
L1
L3
L4
C3
P1
H1
1.00
0.75
0.50
0.25
0.00
Normalized expression of DMR cluster target geneset

Figure 14: DMR programming modules are associated with matching gene expression modules.
(A) UMAP embedding of single-cell RNA sequencing (scRNA-seq) data for 8495 hematopoietic cells. An unpublished in-house scRNA-seq dataset broadly covering the murine hematopoietic system was generated and provided by the Lipka lab (Methods, section 4.4.1). The data were generated using the 10X Genomics platform for hematopoietic cells collected by FACS in three tiers: i) 1070 Lin$^-$Sca-1$^+$c-Kit$^+$ (LSK) cells, ii) 3441 Lin$^-$c-Kit$^+$ (LK) cells, and iii) 3984 total bone marrow cells. The plot shows a UMAP embedding of the data, with individual cells colored by their FACS tier. (B) Annotation of hematopoietic cell types in the scRNA-seq dataset. Single-cell RNA-seq analysis was performed using the standard scanpy workflow [145], extended to use the sctransform algorithm for gene expression normalization [146] and the PARC community detection algorithm [150] for clustering. Cell types were annotated manually by inspecting the expression of established hematopoietic marker genes. Single cell clusters and their annotated cell type labels are shown on the UMAP embedding of the data.
(C) Enrichment of expression-based cell type marker genes within the target genes of each DMR cluster. The top 50 expression marker genes for each cell type detected in the scRNA-seq dataset were computed using the Wilcoxon rank-sum test and the Benjamini-Hochberg (BH) method [153] for multiple testing correction (FDR < 0.01%). For each DMR cluster, all genes that had at least one DMR from the DMR cluster annotated to them were considered to be potential target genes of that DMR cluster. DMR-to-gene annotation was performed using an innovative proximity-based algorithm (Figure 12). Only DMRs within ±15 kb of the TSS of their annotated gene were considered. For each DMR cluster, enrichment of the expression-based cell type marker gene sets within its DMR cluster target gene set was computed. Enrichments were computed against the background of all other DMR clusters using Fisher's exact test. Multiple testing correction by adjusting p-values to q-values was performed using the BH method [154, 155]. The heatmap shows the significance of the cell type marker gene set vs. DMR cluster target gene set associations by encoding the -$\log_{10}$(q-values) of the corresponding tests.
(D) Average expression of the DMR cluster target gene sets of representative DMR clusters. The potential DMR cluster target genes were defined as for (C). For each DMR cluster, average gene expression across all potential DMR cluster target genes was computed within each cell. For this purpose, gene expression vectors were normalized to equally weigh all genes. Min-max normalized average target gene set expression levels are shown on a UMAP embedding of the scRNA-seq data for various representative DMR clusters.

- The *Klf1* gene encoding the Krüppel-like factor 1 transcription factor, which plays a key role during erythropoiesis [156], was associated with promoter DMRs from the CFU-E specific E2 and E4 clusters.

- The *Elane* (neutrophil elastase), *Mpo* (myeloperoxidase), and *Ctsg* (cathepsin G) genes, which encode enzymes that play a key role in early myelopoiesis and are expressed in single-cell transcriptome-based myeloid progenitor cell types [29], were associated with promoter DMRs from the pan-myeloid M1 cluster.

- The *S100a9* [157] and *Cd177* [158] genes, which encode surface markers that are specifically expressed in neutrophils, were associated with promoter DMRs from the neutrophil-specific M4 cluster.

- The *Cd74* gene, which encodes a surface marker expressed by the cDC1 and cDC2 populations [159], was associated with promoter DMRs from the cDC-specific D2 cluster.

- The *Siglech* gene encoding a lectin receptor that is primarily expressed in pDCs [160], was associated with promoter DMRs from the pDC-specific D3 DMR cluster.

- The *Lck* gene encoding the lymphocyte-specific protein tyrosine kinase, which plays a crucial role in the development of both T cells and B cells [161, 162], was associated with promoter DMRs from the pan-lymphoid cluster L1.

- The *Cd79a* and *Cd79b* [163] genes that encode B cell receptor components, were associated with promoter DMRs from the B cell-specific L3 cluster.

- The *Prg2* (encoding Proteoglycan 2) and *Car1* (encoding Carbonic anhydrase 1) genes, which are markers for eosinophils and erythroid cells respectively [29], were both associated with promoter DMRs from the C7 cluster marking both eosinophils and CFU-Es.

- The *Spi1* gene encoding the versatile PU.1 transcription factor, which among other functions plays a critical role for the commitment of multipotent progenitors to the myeloid lineage and subsequent differentiation towards monocytes and dendritic cells [164], was associated with promoter DMRs from the C3 cluster marking both cDCs and monocytes.

- The *Meis1* (myeloid ecotropic viral integration site 1) [165] and *Mllt3* (myeloid/lymphoid or mixed-lineage leukemia; translocated to, 3) [166] genes, which are involved in hematopoietic stem cell maintenance, were associated with promoter DMRs from the HSC-specific H1 DMR cluster.

Next, I determined potential target gene sets for each DMR cluster and tested the enrichment of the scRNA-seq-based hematopoietic cell type marker gene sets within these DMR cluster target gene sets. The target gene sets for each DMR cluster were computed using the proximity-based DMR-to-gene annotations introduced in section 2.2.3. The target gene set for each DMR cluster was defined as the set of all genes that had at least one DMR from the DMR cluster annotated to them. DMR-to-gene annotations with large distances between the DMR and the putative target genes were not considered, because such long-range regulatory associations can only be made with low confidence when using proximity-based DMR-to-gene annotations [51, 142]. Specifically, only DMRs which were located within ±15 kb of the TSS of the gene they were annotated to were considered. Next, for the target gene sets of each DMR cluster, enrichment of the cell type expression marker gene sets was computed against the background of all other DMR clusters, using Fisher's exact test. P-value adjustment into q-values was performed using the BH method [154, 155]. The individual cell type expression marker gene sets were specifically enriched within the target gene sets of DMR clusters with matching regulatory patterns (Figure 14C). The target genes of DMR clusters specifically marking individual hematopoietic populations were enriched in expression markers for the corresponding cell type. For example, the DMR clusters E1-E4 (all exclusively marking the CFU-E population) showed association exclusively with the erythroid progenitor and erythroblast cell type marker gene sets. The population specific M4 DMR cluster (marking neutrophils), M5 DMR cluster (marking eosinophils), D3 DMR cluster (marking pDCs), D4 DMR cluster (marking cDC1), L2 and L4 DMR clusters (marking T cells) and L3 DMR cluster (marking B cells) similarly showed significant association specifically with the cell type marker gene set corresponding to their marked population. Moreover, the target genes of DMR clusters with broader population specificity were enriched with cell type marker genes for corresponding progenitor cell types. For example, the DMR cluster target genes of the M1 (pan-myeloid) DMR cluster were most strongly enriched with the monocyte and neutrophil progenitor cell type marker genesets. Finally, the target genes of the H1 and H2 gain of methylation clusters were both characterized by enrichment of cell type marker genes for the HSC, ST-HSC, and MPP cell types. In addition, the H1 DMR cluster was associated with Megakaryocyte cell type marker genes. This is in line with several reports that the default lineage bias of immunophenotypic HSCs may default to the megakaryocyte lineage [167]. Taken together, the potential target genes of the individual DMR clusters were each enriched in scRNA-seq-based cell type markers for matching progenitor and mature hematopoietic cell types.

This enrichment analysis provided a high-level characterization of the expression patterns of the potential target genes for each DMR cluster. However, the classification of gene expression patterns into cell type marker genes strongly simplifies potentially complex expression patterns. Cell type marker genes as conventionally defined [145, 168] are not necessarily specifically expressed in the cell type they mark - they may show additional expression at similar or reduced levels in other cell types. Therefore, I next investigated whether the gene expression patterns of the potential target genes for each DMR cluster matched the regulatory pattern of the DMR cluster in finer detail. For this purpose, I computed for each DMR cluster the average gene expression across all its potential target genes for each cell in the scRNA dataset. For the computation of this average, gene expression vectors were normalized to equally weigh all genes. I then projected the resulting DMR cluster target gene set scores on a UMAP embedding of the scRNA-seq data (Figure 14D). This gene set score quantitatively reflects enrichment of elevated gene expression levels across the DMR cluster target genes.

In line with the cell type marker gene enrichment analysis, for each DMR cluster, the average DMR cluster target gene set expression was strongest in the populations with the strongest hypomethylation (the marked populations) in that DMR cluster. Additionally, the gene set expression score revealed a clearly correlated relationship between partial hypomethylation observed in other populations besides the marked populations in a DMR cluster and partial expression of the target genes of this DMR cluster in matching cell types. Various representative examples are shown in Figure 14D. The highly population-specific hypomethylation observed for example in the E4 DMR cluster (in the CFU-E population) was matched by highly specific expression of the E4 DMR cluster target genes in the erythroid progenitor and erythroblast cell types within the scRNA-seq dataset. Similar correlation was observed for other population-specific DMR clusters such as the the M4 cluster (for the neutrophil population), the D3 cluster (for the pDC population), the L3 cluster (for the B cell population), and the L4 cluster (for the T cell population), as well as the H1 DMR cluster (for the HSC population). This correlation was also observed for DMR clusters with less specific regulatory patterns (i.e., with broadly occurring hypomethylation across multiple populations). One example is the M1 DMR cluster, which exhibited pan-myeloid hypomethylation as well as partial hypomethylation in dendritic cell populations. The target genes of this cluster were, on average, highly expressed in the neutrophil and monocyte scRNA-seq cell clusters, and partially expressed in the myeloid progenitor and dendritic cell type scRNA-seq cell clusters. In summary, the hematopoietic expression pattern of each DMR cluster target gene set, observed across the scRNA-seq cell clusters, closely matched the methylome programming pattern of the DMR cluster. This suggests that the target genes for each DMR cluster are enriched in genes whose hematopoietic expression pattern matches the DMR programming pattern represented by the cluster.

2.3.3 Co-regulation within CREs by DMR programming and enhancer establishment

While, prior to this work, only an incomplete picture of methylome remodeling during hematopoiesis was available, the chromatin state dynamics during hematopoiesis have received more attention. An important atlas of hematopoietic enhancers was presented by Lara-Astiaso et al. [107]. This atlas comprises nine population- and lineage-specific enhancer clusters. I next investigated whether these enhancer clusters significantly overlapped with the DMR clusters identified in this study. Because Lara-Astiaso et al. only provided the raw data (H3K4me1 read counts per enhancer region) underlying the enhancer clustering described in their study, I repeated the clustering analysis as described in the paper. This resulted in the identification of enhancer clusters which closely resembled the described enhancer clusters. These enhancers clusters were therefore named using the same names which were proposed in the study to characterize their activity patterns: i) ”Erythroid”; ii) ”Erythroid+Progenitors”, primarily characterized by enhancer activity across the erythroid lineage; iii) ”T/NK cells”; iv) ”B cells”; v) ”Myeloid cells”, primarily characterized by enhancer activity in monocytes and granulocytes; vi) ”Progenitors”, primarily characterized by enhancer activity in LT-HSCs, ST-HSCs, MPPs, CMP and CLPs; vii) ”Lymphoid+Progenitors”, primarily characterized by enhancer activity across the lymphoid lineage; viii) ”Myeloid+Progenitors”, primarily characterized by enhancer activity across the myeloid lineage; and ix) ”Common”, characterized by pan-hematopoetic enhancer activity.

Next, I tested whether these enhancer clusters were enriched in each DMR cluster (Figure 15). Enrichments were tested against the background of all other DMR clusters using Fisher’s exact

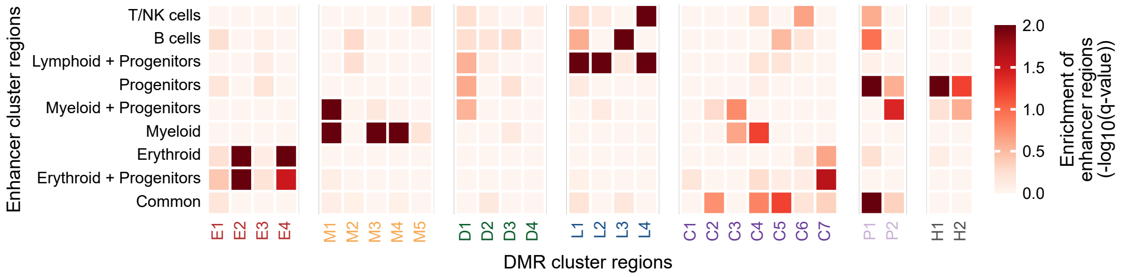

Figure 15: Co-regulation within CREs by DMR programming and enhancer establishment. An atlas of population- and lineage-specific hematopoietic enhancer clusters was recomputed based on data and methods from Lara-Astiaso et al. [107]. For each DMR cluster, enrichment of overlaps with each enhancer cluster was tested. Tests for each DMR cluster were performed against the background of all other DMR clusters using Fisher's exact test. P-value adjustment into q-values was performed using the BH method [154, 155]. The heatmap shows the significance of the DMR cluster vs. enhancer cluster overlaps by encoding the -$\log_{10}$(q-values) of the corresponding enrichment tests.

test. P-value adjustment into q-values was performed using the BH method [154, 155]. The enhancer atlas does not contain enhancers for dendritic and eosinophil cells. Consequently, no significant association between the DMR clusters specifically marking these populations and any enhancer cluster was found. Many of the remaining population- and lineage-specific DMR clusters were strongly associated with enhancer clusters with matching specificity. The E2 and E4 DMR clusters were strongly enriched in enhancers from the ”Erythroid” and ”Erythroid+Progenitors” enhancer clusters, while the E1 and E3 DMR clusters did not show a significant association, possibly due to the relative small size of these DMR clusters, or a distinct role of these DMR clusters which may be related to the partial hypomethylation in T cells (E3) and cCDs (E1) observed in these DMR clusters. The myeloid M1, M3 and M4 clusters were all enriched in enhancers from the ”Myeloid” cluster. In addition, the pan-myeloid M1 cluster, which showed the strongest hypomethylation in myeloid progenitor populations, was also enriched in enhancers from the ”Myeloid+Progenitors” cluster. The B cell (L3) and T cell (L4) specific clusters were enriched in enhancers with corresponding specificity (from the ”B cells” and ”T/NK cells” enhancer clusters). The H1 and H2 DMR clusters were enriched in enhancers from the ”Progenitors” enhancer cluster, characterized primarily by the specific activity of these enhancers in the HSPC compartment. DMR clusters exhibiting hypomethylation across multiple lineages also demonstrated matching associations with enhancer clusters as far as possible given the scope of the enhancer atlas. The monocyte and cDC specific C3 DMR cluster was enriched in ”Myeloid+Progenitors” and ”Myeloid” enhancers (dendritic cell-specific enhancer clusters were not included in the dataset), while the erythroid and eosinophil specific C7 DMR cluster was enriched in ”Erythroid+Progenitors” as well as ”Erythroid” enhancers (eosinophil-specific enhancers were not available). The P1 DMR cluster was strongly associated with the ”Progenitors” and the ”Common” enhancer clusters which are characterized by broad activity across the hematopoietic system starting from early progenitors. In summary, the DMR clusters were enriched in overlaps with enhancer clusters with matching lineage- and population-specificity.

2.4 DMR seeding and expansion during hematopoietic differentiation

2.4.1 DMRs expand progressively during hematopoietic differentiation

The DMR clustering analysis presented in section 2.3.1 has demonstrated that all detected DMR clusters exhibited progressively changing DMR methylation levels across multiple hematopoietic populations (Figure 13). Progressive loss of DMR methylation could indicate DMR deepening (i.e., homogeneously decreasing CpG methylation levels for a fixed set of CpGs) and/or DMR expansion (i.e., additional, newly-regulated CpGs show significant hypomethylation in a second population). Of course, analogous logic applies to DMR clusters characterized by progressive gain of methylation.

To investigate the mechanism underlying the observed progressive DMR methylation changes, I first assessed how the profile of the DNAme levels within the DMRs of each DMR cluster changed between populations Figure 16. These profiles describe the DNAme levels across the individual DMCpGs within each DMR and are briefly referred to as DMR profiles in the following. To enable the comparison of DMR profiles across DMRs of varying sizes, the DMRs and their flanking regions (±200 bp) were segmented into bins. Then a methylation level vector was computed for each DMR, containing the average CpG methylation levels across these bins along the plus strand. These methylation level vectors thus captured the DMR profiles of individual DMRs in a size independent manner. An initial inspection of the data suggested that many DMRs expanded asymmetrically across a series of populations: within such DMRs, DMR expansion occurred primarily along either the plus or the minus strand (with each case being observed in about half of the DMRs). To align all the methylation level vectors of the individual DMRs by their DMR expansion direction, I reversed the methylation level vectors of all DMRs where the major direction of DMR expansion occurred along the minus strand (Methods, section 4.6.2). Taken together, I have computed DMR profiles for all DMRs in the hematopoietic DMR/DMCpG atlas, introducing an innovative concept for DMR profile alignment. This alignment is crucial for the correct interpretation of trends across sets of DMR profiles, as further detailed in the discussion (section 3.5.3).

As an illustrative example for the DMR profile changes observed during hematopoietic populations, Figure 16A shows the DMR profiles of the individual D1 cluster DMRs in the exemplary HSC, MPP3, CMP CD55^{-}, CDP, cDC1, GMP and monocyte populations. Moreover, Figure 16B presents the average DMR profile in the D1 cluster for each of these populations. For the computation of the average DMR profiles, DMRs were stratified by DMR size, to allow for the comparison of the DMR profiles between DMRs of different sizes. Figure 16B additionally shows the average DMR profiles for the cDC2 and pDC populations, to complete the picture of the DMR states in the mature dendritic cell populations. As previously described (section 2.3.1), the D1 DMR cluster was characterized by strong DMR hypomethylation in all mature dendritic populations and showed progressively decreasing DMR methylation levels across the progenitor populations GMP, CMP CD55$^-$, MDP, and CDP, as well as outside of the dendritic lineage across the cMoP and monocyte populations. The DMR profiles of these D1 cluster DMRs now revealed that this progressive loss of DMR methylation could to a considerable degree be attributed to progressive DMR expansion during differentiation. This DMR expansion occurred asymmetrically in a substantial fraction of the DMRs. The lowest DMR methylation levels in the D1 cluster were observed in the cDC1, cDC2 and pDC populations. This corresponded to strong, relatively homogeneous hypomethylation across the full DMR intervals in these populations. The progressively decreasing, intermediate DMR methylation levels occurring in progenitor or sibling populations of the mature dendritic populations were the result of averaging over a relatively strong hypomethylation in a progressively expanding subpart of the DMRs together with high methylation levels in the remaining subparts of the DMRs. Interestingly, for a significant fraction of the DMRs, methylation loss in small DMR subparts was already apparent within the HSC population itself, suggesting initial seeding of some DMRs already within this population. In addition to this progressive DMR expansion, a complementary DMR deepening effect was also observed, i.e., the methylation levels decreased progressively for the same CpGs across multiple populations. The same pattern of combined DMR expansion and deepening was observed across DMRs of all sizes, becoming more pronounced with larger DMR sizes (Figure 16B). In summary, initial hypomethylation in the D1 cluster DMRs often emerged in progenitor cells, including within the HSC population in a substantial fraction of the DMRs. This initial DNAme programming was characterized by hypomethylation limited to only a small subregion of the full DMR interval (referred to as ”seed” region in the following). This seed region was then progressively expanded across downstream populations.

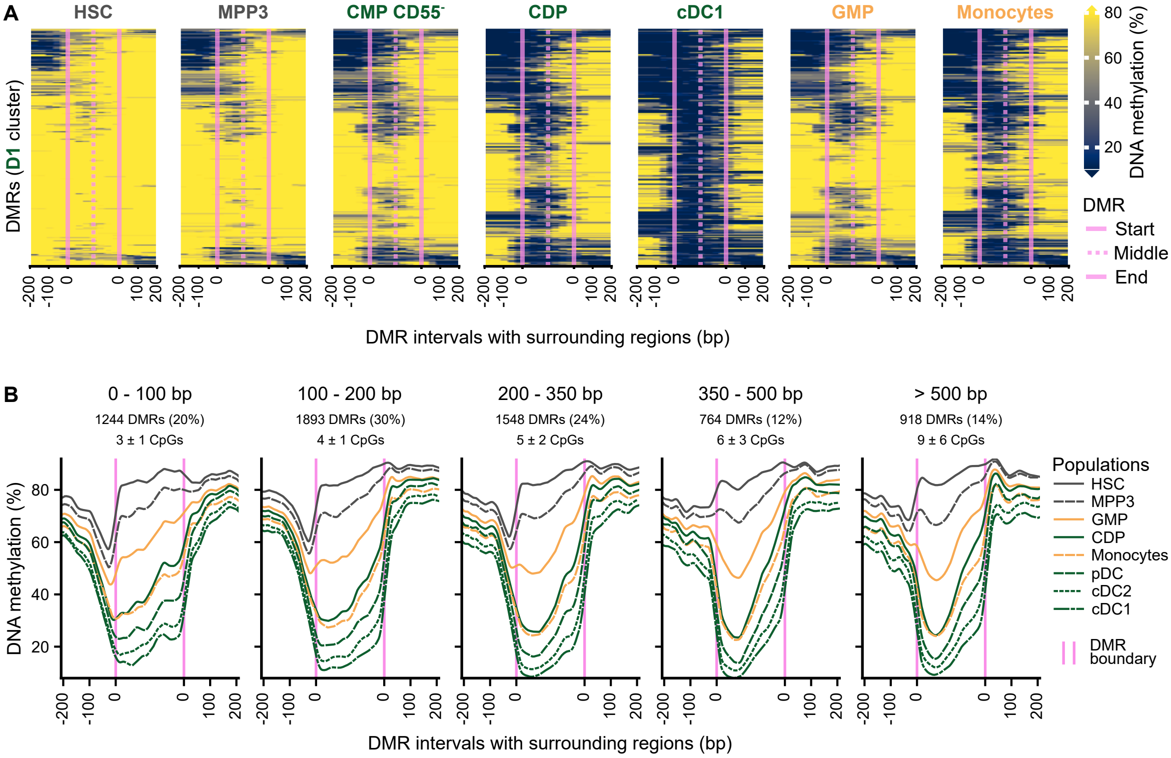

Figure 16: DMR regions expand progressively during differentiation. DNA methylation profiles in and around (±200 bp) the D1 cluster DMRs (referred to as DMR profiles) were computed. DMRs were binned into 21 bins with an average size of 10 bp, and flanking regions were similarly binned into 10 bp bins. For each DMR, a methylation level vector containing the average methylation levels in these bins was computed. Many DMRs expanded asymmetrically from a hypomethylated seed region arising in progenitor populations, with expansion occurring more strongly along the plus or the minus strand in about half of the DMRs, respectively. To avoid averaging out this expansion behavior in aggregate views on the data, I aligned all methylation level vectors by their DMR expansion direction prior to computing the mean vectors. To achieve this, I reversed the methylation level vectors of all DMRs for which expansion occurred along the minus strand. (A) Heatmaps of individual DMR profiles within the D1 DMR cluster, for the populations showing progressive DMR methylation level loss in this cluster. Shown are DMR profiles for 500 randomly sampled, representative DMRs. Missing values (i.e., bins without CpGs in a given DMR) were interpolated by convolution with a gaussian kernel for the heatmap display. (B) Average DMR profiles for the D1 cluster DMRs, stratified by DMR size. Profiles are shown for the same populations as in A, complemented with the remaining mature dendritic cell populations.

2.4.2 A map of DMR expansion states across the hematopoietic system

I next classified the DMR expansion state of each DMR in each population into four discrete states: unregulated, seeded, intermediate and completed. This provided a detailed DMR expansion state map across the hematopoietic system. Specifically, for each DMR in each population, I first determined the number of regulated DMCpGs. DMCpGs were considered regulated if they exhibited an absolute methylation level shift of at least 30% compared to the HSC methylation level. For the MPP1-5 populations, this threshold was lowered to a shift of 20%. The threshold was reduced for the MPP1-5 populations, because, due to their high heterogeneity, a significant shift of the frequency of DNAme at a given CpG may only occur in a population subset and still be biologically meaningful. For each DMR, I noted the maximal number of regulated DMCpGs observed in any population. The DMR state for each population was determined based on the percentage of regulated DMCpGs relative to the maximum observed count: DMRs were thus classified as unregulated (0% regulated CpGs), seeded (< 45%), intermediate (< 81%), or completed (≥ 81%).

The DMR expansion state classification globally quantified the DMR expansion patterns suggested by the visual assessment of DMR expansion in the D1 DMR cluster (Figure 16). Figure 17 presents the proportion of these DMR expansion states in the different DMR clusters for all populations. For the D1 DMR cluster, this aggregate view of the DMR expansion state classification further characterizes how DMR expansion proceeds. Besides the MPP3 population, also the MPP2 and MPP4 populations display a considerable fraction of DMRs in a partially expanded (seeded or intermediate) state. The fraction of DMRs in a partially expanded DMR state then increases across the CD55^{-}, GMP, MDP, and CDP populations, shifting from seeded towards intermediate and completed states. In both the cDC1 and cDC2 populations, all DMRs are fully completed, and almost all DMRs are fully completed in the pDC population. Outside of the classical dendritic differentiation trajectory, seeded and intermediate DMR states are most prominent in the cMoP and monocytes. However, small amounts of partial DMR expansion (predominantly restricted to seeded DMR states) are visible in multiple other populations, including mature erythroid, myeloid, and lymphoid populations. This example demonstrates that the proposed DMR expansion state classification can be used to assess patterns of DMR expansion across the hematopoietic system. Thus the DMR expansion state classification provides a valuable, novel layer extending the hematopoietic DMR/DMCpG atlas.

Similar patterns of DMR expansion characterized all of the DMR clusters, varying only in the level of population specificity. For example, the D4 DMR cluster exhibited a highly specific regulatory pattern, characterized by strong hypomethylation in the cDC1 population and partial hypomethylation observed mainly in the CDP, cDC2, cMoP and monocyte populations. Correspondingly, completely expanded DMRs were almost exclusively observed in the cDC1 population, and partially expanded DMR states mainly occurred in the partially hypomethylated populations (Figure 17). In contrast, the C3 DMR cluster had a broader regulatory pattern, with the strongest hypomethylation observed simultaneously in cDC1, cDC2 as well as monocytes and partial hypomethylation observed across all myeloid and dendritic populations. Correspondingly, the majority of the C3 cluster DMRs were completely expanded in each of the monocyte, cDC1, and cDC2 populations, while a significant fraction of mostly partially expanded DMRs was observed across all other myeloid and dendritic progenitor and sibling populations as well as in the MPP2/3/4 populations. This indicated broad occurrence of partial DMR expansion across the hematopoietic system in this DMR cluster (Figure 17). In summary, methylome remodeling along various trajectories of the hematopoietic differentiation system appears to be accompanied by progressive DMR expansion in many DMRs. This is likely a main mechanism underlying the progressive DMR methylation level changes characterizing all DMR clusters. In addition, for most DMR

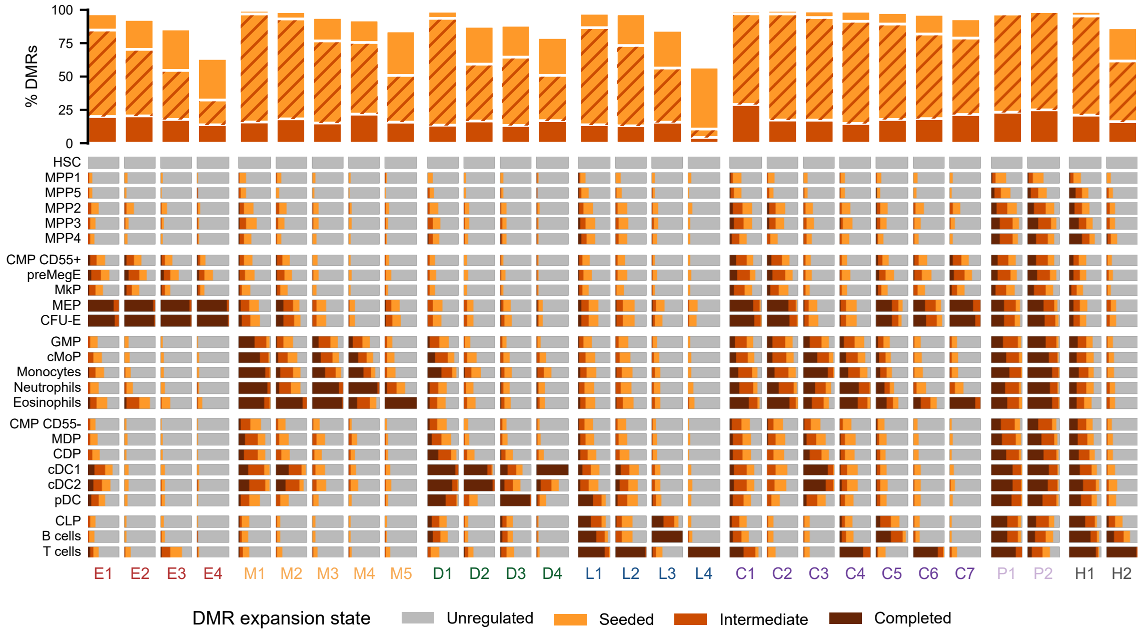

Figure 17: A DMR expansion state layer for the hematopoietic DMR/DMCpG atlas. The DMR expansion state of each DMR in each population was classified as unregulated, seeded, intermediate or completed. Specifically, for each DMR in each population, first the number of regulated DMCpGs was determined, defined as DMCpGs with a methylation level shift of at least 20% compared to the HSC reference level for the MPP1-5 populations and of at least 30% for all other populations. For each DMR, I noted the maximal number of regulated DMCpGs observed in any population. The DMR state for each population was then determined based on the percentage of regulated DMCpGs relative to the maximum observed count: DMRs were thus classified as unregulated (0% regulated CpGs), seeded (< 45%), intermediate (< 81%), or completed (≥ 81%). DMRs with at least five regulated DMCpGs were considered to be in an intermediate expansion state, even if these five DMCpGs represented less than 45% of all DMCpGs in the DMR. The heatmap shows the proportion of these DMR states within each DMR cluster for each population. The heatmap annotation indicates the percentage of DMRs in each DMR cluster for which a seeded and/or an intermediate DMR expansion state was observed. The combined occurrence of seeded and intermediate states is indicated by a hatch pattern combining the colors indicating seeded and intermediate DMR expansion states. Each DMR reaches the completed DMR expansion state at least once by definition, therefore this state is not considered for this annotation.

clusters, populations conventionally assumed to reside outside of the primary differentiation trajectory associated with the DMR cluster also exhibited partial DMR expansion.

Next, I used the DMR expansion state classification to quantify the proportion of progressively expanding DMRs observed within each DMR cluster (Figure 17, top panel). Expanding DMRs were defined as DMRs exhibiting a seeded and/or an intermediate DMR expansion state. Of course, this state occurred in addition to the unregulated and completed DMR expansion states, which each DMR exhibited by definition. Across all DMR clusters, a high fraction of hematopoietic DMRs showed DMR expansion (in total 109,914 DMRs, 89% of all DMRs), with an average fraction of expanding DMRs across the individual DMR clusters of 91.64 ± 10.63% (mean ± s.d.). Often, both seeded and intermediate states were observed for the same DMR, representing progressive expansion with multiple intermediate steps (67,934 DMRs, 55.43% of all DMRs). The percentage of expanding DMRs in each DMR cluster was inversely correlated with the population specificity of the regulatory pattern of the DMR cluster. For example, the pan-myeloid M1 cluster exhibited 99.6% expanding

DMRs, in contrast to only 84.2% for the neutrophil-specific M5 cluster. This correlation was particularly pronounced when only the intermediate expansion state, but not the seeded DMR expansion state, was considered (97.0% and 50.8% of the DMRs in the M1 and M5 cluster exhibited an intermediate DMR expansion state, respectively). Thus, for many highly population-specific DMR clusters, programming in large subparts of the DMR intervals appears strongly associated with fate commitment. These subparts thus may serve as population marker regions. Conversely, for less population-specific DMR clusters, programming in large subparts of the DMRs may occur prior to fate commitment, as suggested by the broader occurrence of hypomethylation (or hypermethylation for gain of methylation clusters) in these DMR subparts. Still, also for these DMRs, other DMR subparts appear to be exclusively hypomethylated in specific populations (namely the populations marked by their DMR cluster). Taken together, progressive DMR expansion appears to be an almost ubiquitous pattern of methylome reprogramming within all DMR clusters and often involves multiple distinguishable expanding steps within a single DMR.

2.4.3 Widespread methylome programming in MPP populations

I next used the DMR expansion state classification to investigate the regulation of DMRs in the MPP2, MPP3 and MPP4 populations. All DMR clusters, including those with a highly population-specific regulatory pattern, contained a considerable number of DMRs carrying hypomethylation (or hypermethylation for the gain of methylation clusters) in cells of the MPP2, MPP3 and MPP4 populations. However, DNAme programming was predominantly restricted to subparts of the DMR regions (Figure 18). This spatial restriction was seen in all DMR clusters, and it was particularly pronounced in the lineage and multi-lineage-specific clusters (i.e., the E, M, D, L, and C clusters). In these clusters, on average only 2% of all DMRs showed completed DMR expansion in the MPP2-4 populations, while 20% of the DMRs showed spatially restricted methylome programming, with 6% of the DMRs in an intermediate and 14% in a seeded DMR expansion state. Regulation restricted to DMR subregions was still the dominant form of DNAme programming in the MPP populations for the P1, P2, H1 and H2 DMR clusters, although the spatial restriction was less pronounced. An average across the MPP2-4 populations, 27% of all DMRs in the pan-hematopoietic (P1 and P2) clusters exhibited a completed DMR state, while 60% of the DMRs showed spatially restricted methylome programming, with 38% in an intermediate and 22% in a seeded DMR expansion state. This suggests that, as a shared feature across the widely differing DMR clusters, spatially restricted regulation within subparts of DMRs may play a role in early fate priming. The percentage of DMRs within each DMR cluster exhibiting such regulation in the MPP populations was negatively correlated with the population-specificity of the DMR cluster. In the pan-lymphoid L1 DMR cluster, the average percentage of regulated DMRs (DMRs in a seeded, intermediate or completed DMR expansion state) across the MPP2-4 populations was 45% (2006 DMRs). In contrast, only 9% (555 DMRs) of the DMRs in the T cell specific L4 cluster were on average regulated in the MPP2-4 populations. The multi-lineage (C and P) clusters showed even higher percentages of regulated DMRs, reaching 55% of all C1 cluster DMRs (2348 DMRs) on average across the MPP2-4 populations. In summary, DNAme programming in the MPP populations was predominantly restricted to subregions of larger DMR intervals. All DMR clusters showed such DMR seeding in a fraction of their DMRs. The percentage of seeded DMRs in the DMR clusters was correlated with the population specificity of the DMR clusters.

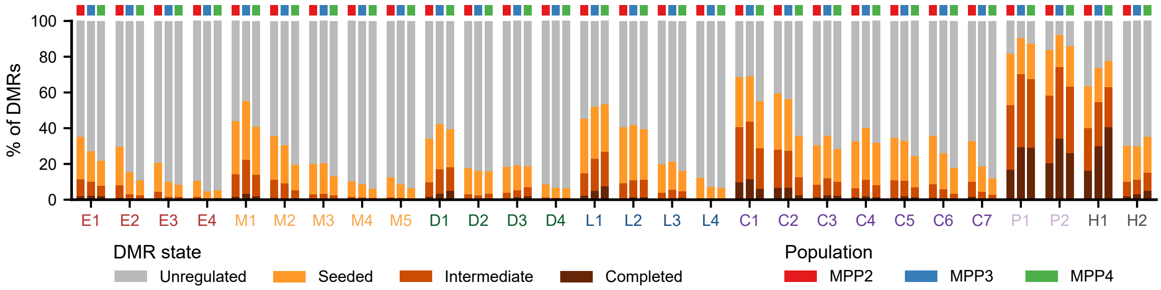

Figure 18: Seeding and expansion of DMRs in MPP populations occurs in all DMR clusters. Bars show the percentages of the DMR expansion states observed in the MPP2, MPP3 and MPP4 populations for all DMR clusters. Details of the DMR expansion state classification and a high-level overview of the DMR expansion states across the full hematopoietic system are presented in Figure 17. Here, an enlarged view of the same data focused on the MPP2, MPP3 and MPP4 populations is shown.

Recent studies have indicated that the MPP2, MPP3, and MPP4 populations likely represent overlapping, heterogeneous samplings of the early hematopoietic differentiation state continuum. Furthermore, these studies have suggested that the MPP2-4 populations may be differentially enriched in progenitor cells with different fates, leading to population-level lineage output biases towards erythroid and megakaryocytic fates (MPP2), myeloid fates (MPP3), and lymphoid fates (MPP4) [12–14, 19, 20, 122, 134] (for a more nuanced view on the MPP1-5 populations, see Introduction, section 1.1.4). In line with these known biases, the MPP2 population exhibited considerably higher percentages of regulated DMRs in clusters associated with erythroid differentiation than the MPP3 and MPP4 populations. This included the erythroid specific E1-4 DMR clusters as well as the C6 and C7 DMR clusters, which were characterized by maximum hypomethylation in both the CFU-E and Erythroid populations. The MPP3 and MPP4 populations showed the highest percentage of DMR regulation in the pan-myeloid M1 DMR cluster and the pan-lymphoid L1 DMR cluster, respectively, which was again in line with their suggested fate potential biases. However, the differences between the MPP2, MPP3 and MPP4 populations became less pronounced and less aligned with the reported fate potential biases when regarding the more population specific myeloid and lymphoid clusters.

Taken together, these observations suggest that the reported fate potential biases of the MPP2, MPP3 and MPP4 populations may be in part associated with differences of the DMR expansion states within lineage-specific DMR clusters between these populations.

2.4.4 Methylome programming often starts with small seed regions

I next sought to quantitatively characterize the size of the DMR subregions in which early DNAme programming in the MPP populations occurred (Figure 19). To measure the size of these subregions, I counted, within each DMR, the number of DMCpGs exhibiting regulation in the MPP populations. As described previously, DMCpGs were considered as regulated in a MPP population if the absolute methylation level difference compared to the HSC population was at least 20%. The regulated DMR subregions were small, both in terms of the number of regulated DMCpGs and with regard to the percentage of regulated DMCpGs (with respect to the full number of DMCpGs in the DMR). As an example, I considered the set of DMRs exhibiting any regulated DMCpG in the MPP4 population, which comprises 37131 DMRs representing 30% of all DMRs in the atlas (Figure 19). Of these DMRs, 79% showed regulation of less than half of their total number of DMCpGs. In absolute terms, the number of regulated DMCpGs in these DMRs was mainly limited to 1 CpG (54% of the DMRs) or 2 CpGs (23% of the DMRs). Similar patterns of DMCpG occurrence were also observed for the MPP2 and MPP3 populations. DMR regulation was even more spatially restricted in the MPP1 and MPP5 populations. Taken together, these findings suggest that methylome programming in MPP populations primarily occurs within small genomic intervals, typically consisting of only one or two CpGs. These small genomic intervals appear to act as seed regions for DMRs which are expanded in the course of differentiation.

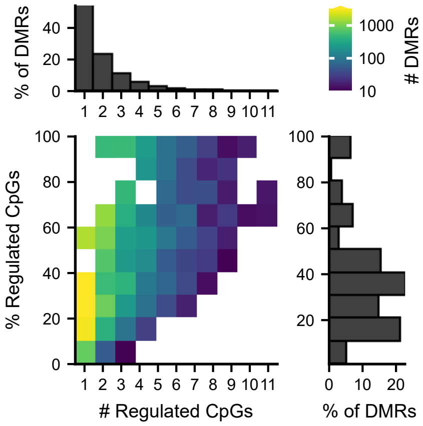

Figure 19: DMR regulation in MPP populations is restricted to small seed regions. Distribution of the size of the regulated DMR subregions in the MPP4 population. The size of the regulated DMR subregions was measured by the number and percentage of regulated DMCpGs (relative to the total number of DMCpGs in the DMRs). DMCpGs were considered regulated in the MPP4 population if they exhibited a methylation level shift of at least 20% compared to the HSC population. In total, 37131 DMRs representing 30% of all DMRs in the atlas exhibited at least one DMCpG regulated in the MPP4 population. Shown are the joint and marginal distributions of the numbers and percentages of DMCpGs in these DMRs.

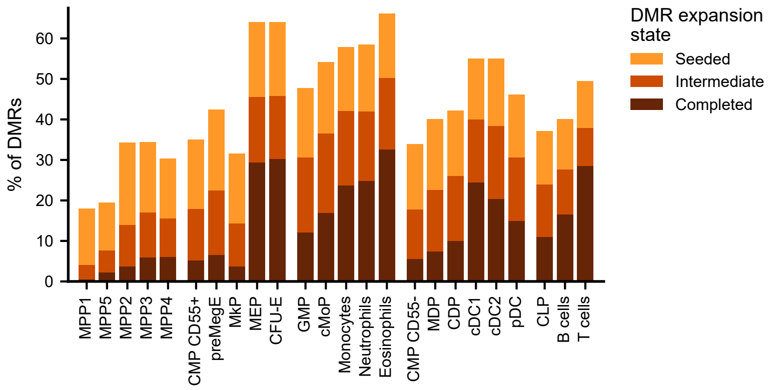

Figure 20: Mature populations exhibit a mixture of DMR expansion states across the hematopoietic DMRs. Barplots show, for each population, the percentage of hematopoietic DMRs in a seeded, intermediate or completed DMR expansion state. Details of the DMR expansion state classification and a high-level overview of the DMR expansion states across the full hematopoietic system are presented in Figure 17.

2.4.5 Mature cells exhibit hypomethylated seed regions in DMRs associated with alternative fates

The DMR expansion state classification also provided valuable insights into the methylome state of mature hematopoietic cells. Figure 20 shows, for each population, the percentage of hematopoietic DMRs exhibiting a seeded, intermediate or completed DMR expansion state. Each mature population was primarily characterized by a set of marker DMRs (Figures 13 and 17). These DMRs reached their maximum expansion and hypomethylation specifically in this population or in a set of populations including this population. For example, neutrophils were characterized by DMRs from the M4 DMR cluster, which exclusively marked this population, as well as by DMRs from other DMR clusters marking multiple populations, such as the M1 (pan-myeloid) and P2 (pan-hematopoietic) DMR clusters. In addition, each mature population exhibited a seeded or intermediate DMR expansion state in a considerable fraction of the remaining DMRs. These DMRs reached their maximum hypomethylation and full DMR expansion in other mature populations. For example, the neutrophil population exhibited a considerable fraction of such DMRs across multiple clusters, such as i) the M2 and M5 DMR clusters (where the maximum DMR expansion was observed in the eosinophil population); ii) the D1 cluster (maximum DMR expansion in the dendritic cell populations); and iii) the C3 cluster (maximum DMR expansion in the cDC1, cDC2 and monocyte populations). Taken together, each mature population exhibited a complex landscape of DMR expansion states across the overall set of hematopoietic DMRs. In each mature population, this landscape was made up of a mixture of i) DMRs in a completed DMR expansion state; ii) DMRs in a seeded or intermediate DMR expansion state, which were further extended in other mature populations; and iii) of fully unregulated DMRs which were only programmed in other mature populations. This suggests that many individual mature hematopoietic cells carry hypomethylated seeds in a significant number of DMRs whose primary role during differentiation appears to relate to the development and function of alternative cell types.

2.5 Hierarchical DNA methylation programming at the DMR and DMCpG level

2.5.1 Rationale: DMR expansion as the result of DMCpG-resolved programming within DMR intervals

In section 2.4, progressive DMR expansion was identified as a common mechanism of DNAme programming during hematopoietic differentiation. The progressive expansion of a DMR interval indicates the progressive addition of newly regulated DMCpGs to an existing DMR site over the course of differentiation. This entails that, within a DMR that exhibits DMR expansion, the programming of different DMCpGs starts at different stages of differentiation. Thus, in this scenario, different DMCpGs within the same DMR interval would be subject to substantially different DMCpG programming. I hypothesized that the apparently heterogeneous DMCpG programming within DMR regions, observed through the widespread existence of expanding DMRs, was not the result of stochastic patterns arising during DNAme remodeling. Instead, I reasoned that the observed differences between DMCpGs within individual DMR regions could represent the activity of systematic mechanisms acting to alter DNAme at the level of individual DMCpGs. Various effector proteins related to DNAme programming are known to be able to effect narrow DNAme changes, such as pioneering transcription factors [49, 60] (see also Introduction, section 1.3). Such effector proteins could underlie DMCpG-level programming modules, which could systematically regulate target DMCpGs across the genome. This could, for example, involve binding of a given pioneering transcription factor at many distinct binding sites. Thus, such DMCpG-level programming modules would be likely to act concurrently across many DMRs at once. Under this model of DMCpG-resolved programming within DMR intervals, similar patterns of DMCpG programming should recur across many individual DMCpGs, contained in various DMRs across the genome.

2.5.2 Characterization of lineage- and population-specific DMCpG programming modules

I next sought to identify clusters of DMCpGs sharing the same DMCpG programming pattern, independent of whether they were located in spatial proximity to each other. For this purpose, I pooled the individual DMCpGs across all hematopoietic DMRs. Then I clustered the individual DMCpGs as explained previously for the clustering analysis of the DMR regions (section 2.3.1). Briefly, to ensure that the DMCpG clustering analysis was not biased by the bulk progenitor populations in our dataset, clustering was based on only the data from the HSC population and the mature populations (the CFU-E, monocyte, eosinophil, neutrophil, B cell, T Cell, cDC1, cDC2, and pDC populations). DMCpG clustering was performed with the unsupervised Leiden community detection algorithm [143], using the correlation distance. This resulted in the identification of 30 DMCpG clusters (Figure 21A). CpG cluster sizes varied widely (Figure 21B), ranging from 4624 DMCpGs (/c5/ cluster) to 43207 DMCpGss (/e3/ cluster), with an average cluster size of 19 792 ± 9424 DMCpGs (mean ± s.d.). The scale invariance of the correlation distance was not of concern for this dataset because all DMCpGs exhibited considerable methylation level shifts during hematopoiesis (Figure 21C): all individual DMCpGs showed a methylation level shift compared to the HSC population of at least 30%. Furthermore, the strong DMCpG cluster compactness and separation observed for z-score transformed DMR methylation levels (Figure 21A) were largely retained when comparing the DMCpG methylation levels directly (Figure S4). Taken together, 30 DMCpG clusters were identified, grouping DMCpGs by DMCpG programming pattern independent of their genomic adjacency.

The individual DMCpG clusters grouped DMCpGs characterized by distinct DMCpG programming patterns, with similar properties as the DMR programming patterns characterizing the DMR clusters (Figure 13A). Therefore, a nomenclature analogous to that used for the DMR clusters was applied. To distinguish the DMR clusters from the CpG clusters, uppercase names were used for DMR clusters (e.g., H1), while lowercase names in italic were used for the CpG clusters (e.g., *l1*). Two of the DMCpG clusters (*h1* and *h2*) were characterized by DMCpG programming patterns associated with gain of methylation compared to the HSC population. The remaining 28 DMCpG clusters captured DMCpG programming associated with loss of methylation compared to the HSC population. In total, 53268 DMCpGs (9% of all DMCpGs) were part of the gain of methylation DMCpG clusters, while 540521 DMCpGs (91% of all DMCpGs) were part of loss of methylation DMCpG clusters. The programming patterns of the loss of methylation DMCpG clusters were characterized by identifying a set of marked populations for each DMCpG cluster. The marked populations for a DMCpG cluster were defined to consist of the population with the lowest average DMCpG methylation level in that cluster and of all populations with an average DMCpG methylation level within 15% of that value. In total, 18 DMCpG clusters exclusively marked populations from a single lineage. Their cluster names were prefixed to indicate their lineage-specificity, including three DMCpG clusters specifically marking the CFU-E population (*e1*-*e3*), five DMCpG clusters specifically marking myeloid populations (*m1*-*m5*), four clusters specifically marking dendritic cell populations (*d1*-*d4*) and five clusters specifically marking lymphoid populations (*l1*-*l5*). Nine DMCpG clusters marked populations across multiple lineages (cross-lineage clusters *c1*-*c9*). Finally, two DMCpG clusters marked more than seven populations across at least three lineages (pan-hematopoietic clusters *p1* and *p2*). In summary, the DMCpG clusters were characterized by DMCpG programming patterns with different levels of lineage-specificity, ranging from population-specific across lineage-specific to pan-hematopoietic

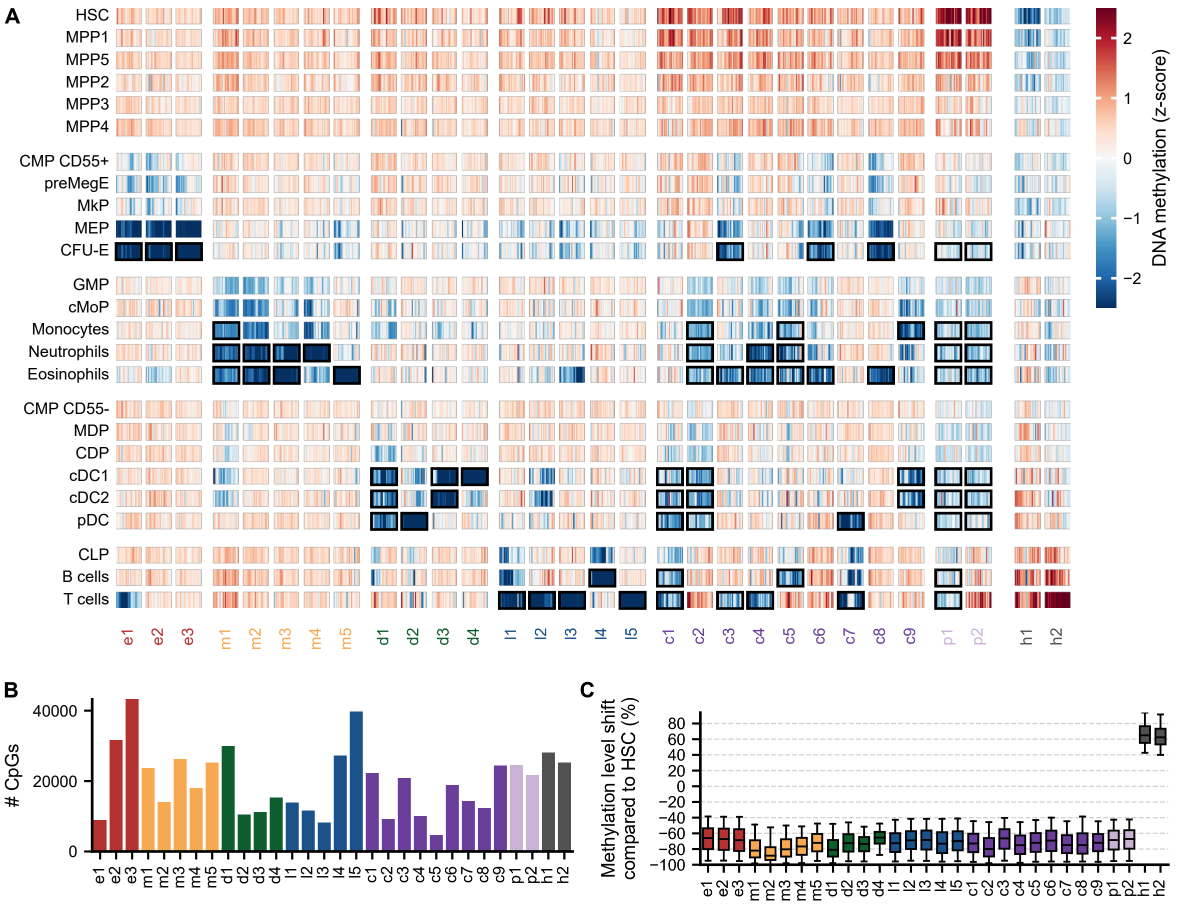

Figure 21: Clustering analysis reveals population- and lineage-specific DMCpG programming modules. (A) Heatmap showing z-score transformed methylation levels of 200 randomly selected DMCpGs for each of the 30 identified DMCpG clusters. Clustering was performed on the DMCpG methylation levels of the mature populations and the HSC population, using Leiden clustering with the correlation distance. To distinguish the cluster names of DMR and DMCpG clusters, DMR cluster names are written in uppercase (e.g., H1), while DMCpG cluster names are written in lowercase and additionally set in italic within text (e.g., /h1/). The /h1/ and /h2/ DMCpG clusters were characterized by low methylation levels in the HSC population and subsequent gain of methylation in downstream populations. All other clusters were characterized by loss of methylation compared to the HSC population. For each of these clusters, black rectangles indicate the population with the lowest average DMR methylation level in the cluster as well as all mature populations with an average DMR methylation level within 15% of that value (referred to as the marked populations for that cluster). Clusters were grouped according to the lineage-specificity of their marked populations: within the cluster names, the prefixes E, M, D, and L indicate erythroid, myeloid, lymphoid, or dendritic cell lineage-specific clusters, respectively; the prefix C (for cross-lineage) indicates DMCpG clusters marking populations across multiple lineages; the prefix P indicates pan-hematopoietic DMCpG clusters, marking seven or more hematopoietic populations across three (/p2/) or four (/p1/) lineages. The DMCpG clusters were ordered by increasing population-specificity of their programming pattern within each group of DMCpG clusters, as indicated by the ordinal number within the cluster name.
(B) DMCpG cluster sizes, ranging from 4624 DMCpGs to 43207 DMCpGs.
(C) Distribution of the DMCpG methylation level shifts compared to the HSC population for each DMCpG cluster. The methylation shift for each DMCpG was computed as the methylation level difference between the HSC population and the population with the most different methylation level at this DMCpG site. Whiskers indicate the second and 98th percentile of the DMCpG methylation level shifts for each DMCpG cluster.

programming patterns. The lineage-specificity of each DMCpG cluster was encoded in the cluster name.

As observed for the DMR clusters, within each DMCpG cluster group, individual clusters differed in how broadly methylation loss (or methylation gain for the gain of methylation clusters) was observed across the full spectrum of the mature populations. Analogously to the nomenclature of the DMR clusters, a hypomethylation-specificity score for each DMCpG cluster was computed, by first calculating the average DMCpG methylation levels for all mature populations in this cluster, and then calculating the overall mean of these values, weighted such that each lineage (erythroid, myeloid, lymphoid, dendritic cells) contributed equally to the mean. The specificity of the regulatory patterns was then denoted in the cluster names: the order of the DMCpG clusters within each group reflects the order of their hypomethylation-specificity scores, either in ascending sorting order for the loss of methylation clusters, or in descending sorting order for the gain of methylation clusters. The cluster names thus reflect the observed specificity of the methylation loss or gain observed in the DMCpG clusters.

2.5.3 A novel, hierarchical approach for the annotation of DNA methylation programming patterns

The hematopoietic DMR/DMCpG atlas presented in this thesis provides a genome-wide map of 122561 hematopoietic DMRs, and it pinpoints the location of a total of 593789 DMCpGs contained within these DMRs. A central aspect of this atlas is the compilation of a comprehensive set of annotations at both the DMR and the DMCpG level. In section 2.3.1, DMR clustering analysis was used to annotate each hematopoietic DMR with a DMR cluster membership. This annotation indicated characteristic programming patterns for all individual DMRs. These DMR clusters were then, in turn, annotated with associations to lineage- and population-specific gene and enhancer sets. In this section, DMCpG clustering analysis was used to annotate each individual DMCpG with a DMCpG cluster membership. This annotation indicated characteristic programming patterns for all individual DMCpGs. Together, these annotations build a dual-layer atlas of DNAme programming. These annotations provide a novel, hierarchical view of DNAme programming during hematopoietic differentiation.

2.5.4 Terminology: DMR and DMCpG programming

The hematopoietic DMR/DMCpG atlas presented in this study maps and annotates both hematopoietic DMRs as well as the individual DMCpGs within these DMRs. To be able to clearly address the DMR and DMCpG layers in this atlas, a summary of the terminology developed in the previous sections of this thesis is collected here. The term DNAme programming comprises any changes at a specific genomic site of arbitrary size, to the DNAme level of a population or the DNAme state of a cell. Different kinds of DNAme programming are specifically addressed in this thesis. The term "DMR programming" (abbreviated as DMR-PP in the following) refers to the methylation-dependent regulation of genomic intervals as a whole, which typically have a cis-regulatory function. The term DMR programming may refer to the establishment of new DMRs or the regulation of the regulatory activity of DMRs acting as cis-regulatory elements. In other words, DMR programming refers to DNAme programming where the smallest unit of information encoding is a single DMR interval. The regulatory activity of a DMR as a whole is conventionally measured using aggregate statistics such as the mean DMR methylation level across all DMCpGs within the DMR. The DMRs within each DMR cluster identified in this thesis share a characteristic DMR methylation level profile across the 25 hematopoietic populations in the dataset, referred to as the "DMR programming pattern" of these DMRs. On the other hand, the term "DMCpG programming" refers to the encoding of regulatory information at the level of individual DMCpGs, through molecular mechanisms which specifically target narrow regions around individual CpG sites. Analogous to the terminology for DMRs, the DMCpGs within each DMCpG cluster share a characteristic profile of DMCpG methylation levels across the 25 hematopoietic populations in the dataset, which is referred to as the "DMCpG programming pattern" (abbreviated as DMCpG-PP in the following) of these DMCpGs. A given DMR may contain several DMCpGs that belong to different CpG clusters. To briefly refer to this scenario, the formulation that the DMR "exhibits multiple DMCpG-PPs" is used in the following.

2.5.5 DMCpG-resolved programming within DMR intervals is a ubiquitous mechanism

In total, 88493 DMRs (72% of all DMRs) exhibited more than one DMCpG-PPs (Figure 22). Most DMRs with multiple DMCpG-PPs only exhibited two (50688 DMRs, 41% of all DMRs) or three (24555 DMRs, 20% of all DMRs) different CpG-PPs. More diverse combinations of DMCpG-PPs within individual DMRs were rare (only 13250/11% of all DMRs demonstrated four or more programming patterns). The high frequency of DMRs exhibiting more than one DMCpG-PP suggests that heterogeneous programming of DMCpGs within DMR intervals may be a ubiquitous mechanism of hematopoietic DNAme remodeling.

I next investigated which DMCpG-PPs occurred most frequently in each DMR cluster. I considered each DMR cluster separately. For each DMR cluster, I screened for the presence of each DMCpG-PP. To quantify the presence of a given DMCpG-PP in a given DMR cluster, I calculated the percentage of all DMRs in the DMR cluster containing the DMCpG-PP at least once (Figure 23). For each DMR cluster, only a limited number of DMCpG-PPs were present in a substantial percentage of its DMRs, while the other DMCpG-PPs were not or rarely observed. The set of frequent DMCpG-PPs was different for each DMR cluster. Moreover, many DMCpG-PPs were observed in significant frequencies across multiple DMR clusters.

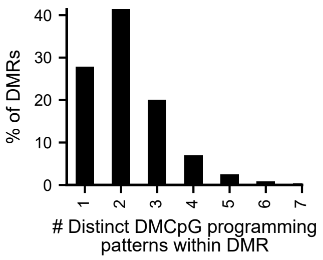

Figure 22: Heterogeneous programming of DMCpGs within DMR intervals is a common mechanism behind hematopoietic DNA methylation remodeling. The histogram shows the frequency of DMRs exhibiting more than one DMCpG programming pattern.

This suggested that the same DMCpG-PP can play a role in the programming of DMRs with distinct DMR-PPs. Taken together, each DMR cluster was associated with a distinctive set of frequently occurring DMCpG-PPs.

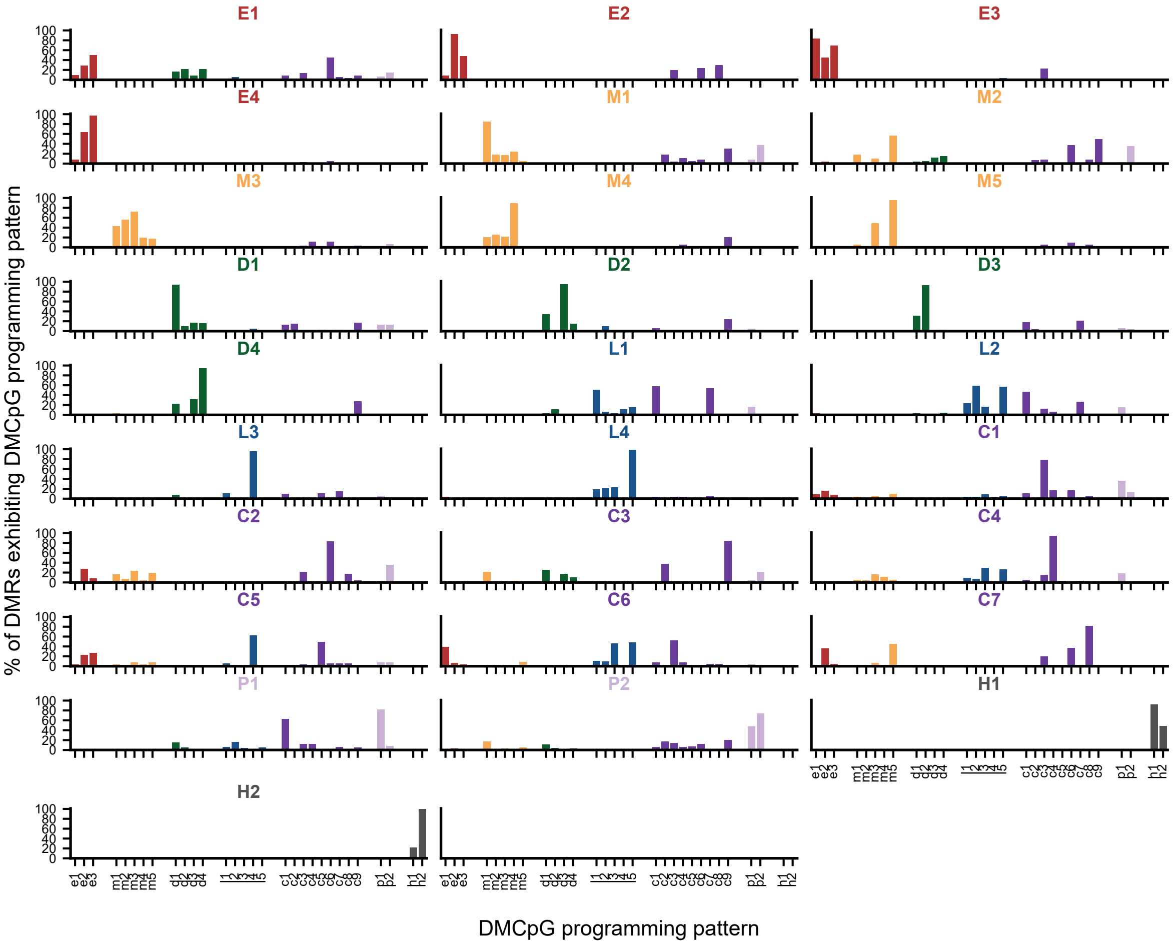

Figure 23: Each DMR cluster is characterized by a distinctive set of frequently occurring DMCpG programming patterns. For each DMR cluster, bar plots show the percentage of DMRs containing at least one DMCpG characterized by each DMCpG programming pattern.

To characterize how the DMCpG-PPs which frequently occurred in a given DMR cluster related to each other, I created a complex heatmap-based visualization, leveraging the capabilities of the codaplot package [SOFT3]. Figure 24 juxtaposes i) the z-score transformed mean DMR methylation levels across all populations for each DMR cluster; ii) the z-score transformed mean DMCpG methylation levels across all populations for each DMCpG cluster; and iii) for each DMR cluster, the percentage of DMRs exhibiting each DMCpG-PP. The figure highlights that each loss of methylation DMR cluster was associated with a series of DMCpG-PPs characterized by progressively increasing population-specificity. For example, the D3 DMR cluster specifically marked the pDC population and showed partial hypomethylation in the cDC1, cDC2, and, to a lesser extent, in the lymphoid populations. 93% of the D3 cluster DMRs contained DMCpGs exhibiting the *d2* DMCpG-PP. The *d2* DMCpG-PP marked the pDC population and was overall highly similar to the D3 DMR-PP. In addition, 31% of the D3 cluster DMRs contained the *d1* DMCpG-PP, which was characterized by strong hypomethylation across all mature dendritic cell populations as well as the CDP population. Moreover, 18% of the D3 cluster DMRs exhibited the *c1* DMCpG-PP, which was characterized by broad hypomethylation across the dendritic and lymphoid lineages. Finally, 6% of all D3 DMRs exhibited the *p1* DMCpG-PP, which was characterized by broad hypomethylation across the entire hematopoietic system downstream of the HSPC compartment. The different DMCpG-PPs occurring in the D3 DMR cluster thus all shared the strong loss of methylation in the pDC population. While the *d1* DMCpG-PP exclusively marked the pDC population, the other DMCpG-PPs marked additional populations with increasingly broad specificity. Collectively, this example highlights a pattern of combining a series of CpG-PPs with increasingly narrowing population specificity within individual DMR regions, which was exemplary for many other DMR clusters (Figure 24).

While substantially different DMCpG-PPs were combined within the D3 cluster DMRs, the overall DMR-PP of the D3 DMR cluster was highly similar to the *d2* DMCpG-PP. This suggested that the majority of the DMCpGs within D3 cluster DMRs exhibited the *d2* DMCpG-PP: the DMR methylation level was computed as the mean methylation level across all DMCpGs in a DMR; therefore, a predominant occurrence of the *d2* DMCpG-PP would shape the DMR methylation level even in the presence of other CpG-PPs in lower frequency. As an initial analysis of this hypothesis, all DMCpGs within each DMR cluster were pooled, and then the overall percentage of all DMCpGs within each DMR cluster exhibiting each individual DMCpG-PP was computed. This captured the relative frequency of each DMCpG-PP within each DMR cluster. These relative frequencies are shown in Figure S5, which demonstrates that the predominant majority of all DMCpGs within the D3 DMR cluster indeed exhibited the *d2* DMCpG-PP, as hypothesized. Figure S5 furthermore demonstrates similarly high prevalence of a single specific DMCpG-PP for many other DMR clusters. In summary, while many DMRs exhibited two or three different DMCpG-PPs in total, this initial analysis suggested that many DMRs had a single, predominant DMCpG-PP, exhibited by the majority of their contained DMCpGs.

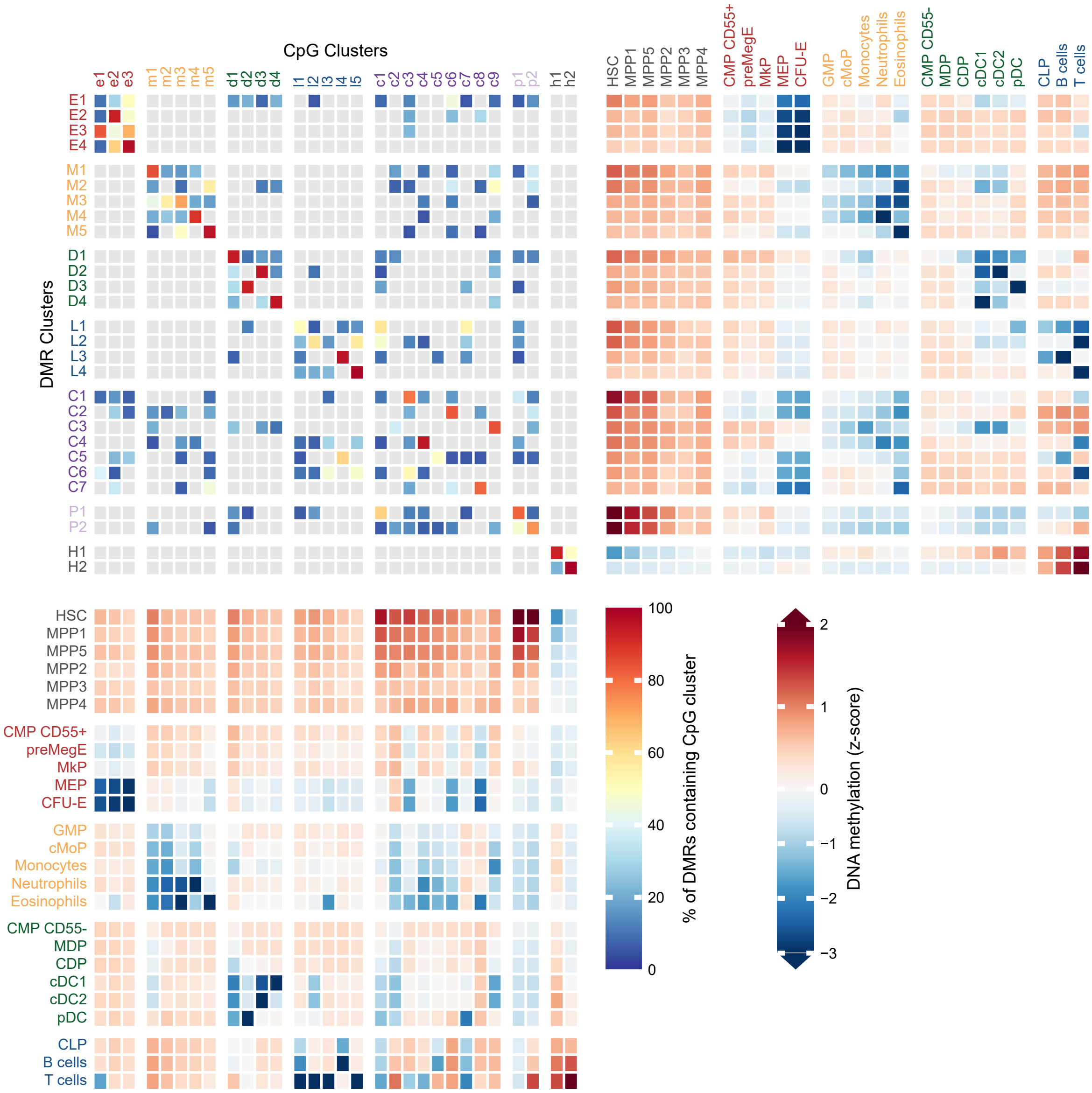

Figure 24: Within each DMR cluster, DMRs exhibit a characteristic series of DMCpG programming patterns with increasingly narrowing population specificity. The complex heatmap juxtaposes i) the z-score transformed mean DMR methylation levels across all populations for each DMR cluster; ii) the z-score transformed mean DMCpG methylation levels across all populations for each DMCpG cluster; and iii) for each DMR cluster, the percentage of DMRs exhibiting each DMCpG programming pattern. This visualization is complemented by Figure S5, which shows the percentage of the DMCpGs within each DMR cluster which are characterized by each DMCpG programming pattern.

2.5.6 Correlation between the extent of early DMCpG programming and the breadth of DMCpG programming across the mature cell types

The different DMCpG clusters were characterized by distinct DMCpG-PPs, ranging from pan-hematopoietic over lineage-specific to highly population-specific programming patterns (Figure 21A). Individual DMRs appeared to often exhibit multiple DMCpG-PPs with progressively narrowing population specificity (Figures 23 and 24). I next sought to understand how such DMCpG-PP combinations shaped the DMRs over the course of differentiation. The DMCpG and DMR clustering analyses were performed while considering only information from the HSC population and the mature populations. Consequently, the DMCpGs within each DMCpG cluster exhibited highly similar methylation level profiles across the mature populations (Figure 21A, Figure S4). On the other hand, heterogeneity within the progenitor populations was explicitly allowed by the clustering strategy. The heatmap visualizations presented in Figure 21A and Figure S4 provided initial insights into the methylation level profiles of the progenitor populations for each DMCpG cluster. These high-level visualizations suggested that many DMCpG clusters contained a significant percentage of DMCpGs, demonstrating substantial methylation shifts in progenitor populations. The visualization further suggested that less population-specific DMCpG clusters showed more extensive programming across the progenitor populations. To systematically quantify these initial observations, I considered each DMCpG cluster separately and computed for each population the percentage of regulated DMCpGs. Regulated DMCpGs were defined as DMCpGs exhibiting a methylation level shift compared to the HSC population of at least 20% in the MPP1-5 populations and of at least 30% in the downstream populations, as introduced in section 2.4.2. Figure 25A shows, for all DMCpG clusters, the percentage of regulated DMCpGs across all populations. This analysis confirmed the anticorrelation (observed across the individual DMCpG clusters) between the extent of programming in progenitor cells (both with regard to the onset of programming and to the fraction of programmed DMCpGs) and the specificity of programming across the mature populations. In summary, the extent of DMCpG programming within progenitor populations and the specificity of DMCpG programming across the mature populations observed across the individual DMCpG clusters were anticorrelated: the more extensive DMCpG programming occurred at early progenitor stages, the less specific was hypomethylation restricted to specific mature populations (or hypermethylation for the gain of methylation clusters).

During the initial characterization of the DMCpG clusters, each DMCpG cluster was annotated with a set of marked populations, defined as the populations where markedly strong hypomethylation was observed (section 2.5.2). To complement these annotations, each DMCpG cluster was now further annotated with an extended set of “programmed populations”, defined to include all populations which demonstrated at least 70% regulated DMCpGs across all DMCpGs within the DMCpG cluster (Figure 25B). This threshold was lowered to 25% regulated DMCpGs for the MPP1-5 populations to account for their expected high heterogeneity, analogous to how the MPP1-5 populations were treated in similar categorization tasks (cf. for example section 2.4.2). The set of programmed populations for a DMCpG cluster thus comprised all marked populations of the cluster as well as additional populations that showed non-maximal but biologically relevant programming.

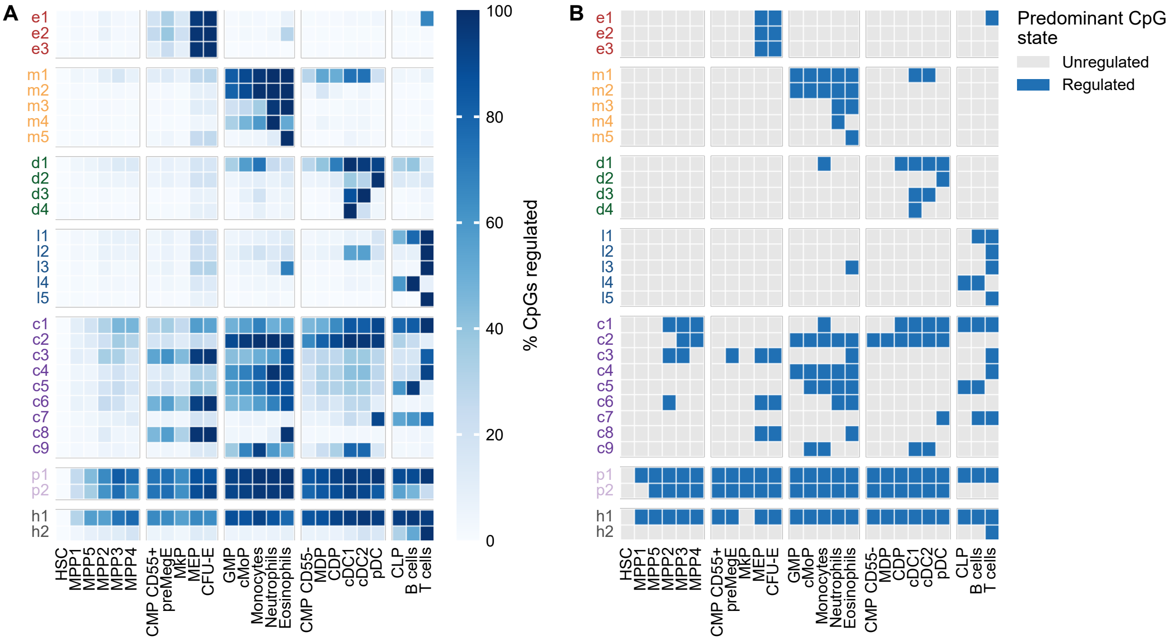

Figure 25: The extent of early DMCpG programming and the breadth of DMCpG programming across the mature cell types are correlated.
(A) Each DMCpG cluster was considered separately, and the percentage of regulated DMCpGs for each population was computed. Regulated DMCpGs were defined as DMCpGs exhibiting a methylation level shift compared to the HSC population of at least 20% in the MPP1-5 populations and of at least 30% in the downstream populations, as introduced in section 2.4.2. The heatmap shows the percentage of regulated DMCpGs across all populations for each DMCpG cluster.
(B) Each DMCpG cluster was annotated with a set of programmed populations, defined to include all populations which demonstrated at least 70% regulated DMCpGs within the DMCpG cluster. This threshold was lowered to 25% regulated DMCpGs for the MPP1-5 populations to account for their expected high heterogeneity. This annotation complemented the set of maximally hypomethylation populations (the marked populations), which was previously determined for each DMCpG cluster (Figure 21). The set of programmed populations for a DMCpG cluster comprised all its marked populations as well as additional populations that showed non-maximal but biologically relevant programming. The heatmap indicates the set of regulated populations for each DMCpG cluster. DMCpG clusters with less population-specific programming patterns (Figure 21) demonstrated broader sets of regulated populations, including a larger set of regulated progenitor populations.

2.5.7 Successive programming of distinct DMCpGs within DMRs underlies DMR expansion

The findings presented in section 2.5.5 established that the combination of multiple DMCpG-PPs within individual DMRs was a widespread mechanism observed across all DMR clusters. This generally involved a series of DMCpG-PPs characterized by an increasingly narrow specificity for certain mature populations, and thus as demonstrated above by an increasingly late onset of programming during differentiation (section 2.5.6). The widespread occurrence of this DMCpG programming mechanism provided a direct and sufficient explanation for the widespread occurrence of DMR expansion during hematopoiesis, as demonstrated in the following. To exemplify how this DMCpG programming mechanism underlies DMR expansion, Figure 26 depicts the DMCpG methylation levels across a DMR proximal to the TSS of the *Elane* gene, which encodes the neutrophil elastase ELANE. This complex locus plot was generated with the codaplot packages developed as a part of my doctoral work [SOFT3]. This DMR belonged to the M4 DMR cluster and exhibited the following three DMCpG-PPs: i) the pan-hematopoietic *p2* DMCpG-PP (2 DMCpGs), the pan-myeloid *m1* DMCpG-PP (4 DMCpGs) and the neutrophil specific *m4* DMCpG-PP (12 DMCpGs) (cf. Figure 21A, Figure 24). The DMCpGs characterized by the *p2* DMCpG-PP represented small, hypomethylated seed regions demonstrating methylation loss already in the MPP populations. These seed regions were strongly hypomethylated in all downstream populations as well, except for the lymphoid lineage. The *m1* DMCpGs showed considerable hypomethylation in the myeloid progenitor populations GMP and CMP, as well as in the cDC1 and cDC2 populations, in addition to strong hypomethylation across the neutrophils, eosinophils, and monocytes. These DMCpGs underlay a considerable expansion of the DMR in the aforementioned populations. Finally, the *m4* DMCpGs showed considerable DMCpG regulation specifically in neutrophils. Therefore, the DMR demonstrated its maximal expansion exclusively in the neutrophil population. This Elane promoter DMR was selected as a representative showcase: it illustrates how the combination of DMCpG-PPs with increasing population specificity within a DMR directly results in the expansion of the DMR over the course of differentiation. In summary, the widespread occurrence of DMR expansion in the course of hematopoietic differentiation was the direct result of the ubiquitous role of DMCpG-resolved programming within individual DMRs. Different DMCpG-PPs within a DMR captured distinct programming steps which appeared to occur at subsequent stages of differentiation.

Of note, because the DMR contained only a few *p2* DMCpGs (2 DMCpGs out of 21 DMCpGs within the DMR), the seeding within the MPP compartment, the erythroid lineage, and the dendritic cell lineages was largely obscured when characterizing the DMR through its DMR methylation level (defined as the mean DMCpG methylation level across the DMR). Similarly, the pan-myeloid hypomethylation characterizing the *m1* DMCpG-PP was largely obscured because only 5 DMCpGs exhibited this DMCpG-PP. This was reflected in the M4 DMR cluster membership of the DMR: the D4 DMR cluster was characterized by neutrophil-specific hypomethylation with little evidence of relevant levels of hypomethylation outside of the myeloid lineage.

Figure 26 also provides an exemplary comparison between two different ways of classifying DMR expansion states. The first strategy, based on the inspection of the individual methylation levels at each DMCpG in a DMR, was introduced in section 2.4.2. Here, I assessed whether the information about the DMCpG-PPs contained within a DMR was sufficient to effectively distinguish DMR expansion states - without any consideration for the individual DMCpG methylation levels in a DMR. Within the methylation-level-based expansion state classification method, a DMCpG was considered regulated in a given population if its methylation level shift in that population compared to the HSC population was at least

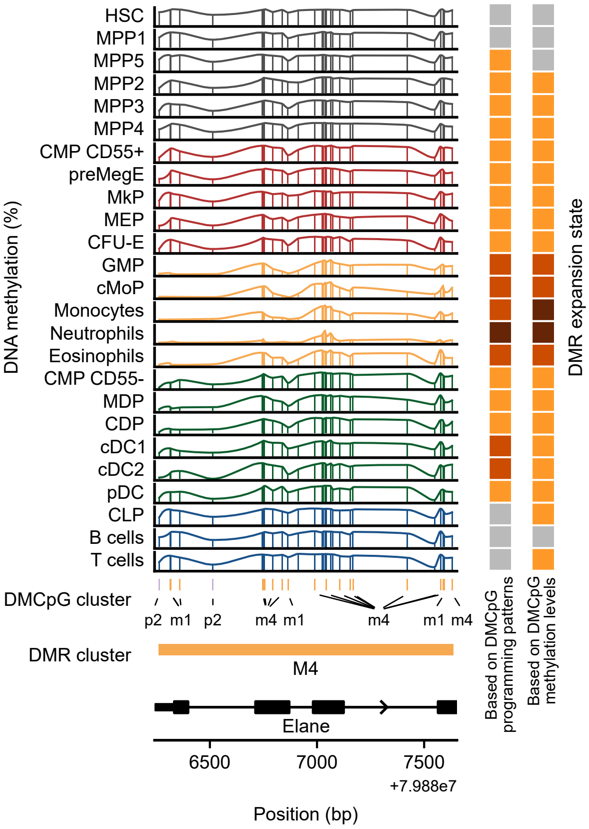

Figure 26: Exemplary case study: subsequent pan-hematopoietic, pan-myeloid, and neutrophil-specific DMCpG programming underlies the expansion of an *Elane* promoter DMR. Locus plot shows a DMR located 700 bp upstream of the transcription start site of the *Elane* gene. Barplots show the DNAme levels for all CpGs within the DMR, and spline lines connecting the bars are displayed to facilitate the comparison of the DNAme profiles between populations. The first bottom annotation indicates which CpGs were identified as DMCpGs as well as the DMCpG programming patterns for each of these DMCpGs. Further annotations below indicate that the DMR belonged to the M4 DMR cluster and overlapped the 5'-UTR and multiple exons of the canonical *Elane* transcript. Right annotations indicate the DMR expansion state classifications for each population, computed once with the DMCpG programming pattern-based classification approach and once with the DMCpG methylation level-based classification approach.

30% (or 20% for the MPP1-5 populations). For the DMCpG-PP-based approach, only the different DMCpG-PPs exhibited by each individual DMCpG within a DMR were used. A population was considered regulated at a DMCpG if the population was among the regulated populations for the DMCpG-PP characterizing the DMCpG (Figure 25B). The rest of the two classification approaches was identical. Briefly, for each DMR in each population, I first determined the number of regulated DMCpGs. For each DMR, I then noted the maximal number of regulated DMCpGs observed in any population. The DMR state for each population was determined based on the percentage of regulated DMCpGs relative to the maximum observed count: DMRs were thus classified as unregulated (0% regulated CpGs), seeded (< 45%), intermediate (< 81%), or completed (≥ 81%). DMRs with at least five regulated DMCpGs were considered to be in an intermediate expansion state, even if these five DMCpGs represented less than 45% of all DMCpGs in the DMR. The DMR state expansion map resulting from the new, DMCpG-PP-based classification is shown in Figure 27.

The previously introduced methylation level-based classifications are shown in Figure 17. These visualizations demonstrate that the two classifications yielded highly similar DMR expansion state landscapes across the hematopoietic system. The predominant agreement between the two DMR expansion state classifications was further confirmed by computing a confusion matrix providing a DMCpG cluster-stratified comparison of the two classification approaches (Figure S7). In summary, the progressive expansion of individual DMRs during hematopoietic differentiation could be effectively tracked and classified based solely on the information about the DMCpG-PPs present within the DMRs. This further supported a model of DMCpG programming involving heterogeneous programming of different DMCpGs within individual DMRs, occurring successively over the course of differentiation.

The DMCpG-PP-based DMR expansion state map (Figure 27) differed in a few interesting aspects from the methylation level-based map (Figure 17). It appeared to provide a more differentiated picture of DMR seeding in the MPP1-5 populations. The DMCpG-PP-based classification method leverages the cluster membership information from the DMCpG clustering analysis, which aggregated information across multiple mature populations with strong methylation level shifts, conferring robustness against sampling noise in individual populations. In contrast, the methylation level-based method inspected each population separately at each DMCpG within a DMR. This difference may underlie a more robust classification with the DMCpG-PP-based classification method. Remarkably, the DMCpG clustering analysis was solely based on the data from the HSC and the mature populations: information from progenitor populations was not used for the clustering. However, the DMCpG clusters still appeared to capture distinct patterns of DMCpG programming between the MPP populations. This further supported the observation of a close relationship between the extent of early DMCpG programming at a given DMCpG and the subsequent programming pattern across the mature populations at that DMCpG. Taken together, these findings suggest that the classification of DMR expansion states based solely on the information about the DMCpG-PPs contained within each DMR region may provide a highly robust and systematic way of tracking DMR expansion during hematopoietic differentiation. To my knowledge, this work introduces the concept of systematically tracking DMR expansion state in addition to DMR methylation levels for the first time, and provides the first evaluation of systematic approaches for robustly determining these DMR expansion states.

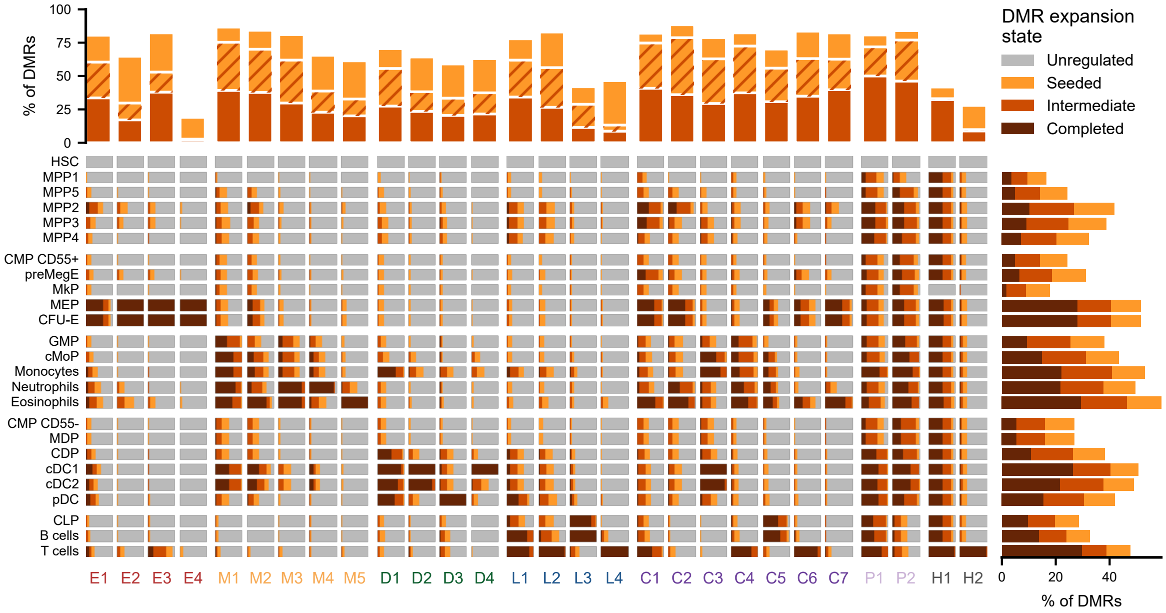

Figure 27: The progressive expansion of individual DMRs during hematopoietic differentiation can be effectively tracked based solely on the information about the DMCpG-PPs present within the DMRs. The DMR expansion state of each DMR in each population was classified as unregulated, seeded, intermediate, or completed, based solely on the information about the DMCpG-PPs exhibited by each DMR. The heatmap shows the proportion of these DMR states within each DMR cluster for each population. The top annotation indicates the percentage of DMRs in each DMR cluster for which a seeded and/or an intermediate DMR expansion state was observed. The combined occurrence of seeded and intermediate states is indicated by a hatch pattern combining the colors indicating seeded and intermediate DMR expansion states. Each DMR reaches the completed DMR expansion state at least once by definition. Therefore this state is not considered for this annotation. The right annotation shows, for each population, the percentage of hematopoietic DMRs in a seeded, intermediate, or completed DMR expansion state. To compute the DMR expansion states for each DMR in each population, first, the number of regulated DMCpGs was computed. A DMCpG was considered regulated in a population if that population was among the programmed populations for the DMCpG-PP characterizing the DMCpG (Figure 25). For each DMR, I then noted the maximal number of regulated DMCpGs observed in any population. The DMR state for each population was determined based on the percentage of regulated DMCpGs relative to the maximum observed count: DMRs were thus classified as unregulated (0% regulated CpGs), seeded ($< 45\%$), intermediate ($< 81\%$), or completed ($\geq 81\%$). DMRs with at least five regulated DMCpGs were considered to be in an intermediate expansion state, even if these five DMCpGs represented less than 45% of all DMCpGs in the DMR. An initial strategy for the classification of DMR expansion states was introduced previously within this thesis (Figure 17). This strategy was based on the inspection of the individual methylation levels at each DMCpG in a DMR. An exemplary comparison between the results of the DMCpG methylation level-based, and the DMCpG-PP-based DMR expansion state classifications is presented in Figure 26. The comparison between the two DMR expansion state classifications is further characterized by a confusion matrix (Figure S7).

2.5.8 Typical mechanisms of DMCpG-resolved programming within DMRs

Characterization of DMCpG programming in the exemplary C3, D2 and M1 DMR clusters

Each DMR cluster was associated with a characteristic set of several frequently occurring DMCpG-PPs (Figure 23). However, most individual DMRs contained only one (34068 DMRs, 28% of all DMRs), two (50688 DMRs, 41%) or three (24555, 20%) DMCpG patterns

(Figure 22). I next asked how the DMCpG-PPs associated with each DMR cluster were combined in individual DMR regions.

I first focused on an exemplary subset of the DMR clusters. Figure 28 illustrates the co-occurence of the *m1*, *d3*, *c9*, *c2*, and *p2* DMCpG-PPs within individual DMRs of the M1, D2, and C3 DMR clusters. A characteristic mechanism of DMCpG programming was observed across all three DMR clusters. First, each DMR cluster was associated with one DMCpG-PP occurring in virtually all of its DMRs. For each DMR cluster, this highly prevalent DMCpG-PP was very similar to the overall DMR-PP of the DMR cluster. For example, the M1 DMR cluster marked the monocyte, neutrophil, and eosinophil populations. In virtually all of its DMRs, the *m1* DMCpG-PP was observed at least once. The *m1* DMCpG-PP was highly similar to the overall DMR-PP of the M1 DMR cluster: it also marked the monocyte, neutrophil, and eosinophil populations. As discussed in section 2.5.4, the DMR-PP describes the profile of the DMR methylation levels across the hematopoietic populations. This DMR methylation level was defined as the mean methylation level across all DMCpGs in a DMR. Thus the high similarity between the overall M1 DMR-PP and the *m1* DMCpG-PP suggested that the majority of the DMCpGs within the M1 clusters DMRs exhibited the *m1* DMCpG-PP. This was in line with initial findings suggesting the existence of one or two main DMCpG-PPs for each DMR cluster, which occurred in the predominant majority of the DMRs of each respective DMR cluster (Figures 23 and 24). In the following I refer to the most prevalent DMCpG-PP exhibited by a DMR cluster as its “main” DMCpG-PP. Note that the agreement between the names of the *m1* DMCpG-PP and the *M1* DMR-PP is coincidental. For example, *d2* (marking the pDC population) is the main DMCpG-PP for the D3 DMR cluster (which consequently also marks the pDC population). In summary, each of the M1, D2, and C3 DMR clusters exhibited regulation by a characteristic main DMCpG-PP occurring in virtually all of their DMRs, which predominantly shaped the DMR methylation levels.

In addition to the presence of a main DMCpG-PP across all cluster DMRs, each of the C3, D2, and M1 DMR clusters exhibited one or two other DMCpG-PPs in a subset of its DMRs. These DMCpG-PPs typically exhibited strong hypomethylation in an extended set of populations beyond the populations marked by the DMR cluster. For example, the pan-myeloid M1 DMR cluster exhibited the pan-hematopoietic *p2* DMCpG-PP in 36% of its DMRs, which is characterized by strong hypomethylation in the erythroid, myeloid, and dendritic cell lineages. Interestingly, the *p2* DMCpG-PP also occurred in the C3 DMR cluster (21% of its DMRs). This highlights that one DMCpG-PP usually contributed to multiple distinct DMR clusters. In summary, DMCpG programming within the C3, D2, and M1 DMR cluster intervals often involved the combination of a main DMCpG-PP (occurring in virtually all DMRs of the DMR clusters and across most DMCpGs within each DMR) together with less specific DMCpG-PPs marking an extended set of populations.

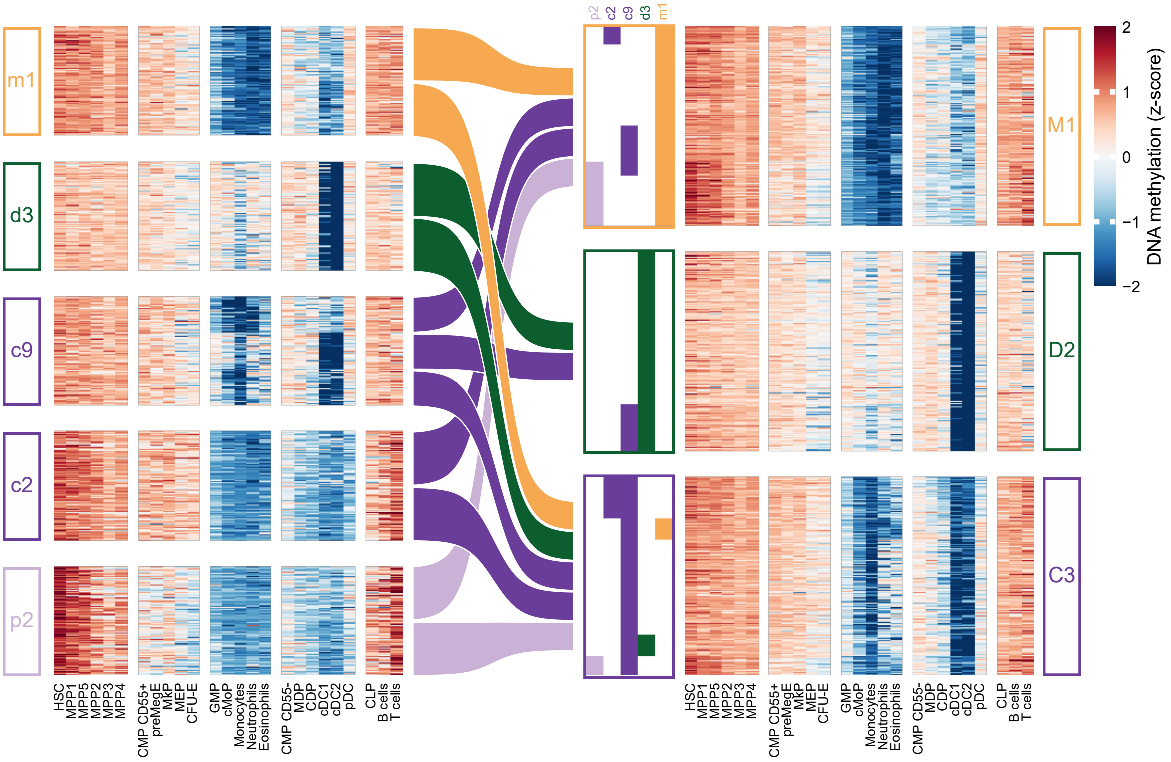

Figure 28: Characteristic combinations of DMCpG clusters underlie seeding and expansion of the DMRs in the C3, D2, and M1 clusters. Left heatmaps show the z-score transformed DNA methylation levels of 500 randomly sampled DMCpGs from selected DMCpGs clusters. These DMCpG clusters were part of the characteristic set of frequently exhibited DMCpG clusters of the C3, D2, and M1 DMR clusters. Flow lines indicate the DMR clusters to which each DMCpG cluster frequently contributed. The flow lines are aligned with the DMCpG methylation level heatmaps, indicating which of the DMCpGs contribute to each DMR cluster. The shown random DMCpGs were selected such that the same number of DMCpGs contributing to each DMR cluster was chosen. Heatmaps at the right show the z-score transformed DMR methylation levels of 750 randomly chosen DMRs for each DMR cluster. The heatmap annotation indicates for each DMR which DMCpG clusters were exhibited by the DMR. The heatmap annotation indicates that the DMRs typically exhibited one or two of the selected DMCpG clusters. The co-occurring DMCpG clusters were characterized by narrowing population specificity. Such combinations of DMCpG clusters result in DMR seeding followed by DMR expansion (Figures 25 to 27). Nevertheless, the heatmaps at the right indicate that the overall DMR methylation levels, computed as the mean DNAme level across all DMCpGs in the DMR, were mainly shaped by only the *m1*, *d3* and *c9* DMCpG clusters for the M1, D2, and C3 DMR clusters respectively. This suggested that these DMCpG clusters were exhibited by the majority of the DMCpGs within the DMRs, in line with previous findings (Figure 24). This visualization includes only a subset of the most frequent DMCpG programming patterns exhibited by the C3, D2, and M1 DMR clusters; a more complete picture is provided in Figure 29.

2.5.9 High compactness and clear separation of DMR and DMCpG clusters independent of DMR buildup or DMCpG location

Figure 28 furthermore highlighted that all DMCpGs within the *p2*, *c2*, *c9*, *d3*, and *m1* DMCpG clusters had highly similar DMCpG methylation level profiles across the 25 hematopoietic populations. Thus, the DMCpGs from one DMCpG cluster were all characterized by the same DMCpG-PP, independent of the DMR cluster in which the individual DMCpGs resided. Minor heterogeneities of the DMCpG methylation levels were observed within the *c9* DMCpG cluster, which could potentially be subpartitioned for an even finer DMCpG-PP mapping. However, the *c9* cluster homogeneously grouped DMCpGs characterized by a combination of specific hypomethylation in the monocyte and cDC1/cDC2 populations, which constituted a sufficient level of resolution for the purposes of this study. Moreover, all individual DMRs within the C3, D2, and M1 DMR clusters also exhibited highly similar DMR methylation level profiles across the 25 hematopoietic populations, confirming that each DMR cluster was homogeneously characterized by a specific DMR-PP. The presence of different DMCpG-PPs within individual DMRs only led to subtle differences in these DMR methylation level profiles. These subtle modifications could be of interest for highly resolved, secondary analyses, but they did not alter the overall DMR-PP of the respective DMR clusters. Furthermore, the homogeneity of the DMCpG methylation level profiles within all DMCpG clusters was systematically demonstrated in Figure S8, which shows the average, z-score transformed DMCpG methylation levels within each DMCpG cluster, stratified by the DMR cluster in which the DMCpGs reside. The homogeneity of the DMR methylation levels profiles within all DMR clusters is demonstrated in Figure S9, which shows the mean, z-score transformed DMR methylation levels in each DMR cluster, stratified by the most frequent DMCpG co-occurrence patterns in the DMR cluster. Taken together, these analyses demonstrated the compactness and separation of the DMR cluster and the CpG clusters and precluded that the observations concerning the composition of the hematopoietic DMRs were artifacts of incomplete or incorrect DMR or DMCpG clustering.

2.5.10 Systematic quantification of DMCpG programming pattern co-occurrence within DMR regions

I next sought to systematically characterize the co-occurrence of different DMCpG-PPs within individual DMRs. Each DMR cluster was considered separately for the following analysis. First, the most frequent DMCpG-PPs (cf. Figure 23) were identified: only DMCpG-PPs occurring in at least 10% of the DMRs were considered. If more than six DMCpG-PPs met this threshold, the six most frequent DMCpG-PPs were collected. Only these highly frequent DMCpG-PPs were considered for the next quantification steps. Each DMR was annotated with the combination of DMCpG-PPs it exhibited. Additionally, within each DMR, I noted the relative frequencies of all exhibited DMCpG-PPs. Taken together, this analysis provided an additional layer of information for the DMR/DMCpG-atlas, detailing the combination of DMCpG-PPs occurring within each individual DMR.

When this information was aggregated across the DMRs within each DMR cluster, the most frequent combinations of DMCpG-PPs within each DMR cluster could be identified. For this purpose, the following quantification steps were performed separately for each DMR cluster: i) all unique combinations of DMCpG-PPs arising in at least one DMR were gathered; ii) the frequency of each DMCpG-PP combination was counted; and iii) additionally, for each combination of DMCpG-PPs, I computed the average relative frequency of each individual DMCpG-PP in the combination (as the mean relative frequency observed across all DMRs exhibiting the DMCpG-PP combination). Figure 29A demonstrates a part of the gathered data by illustrating the most frequent DMCpG-PP combinations for the M1, D2, and C3 DMR clusters. This analysis quantified and extended the previous qualitative observations for the same DMR clusters presented in Figure 28. For example, the most frequent combination of DMCpG-PPs in the D2 DMR cluster consisted only of the *d3* DMCpG-PP; consequently, the *d3* DMCpG-PP was observed for 100% of the individual DMCpGs in the corresponding DMRs. The second most frequent DMCpG-PP combination consisted of the *d1* and the *d3* DMCpG-PPs; on average across all DMRs exhibiting this combination, 41% of the DMCpGs within an individual DMR exhibited the *d1* DMCpG-PP and 59% exhibited the *d3* DMCpG-PP. Figure 29B illustrates the most frequent DMCpG-PP combinations for further exemplary DMR clusters, including the erythroid-specific E2, the T cell-specific L4, the pan-hematopoietic P1, and the gain of methylation H2 DMR clusters. Further examples are shown in Figure S6. In summary, as a common mechanism observed for all DMR clusters, individual DMRs often exhibited a combination of two or three DMCpG-PPs. One of these DMCpG-PPs typically covered the majority of the CpGs in the DMR and marked a limited set of populations, while the other DMCpG-PPs represented broadened programming across an extended set of populations.

The quantitative analysis of DMCpG-PP co-occurrence also revealed another mechanism of DMCpG programming within DMR regions. In some cases, DMRs marking multiple populations exhibited (on a minority of the contained DMCpGs) DMCpG-PPs specifically marking a subset of the overall marked populations. For example, the pan-myeloid M1 DMR cluster contained DMRs exhibiting the neutrophil-specific *m4* DMCpG-PP on a minority of their DMCpGs. This indicated that in these DMRs, a small fraction of the DMCpGs were specifically programmed in neutrophils, while most other DMCpGs marked the monocytes, eosinophil, and neutrophil populations. This could represent fine-tuning of the overall DMR state in specific populations on a minority of the DMCpGs in a DMR.

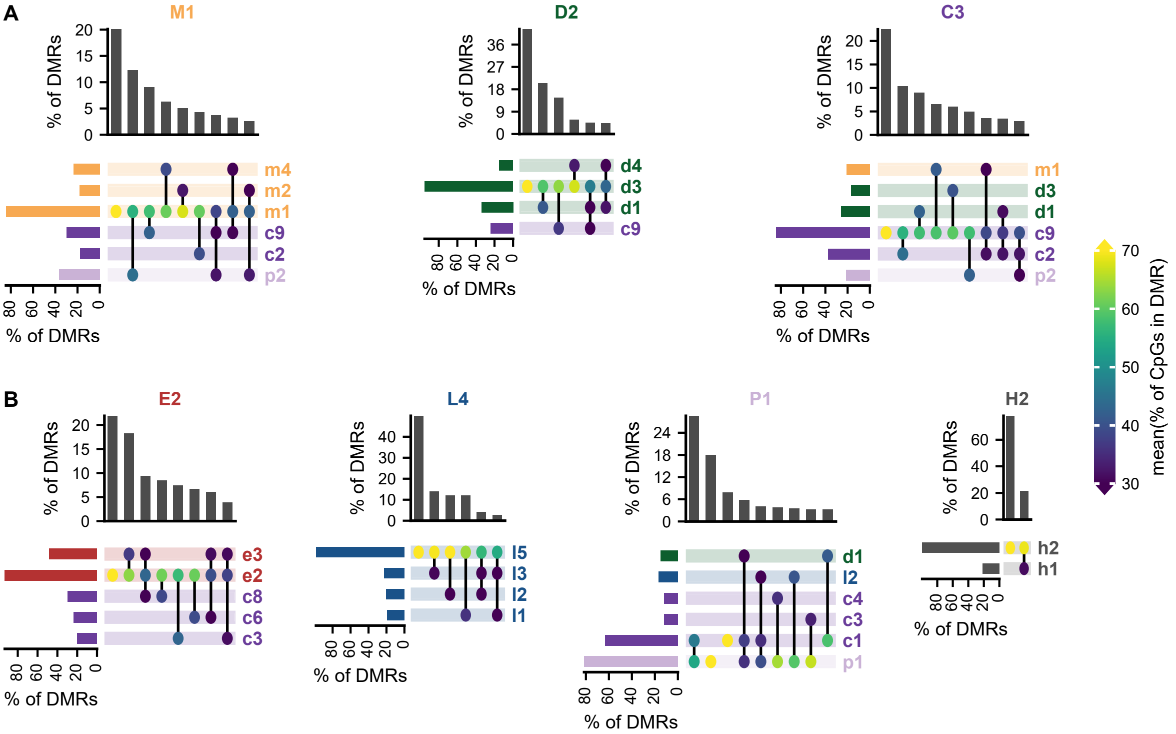

Figure 29: Programming within DMRs typically involves a predominant DMCpG-PP shaping the population-specificity of the DMR, in combination with less frequent, less population-specific DMCpG-PPs which indicate preceding DMR seeding and expansion steps during differentiation. Each DMR was annotated with the combination of DMCpG-PPs it exhibited. For this purpose, the most frequent DMCpG-PPs for each DMR cluster were identified (cf. Figure 23): only DMCpG-PPs occurring in at least 10% of the DMRs of a DMR cluster were considered, and if more than six DMCpG-PPs met this threshold, only the six most frequent DMCpG-PPs were collected. Next, each DMR was annotated with the combination of these main DMCpG-PPs it exhibited, and the relative frequencies of all exhibited DMCpG-PPs within the DMR were noted. Finally, this information was aggregated across the DMRs within each DMR cluster to identify the most frequent combinations of DMCpG-PPs within each DMR cluster. For each combination of DMCpG-PPs observed within a DMR cluster, the average relative frequency of each individual DMCpG-PP in the combination was computed. Finally, the properties of the most frequent DMCpG-PPs for each DMR cluster were detailed using UpSet plots.
(A) Upset plots illustrating the most frequent DMCpG-PP combinations exhibited by the M1, D2, and C3 DMR clusters; a detailed inspection of the structure of the DMRs in these clusters was given in Figure 28.
(B) Upset plots illustrating the most frequent DMCpG-PP combinations exhibited by representative erythroid lineage-specific, lymphoid lineage-specific, pan-hematopoietic, and gain of methylation DMR clusters.
UpSet plots for further representative DMR clusters are shown in Figure S6.

Many DMCpG-PP combinations were characterized by the co-occurrence of a lineage- or population-specific DMCpG-PP together with one or two less specific DMCpG-PPs, which may indicate stepwise seeding and expansion of the DMR during differentiation (Figures 26 and 27). Another common co-occurrence pattern was the combination of a DMCpG-PP marking multiple populations together with a less frequent, highly population-specific DMCpG-PP, which may indicate tuning of DMR states in individual populations.

2.5.11 Epigenetic memory of early alternative fate exploration is maintained throughout differentiation in the form of partially expanded DMR states

As shown above, the extent of DMCpG programming in progenitor populations and the breadth of DMCpG programming among the mature populations were correlated across all DMCpG-PPs. So far, this correlation was introduced qualitatively to capture how DMCpG programming underlay the expansion of DMRs. Next, I systematically quantified this correlation. Each CpG cluster had previously been characterized by a hypomethylation specificity score, which described how specifically hypomethylation was restricted to certain mature populations (section 2.5.2). Moreover, the CpG clusters within each (E, M, D, L, C, P, H) group were ordered according to this score (section 2.5.2). In the following, I will refer to this score as "mature programming" score because it quantified the breadth of programming across the mature populations: the higher the mature programming score, the more broadly hypomethylation occurred across all mature populations. Of course, the situation was reversed for the gain of methylation DMCpG clusters *h1* and *h2*. Here, higher mature programming scores, which indicated broader hypomethylation across the mature populations, signified less broad occurrence of gain of methylation programming. To complement this mature programming score, I next computed the mean hypomethylation across the MPP1-5 populations for each DMCpG cluster as a measure of the extent of DMCpG programming during early hematopoietic differentiation. I will refer to this score as the "early programming score" in the following: the higher the early programming score, the higher the extent of programming during early hematopoietic differentiation (and vice versa for the gain of methylation clusters). The mean early programming and mature programming scores for the individual CpG clusters were remarkably correlated (Figure 30A, Pearson correlation coefficient (PCC) = 0.97, p-value from permutation test with 1e5 permutations = 1e-5). I next pooled all individual DMCpGs from all DMRs and re-assessed the correlation at the level of individual DMCpGs (Figure 30B). The correlation was again striking (PCC = 0.85, p-value using 10,000 randomly sampled DMCpGs and 1e6 permutations = 1e-6). To assess whether the amount of correlation differed between clusters with low and high mature programming scores, I stratified the correlation analysis by DMCpG cluster (Figure S10). The correlation of the early programming and mature programming scores measured across individual DMCpGs was strong and highly significant within all individual DMCpG clusters after Benjamini-Hochberg correction. Taken together, there was a clear, globally present, near-linear correlation between the extent of DMCpG programming during early hematopoietic differentiation and the breadth of hypomethylation across the mature populations (or of hypermethylation for the gain of methylation clusters).

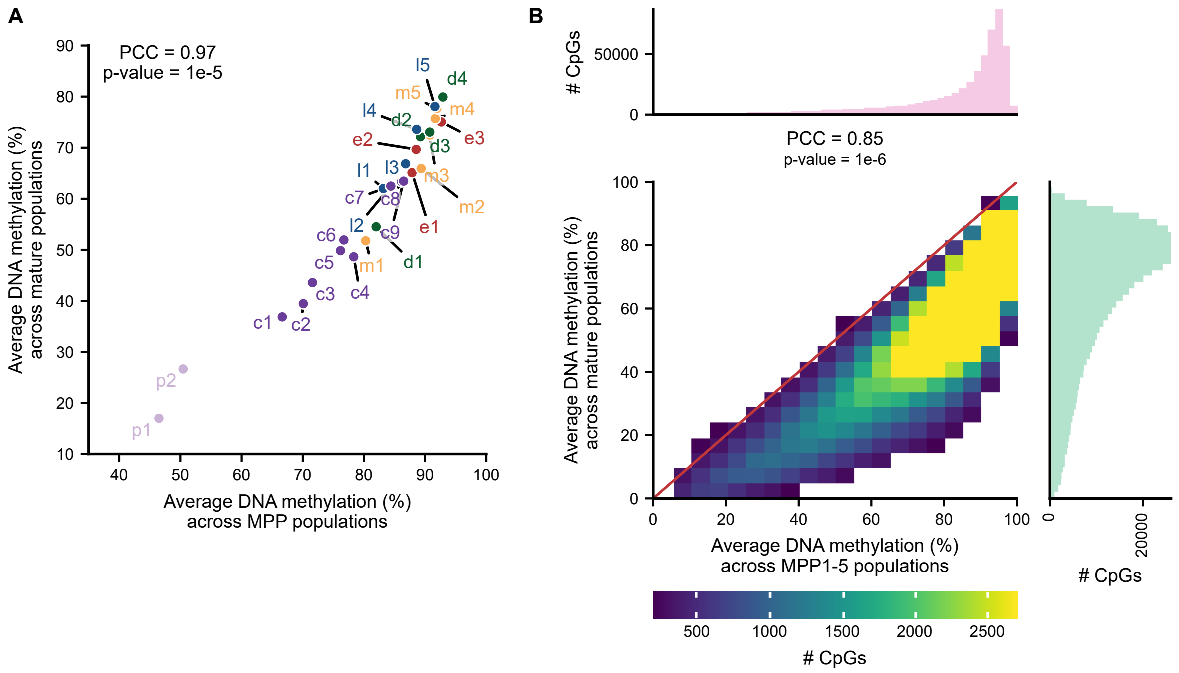

Figure 30: Strong correlation between the extent of early DMCpG programming and the breadth of hypomethylation across the mature populations.
(A) Scatter plot showing the relationship between the mean DNAme level in the MPP1-5 populations and the mean DNAme level across the mature populations for each loss of methylation DMCpG cluster. The significance of the Pearson correlation was assessed with a permutation test using 10,000 permutations. The mean DNAme level across the mature populations measures the population-specificity of each DMCpG cluster: the lower the average DNAme level, the more mature populations show at least partial programming in the DMCpG cluster. This score has been introduced previously in this thesis: the CpG clusters within each (E, M, D, L, C, P, H) group were ordered according to this score (Figure 21). The mean DNAme level in the MPP populations is a measure for the extent of DMCpG programming during early hematopoietic differentiation.
(B) Joint and marginal distributions of the mean DNAme levels across the mature populations and the mean DNAme levels across the MPP1-5 populations of all individual DMCpGs. The significance of the Pearson correlation was assessed with a permutation test, using 10,000 randomly sampled DMCpGs and 1,000,000 permutations.
To assess whether the strength of the correlation differed between DMCpG clusters, the correlation analysis was further stratified by DMCpG cluster (Figure S10).

This correlation was particularly remarkable because it represented a significant indication of a widespread role of epigenetic memory in shaping the DNA methylome of hematopoietic cells when viewed in the context of the preceding findings. The individual DMR clusters were homogeneously characterized by distinct programming patterns, ranging from pan-hematopoietic to highly population-specific programming (cf. section 2.3.1, section 2.5.9). The lineage- and population-specific DMR clusters were strongly enriched in gene sets, and enhancer sets with matching specificity (cf. section 2.3.2, section 2.3.3). The DMRs from lineage- or population-specific clusters predominantly achieved full DMR expansion in their marked populations (section 2.4.2). Taken together, these findings suggested that these lineage- and population-specific DMR clusters were predominantly involved in supporting differentiation towards their marked populations.

Nevertheless, many DMRs from these DMR clusters exhibited partial DMR expansion in progenitor populations and mature populations representing alternative cell fates, i.e., fates to which differentiation did not appear to be supported by these DMRs (section 2.4.2). For example, the myeloid M1 DMR cluster appeared strongly tied to the differentiation towards the monocyte, neutrophil, and eosinophil cell fates (Figures 13 and 14). At the same time, a considerable fraction of M3 cluster DMRs exhibited partial DMR expansion already in the MPP3 population, which was maintained in the populations of the erythroid and the dendritic cell lineages (Figure 27). Furthermore, this phenomenon of widespread DMR seeding shaped the methylomes of the mature population. Each mature population was characterized (Figure 27) by a combination of i) DMRs that were maximally hypomethylated and fully expanded in this population, and ii) other DMRs that were more strongly hypomethylated and expanded in other populations (or hypermethylated for the gain of methylation clusters). Together, these findings suggested that mature hematopoietic cells likely possess a considerable number of partially expanded DMRs representing epigenetic memory of DNAme programming associated with alternative, discarded fate explorations.

Under this hypothesis, DMCpG-PP programming associated with fate exploration would occur in non-lineage-restricted progenitor cells. Such DNAme changes would then be at least partially maintained throughout differentiation. This model thus requires the assumption that fate exploration-associated DMCpG programming in progenitor cells can be stably maintained if the cell fate exploration is abandoned in favor of an alternative fate commitment. One direct expectation from this model is a strong correlation, at the level of individual DMCpGs, between the extent of DMCpG-PP programming in early stages of differentiation and the breadth of programming observed across the spectrum of mature hematopoietic cell types: a memory of fate exploration-associated DMCpG-PP programming in a progenitor cell would be maintained in all its progenitor cells; the earlier and the broader such DMCpG programming would occur within the early hematopoietic differentiation landscape, the more mature cells would inherit epigenetic memory of that programming. The observed global, DMCpG-level correlation between the extent of early programming and the breadth of programming across the mature populations was entirely in line with this expectation. It thus provides experimental evidence supporting a model where early fate exploration in progenitor cells is accompanied by DMCpG-resolved DNAme remodeling, which is at least partially maintained as epigenetic memory in mature cells, even if the fate exploration was abandoned in favor of other fates.

2.5.12 Each DMCpG programming pattern is associated with specific transcription factors

The ubiquitous role of DMCpG programming within the hematopoietic DMRs raised the question of the biological mechanism associated with this highly resolved DNAme remodeling. DNAme has a multi-faceted, direct relationship to transcription factor (TF) binding. DNAme at a transcription factor binding site (TFBS) can either promote or inhibit transcription factor binding [55, 56]. Furthermore, pioneering TFs can establish permissive seed regions within cis-regulatory elements, which may involve the removal of DNAme [49, 60]. I, therefore, hypothesized that heterogeneous DMCpG programming within DMRs may be associated with the activity of distinct TFs.

A novel paradigm for transcription factor enrichment analysis

I obtained a comprehensive catalog mapping the genomic locations of various archetype transcription factor binding motifs (TFBMs) across the murine genome [169]. Each archetype TFBM was associated with a cluster of individual TFBMs for multiple TFs. The TFBMs in each cluster were highly similar to each other. The archetype TFBM for each TFBM cluster was computed as a consensus motif obtained by aligning all associated individual TFBMs. The high TFBM similarity within each cluster of TFBMs made the computational distinction of genomic locations for individual TFBMs unreliable. Instead, the genomic locations of the archetype TFBMs should be viewed as potential binding sites for any TF associated with the archetype TFBM. Exemplary archetype TFBMs include i) the GATA archetype motif, comprising individual TFBMs for a group of transcription factors including GATA1-6 and TAL1; and ii) the CCAAT/CEBP archetype motif, comprising individual TFBMs for a group of TFs including CEBPA, CEBPB, CEBPD, and others. I next pooled all DMCpGs across all DMRs in each DMR cluster, obtaining one set of DMCpGs per DMR cluster. Then I grouped these DMCpGs according to their DMCpG-PPs. For this purpose, I only considered the (up to six) most frequent DMCpG-PPs for each DMR cluster (identified in section 2.5.8). This resulted in up to six distinct DMCpG sets per DMR cluster. For each such DMCpG set, I only allowed one randomly chosen DMCpG from each individual DMR. I reasoned that DMCpGs within a single DMR were more likely to introduce dependency structures into the enrichment tests, which would skew the testing results. To make statistical testing across all DMCpG sets comparable, I considered only DMCpG sets with at least 650 DMCpGs. DMCpGs sets containing more than 650 DMCpGs were downsampled to contain exactly 650 DMCpGs. I then screened each group of DMCpGs for associations with the archetype TFBMs, using Fisher's exact test to test each DMCpG group against the background of all other DMCpG groups. P-value adjustment into q-values was performed using the BH method [154, 155]. Taken together, this workflow represents a novel paradigm for TF enrichment testing: testing is neither performed at the DMR nor at the DMCpG level; instead, DMCpGs are grouped by two hierarchical annotations: the DMR-PP and the DMCpG-PP, and enrichment testing is performed on these groups. This enables testing for TFBM enrichments within systematically defined subregions of DMRs.

DMCpG-resolved programming within DMRs through different transcription factors

Figure 31 shows the enriched (archetype) TFBMs for all DMCpG sets identified within 13 representatively selected DMR clusters. The figure comprises the enrichment statistics for 30 TFBMs, representing key hematopoietic TFs. Each DMR cluster was characterized by a specific, limited set of highly enriched TFBMs. For many DMR clusters, one or more TFBMs were enriched across all DMCpG groups, independent of their respective DMCpG-PPs. At the same time, many DMR clusters also exhibited DMCpG-PP-specific enrichments of TFBMs. These TFBMs were specifically enriched in one or more DMCpG-groups within the DMR cluster, indicating association with specific DMCpG-PPs. For example, the C3 DMR cluster marked the cDC1, cDC2, and monocyte populations. Highly significant enrichment of the SPI TFBM (associated with SPI1, SPIB, and SPIC) was observed across all DMCpG-groups in the C3 cluster DMRs. However, the myeloid TFBMs CCAAT/CEBP (associated with CEBPA/B/D and other TFs) and CREB/ATF/3 (associated with ATF4, CEBPG, and DDIT3) were predominantly enriched in DMCpGs exhibiting the *m1* DMCpG-PP (q-value 0.06, log-odds 0.79). In contrast, no strong enrichment of the same myeloid TFBMs was found in other DMCpG groups within the C3 DMR cluster. For example, the enrichment test within the DMCpGs exhibiting the pan-dendritic *d1* DMCpG-PP yielded a q-value of 1.0 (log-odds -0.16). As another example, the eosinophil-specific M5 DMR cluster contained only two frequent DMCpG-PPs: the *m3* DMCpG-PP, which marked both neutrophils and eosinophils, as well as the *m5* DMCpG-PP, which marked eosinophils with high specificity. Within the M5 DMR cluster, the *m3*-DMCpGs were associated with myeloid CCAAT/CEBP TFBMs (q-value 1e-29). The *m5*-DMCpGs, on the other hand, showed little association with these TFBMs (q-value 0.93, log odds 0.24 for the CCAAT/CEBP TFBM) and were instead enriched in GATA binding sites (q-value 1e-3, log odds 0.7). The GATA TFBM was, in turn, not enriched in the *m5*-DMCpGs (p-value 0.65, log-odds -0.31). In summary, each DMR cluster was characterized by a specific set of enriched (archetype) TFBMs. Within each DMR cluster, some TFs showed significant associations across all DMCpGs, independent of their respective DMCpG-PP. Other TFBMs, however, were predominantly enriched in one or more DMCpG-subsets of the DMR cluster, characterized by specific DMCpG-PPs. This suggests that DMCpGs residing within the same DMR that exhibit different DMCpG-PPs may be preferentially regulated by different TFs.

Many (archetype) TFBMs were enriched in DMCpG exhibiting certain DMCpG-PPs independently of the DMR clusters in which these DMCpGs resided (Figure S11). For example, the myeloid CCAAT/CEBP and CREB/ATF/3 TFBMs were enriched in DMCpGs exhibiting the *m1* DMCpG-PP, independently of whether these DMCpGs were located within the C3, M4, or M1 DMR clusters. Another example is the erythroid-specific *e2* DMCpG-PP which is exhibited by the erythroid-specific E4 DMR cluster but also by the C2 DMR cluster, which marks CFU-Es and eosinophils and shows broad hypomethylation across the erythroid and myeloid lineages. The DMCpG-PP is enriched in GATA and MECOM TFBMs in both

DMR clusters. Other DMCpG-PPs in the C2 DMR cluster, such as the myeloid *m3* and *m5* DMCpG-PPs, are not significantly enriched in GATA or MECOM sites. Collectively, these observations support a model where the activity of certain TFs influences the DNAme state at distinct DMCpG sites across the genome, which may be located in DMRs belonging to different DMR clusters.

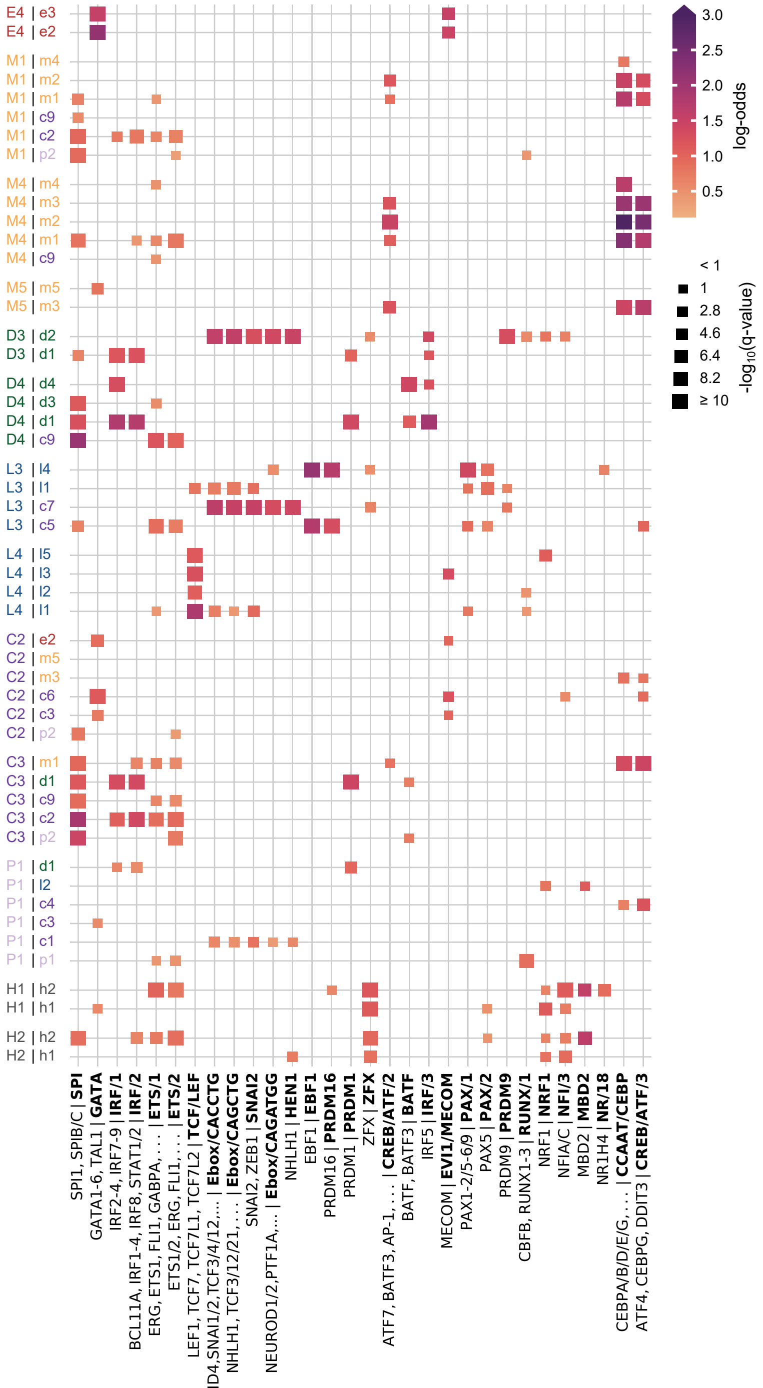

Figure 31: DMR subregions exhibiting different DMCpG programming patterns are enriched in distinct transcription factor binding motifs. DMCpGs were grouped hierarchically, first by the DMR cluster in which they occurred, then by the DMCpG cluster to which they belonged. For this purpose, only the (up to six) most frequent DMCpG-PPs for each DMR cluster were considered. This novel approach enabled testing for TFBM enrichments within systematically defined subregions of the DMRs. Each group of DMCpGs was screened for enrichments of archetype transcription factor binding motifs (TFBMs). Each archetype TFBM represented potential binding sites for multiple transcription factors, related through highly similar TFBMs [169]. Exemplary TFs associated with each archetype TFBM are indicated in the tick labels. The heatmap shows the enrichment testing results for 30 archetype TFBMs representing key hematopoietic TFs, within 13 representatively selected DMR clusters. The color of the rectangles encodes the log-odds score, and the size of the rectangles encodes the -$\log_{10}$(q-values) of the enrichment tests. Enrichment testing was performed using Fisher's exact test, comparing each DMCpG set against the background of all other DMCpG sets. P-value adjustment into q-values was performed using the BH method [154, 155]. To make statistical testing across all DMCpG sets comparable, only DMCpG sets with at least 650 DMCpGs were considered, and DMCpGs sets containing more than 650 DMCpGs were downsampled to contain exactly 650 DMCpGs. An alternative heatmap visualization of the same data, sorted by DMCpG cluster instead of by DMR cluster, is provided in Figure S11.

2.6 Highly resolved, hierarchical DMR/DMCpG programming in single cells

2.6.1 Engineering of a single-cell bisulfite sequencing analysis pipeline to generate high-quality methylome maps for 312 HSPCs

A key result of this thesis is the generation of a comprehensively annotated dual-layer DMR/DMCpG atlas of DNA methylome remodeling during hematopoietic differentiation. Another key result is the proposal of a new paradigm for how DNAme programming should be modeled: as a hierarchical process occurring at the level of DMRs and at the level of individual DMCpGs within DMRs. Together, these findings open many powerful possibilities for further advancing our understanding of the functional role of DNAme programming during hematopoietic differentiation and of the mechanisms by which information is encoded into the DNA methylome in differentiation systems. The analysis of DNAme programming in single cells is a highly relevant use case that could benefit from leveraging the DMR/DMCpG atlas and its underlying concepts. Therefore, I attempted to leverage the DMR/DMCpG atlas to investigate the DNAme state manifold in the HSPC compartment at the single-cell level.

FACS was used to isolate single cells in two tiers, ultimately generating single-cell methylomes for 74 LSK CD150$^+$ cells and 230 LSK cells that passed all quality control criteria. This dataset was supplemented with eight FACS-sorted immunophenotypic HSCs from a pilot experiment. Genome-wide (albeit sparse) single-cell methylome profiling was performed following the single-cell bisulfite sequencing (scBS-seq) protocol published by Clark et al. [74, 76]. All scBS-seq wet-lab experiments were carried out by members of the Section Translational Cancer Epigenomics (Mark Hartmann, Sina Stäble, and Maximilian Schönung), with support from Dr. Dieter Weichenhan (Div. Cancer Epigenomics), Julia Knoch (Div. Experimental Hematology), and the Single-cell Open Lab facility at the German Cancer Research Center. Index sort information was recorded for all LSK and LSK CD150$^+$ single cells and used for in silico gating to annotate each cell as either HSC or MPP1-5. The in silico gating was performed by Sina Stäble.

I was entirely responsible for the bioinformatical processing of the generated NGS data. To perform alignments, methylation calling, and quality control for the scBS-seq samples, I developed a comprehensive snakemake workflow. I have published this workflow as an open-source package [SOFT5]. Briefly, adapter and quality-trimmed sequencing reads were aligned with Bismark [170]. To deal with chimeric reads, read pairs were first aligned through paired-end alignment; unmapped read pairs were then subjected to single-end alignments to rescue the mappable read portions from chimeric reads. Methylation calling was subsequently performed with MethylDackel [171]. Using this workflow, sparse genome-wide

DNAme maps were generated for 312 HSPCs in total, with a median autosomal CpG coverage (Figure 32) of 528 149 CpGs, and values ranging from 124 150 CpGs to 2 387 069 CpGs (IQR = 375 581 CpGs to 690.993 CpGs). In summary, I have developed a comprehensive alignment, methylation calling, and quality control workflow and used it to generate a dataset of genome-wide, high-quality DNAme maps for 312 HSPCs.

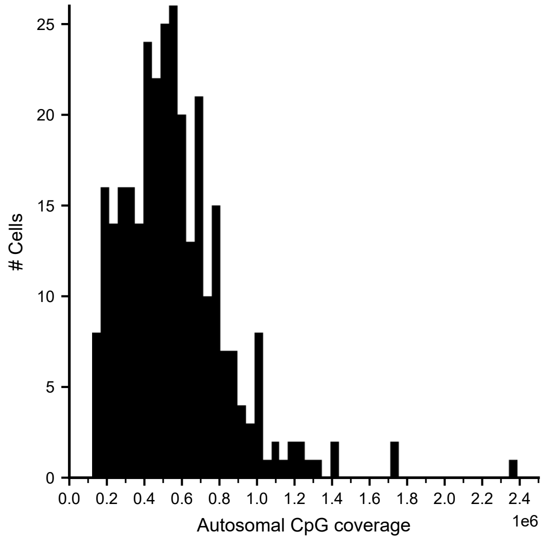

Figure 32: Histogram showing the autosomal CpG coverage for 312 HSPCs. LSK CD150^{+} and LSK cells were isolated by FACS. Whole-genome single-cell bisulfite sequencing was performed using the scBS-seq protocol published by Clark et al. [74, 76]. All wet-lab experiments were performed by collaboration partners. I developed a comprehensive workflow for the alignment, methylation calling, and quality control of scBS-seq data [SOFT5]. This workflow was used to generate a dataset of high-quality genome-wide methylome maps. Due to experimental limitations, these genome-wide methylome maps are sparse. For each cell, only a random fraction of all autosomal CpGs was measured. The histogram shows the total number of CpG dinucleotides with at least one methylation call for all cells.

2.6.2 Dual-layer DMCpG sets: novel features for scBS-seq analysis with unprecedented resolution capabilities

In this study, I observed that different DMCpGs within a DMR can be distinctly regulated at subsequent progressive differentiation stages (section 2.5.7). Additionally, the DMR subregions programmed in the early bulk progenitor populations MPP1-5 often consisted of small seed regions, typically containing just one or two DMCpGs, thus covering only a small fraction of the whole DMR regions. Consequently, mapping DNAme programming at the level of DMRs for single HSPCs appears likely to average out and obscure such DMR programming. To address this issue, I leveraged the integrated information about DMR and DMCpG programming provided by the dual-layer DMR/DMCpG programming atlas. This allowed for the systematic and robust quantification of DNAme programming, even in small DMR subregions. My approach involved identifying the DMCpG clusters to which most of the DMCpGs from a given DMR cluster belong and then grouping the DMCpGs within each DMR cluster based on their membership in these DMCpG clusters (as used previously for hierarchical TF enrichment analysis in section 2.5.12). This provided a partitioning of the hematopoietic DMCpGs into 110 sets defined by residence in the DMRs of a particular DMR cluster and membership in a particular DMCpG cluster, which I refer to as hierarchical DMCpG sets in the following. As a shorthand, I refer to particular hierarchical DMCpG sets by combining the DMR and DMCpG cluster names with a " | " symbol. For example, the P2 | *p2* hierarchical DMCpG set contains all DMCpGs from the P2 DMR cluster that belong to the p2 DMCpG cluster. Every single cell exhibited sufficient coverage across all individual hierarchical DMCpG sets to estimate the fraction of methylated DMCpGs within each set with an acceptable level of uncertainty (Figure 33). In summary, I propose to quantify the DNAme states of single hematopoietic cells by mapping the average methylation levels in 110 distinct hierarchical DMCpG sets derived from the DMR/DMCpG atlas.

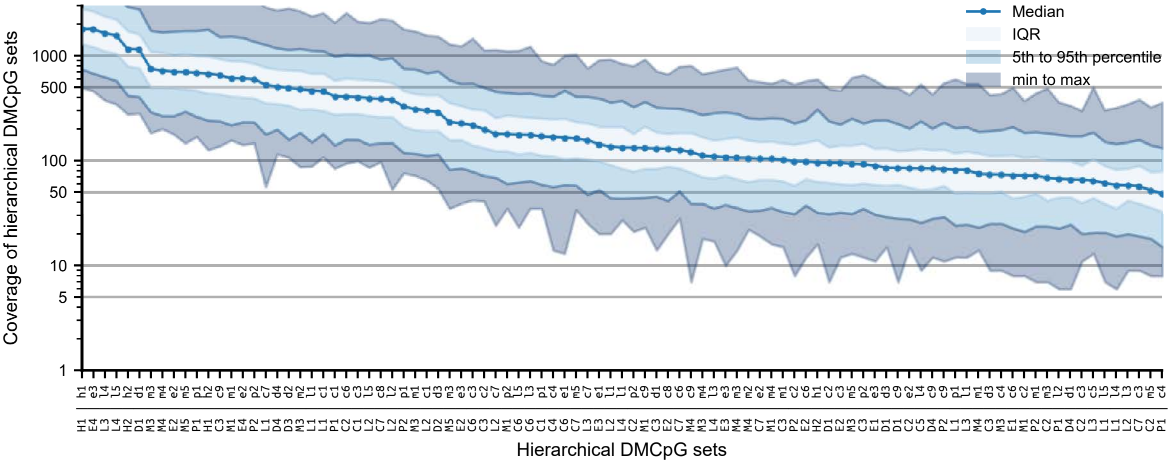

Figure 33: All hierarchical DMCpG sets exhibit sufficient coverage for robust methylation level estimation. The solid line with markers indicates the Median number of CpGs with at least one methylation call computed across all single HSPCs. Colored ribbons indicate the interquartile range (IQR), the range from the fifth to the 95th percentile, and the range from the minimum to the maximum observed coverage across the HSPCs.

A potential simplification of this analysis approach would be to create one set of DMCpGs per DMCpG cluster, ignoring the information about the DMR clusters in which the DMCpGs reside. Although the DMCpG clusters group DMCpGs with highly similar programming patterns regardless of their genomic location, this strategy has multiple caveats. First, the DMCpG clustering analysis intentionally grouped DMCpGs exclusively based on their methylation levels across mature cell populations. The programming patterns of DMCpGs in bulk progenitor populations were not considered to prevent bias in the DMR/DMCpG atlas due to surface marker-based definitions of FACS progenitor populations. As a result, DMCpGs from the same DMCpG cluster but residing in different DMR clusters could behave differently in the same progenitor cell. Second, the granularity of the DMCpG clustering was restricted to a manageable number of clusters. As discussed previously (section 2.5.9), DMCpG clusters could be further partitioned based on the DMR cluster in which individual DMCpGs resided, generating subgroups of DMCpGs with highly similar programming patterns that still varied slightly concerning the programming of a few populations with intermediate methylation levels. This indicated that some of the DMCpG clusters contained a low level of heterogeneity in their programming patterns, even across mature populations, particularly in populations without strong hypomethylation for a given DMCpG cluster. In summary, aggregating methylation levels across all DMCpGs belonging to the same DMCpG cluster for progenitor cell analysis could mask differential programming between different DMR clusters within individual cells, even though the DMCpGs within each DMCpG cluster shared highly similar programming patterns across various mature hematopoietic populations.

2.6.3 Mapping the structured continuum of single-cell DNA methylome states in the HSPC compartment

Evidence that differentiation starts from a lineage-naive DNA methylome state and is initiated by multi-lineage seeding

To compare the DNA methylome states of single cells within the HSPC compartment, each cell was characterized by its average methylation levels across the hierarchical DMCpG sets, following the rationale outlined above. The analysis revealed substantial, structured heterogeneity of the single-cell DNA methylome states within the HSPC compartment (Figure 34). Hierarchical clustering with Ward's method in combination with the cutreeHybrid partitioning algorithm [172] identified 12 single-cell clusters. One single-cell cluster stood out by exhibiting minimum methylation levels in the HSC-specific DMR clusters H1 and H2, and consistently high methylation levels in all lineage-specific DMR clusters. This cluster contained 19 cells predominantly composed of immunophenotypic HSCs (17 HSCs, one MPP1, one MPP2 cell). The cluster appeared to capture the most primitive stem cells in our dataset and was called the Apex_HSC cluster. Below the Apex_HSC single-cell cluster, multiple clusters characterized by complex DNA methylome states emerged, exhibiting concurrent seed hypomethylation across DMR clusters associated with differentiation towards multiple hematopoietic lineages. In summary, this study provides evidence that methylome programming in the HSPC compartment may start from a lineage-naive apex HSC state, followed by an initial concurrent accumulation of seed methylation in DMRs associated with differentiation towards multiple distinct lineages.

Continuous accumulation of hypomethylation in lineage-specific DMR clusters, progressing from multi-lineage seeding through oligo-lineage and lineage-specific DMR expansion

A closer examination of the single-cell clusters below the Apex_HSC cluster led to a tentative labeling of each single-cell cluster. The Early_MPP cluster, closest to the Apex_HSC cluster, exhibited a comparable amount of seed hypomethylation concurrently in the P1 | *p1*, P2 | *p2*, C2 | *p2*, M1 | *p2*, D1 | *p1*, and L1 | *p1* hierarchical DMCpG sets. This appeared to represent the initial seeding of DMR clusters associated with the erythroid, myeloid, lymphoid, and

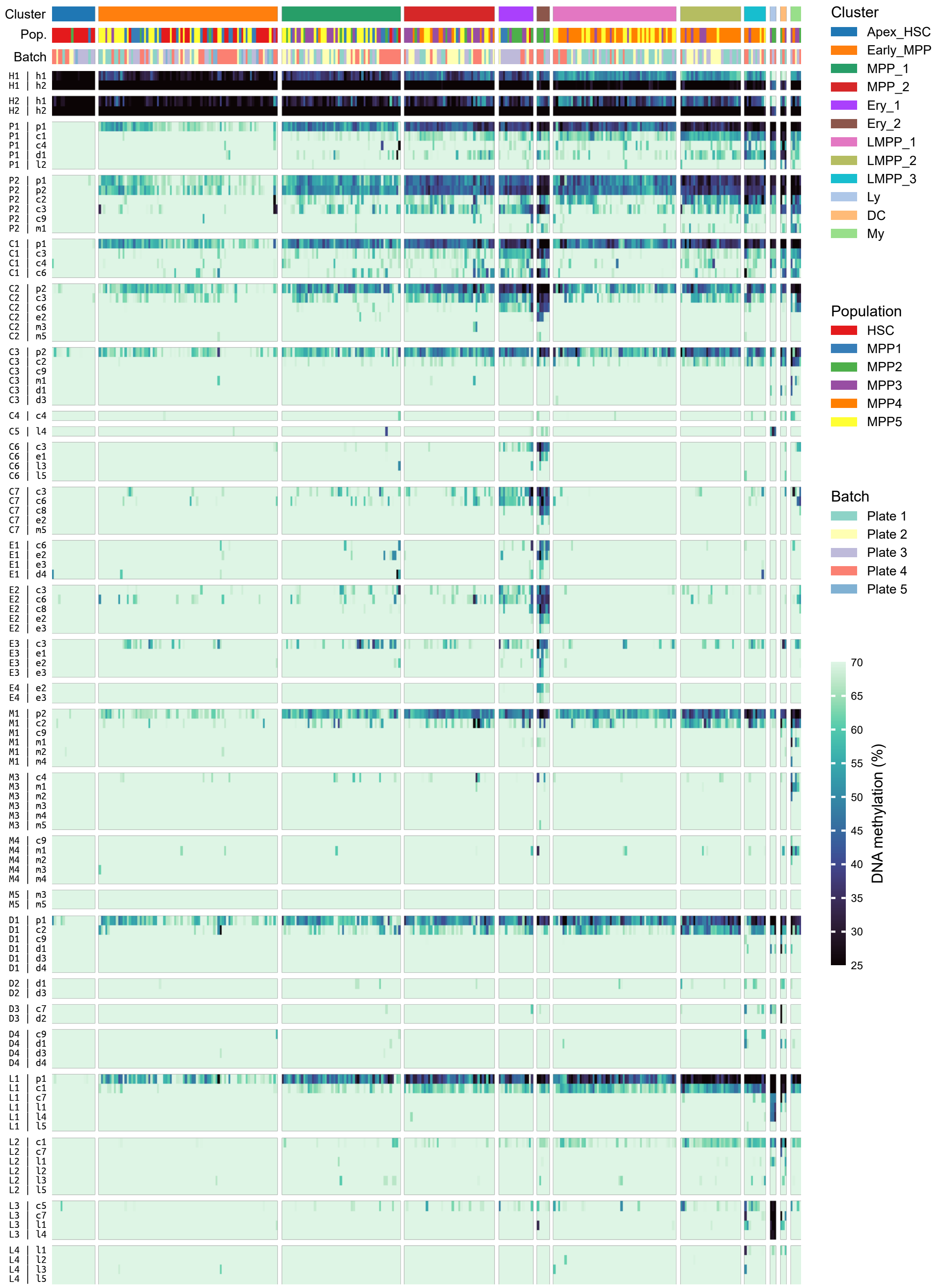

Figure 34: Structured heterogeneity of DNA methylome states in the HSPC compartment reveals a prominent role for multi-lineage priming upon exit from a lineage-naive apex HSC state. The heatmap shows the average methylation levels across 110 hierarchical DMCpG sets for 312 HSPCs. The hierarchical DMCpG sets belonging to each DMR cluster are ordered in increasing order of lineage- and cell type-specificity (c.f. Figure 25). The heatmap is annotated with clustering, surface marker, and batch information. Cells were clustered with hierarchical clustering in combination with the cutreeHybrid algorithm [172] for partitioning. Index sort information was used to annotate each cell based on its FACS gate residence as HSC or MPP1-5. This annotation was performed by Sina Stäble. The plate-based scBS-seq experiments were performed on five separate plates indicated as a batch variable.

dendritic cell lineages. Programming within DMR regions was confined to the seed regions identified by the *p1* and *p2* DMCpG clusters and did not involve further DMR expansion. Single-cell clusters representing further progressed differentiation stages were characterized by the progressive accumulation of hypomethylation in lineage-specific DMR clusters, which gradually extended to additional DMCpG clusters beyond the *p1* and *p2* clusters, indicating gradual DMR expansion. All of these single-cell clusters exhibited substantial seed hypomethylation across the P1, P2, C2, M1, D1, and L1 DMR clusters, which gradually extended beyond the earliest seed regions indicated by the *p1* and *p2* DMCpG clusters. This suggested that cells carry out a substantial amount of multi-lineage programming during early differentiation.

However, the clusters could be distinguished by diverging levels of methylation loss across different lineage-specific DMR clusters. For example, three LMPP clusters (LMPP_1/2/3) were characterized by increasingly high methylation levels in the H1 $|$ *h1* and H2 $|$ *h1* hierarchical DMCpG sets, which was expected to be associated with lymphoid differentiation (Figures 13 and 21). Furthermore, cells in these clusters prominently exhibited increasing LOM in the L1 $|$ *c1*, M1 $|$ *c2*, and D1 $|$ *c2* hierarchical DMCpG sets, indicating the expansion of lymphoid, myeloid, and dendritic cell lineage-specific DMRs. In contrast, the MPP_1 and MPP_2 single-cell clusters were characterized by stronger LOM in the erythroid-eosinophil-specific C7 DMR cluster and the erythroid-specific E3 DMR cluster, suggesting a stronger association with erythroid lineage output. In summary, the clustering analysis revealed structured heterogeneity in the HSPC compartment, even among the earliest progenitor stages, characterized by initially balanced multi-lineage DMR seeding followed by a gradual bias towards further accumulation in specific lineages.

The analysis also revealed small clusters of cells characterized by strong DMR expansion in DMR clusters associated with a single lineage, such as the Ery_1 and Ery_2, Ly, My, and DC single-cell clusters. Importantly, all cells with apparently unilineage-associated DNA methylome states still exhibited strong seed hypomethylation across the M1, L1, D1, and C3 DMR clusters, i.e., across all other lineages. A PCA analysis of the single-cell DNA methylome states provided initial evidence for a continuous DNA methylome state manifold in the HSPC compartment (Figure 35). Along this manifold, separate trajectories could be tentatively envisioned from the Apex_HSC and the Early_MPP clusters through the MPP_1/2 and Ery_1/2 clusters to erythroid fates or through the LMPP1/2/3 to lymphoid fates; myeloid and dendritic differentiation trajectories could in part overlap with either erythroid or lymphoid differentiation trajectories.

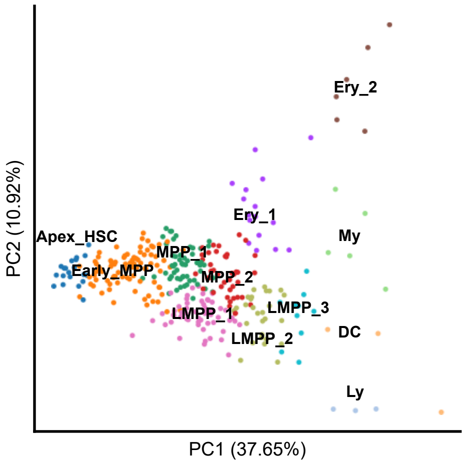

Figure 35: PCA analysis suggests lineage-independent DNA methylation programming upon exit from the apex HSC state, followed by an increasingly biased accumulation of lineage-specific DNA methylation programming. PCA analysis was based on the average methylation levels of the 110 hierarchical DMCpG sets (Figure 34). Single-cell cluster membership is indicated through colored markers and labels.

2.6.4 Staggered activation of DMCpG programming modules underlies progressive DMR expansion in single cells

Early hematopoietic differentiation in single cells seemed to involve the progressive accumulation of hypomethylation within DMR regions, indicating the progressive expansion of these DMR regions (Figure 34). This appeared to be driven by temporally staggered programming of the DMCpGs belonging to different DMCpG clusters. To further examine this mechanism, I leveraged the observation that the *p1* and *p2* DMCpG clusters exhibited relatively continuous methylation loss along all putative differentiation routes (Figure 34, Figure 35). Consequently, I reasoned that the methylation level in these clusters could act as an approximate DNAme-based differentiation pseudotime statistic. I ordered cells along three putative differentiation routes: i) an erythroid differentiation route (Apex_HSC, Early_MPP, MPP_1, Ery_1, Ery_2, Figure 36A); ii) a myeloid differentiation route (Apex_HSC, Early_MPP, MPP_1, MPP_2, LMPP_3, My, Figure 36B); and iii) a lymphoid differentiation route (Apex_HSC, Early_MPP, LMPP_1, LMPP_2, Ly, Figure 36C). For each route, I inspected the accumulation of hypomethylation in the M1, C3, and L1 DMR clusters, which are myeloid-, erythroid/eosinophil-, and lymphoid-specific, respectively. Along each route, I ordered cells by the methylation level of the hierarchical DMCpG sets L1 | *p1*, M1 | *p2*, and C2 | *p2*, respectively. For each of these DMR clusters, I tracked the accumulation across those DMCpG clusters which exhibited the greatest extent of programming along the inspected trajectories within the DMR cluster. In all cases, the programming of individual DMCpG clusters occurred in a staggered manner along the differentiation trajectories, with substantial time lags between the initiation of programming for each subsequent DMCpG cluster. The programming order of the DMCpG clusters correlated with increasing population specificity of the DMCpG clusters. For example, programming within the L1 DMR cluster began with the pan-hematopoietic *p1* DMCpG cluster. Subsequently, programming of the dendritic-lymphoid specific *c1* DMCpG cluster, the pDC-lymphoid specific *c7* DMCpG cluster, and finally the lymphoid-specific *l1* DMCpG cluster started at later stages of differentiation, each time after a substantial lag. The staggered programming of distinct DMCpGs within the same DMR cluster, beginning with DMCpGs belonging to the least specific DMCpG cluster and proceeding with DMCpGs belonging to clusters with increasing population specificity, was in accordance with expectations from the bulk population data analysis.

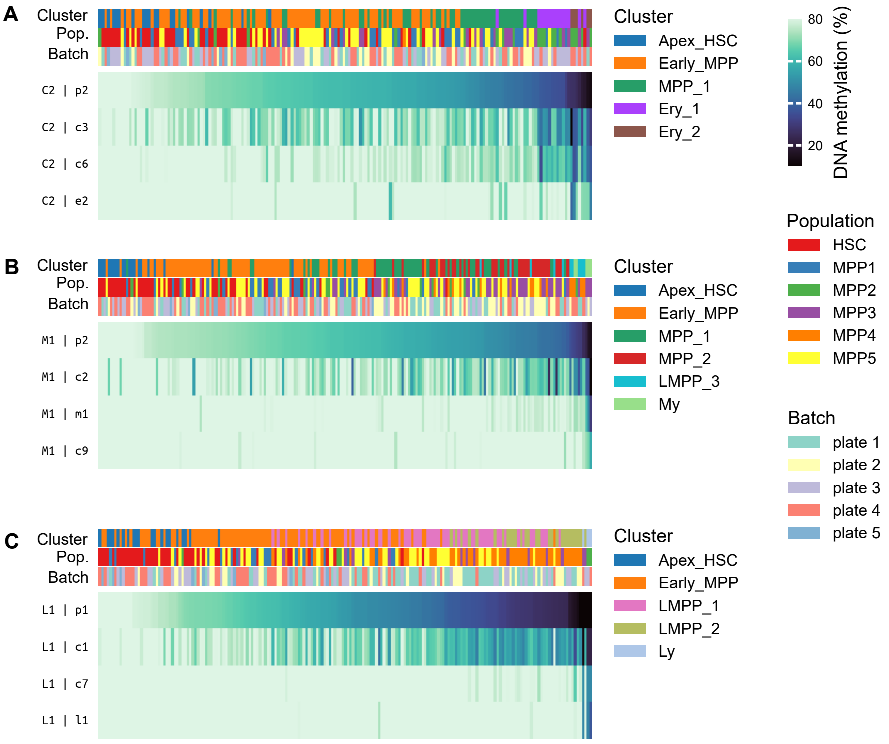

Figure 36: Staggered programming of DMCpGs belonging to different DMCpG clusters along differentiation trajectories. Single-cell clusters located along potential differentiation trajectories towards lymphoid-, myeloid- and erythroid-specific DNA methylome states were selected. Along each trajectory, cells from these single-cell clusters were ordered by their methylation level in the P1 | *p1* hierarchical DMCpG set, which provided an approximate, DNA methylation-based differentiation pseudotime. Heatmaps show programming in the lymphoid-specific L1 DMR cluster (A), the myeloid-specific M1 DMR cluster (B), and the erythroid/eosinophil-specific C2 DMR cluster (C) for the corresponding differentiation trajectories. For each of these DMR clusters, the methylation levels for the DMCpG clusters exhibiting the greatest extent of programming are shown.

Once programming of the DMCpGs within a cluster commenced, the fraction of unmethylated DMCpGs gradually increased along differentiation, i.e., I did not observe simultaneous demethylation of all DMCpGs of the same DMCpG cluster at a specific stage of differentiation. This observation suggests that the DMCpG clustering analysis grouped DMCpGs regulated at comparable differentiation stages but not programmed simultaneously in a monolithic DNA methylome remodeling operation. Instead, individual DMCpGs within a DMCpG cluster appeared to be progressively unmethylated along extensive segments of the differentiation continuum.

These findings provide further evidence that DMRs do not randomly accumulate hypomethylation across all DMCpGs they contain. Instead, DMCpGs within DMRs are systematically and heterogeneously programmed, forming clusters of thousands of DMCpGs across many DMRs that appear to be regulated as coherent programming modules. Notably, DMCpGs within these DMCpG clusters were not all programmed simultaneously but exhibited a continuous increase in the fraction of unmethylated DMCpGs during differentiation. These different DMCpG programming modules were activated in a staggered fashion throughout the differentiation process. This involved seeding of DMR regions with DMCpG clusters with broad population specificity followed by DMR expansion through the staggered activation of programming of DMCpG clusters with increasingly narrowing population specificity.

2.6.5 Highly-resolved characterization of cell type-specific DNA methylome states in mature hematopoietic cell types

The characterization of single-cell DNA methylome states through the methylation levels of the hierarchical DMCpG sets has enabled a detailed mapping of the structured continuum of single-cell DNA methylome states in the HSPC compartment. I next sought to determine if the same analysis approach could distinguish and deconvolve the DNA methylome states of different mature cell types. For this purpose, single-cell whole-genome DNAme maps were generated for 35 B cells, 32 T cells, 35 CFU-Es, and 34 monocytes, as described previously (section 2.6.1). All scBS-seq wet-lab experiments were carried out by members of the Section Translational Cancer Epigenomics (Mark Hartmann, Sina Stäble, and Maximilian Schönung), with support from Dr. Dieter Weichenhan (Div. Cancer Epigenomics), Julia Knoch (Div. Experimental Hematology), and the Single-cell Open Lab facility at the German Cancer Research Center. I was responsible for NGS read alignment and methylation calling, using a workflow developed by me for this purpose, as described previously (section 2.6.1). Each cell was characterized by average methylation levels across the hierarchical DMCpG sets. Hierarchical clustering of the cells resulted in the complete separation of the four cell types (Figure 37), indicating that distinct DNA methylome states characterized B cells, T cells, CFU-E, and monocytes. This included i) homogeneous, strong hypomethylation and full DMR expansion for all cell type-specific DMR clusters within each cell type (e.g., for the L3 DMR cluster for B cells, the L2 and L4 DMR clusters for T cells, and the E1-E4 DMR clusters for CFU-Es); and ii) homogeneous, strong hypomethylation of multi-cell type-specific DMR clusters in all corresponding cell types, involving in each case strong, but incomplete DMR expansion because individual DMCpG clusters were specifically programmed in single cell types. For example, DMRs from the lymphoid lineage-specific L1 DMR cluster were almost entirely expanded in both B cells and T cells. However, DMCpGs from the B cell-specific *l4* DMCpG cluster remained highly methylated in T cells, while DMCpGs from the T cell-specific *l5* DMCpG cluster remained highly methylated in B cells. As another example, DMRs from the monocytes/cDC-specific C3 DMR cluster were almost entirely expanded in monocytes, except for DMCpGs belonging to the cDC-specific *d3* DMCpG cluster. In summary, B cells, T cells, CFU-E, and monocytes were clearly distinguished by markedly strong, cell type-specific DMR hypomethylation and DMR expansion states in certain DMR clusters. Such markedly strong DMR programming involved complete DMR expansion in cell type-specific DMR clusters and near-complete DMR expansion in multi-cell-type specific DMR clusters, where programming of specific DMCpGs was reserved for specific cell types. Together, these findings provide further evidence for the DMCpG-resolved nature of DNAme programming across all stages of differentiation.

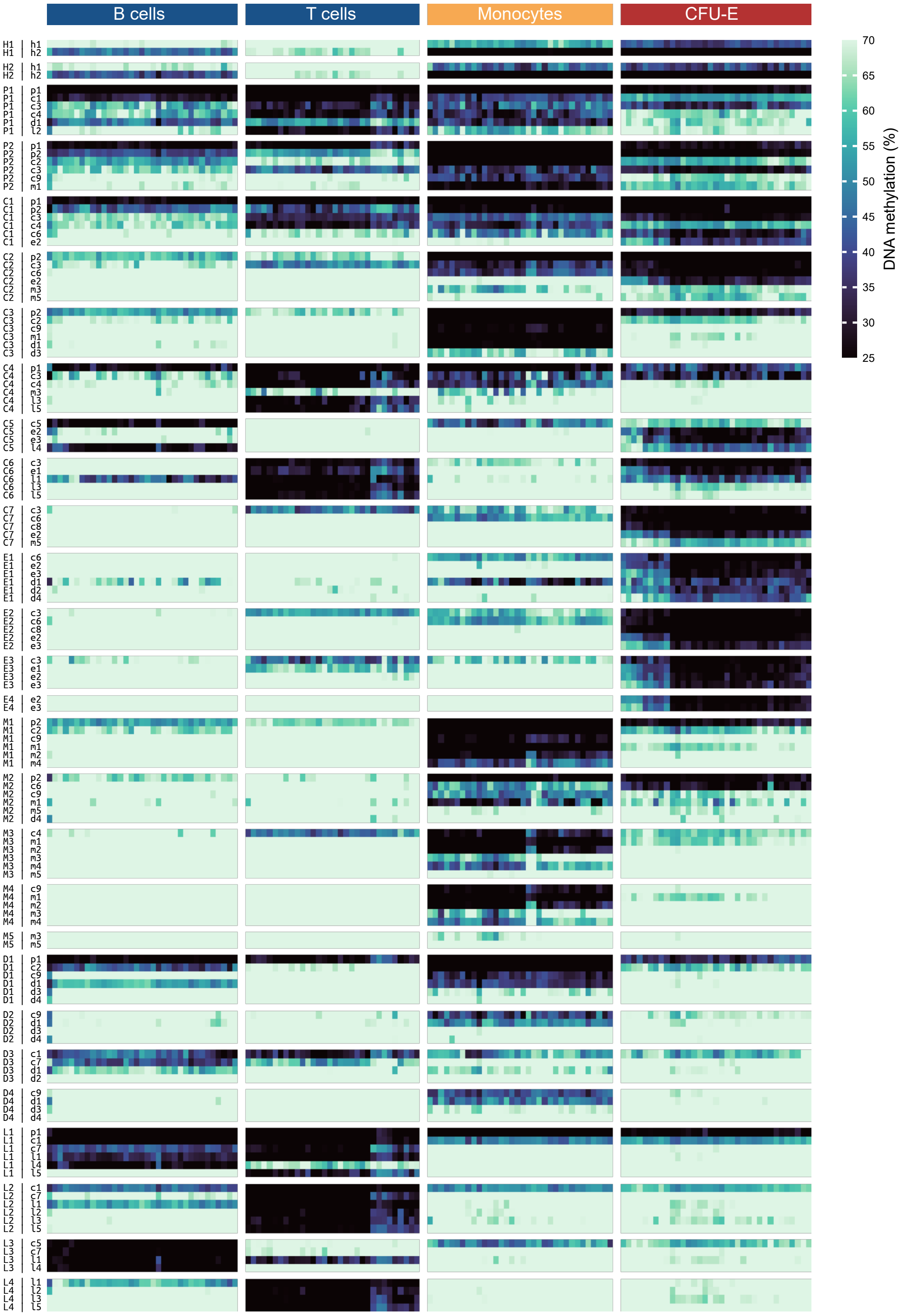

Figure 37: Single mature cells exhibit completed DMRs in cell type-specific DMR clusters and partially expanded DMRs in DMR clusters associated with alternative fates. The heatmap shows the average methylation levels across 110 hierarchical DMCpG sets for 35 B cells, 32 T cells, 34 monocytes, and 35 CFU-Es. The hierarchical DMCpG sets belonging to each DMR cluster are ordered in increasing order of lineage- and cell type specificity.

2.6.6 Exploring partial DMR expansion and the origins of widespread epigenetic memory in single mature cells

The analysis of bulk population data has suggested that mature cell types could contain many DMRs in partially expanded states within DMR clusters associated with alternative fates, representing epigenetic memory of DNAme programming in progenitor cells before fate commitment (section 2.5.11). I next investigated whether such partial DMR expansion was observable in single mature cells. In single cells of each mature cell type, systematic and substantial hypomethylation of DMR seed regions was observed across DMR clusters associated with alternative fates. In such cases, hypomethylation was specifically restricted to DMR seed regions composed of DMCpGs belonging to specific DMCpG clusters. DMRs were systematically not expanded beyond these seed regions, as evidenced by the high methylation levels of DMCpGs in the DMRs of the same DMR clusters that belonged to other DMCpG clusters. For example, all mature cell types exhibited multi-lineage seeding in the myeloid lineage-specific M1, the dendritic cell-lineage specific D1, and the lymphoid lineage-specific L1 DMR clusters, as indicated by significant hypomethylation of the *p1* and/or *p2* DMCpGs within these DMR clusters. The *p1* and *p2* DMCpG clusters were found to be the first DMCpG clusters programmed in a balanced multi-lineage seeding effort during the initial exit from the apex HSC state through the single-cell analysis of the HSPC compartment presented above (section 2.6.4). This shared hypomethylation of seed regions across multiple lineage-specific DMR clusters thus appeared to be epigenetic memory of early DNAme programming, retained independently of fate decisions during later differentiation.

Similar mechanisms may also shape the methylome of mature cell types at later stages of differentiation. For example, B cells showed hypomethylation in the T cell-specific L4 DMR cluster exclusively at lymphoid-lineage specific *l1* DMCpGs, but not at the T cell-specific *l2*, *l3*, or *l5* DMCpGs. On the other hand, T cells showed hypomethylation in the B cell-specific L3 DMR cluster specifically at the lymphoid-lineage specific *l1* DMCpGs, but not at the B cell-specific *c5* or *l4* DMCpGs. One possible explanation for this pattern is that the L3 and L4 DMR clusters could be seeded in parallel in a shared progenitor stage of B and T cells. As another example, CFU-Es exhibited partially expanded DMRs in the eosinophil-specific M2 DMR cluster, with hypomethylation of the *p2* and *c6* DMCpGs, but not the myeloid-specific *m1* and *m5* DMCpGs. This could be explained by a shared progenitor stage for erythroid and eosinophil fates.

For all seeded or partially expanded DMRs, the DMR programming states were relatively homogeneous across all cells from the same cell type. This suggests that the individual cells observed for each mature cell type may have been produced on similar differentiation routes. In summary, all mature hematopoietic cells carried strong hypomethylation in seed regions of DMRs of DMR clusters associated with alternative fates. This suggests that partially expanded DMRs in mature cells could represent epigenetic memory of parallel seeding of DMRs associated with differentiation towards multiple cell types, occurring hierarchically at different stages of differentiation. Such multi-cell type DMR seeding appeared to start with balanced multi-lineage seeding upon exit from an apex HSC state. This appeared to be followed by gradually more lineage-specific accumulation of hypomethylation over the course of further differentiation, resulting in progressive DMR expansion.

2.7 Publications, manuscripts and open source software packages

2.7.1 Manuscripts in preparation

A manuscript publishing the main results from this thesis is in preparation [PLANNED1].

This thesis proposes several extensions to the classical paradigm of regional DNA methylation programming, summarized in a new paradigm of hierarchical DNA methylation programming. Based on this new paradigm, this thesis develops novel computational strategies for the analysis of DNA methylation data. A review discussing the application of these ideas in future studies of the role of DNA methylation in health and disease is planned [PLANNED2].

2.7.2 Publications of data analysis contributions in collaboration projects

During my thesis, I have developed novel concepts concerning the structure and information content of the DNA methylome, and I have developed various computational tools for the performant and integrative analysis of multi-omics data comprising a DNA methylation layer. I have applied these insights and tools in several collaboration projects, confirming the applicability of the software packages and data analysis ideas developed in this thesis. Where applicable, these projects are also indicated at related sections in this thesis. This included projects i) with a focus on statistical and computational method development [OWN4–OWN6]; ii) with a focus on the analysis of the role of DNA methylation in the hematopoietic system in health, age, and disease [OWN1–OWN3, OWN7, OWN8]; and iii) two other projects to which I provided computational support [OWN9] and support during method development [OWN10].

2.7.3 Software packages for complex data visualizations in Python

The predominant majority of all data analysis tasks performed in this thesis were carried out within the Python ecosystem. However, while the Python ecosystem excels in data analysis and machine learning capabilities, it offers a less complete visualization ecosystem compared to other languages, such as R. Specifically, packages for illustrating complex multimodal and multidimensional data and specialized packages for the visualization of bioinformatics data, such as locus plots are missing.

Codaplot - flexible, multi-layered and modular complex heatmaps within the Python ecosystem

To address this gap, I have created the codaplot Python package. The codaplot package provides novel features compared with similar packages written in other languages, such as the ComplexHeatmap R package [173]. For example, codaplot simplifies the creation of multi-layered heatmaps or of plots with arbitrary spacing between groups of observations or samples. Almost all complex data visualizations in this thesis have been created completely within codaplot or with substantial use of codaplot, demonstrating its versatility (cf. Figure 13A, Figure 31, Figure 34, and Figure 28). The codaplot package is tightly integrated with the Matplotlib ecosystem, and thus capable of simple composition with any other Matplotlib-based library, while previous libraries were mainly useful for the generation of standalone complex heatmap visualizations [173]. The codaplot package is published as open source python package [SOFT3]. A manuscript describing the codaplot package is in preparation [PLANNED3].

locplot - genomic region plotting with tight Matplotlib integration

Different libraries for the visualization of multi-modal data within genomic regions exist, such as Gviz [174]. In contrast, the Python ecosystem still has a need for improved visualization capabilities for track-based genomic region data. Initial advances have been made, for example with the pyGenomeTracks package [175], but the currently available solutions are mainly intended for standalone use. They are not primarily designed as toolbox libraries for free customized use within the Matplotlib ecosystem, limiting the possible locus plot visualizations which can be achieved with these libraries. To address the need for a flexible genomic region plotting library in Python, I have developed a collection of low-level genomic region plotting functions which can be used in any Matplotlib figure. All locus plots in this thesis have been created using these tools, demonstrating the capabilities of the toolbox (Figure 26 and Figure 11A,B). I plan to make these tools available as a Python package code tentatively named locplot, and a manuscript describing this package is planned [PLANNED4].

2.7.4 Automated, multidimensional M-bias filtering with bistro

The T-WGBS data used for this study exhibited substantial and strongly varying M-bias, which varied significantly within individual samples depending on the fragment length, as reported before [81]. Many methylation callers provide basic functionality for trimming read ends to remove read positions which could be affected by M-bias [171]. However, to my knowledge, no methylation calling algorithm capable of precisely detecting and removing read positions affected by M-bias in a fragment length-dependent manner exists. Therefore, I have engineered a novel methylation caller capable of removing read positions affected by M-bias or gap repair nucleotides, as introduced by tagmentation reactions, only where it is truly needed, in a fragment dependent length-dependent manner. This precise M-bias filtering avoids the unnecessary exclusion of usable read positions and thereby retains more coverage, while removing more of the truly biased read positions. The bistro package was used to process the full T-WGBS dataset comprising multiple replicates for 25 hematopoietic populations presented in this thesis. The application of bistro provided methylation calls with sufficient quality for the CpG-resolved analyses presented in this work. A manuscript describing bistro is planned [PLANNED5].

2.7.5 Other software packages

I have open-sourced several utility software packages developed in the context of this thesis, without plans for publication, such as a comprehensive snakemake workflow for the analysis of scBS-seq data [SOFT5] and a Python package for handling large methylation profiling data sets [SOFT2].

Chapter 3

Discussion

3.1 A comprehensive atlas of DNA methylome remodeling during hematopoietic differentiation

3.1.1 Generation of genome-wide, high coverage DNA methylation maps for 25 hematopoietic populations

Before this work, a dataset enabling the genome-wide analysis of hematopoietic methylome remodeling across the full hematopoietic system was missing in both mouse and human, as detailed in the introduction (section 1.5.3). For this thesis, I have compiled a uniformly processed catalog of high coverage methylation maps for 25 hematopoietic populations. The T-WGBS sequencing data for nine of these 25 populations were generated in previous studies [14, 122, OWN1]. For the remaining populations, new sequencing data were generated by external collaborators (Methods, section 4.1.1). I have re-processed the previously published samples together with the newly generated samples, while optimizing the quality of the methylation calls for both the new and the previously published samples. For this purpose, I have developed a novel methylation calling software optimized for T-WGBS data [SOFT1]. These methylation calls were performed based on re-alignments of all data performed at the Omics IT and Data Management Core Facility at the DKFZ. This newly compiled dataset provides comprehensive coverage of the hematopoietic system, but is still partially limited through incomplete coverage of certain lineages. For example, the dataset is missing a megakaryocyte and a natural killer cell population. Still, the broad coverage of the HSPC compartment and the erythroid, myeloid, lymphoid and dendritic cell lineages provides unprecedented possibilities for mapping the methylome dynamics during hematopoietic differentiation.

3.1.2 A novel, dual-layer atlas capturing hierarchical DMR and DMCpG programming, with rich annotations

My thesis provided the bioinformatics analysis which translated this large-scale dataset into the (to my knowledge) first comprehensive, genome-wide atlas of hematopoietic methylome remodeling. The atlas maps DNAme changes using a novel dual-layer architecture that hierarchically maps DMR and DMCpG programming (discussed in section 3.6.4). This architecture is motivated by a novel model of hierarchical DNAme programming proposed in this thesis (discussed in section 3.6.3). The dual-layer DMR/DMCpG atlas provides multiple, complementary statistics quantifying the DNAme programming states of individual DMRs or DMCpGs across all populations. The atlas was further equipped with a rich set of annotations for both the DMR and the DMCpG layers. Many of these statistics and annotations were generated using novel strategies for the analysis of DNAme developed in this thesis. A selection of the information provided for the DMR and DMCpG layers is summarized in the following.

- DMR layer
  - Differentiation between DMCpGs and non-DMCpGs within each DMR.
  - Signal-to-noise ratio optimized DMR methylation levels, computed while excluding uninformative CpGs contained within DMR regions.
  - DMR expansion states, computed with two separate, novel classification strategies.
  - Gene annotations and genomic region annotations leveraging an innovative proximity-based gene annotation method.
  - DMR clustering information, associating each DMR with one of 28 characteristic DMR programming patterns signifying distinct functional roles during differentiation.
  - Significantly enriched lineage- and cell type-specific expression markers.
  - Significantly enriched, lineage- and cell type-specific enhancer overlaps.
  - Information about the individual DMCpG-PPs occurring within each individual DMR, allowing for subregional-stratification of DMRs or for subpartitioning of DMR clusters.
- DMCpG layer

- Autosome-wide tests for global null hypothesis of differential methylation during hematopoiesis.
- Autosome-wide tests for differential methylation between each population and the HSC population.
- Annotation of the parent DMR of each DMCpG contained within the DMR regions, identifying clusters of spatially adjacent DMCpGs.
- DMCpG clustering information, associating each DMCpG with one of 28 characteristic DMCpG-PPs. This allows grouping DMCpGs by their DMCpG-PP, independently of their genomic proximity.
- Annotation of significantly enriched TFBMs.
- Programming scores capturing how early during hematopoietic differentiation each DMCpG is likely to be regulated.

3.2 Integrated DMCpG and DMR calling with FDR control at the DMCpG level

3.2.1 Technical considerations: FDR control and dispersion estimates in a multi-group setting with few replicates

The bulk WGBS dataset used for DMR/DMCpG calling comprised 25 hematopoietic populations in two to five replicates. The task of identifying differentially methylated sites based on this large scale, multi-group bulk dataset presented multiple technical and conceptional challenges.

First, the high genome-wide coverage allowed comprehensive testing at almost all mappable CpG sites on the murine autosomes. The resulting high number of tests required adequate control of the FDR. FDR control at the DMCpG level of the atlas was critical for this study, which aimed to analyze the information content of the DNA methylome at CpG resolution. Supplementary FDR control at the DMR level would have further enhanced the interpretability of the analysis, but was of lower priority. Multiple statistical methods capable of identifying DMCpGs with FDR control [82] or DMRs with FDR control [87] exist. However, none of these methods provides simultaneous control of the FDR at both the DMCpG and the DMR level. Therefore, for the presented analysis, FDR control at the DMCpG level was prioritized, and the statistical procedure developed for this study was built on the statistical methods from the DSS toolbox [82], which allow FDR control at the DMCpG level. The idea of applying FDR control at the DMR level is partially motivated by the aim to avoid loss of sensitivity due to unnecessarily high numbers of tests [87]. While this trade-off certainly existed, it appears to be mainly a concern for lower coverage data or situations with small methylation level shifts: the high coverage of our data and strong methylation level shifts during hematopoiesis still allowed the detection of a large number of DMCpGs, in line with the expected number of hematopoietic DMCpGs (see below). I therefore concluded that a DMCpG-focused strategy was feasible with acceptable loss of sensitivity. Consequently, the DMR definition applied on top of the FDR-controlled DMCpG calls represents a heuristic definition capturing clusters of DMCpGs satisfying certain adjacency criteria. The DMR definition was parametrized to favor focal DMRs over large DMR blocks with intermittent unregulated CpGs, following the common convention in studies of differentiation systems [87, 176]. Taken together, the DMR/DMCpG atlas is based on FDR-controlled DMCpG mapping, favored over FDR-controlled DMR mapping. The definition of which spatial DMCpG clusters constitute DMRs follows common conventions, but it still necessarily includes parametrization regarding DMR size and density that are to a certain degree arbitrary.

Second, this dataset provided high genome-wide coverage, but few replicates per population. The low replicate number required consideration of the stability of the dispersion estimates underlying the detection of differential methylation at individual CpG sites. Basing the identification of DMCpGs on pairwise HSC vs. other DMCpG tests allowed leveraging the robust shrinkage-based dispersion estimation of the pairwise DMCpG hypothesis test provided within the DSS framework [82]. This ensured robust statistical inference with the low replicate numbers in our dataset (between 2 and 5 replicates, median of 3 replicates). Alternatively, a generalized least squares (GLS)-based beta-binomial modeling approach could have been used to directly test for DMCpGs in a multi-group setting. An implementation of this approach has been proposed by Park et al. [83] and is provided as part of the DSS library. This approach could in theory provide more statistical power due to its more direct testing strategy. However, its dispersion estimation was reported by the authors to be suboptimal with low replicate counts, making it less suitable for the dataset of this study. Thus, a procedure combining pairwise DMCpG testing employing shrinkage-based dispersion estimation with subsequent global null testing was chosen in favor of modeling based multi-group differential methylation calling due to the low replicate count.

3.2.2 An innovative procedure for integrated DMR and DMCpG calling and filtering

Following these considerations, I have developed and applied an innovative multistep procedure for the integrated identification of DMR and DMCpGs Figure 9, with FDR control on the level of individual DMCpGs. Briefly, candidate hematopoietic DMRs were identified by performing pairwise DMR calling between the HSC population and all other populations using the DSS DMR detection algorithm [82] and then merging these DMR intervals.

Hematopoietic DMCpGs were identified autosome-wide (FDR $\leq$ 1%, BKY method) and filtered based on a minimal significant methylation level shift of 20% compared to the HSC population in at least one population. Next, candidate DMR regions were filtered to only include those containing at least two DMCpGs and showing a methylation level shift of at least 30% compared to the HSC population in at least one population. These DMRs were trimmed to end with DMCpGs. No smoothing was applied at any step of the workflow. This choice, the parameterization of the DMRs (minimal DMR size of 3 CpGs and 2 DMCpGs, at most 50% of not differentially methylated CpGs within a DMR interval) and the DMCpG trimming step allowed for the detection of highly focal DMRs. Separately, pairwise autosome-wide DMCpG tests between the HSC population and each other population were performed (FDR between 5% and 0.5% depending on the population) to gain additional annotations for the DMR/DMCpG atlas.

3.2.3 Advantages and limitations: a focal DMR model and a sensitivity tradeoff

Taken together, the testing procedure developed for this study provided the following advantages over alternative approaches: i) FDR control at the DMCpG level; ii) the ability to detect very small, highly focal DMRs, even when occurring in close spatial proximity; iii) highly conservative testing at the DMCpG level combined with strict signal-to-noise ratio filtering to obtain an atlas of DMR/DMCpG programming with high precision, containing features exhibiting strong DNAme shifts; iv) high depth of annotation through pairwise HSC-vs-other DMCpG tests, global DMCpG tests and DMR annotations. These advantages come with disadvantages. First, the conservative, multistep DMCpG calling approach combined with the large number of autosome-wide DMCpG tests and strict FDR control represents a trade-off of precision in favor of sensitivity. Thus, the DMR/DMCpG atlas is not presented as a definitive map of all sites of DNAme programming during hematopoiesis. Rather, it was compiled with the intention of identifying a sufficient number of DMCpGs and DMRs to allow a variety of downstream applications which require highly precise DMCpG calling, but do not require an exhaustive mapping of DMCpG sites. One exemplary application with these requirements is the use of the DMR/DMCpG atlas as a reference for single-cell methylome analysis. A second disadvantage is that the highly focal DMR calling strategy may lead to oversegmentation of methylation-dependent regulatory elements into multiple DMRs. This increases the signal-to-noise ratio for each individual DMR, but may lead to unwanted dependency structures during statistical enrichment testing and similar applications. However, this could be addressed without further analysis efforts by simply clustering adjacent DMRs in the atlas. Collectively, the integrated DMR/DMCpG testing strategy favored DMCpG with high precision and focal DMR annotations over exhaustive sensitivity of DMCpG detection and the detection of larger, methylation-dependent cis-regulatory elements.

3.2.4 High coverage data enable conservative, autosome-wide DMCpG detection while retaining high sensitivity

In total, the autosome-wide DMCpG testing identified 1,136,816 DMCpGs (5.6% of all autosomal DMCpGs) at an FDR ≤ 1%. Of these, 584071 DMCpGs occurred within 122,613 DMRs. 84.3% of all CpGs contained within DMRs were identified as DMCpGs. The detected numbers of autosomal DMCpGs and DMRs were in line with expectations, [129, 176–178]. I have developed and implemented a novel proximity-based DMR annotation algorithm, which yielded DMR-to-gene annotations for 97074 DMRs (79% of all DMRs, Figure 12C). 10417 DMRs (8.5% of all DMRs) were located in promoter regions, while the most frequent genomic position was within introns. These genomic locations of the DMRs were in line with previous findings [176]. The distances between the individual DMRs and their closest TSSs formed a distribution centered close to zero, with a strong decay towards larger distances (Figure 12A), as reported previously [176]. Taken together, these findings indicate that, despite its sensitivity tradeoffs, the applied testing procedure retained sufficient power to detect large parts of the methylome remodeling occurring during murine hematopoiesis.

3.3 Hierarchical hypomethylation dynamics during hematopoiesis

3.3.1 Methylome remodeling during hematopoiesis predominantly involves unidirectional loss of methylation

The important role of cell type-specific methylome remodeling during hematopoiesis has been thoroughly established [93, 118, OWN2, 176, 178]. However, few studies have described multi-cell type-specific methylation changes during hematopoiesis [OWN1, 130, OWN2, 179]. The biological function and detailed programming patterns of such multi-cell type-specific programs have not been systematically explored yet. One reason is that the functional associations of DNAme programming during hematopoiesis have mostly been analyzed through separately treated sets of cell type specific DMRs, and not through the characterization of clusters of DMRs which show co-regulation across multiple cell types [93, 118, 129, 130, 176, 178, 179]. However, functional characterization of such DMR clusters may be the more informative strategy for many research questions, as further discussed in section 3.3.4.

I therefore performed the first systematic characterization of 28 DMR clusters exhibiting co-regulation across the hematopoietic system. Hematopoietic DMR programming was predominantly characterized by unidirectional loss of methylation (LOM) compared to the methylation level of the HSC population (Figure 13A). While 26 distinct DMR clusters characterized by a range of such LOM patterns were detected, only two DMR clusters characterized by gain of methylation (GOM) were detected. In total, 11 166 DMRs (9.1% of all DMRs) were part of the GOM clusters, while 111 395 DMRs (90.9% of all DMRs) were part of LOM clusters. This high proportion of hypomethylation-based DMR programming was remarkably similar to the findings in another differentiation system: Gascard et al. [178] reported that the ectoderm to breast epithelia differentiation was dominated by DNAme loss (87% of all differentially methylated CpGs).

3.3.2 Capturing hierarchical hypomethylation dynamics through DMR clustering analysis

Population- and lineage specific DMR programming occurred exclusively through the progressive establishment of hypomethylation, except for one DMR cluster which could potentially indicate a role of hypermethylation specifically during the differentiation towards lymphoid populations (further discussed in section 3.4.2). Focusing for each DMR cluster on those populations characterized by markedly strong hypomethylation (the marked populations for each DMR cluster), the hypomethylation DMR clusters appeared to capture a hierarchical system of DMR programming modules (Figure 13A). First, two DMR clusters were characterized by multi-lineage hypomethylation. Of note, while these clusters were named pan-hematopoietic DMR clusters, they did not exhibit fully lineage-independent programming. The P1 cluster marked the myeloid, lymphoid and dendritic cell lineages, but not the erythroid lineage, while the P2 cluster marked all lineages except for the lymphoid lineage. Second, seven DMR clusters marked multiple populations across two of the erythroid, myeloid, lymphoid and dendritic cell lineages. Third, five DMR clusters marked multiple populations within a lineage. Fourth, 12 DMR clusters specifically marked a single population. Importantly, for each DMR cluster, I observed strong statistical associations with gene expression and enhancer modules whose activity patterns matched the hypomethylation patterns of the DMR clusters. I further investigated the characteristic expression patterns of the target genes of each DMR cluster, and found that these target gene sets were specifically expressed in populations characterized by hypomethylation in the DMR cluster. Collectively, these findings supported the interpretation that the DMR clusters captured a hierarchy of DMR programming modules involved in multi-lineage, single-lineage and cell type specification during hematopoietic differentiation.

Of note, the general architecture of this apparent system of hierarchically organized DMR programming modules was similar to a hierarchical system of histone modification programming modules observed during hematopoietic differentiation by Zeller et al. [108]. Furthermore, the programming patterns of several of these DMR clusters were in line with established hematopoietic differentiation routes. For example, the C3 DMR cluster marked both the monocyte and cDC populations, in line with a possible differentiation route producing both monocytes and cDCs [24]. Multiple cross-lineage DMR clusters marked both erythroid and eosinophil populations. This was in line with various findings suggesting shared differentiation trajectories towards the erythroid and eosinophil/mast cell/basophil cell types (reviewed in [180]). Taken together, several multi-cell type-specific DMR programming patterns recapitulated known shared differentiation trajectories. This further supports a model where hierarchically organized DMR programming modules accompany progressive fate restriction. Of note, hierarchical epigenetic fate encoding is fully compatible with a continuous and complex cell fate restriction state continuum, as was suggested by initial results of Zeller et al. [108] and further characterized in this study (section 2.6.3).

Focusing on cell type-specific hypomethylation neglects the important role of various DMR clusters, which hierarchically characterize groups of cell types

My work revealed a hierarchy of DMR programming modules characterized by an important role for multi-cell type-specific DMR programming. In contrast, a recent study that mapped methylation changes across 39 different human cell types and various tissues reported that 97% of all identified differentially methylated 'blocks' (a segmentation-derived genomic interval concept similar to that of a DMR) were specifically unmethylated in one cell type and specifically methylated in all other cell types. This study employed a unique segmentation-based method for detecting differential methylation without statistical inference and false discovery rate (FDR) control, unlike other segmentation-based methods which emphasize the necessity of a post-segmentation FDR control step [87]. It is possible that different methods may pick up distinct kinds of methylome programming, such that the cell type-specific methylome blocks reported by Loyfer et al. [129] represent a distinct layer of DNAme programming complementing the DMR programming layer revealed in my work. There is however a possibility that suboptimal regularization of the segmentation approach applied by Loyfer et al. [129] in combination with the lack of FDR-controlled statistical inference could have led to a preferential segmentation of small genomic intervals characterized by stochastically arising hypomethylation in single samples. It seems less likely that the multi cell type-specific DMR programming patterns found in my study are a technical artifact because the percentage of multi-cell type specific DMRs found in my study closely matches previous reports of the proportion of such DMRs, generated using different technologies. An essential role for multi-cell type-specific methylome programming is for example supported by a WGBS study across 30 diverse human cell and tissue types which found that 40% of the detected DMRs were multi-cell type-specific [176]. Furthermore, an array-based study of methylome changes across multiple hematopoietic cell populations found that 54% of the differentially methylated CpGs were multi-cell type-specific [OWN2]. These proportions are similar to the percentage of 57% of multi-cell type-specific DMRs found in my study. In summary, the occurrence of DMR programming modules with a range of cell type-specificities is consistent with the majority of previous reports, calling into question recent findings of highly cell type-specific DMR programming.

3.3.3 A functional role for intermediate DMR methylation levels occurring during progressive methylation loss

The regulatory pattern of each LOM DMR cluster could be primarily characterized by markedly strong hypomethylation in a specific set of populations (the marked populations). However, all DMR clusters were additionally characterized by progressively increasing levels of hypomethylation across multiple other populations in addition to the maximum hypomethylation in the marked populations (Figure 13A). Thus, besides marking one or more populations through strong DMR hypomethylation, all DMR clusters also exhibited intermediate DMR methylation levels in specific other populations. An important implication of these progressive DMR programming patterns is that, viewed the other way around, for all DMRs exhibiting intermediate methylation levels observed in any population, strong hypomethylation occurred in at least one mature population. Thus, intermediate methylation levels were never the endpoint of DMR programming during differentiation. Instead, they appeared to be intermediate states within DMR clusters characterized by progressive, unidirectional LOM. This suggests that intermediate DMR methylation levels represent functionally intermediate states in the context of progressive DMR programming modules. Of course, the interpretation of intermediate DMR methylation levels in bulk data is challenged by the possibility of various underlying population structures [181–183]. However, various further analyses discussed in detail in the following sections provided substantial support for the interpretation of intermediate DMR methylation levels as representative of functionally intermediate DMR programming states. Before this study, progressive DNAme remodeling has only been systematically explored within narrow systems. Lipka et al. have demonstrated and characterized progressively changing DMR methylation levels within the HSPC compartment [122]. He et al. have demonstrated progressive DMR methylation level loss during embryogenesis [184]. Furthermore, staggered DMR methylation levels across multiple hematopoietic cell types have been visible in the data representations of several published analyses [130, OWN2, 179], but have not directly addressed in these studies. Taken together, my thesis provides the first systematic demonstration that progressive DMR programming is a ubiquitous feature of lineage- and cell type specific DMR programming modules, and leverages several novel analysis strategies to ascertain the functional nature of progressively increasing DMR hypomethylation.

3.3.4 Dissecting functionally distinct hypomethylated regions for improved cell type characterization

The utility of characterizing DMR programming through the identification of co-regulated DMR clusters has been demonstrated through the successful dissection of DMR programming in mouse embryo development [184] and of DMR programming in the HSPC compartment in the mouse [14, 122], but has still not been widely adopted. Instead, many studies focus on the use of DNAme as a cell type marker, e.g., in the context of tissue deconvolution tasks and not as a highly informative mark of coordinated epigenetic programming modules (see section 3.3.1). My thesis highlights that the DMR cluster-based analysis of DMR programming enables a functional stratification of the hypomethylated regions exhibited by each population. It shows that while each population exhibits many strongly hypomethylated regions, these regions originate from multiple distinct epigenetic programs with different functions and cannot be viewed as a monolithic cell type feature, as is commonly done [93, 118, 129, 130, 176, 178, 179]. Functional characterization of DNAme changes should therefore be performed per DMR cluster and not per population. In summary, my thesis provided the first systematic analysis of the DMR programming patterns associated with hematopoietic fate specification and highlighted the importance of this perspective for the correct functional annotation of DNAme changes.

3.3.5 Limitations of enrichment-based DMR cluster characterizations

The functional characterization of the DMR clusters through enrichment analysis and DMR cluster target gene set expression analyses has several limitations. The limitations of these two approaches can be treated together, because the DMR cluster target gene set expression analysis is essentially an enrichment analysis: it identifies the mean expression profile across sets of genes, which is shaped by the most commonly occurring individual gene expression profiles across this set of genes. As a first limitation, when the enrichment analysis or the gene set expression analysis indicate that DMR clusters marking multiple populations are associated with genes expressed in matching cell types, one cannot distinguish between a situation where the DMR cluster is separately associated with specific marker genes for each cell type, or a scenario where the DMR cluster is associated with genes marking multiple cell types at once. Second, the observed enrichments and average DMR cluster target gene set expression profiles only indicate statistical associations of DMR programming with matching programs on other omics levels, i.e., they indicate that many more DMRs than expected by chance overlap with cell type-specific enhancers or cell type marker genes. This does not preclude that a large fraction of the DMRs within each cluster does not contribute to these associations. Various studies have reported that only a small fraction of DMRs can be functionally associated with gene expression changes [185], suggesting that many DMRs could either represent only composite contributions in more complex regulatory systems across multiple omics levels, or have no direct regulatory function. Thus, while the analyses in this study demonstrate that the different DMR clusters are specifically and highly significantly associated with functional lineage specification programs on other omics levels, they do not quantify which fraction of the DMRs directly underlie these associations.

3.4 Specific roles for DNA hypermethylation during hematopoiesis

3.4.1 Loss of stemness is accompanied by lineage-independent gain of methylation

The H1 DMR cluster was characterized by strong hypermethylation in all lineages, and substantial DMR hypermethylation was already present in the HSPC compartment. This DMR cluster appeared to be the only DMR cluster capturing lineage-independent programming. Its association with gene expression and enhancer modules suggested that H1 DMR programming is involved in silencing HSPC marker genes during lineage commitment. The H1 DMR cluster was specifically associated with marker genes for LT-HSCs, ST-HSCs, MPPs, and LMPPs. Additionally, the cluster was specifically enriched in overlaps with enhancer regions characterized by strong activity in LT-HSC, ST-HSC, and MPP enhancers. Furthermore, the H1 DMR cluster target gene set was specifically expressed in HSC and ST-HSC single-cell clusters in an integrative single-cell transcriptome analysis. The H1 DMR cluster appeared to be the only DMR cluster with a role related to the downregulation of stemness during differentiation. This suggested that DNAme-based control of loss of stemness may exclusively occur through hypermethylation programming, while hypomethylation programming is always associated with hierarchical fate restriction, i.e., with lineage-dependent programming. An important role of hypermethylation for the control of exit from the HSC state is in line with the importance of de novo DNAme through DNMT3A for exiting the HSC state, as reported by [114]. Furthermore, it is consistent with the existence of a program involved in the repression of stemness marker genes shared by all hematopoietic lineages, which operates through the global buildup of H3K9me3 [108]. In summary, loss of stemness appears to be exclusively mediated through hypermethylation during hematopoietic differentiation.

3.4.2 Lineage-specific gain of methylation occurs only for lymphoid populations

Lineage-specific DMR programming appeared to occur almost entirely through hypomethylation, except for a single hypermethylation DMR cluster (H2). This H2 DMR cluster exhibited lymphoid-specific hypermethylation, suggesting a specific role during lymphoid differentiation. These findings were in line with finding from Roy et al. [130], who studied lineage-specific DMR programming across myeloid and lymphoid immune cell types and found that in contrast to a widespread role of hypomethylation DMR programming, lineage-specific hypermethylation DMR programming was restricted to adaptive immune cells. Such a strictly limited role of lineage-specific repressive DMR programming appears to contrast with a broad role of the lineage-specific buildup of repressive histone marks during hematopoietic differentiation [108]. This could indicate complementary roles for the DNAme and histone modification layer during epigenetic lineage specification. Despite the lymphoid-specific DMR programming pattern, no evidence of an association between the H2 DMR cluster and lymphoid-specific expression or enhancer modules could be identified. Furthermore, the H2 DMR cluster target gene set was not strongly expressed in lymphoid cells. Instead, the H2 DMR cluster exhibited similar enrichment profiles and a similar cluster target gene set expression profile compared to the H1 DMR cluster (section 3.4.1). This included specific enrichment of LT-HSC, ST-HSC, and MPP expression markers as well as specific enrichment of LT-HSC, ST-HSC, and MPP-specific enhancers. Thus, the associations of the H2 DMRs with specific expression and enhancer modules suggest a role of H2 DMR programming in stemness suppression rather than a role in lymphoid differentiation. Further characterization is still needed to reconcile these multi-omics associations with the lymphoid-specific DMR programming pattern of this cluster. In summary, lineage-specific DMR programming occurred almost exclusively through loss of methylation, except for one lymphoid-specific hypermethylation DMR cluster with an unclear functional role.

3.5 DMR expansion dynamics reveal mechanisms of hierarchical methylome programming

3.5.1 DMR expansion dynamics in differentiation systems could have high information content but are underexplored

Progressive DMR methylation loss observed in bulk data could either indicate DMR deepening (i.e., homogeneously decreasing CpG methylation levels for a fixed set of CpGs) and/or DMR expansion (i.e., additional, newly-regulated CpGs show significant hypomethylation in a series of populations). On the one hand, progressive DMR deepening during differentiation could represent either an increasing likelihood of DNA demethylation or a progressive enrichment of cells characterized by a pre-existing, fully formed hypomethylated state. On the other hand, progressive DMR expansion during differentiation is an unambiguous result of subsequent, functionally distinct DMR programming steps over the course of differentiation. The question of whether DMRs expand and/or deepen in the course of differentiation is thus of high relevance for the interpretation of progressive methylation level changes observed in bulk data - and analogously when single-cell data are aggregated to pseudo-bulks. Recognizing the importance of characterizing DMR expansion dynamics during hematopoietic differentiation, Hodges et al. [119] have undertaken a promising initial study of DMR expansion during hematopoietic differentiation. They could demonstrate that hypomethylated regions commonly expand asymmetrically in the direction of their nearest genes during differentiation. The study further found indications that such DMR expansion in promoter regions may correlate with differential gene expression [119]. However, since this pioneering work, the progressive expansion of DMR regions in differentiation systems has been largely left unexplored. Distantly related research has been focused on the regulation of the size of methylation canyons through DNTM3A and TET enzymes [186, 187], but these studies also did not lead to a general characterization of DMR programming through expansion. Research may have been partially hindered by a lack of computational methods for the quantification of DMR expansion as well as by a lack of sufficiently comprehensive datasets to track DMR expansion along differentiation trajectories. In summary, the quantification of DMR expansion is a promising approach for distinguishing between different mechanisms which could underlie progressively changing DMR methylation levels during differentiation observed in bulk data. However, the use of DMR expansion as a DMR programming mechanism in differentiation systems is largely unexplored, and no established computational methods exist to quantify DMR expansion dynamics.

3.5.2 Novel methods allow tracking of DMR expansion dynamics across bulk populations and single-cell clusters

To address both the current lack of knowledge regarding the role of DMR expansion in hematopoietic differentiation and the need for computational methods to quantify DMR expansion dynamics, I have developed several two novel methods that allow mapping the expansion state of individual DMRs with high resolution across bulk population data. While I have used four possible expansion states throughout this work (unregulated, seeded, intermediate, and completed DMR expansion states), more or less fine-grained classifications would be equally feasible with the proposed strategies, depending on the goal of each analysis and the quality of the data at hand. The first method directly compares the methylation levels of all individual CpGs within a DMR interval to a root population. This method provides highly interpretable DMR expansion state classifications: DMRs are considered as seeded in a given population if at least one CpG exhibits substantial methylation level differences (set to 30% in this study) from the HSC root population. They are classified as intermediate or completely expanded if additional CpGs exhibit such a methylation level shift at later points of differentiation. However, this method is susceptible to sampling noise, as the coverage at individual CpGs may be small, and no information sharing across multiple cell populations is performed.

I have therefore developed an alternative method that requires the construction of a dual-layer DMR/DMCpG atlas (section 3.6.4) as a prerequisite step. This method classifies the expansion state dynamics of a set of DMRs viewed across multiple populations by inspecting the DMCpG programming patterns occurring in each DMR. Because this method leverages DMCpG clustering, it can share information across populations to robustly identify the programming pattern of each DMCpG within a given DMR. The method is thus more robust and can provide more specific DMR expansion state classifications, pinpointing seeded and intermediate DMR expansion states more precisely. However, a comprehensive comparison of both methods in this study suggested that the results of both methods are largely comparable. Therefore, the methylation level-based first method can be recommended for projects without the resources to perform an in-depth DMCpG-level clustering analysis of large WGBS datasets, which may require clustering across 100,000s of CpGs. My work addresses the need for computational approaches to DMR expansion state tracking by introducing simple and advanced methods that can track DMR expansion state dynamics across multiple bulk populations. Of note, these methods could easily be adapted to track DMR expansion states across single-cell methylome clusters by treating the single-cell methylome clusters as pseudo-bulks and adapting the threshold and resolution parameters of the algorithms to the coverage and cell number of the single-cell dataset.

A critical difference sets both proposed methods apart from previous attempts at mapping DMR or canyon expansion. The quantification of DMR expansion by tracking the outermost regulated CpGs of a DMR locus, as previously applied [119], neglects the potential accumulation of newly regulated CpGs within DMR regions. Consequently, I propose to track the DMR size as the count of regulated CpGs independent of their location within the DMR instead of considering only the absolute extent of the DMR. It should be noted that similar to previous attempts, the proposed methods still rely on a methylation level threshold at one stage of the algorithm to distinguish between regulated and unregulated CpG states. This thresholding could, in principle, lead to the artificial detection of progressive DMR expansion across bulk populations in specific scenarios. One such scenario might involve the progressive enrichment (across a series of populations) of a cell type characterized by fully-formed DMRs which exhibit sequence- or genomic context-dependent varying susceptibility to DNA demethylation across the CpGs within the DMRs, as might occur at the boundaries of certain genomic elements such as CpG shores. However, my analyses have demonstrated that DMR expansion can be systematically explained as the effect of subsequent programming of different DMCpGs within a DMR through distinct DMCpG programming modules (further discussed in section 3.5.5). Each module was identified through unsupervised CpG clustering analysis and represented thousands of co-regulated DMCpGs exhibiting recurring programming patterns across many DMRs distributed across the genome. Furthermore, the presence of DMCpGs belonging to different DMCpG clusters within individual DMR regions could be associated with distinct binding sites for different known hematopoietic transcription factors. Taken together, robust evidence indicated that the proposed methods reveal true dynamic DMR expansion during hematopoietic differentiation. This provides a promising case study, and further applications of these methods in other systems could thus be highly interesting.

3.5.3 Neglecting to align for asymmetric DMR expansion causes artificial DMR symmetry and obscures DMR expansion dynamics

As reported by Hodges et al. [119], DMR expansion was observed to often happen asymmetrically in this study, i.e., predominantly in one direction from the DMR seed. This resulted in asymmetrically shaped DMR methylation level profiles when DMRs were viewed in individual bulk populations. This was in contrast to many other studies reporting symmetric U-shaped DMR methylation level profiles which appeared to exhibit relatively constant DMR sizes across samples [188, 189]. Notably, in the data of this study, the expansion direction of the asymmetrically expanding DMRs occurred with roughly equal frequency on the plus and minus strands, resulting in directional, asymmetric DMR profiles in the bulk populations. If average DMR methylation level profiles had been calculated by aggregating over all DMRs without accounting for the occurrence of opposed DMR expansion directions in roughly equal parts, the average DMR methylation level profiles would have been symmetrically U-shaped, and the strongly asymmetric DMR expansion would have been largely averaged out. The asymmetric DMR expansion visible in Figure 16A,B is the result of a novel method for aligning DMRs by their DMR expansion direction, which was applied during the visualization of the data (Methods, section 4.6.2). Consequently, while the systems studied in previous work may have been characterized by symmetric, U-shaped DMR profiles, these U-shaped curves could theoretically also be an artifact of a failure to align asymmetrically shaped DMR methylation level profiles prior to profile averaging.

Of note, even within the study of Hodges et al. [119], both asymmetric and U-shaped DMR profiles were reported. The asymmetric DMR profiles were shown for promoter DMRs, where the DMR expansion direction was explicitly determined for each DMR. The seemingly symmetric DMR profiles were shown for enhancer DMRs, where DMR expansion direction was not tracked. Taken together, I have developed innovative methods for visualizing the expansion of DMR regions which account for the possibility that DMRs may exhibit asymmetric methylation level profiles. Preceding studies that did not account for this possibility may have reported symmetric DMR profiles with reduced visibility of DMR expansion because the overlay of DMRs expanding in different directions obscured the true DMR profiles.

3.5.4 Progressive DMR expansion is a common mechanism of DMR programming during hematopoietic differentiation

I applied the newly developed DMR expansion state classification methods to track DMR expansion dynamics across the hematopoietic system. The resulting map of DMR expansion states of all individual DMRs across all populations provided a valuable, novel layer extending the hematopoietic DMR/DMCpG atlas. Expansion state statistics represent a novel DMR-level statistic, which takes the systematic heterogeneity of DMCpG programming within DMRs into account. Such statistics complement the conventional approach of capturing DMR methylation states via the mean methylation level of the DMRs. The comprehensive DMR expansion state map revealed that all LOM and GOM DMR clusters exhibited a large fraction of progressively expanding DMRs. Progressive DMR expansion often involved multiple distinguishable expanding steps within a single DMR. For each DMR cluster, progressive DMR expansion generally occurred across the progenitors and siblings of the marked populations of the DMR cluster. In addition, for many DMR clusters, some populations with no close relation to the marked populations of the DMR clusters also exhibited partial DMR expansion. In summary, using newly developed computational strategies, my work demonstrated for the first time that progressive DMR expansion is a common mechanism of DMR programming during hematopoietic differentiation.

Progressive DMR expansion was identified as a fundamental mechanism underlying the progressive DMR hypomethylation characterizing all LOM DMR clusters. This finding signifies that intermediate DMR methylation levels, observed to arise in the context of progressive DMR hypomethylation in all LOM DMR clusters, predominantly represent partially expanded DMR states. This entails that DMRs exhibiting an intermediate DMR methylation level in a hematopoietic cell type are likely in a functionally defined, intermediate programming state characterized by hypomethylation of specific DMCpGs within the DMR, while others have not been programmed. In summary, the widespread occurrence of progressively decreasing DMR methylation levels during hematopoietic differentiation was largely explained through DMR expansion, suggesting that intermediate DMR methylation levels represent DMRs in an intermediate expansion state, in which some DMCpGs in the DMR are already strongly hypomethylated, while others remain strongly methylated.

3.5.5 Progressive DMR expansion is the result of systematic, heterogeneous programming of individual DMCpGs within DMRs

Prompted by the observation of ubiquitous DMR expansion as an apparent mechanism of progressive DMR programming, I have characterized systematic patterns of heterogeneous

DMCpG programming within DMRs. To my knowledge, my work provides the first systematic genome-wide analysis of heterogeneous DMCpG programming within DMRs. Heterogeneous DMCpG programming within DMR regions was observed across most DMRs, with 72% (88,493) of all DMRs exhibiting more than one DMCpG-PPs. The different DMCpG-PPs observed within individual DMRs did not represent random patterns arising independently at separate DMCpG sites. Instead, DMCpG clustering analysis revealed 30 distinct DMCpG clusters, which captured a limited set of characteristic DMCpG-PPs that each recurred across thousands of DMCpGs throughout the genome. These DMCpG clusters identified systematic co-regulation of non-adjacent DMCpGs residing in large numbers of DMRs distributed across the genome. The DMCpG clusters captured a hierarchy of multi-lineage-, lineage-, and population-specific LOM programming patterns, as well as one pan-hematopoietic and one lymphoid-specific GOM DMCpG cluster. The DMCpG programming patterns directly mirrored the programming patterns observed for the DMR clusters. Each DMR cluster was associated with a specific, limited set of DMCpG-PPs from which all of the DMRs in the DMR cluster were made up. These findings further indicated that the DMCpG clusters captured robustly identified DNAme remodeling modules. These findings provided robust evidence that DMRs often contained different DMCpGs programmed by clearly distinct DMCpG programming modules.

DMCpG programming within individual DMRs followed a globally recurring pattern: DMRs generally contained a series of (typically two to four) DMCpG-PPs exhibiting a decreasing extent of programming in progenitor populations coupled with an increasing extent of population specificity. This global pattern of DMCpG programming could be identified as the systematic mechanism underlying the progressive expansion of DMRs during differentiation.

These findings suggest a model of progressive DMR programming through successive programming of individual DMCpGs within DMRs. The model posits that the programming of many lineage and cell type specification DMRs begins in progenitor populations before complete fate restriction. This initial DMR programming is confined to a specific, non-random sub-region in each DMR, termed the seed region. Progenitor cells with hypomethylated seeds in DMRs associated with a particular lineage or cell type can continue to differentiate along this trajectory. This involves accumulating further hypomethylation in fate-associated DMRs through the subsequent programming of additional DMCpGs during differentiation. Alternatively, cells may differentiate toward another fate, in which case the partial DMR hypomethylation associated with abandoned fates is at least partially retained as epigenetic memory throughout differentiation.

This model is directly supported by the observed mechanism of combining multiple distinct DMCpG programming patterns within DMRs. The DMCpG programming pattern with the strongest extent of programming in progenitor populations, always coupled with the broadest degree of hypomethylation across mature populations, represents the early DMR seeding event. These DMCpGs are strongly hypomethylated in one or more progenitor populations, indicating a role in DMR seeding in progenitor cells, and broadly hypomethylated across multiple mature cell types, suggesting potential epigenetic memory of abandoned fate exploration. Progressive DMR expansion steps during progressive fate restriction are captured in DMCpG-PPs with increasingly less programming extent in early progenitor populations and increasingly more specific restriction of hypomethylation to particular mature populations.

3.5.6 DMR programming starts in small seed regions, demanding refined data analysis strategies

The seed regions where DMR programming is initiated are predominantly small, often comprising only one or two DMCpGs. This finding carries significant implications for the interpretation of the mean DMR methylation level as a measure of the regulatory state of DMRs. In particular, the mean DMR methylation level can mask information from less common seeding DMCpGs while emphasizing population-specific, later-stage DMCpG programming. Consequently, a DMR exhibiting strong seed hypomethylation but lacking further programming in a population might appear unprogrammed when assessed solely through mean DMR methylation. This happens because the seed hypomethylation is averaged out by the more numerous, still methylated DMCpGs within the DMR. This issue is especially relevant given the widespread partial DMR programming observed in MPP2-4 populations. In the LOM lineage specification clusters, this programming was primarily confined to small subregions of the DMRs. Thus, when analyzing DMR programming in MPP populations using only mean DMR methylation levels, the extensive initial programming of lineage specification programs in early progenitor cells may be partially obscured.

Moreover, the pronounced spatial restriction of programmed DMR subregions in early progenitor populations implies that a significant portion of the methylome remodeling taking place in these cells might be missed if researchers only search for differential methylation in the form of DMRs between early progenitor cell types and the HSC population, because the programmed seed regions are often smaller than the conventional threshold of at least 3 DMCpGs used to define a DMR region. In summary, conventional approaches for DMR calling and quantifying a DMR's regulatory state through its mean methylation level may not adequately capture the initial seeding of DMRs, highlighting the need for advanced strategies to understand these processes fully. Therefore, this study leverages DMCpG-resolved analysis within each DMR to reliably detect DMR seeding.

3.6 From a regional to a hierarchical model of DNA methylation programming

3.6.1 The classical model of regional DNA methylation programming

The DNAme layer is distinct from chromatin accessibility and histone modification layers in its unique ability to encode information at near nucleotide resolution. Despite this, the DNAme layer is conventionally treated as a regional epigenetic layer, which is assumed to have a similar region-based structure as the chromatin accessibility and histone modification layers. This perspective has dominated the field for the past decade [190, 191]. I, therefore, refer to it as the ”classical model” of DNAme programming in the following. According to the classical model, the primary units of DNAme programming are genomic intervals containing multiple CpGs whose methylation state is relatively homogeneously regulated. The CpGs within such genomic intervals act collectively as one regulatory or information-encoding unit. These intervals are conventionally defined to have a particular minimal size in both base pairs and the number of contained DMCpGs. DNAme changes that fall below these thresholds are considered mostly spurious alterations lacking significant information content or regulatory function.

The classical model of DNAme programming is supported by the success of region-centric studies in elucidating key roles of DNAme in the context of differentiation and disease. Furthermore, a strong global autocorrelation between adjacent CpGs is often cited as justification for assuming a regional nature of DNAme programming [192, 193]. As a result of this perspective, whole-genome bisulfite sequencing (WGBS) studies of DNAme commonly aggregate information across regions, using aggregate region-level statistics (e.g., mean methylation levels in DMRs) for machine learning and data exploration. Various definitions of regional units of DNAme programming have been proposed, such as i) DMRs, contiguous genomic regions exhibiting statistically significant differences in methylation levels between two or more biological conditions) [190]; ii) blocks of co-regulated CpGs identified through change point detection in multi-group comparisons [129]; iii) domains characterized by a predominant methylation state [194], such as methylation canyons [186] [194]; and iv) a priori defined genomic regions like promoters or enhancers [93].

The high concordance of the information content among adjacent CpGs is conventionally assumed to be so strong that many DMR calling and segmentation algorithms perform smoothing, reinforcing autocorrelated patterns along the genome and filtering out non-concordant methylation states across adjacent CpGs. This assumption is based on the idea that non-concordant changes between adjacent CpGs often result from sampling noise [82, 87]. This rationale also underlies the widespread use of horizontal information sharing for imputing sparse methylomes, as exemplified by DeepCpG [97].

3.6.2 Findings challenging the classical model of regional DNA methylation programming

Several findings have challenged the classical model of DNAme programming. A key argument in favor of analyzing DNAme as a regional layer is the assumption that spatially adjacent DMCpGs exhibit highly correlated DNAme levels. This assumption is primarily based on studies using bulk population data [192, 193]. Recent technological advances in single-cell bisulfite sequencing (scBS-seq) protocols have enabled the quantification of the concordance of methylation states between adjacent CpGs in single cells. Hui et al. [75] demonstrated that the concordance of adjacent CpG methylation states in single cells rapidly decreases as the distance between CpGs increases, with concordance never exceeding 90% even for directly adjacent CpGs. It is essential to consider that, given the vast number of CpGs in the genome, a 10% rate of non-concordant directly adjacent CpGs represents a significant number of potentially distinctly programmed closely proximal CpGs. In other words, even a high global correlation across millions of CpGs still allows for a significant subset of heterogeneously programmed, directly adjacent CpGs, which challenges the classical model's assumptions about region-wise DNAme programming.

Several findings have challenged the concept that individual CpGs within methylation-dependent regulatory elements act as a single coherent unit, where all CpGs share the same function. Hodges et al. [119] demonstrated that many DMRs expand during hematopoietic differentiation, providing clear initial evidence that different CpGs within DMR regions are distinctly regulated at different stages of differentiation and are thus likely to have different functions and information content. While not systematic global screens, various studies focused on well-characterized individual regulatory elements have reported similar functional heterogeneity of adjacent CpGs. Several studies have shown that closely adjacent individual DMCpGs within specific regulatory elements can have significantly different information content about gene expression [195, 196]. Furthermore, several studies of specific regulatory loci have reported highly focal DNAme changes at single CpG sites that correlated with gene expression, while adjacent sites were not differentially methylated [197] [198, 199]. These findings further support the notion that the DNAme programming of individual CpGs within regulatory regions can be distinct and heterogeneous.

Numerous genome-wide studies have established a robust and multifaceted link between transcription factor (TF) binding and DNAme. These investigations have demonstrated that the binding affinity of many TFs can be influenced by DNAme at the binding site, both in vitro [55] and in vivo [56, 200]. Moreover, TFs can, in turn, cause DNA demethylation at their binding sites, either directly or indirectly. For example, pioneering TFs, including hematopoietic master factors such as Runx1, can induce active DNA demethylation, thereby establishing a permissive environment for the subsequent binding of other TFs [49, 58–60, 201]. Regulatory elements typically contain distinct binding sites for different transcription factors, which cooperate to regulate the regulatory element in a combinatorial fashion [202]. This suggests that interactions between TFs and DNAme are not uniform across DMR regions. Instead, binding of different TFs at specific sites within a DMR is likely to be associated with distinct DNAme changes, implying a frequent role for sub-DMR-resolved DNAme programming. In summary, the focal nature of TF-DNAme interactions suggests that different TFBS-associated CpGs within a DMR are likely to have distinct functions and information content.

Several genome-wide studies investigating the information content of the DNAme layer at sub-DMR resolution have shown promising results. For example, a genome-wide analysis of cell type-specific epiallele patterns within highly focal genomic loci has provided initial evidence that closely adjacent CpGs may be programmed in a cell type-specific manner [203]. As another example, Schlosberg et al. have gathered initial evidence that capturing DNAme programming around transcription start sites through the average methylation levels of TSS-proximal DMRs may fail to incorporate the complexity of highly resolved DNAme programming occurring around the transcription start sites [185, 204, 205]. They demonstrated that capturing DNAme programming through highly resolved methylation signatures performs better than DMR-level approaches at gene expression prediction tasks. These findings suggest that mapping DNAme programming at sub-DMR resolution could be necessary to model the regulatory information encoded into the DNA methylome accurately.

3.6.3 A novel, hierarchical model of DNA methylation programming

This study systematically demonstrates that progressive, large-scale methylome remodeling during hematopoietic differentiation occurs through the non-random accumulation of meaningful methylation changes within DMR regions over the course of differentiation. This process involves distinct subsequent DMCpG programming steps at different DMCpGs within the DMRs, leading to widespread progressive DMR expansion during hematopoietic differentiation.

The programming of individual DMRs is complex and typically involves programming through multiple functionally distinct DMCpG programming modules. Nevertheless, a finite set of only 28 characteristic, progressive DMR programming patterns emerges as a result of these combinations. These DMR programming patterns represent hierarchically related multi-lineage, lineage-specific, and cell-type-specific DNA methylome remodeling programs guiding the establishment and progressive programming of DMR regions during differentiation.

DMCpG programming within DMRs is a ubiquitous mechanism, with 30 distinct programming modules identified. Each module controls thousands of co-regulated DMCpGs residing in numerous DMRs across the genome. These programming modules are associated with the subsequent activity of distinct transcription factors at separate bindings sites within individual DMRs over the course of differentiation.

These findings indicate that the DNAme layer possesses complex information encoding capabilities, which are missed when DMRs are assumed to be the atomic unit of DNAme programming, challenging the classical model of the DNAme layer as a regional layer. Instead, my work provides systematic evidence supporting a novel model of hierarchical DNAme programming, with the single cytosine as its lowest resolution level. In this model, DMRs and DMCpGs represent hierarchically organized units of DNAme programming, with DMRs acting as the atomic unit of DNAme-based gene regulation and CpGs within DMRs serving as the atomic unit capable of effecting methylation-dependent DMR activity modulation: in this model, DMCpGs thus act as integrative switches within DMRs, integrating signals from different transcription factors, which may occur simultaneously or over time. Consequently, methylation changes within DMR regions do not accumulate randomly. Instead, when DMRs contain sets of DMCpGs exhibiting different DMCpG-PPs, these sets of DMCpGs are likely to be systematically regulated at distinct stages of the differentiation process.

Various studies have reported evidence that differential methylation within DMRs arises in a stochastic manner due to random methylation changes at individual CpGs within DMR regions [181]. This notion that individual CpGs within DMR regions are equally likely to randomly switch their methylation state is fully compatible with the classical regional model of DNAme programming. However, a fully stochastic accumulation of DNAme changes within DMR regions is in clear contrast to the newly proposed model of hierarchical DNAme programming. The differing results could potentially be attributed to the fact that different systems were studied; whether DMRs accumulate methylation changes systematically or stochastically may depend on the system at hand. Landan et al. [181] found evidence for stochastic accumulation of methylation changes within DMRs in long-term in vitro cultures of immortalized fibroblasts and through comparing normal and cancerous tissues. It is plausible that distinct methylome remodeling mechanisms are active in these scenarios compared to those active during hematopoietic differentiation studied in this thesis. In line with this reasoning, Shipony et al. [206] found that vastly different methylome programming mechanisms underlie the maintenance of epigenetic memory in embryonic stem cells and in somatic cells, suggesting that the mechanisms of methylome remodeling can differ drastically depending on the studied cell types and conditions.

It is also conceivable that a combination of systematic and stochastic DMCpG programming is concurrently active in many scenarios. My work demonstrates that while many DMRs exhibit multiple DMCpG-PPs, one DMCpG-PP still typically characterizes multiple DMCpGs within each DMR. A stochastic programming order across these CpGs would be compatible with my findings. Given that often a single main DMCpG-PP covers a predominant majority of all DMCpGs within a DMR, a large subregion of such DMRs could indeed accumulate methylation changes in a partially stochastic manner within the differentiation stage where these CpGs characterized by the main DMCpG-PP of the DMR can be expected to be programmed. In conclusion, while the proposed model of hierarchical DNAme programming posits that groups of DMCpGs within individual DMR regions can be expected to be programmed by distinct DMCpG programming modules, often at subsequent stages of differentiation, the order of programming within each group of DMCpGs may be partially stochastic.

Findings of strong global autocorrelation of CpG methylation levels across the methylome [192, 193], as well as findings that the methylome may contain many blocks of adjacent CpGs with tightly coupled methylation states [129, 207], are not in contradiction to the proposed hierarchical model of DNAme programming. As discussed in section 3.6.2, due to the vast number of CpGs in the murine and human genomes, even a strong autocorrelation between a predominant fraction of these CpGs is compatible with the existence of many genomic sites where adjacent CpGs carry non-concordant methylation levels. Indeed, initial experimental evidence from new single-cell studies suggests a high, but not complete, concordance between adjacent CpGs (section 3.6.2). Furthermore, even within a hierarchically programmed DMR, many DMCpGs may share the same DMCpG-PP, and thus exhibit strongly coupled methylation states, while other DMCpGs within the same DMR are subject to independent DMCpG programming. These considerations indicate that strong global autocorrelation and tightly coupled methylation states in many regions of the methylome do not contradict the assumptions of a hierarchical model of DNAme programming.

3.6.4 A new paradigm for the analysis of DNA methylation data

In this work, I introduce the systematic analysis of hierarchical DNAme programming as a novel paradigm for the analysis of DNAme data. This approach involves mapping differential methylation hierarchically at the level of DMRs and of the individual DMCpGs contained within these DMRs. To categorize the programming patterns exhibited by each DMR and DMCpG, clustering analysis is applied to group co-regulated DMRs and DMCpGs. The cluster membership of each DMR and DMCpG associates each with a specific regulatory pattern. This allows viewing the programming of DMRs as a combinatorial process involving multiple distinct DMCpG programming patterns while maintaining information about the number of DMCpGs within a DMR that exhibit each DMCpG-PP.

The necessity for mapping DNAme programming through integrated DMR and DMCpG analysis is based on a compelling rationale. My work has provided ample evidence that DMR programming patterns cannot be fully understood without accounting for the heterogeneous programming of individual DMCpGs within DMRs. For example, I found that progressive DMR expansion is a common mechanism of DMR programming during hematopoietic differentiation, which involves heterogeneous programming of individual DMCpGs at subsequent stages of differentiation. At the same time, my work has provided compelling evidence that the programming pattern of an individual DMCpG cannot be interpreted without the context of its surrounding DMCpGs. For instance, co-regulated DMCpGs characterized by the *p1* DMCpG-PP can - concurrently, in a single-cell - be involved in the seeding of myeloid- and erythroid/eosinophil-specific DMRs and the *p2* DMCpG-PP can be concurrently involved in the seeding of lymphoid- and dendritic cell-specific DMRs.

By leveraging the information about regulatory patterns exhibited by DMRs, methylation-dependent cis-regulatory elements can be identified and linked to target genes and biological functions. The regulatory state of individual DMRs in a particular population and the mechanisms controlling this state (such as specific transcription factors) can be quantified and investigated by analyzing the DMCpG-PPs of the individual DMCpGs within the DMR. This novel approach towards the analysis of DNAme data was the essential foundation for all mechanistic and biological findings concerning the function and information content of DNAme programming during hematopoietic differentiation achieved in this thesis.

It is, however, important to note the limitations of the proposed approach for hierarchical DNAme analysis through integrated DMR and DMCpG clustering analysis. This approach relies on discretizing a complex landscape comprising gradually differing DMR and DMCpG programming patterns. Within this framework, each DMR and DMCpG is assigned the characteristic programming pattern of its respective DMR or DMCpG cluster. However, the programming patterns of individual DMRs or DMCpGs within a cluster are not identical but cover a range of variations around the characteristic mean programming pattern of the cluster. In particular, the clustering analysis was intentionally designed to allow free variability of the extent of programming of progenitor populations in each cluster. The membership of two DMCpGs in the same DMCpG cluster only indicates that they share characteristic properties regarding their regulatory patterns; it does not imply perfect and synchronous co-regulation. In some cases, DMCpGs at the peripheries of two similar DMCpG clusters may be more closely co-regulated with each other than with the medoid DMCpG of their respective clusters. The annotation of different DMCpG-PPs for individual DMCpGs within a DMR allows distinguishing likely heterogeneously regulated DMCpGs within a single DMR. However, this discrimination should not be expected to be perfect for two reasons. First, DMCpGs grouped under the same DMCpG-PP may still be heterogeneously regulated. Second, DMCpGs grouped under different DMCpG-PPs could, in a non-negligible fraction of cases, actually be relatively homogeneously regulated if the DMCpGs are located at the periphery of their respective DMCpG clusters.

In conclusion, the proposed clustering-based dual-layer analysis of DNAme programming offers a discretization-based aggregation of a complex range of programming patterns observed during hematopoiesis. This approach has proven highly useful for many tasks related to elucidating general mechanisms with genome-wide activity and the overall information content of DNAme landscapes concerning hematopoietic fate restriction states. However, developing more targeted and quantitative approaches is necessary for the precise analysis of individual loci of high interest.

Chapter 4

Methods

4.1 Genome-wide DNA methylation profiling using T-WGBS for bulk populations and scBS-seq for single cells

4.1.1 Isolation and T-WGBS of 25 hematopoietic bulk populations

All wet lab experiments were performed by members of the Section Translational Cancer Epigenomics (DKFZ, Division Translational Medical Oncology) led by Daniel Lipka, with support from Dr. Dieter Weichenhan (Division Cancer Epigenomics), Dr. Ruzhica Bogeska & Julia Knoch (Division Experimental Hematology) and Dr. Melinda Czeh (Uni Münster). The isolation and T-WGBS-based DNAme profiling of the surface marker-defined hematopoietic populations is detailed in the doctoral thesis of Sina Stäble [131], who performed a large part of this experimental work as part of her doctoral project.

A brief summary of the experimental data generation is provided in the following. Bone marrow cells were isolated from the femora, tibiae, hips, and spines of sacrificed mice. Lineage-negative bone marrow cells were enriched using the following antibody cocktail: CD5, CD45R, CD11b, CD8a, Ly-6G, Ly-6C, and Ter119. Monocytes, neutrophils, B cells, and T cells were sorted from total bone marrow cells. The following populations were sorted from lineage-negative bone marrow cells: HSC, MPP1, MPP2, MPP3, MPP4, MPP5, GMP, MEP, CMP CD55^{+} , CMP CD55^{-} , CLP, preMegE, MkP, CFU-E, MDP, CDP, and cMoP. The pDC, cDC1, and cDC2 populations were collected from the spleen. The doctoral thesis of Sina Stäble [131] describes the surface marker definitions chosen for the different hematopoietic populations in detail and presents representative sorting schemes for all populations. A tabular overview of the chosen surface marker definitions and references for these definitions are provided in Tables S2 and S3.

Tagmentation-based whole-genome bisulfite sequencing (T-WGBS) was performed as described in [68], using 10 ng to 30 ng of DNA. Paired-end sequencing was performed on the Illumina HiSeq2000 platform with a read length of 125 bp.

Sequencing was performed at and with the support of the Genomics and Proteomics Core Facility at the DKFZ. Data management for the NGS data was provided by the Omics IT and Data Management Core Facility (ODCF) at the DKFZ.

4.1.2 Alignment, methylation calling, and quality control for the T-WGBS samples

Read alignment was performed as described in [68]. Alignments were performed by the Omics IT and Data Management Core Facility (ODCF) at the DKFZ, using a workflow implementation contributed by Matthias Bieg (`https://github.com/DKFZ-ODCF/AlignmentAndQCWorkflows`). Data management was provided by the ODCF using the One Touch Pipeline (OTP) system [132]. Briefly, adaptor sequences were trimmed using Trimmomatic [208]. Because the T-WGBS method is a directional WGBS protocol, sequencing reads could be in silico bisulfite-converted as follows: cytosines were converted to thymines for the first read, and guanines were converted to adenines for the second read. The mm10 reference genome was extended with the PhiX and lambda phage sequences and also in silico bisulfite-converted. In silico bisulfite conversion was performed with methylctools (`https://github.com/hovestadt/methylCtools`). Read alignment was performed with the BWA-MEM algorithm implemented in the BWA package [209]. The algorithm was used with default parameters to align the converted reads to the in silico bisulfite-converted reference genome. After alignment, reads were converted back to their original state. PCR duplicate removal was performed per library using the MarkDuplicates algorithm from the Picard toolbox [210]. For each replicate, the per-library alignments were merged using the samtools merge algorithm [211]. Alignment quality control was performed based on mapping rates (computed using samtools flagstats [211]), insert size distributions, and genome coverage statistics (computed using custom scripts).

To account for the strong and variable presence of M-bias in T-WGBS data, I have developed the bistro software package [SOFT1]. Methylation calling and M-bias trimming were performed with bistro (version 0.2), using the binomp algorithm for automatic M-bias removal. Bistro intelligently removes the gap repair nucleotides on both reads, introduced by the tagmentation reaction. Bistro takes the fragment length into account and removes the first nine base pairs following sequencing primer two if they are present in the sequencing read. Bistro further automatically detects and removes additional read positions exhibiting

M-bias while accounting for fragment-length dependent differences of M-bias. The M-bias profile was fitted individually per sample. Alignment quality filtering was applied with a mapping quality threshold of at least 25, and read positions were filtered with a Phred score threshold of at least 25. All samples showed very high conversion rates, estimated based on the autosomal CH conversion rate (99.43 ± 0.21%, mean ± s.d.), as detailed in Table S5.

4.1.3 Isolation and scBS-seq of LSK cells, LSK CD150$^+$ cells, and cells of different mature cell types

All scBS-seq wet-lab experiments were carried out by members of the Section Translational Cancer Epigenomics (Mark Hartmann, Sina Stäble, and Maximilian Schönung), with support from Dr. Dieter Weichenhan (Div. Cancer Epigenomics), Julia Knoch (Div. Experimental Hematology), and the Single-cell Open Lab facility at the German Cancer Research Center. Surface marker definitions and gating strategies were applied as described in section 4.1.1. Bone marrow cells were isolated from the femora, tibiae, hips, and spines of sacrificed mice. The HSC, LSK, LSK CD150$^+$, and CFU-E populations were sorted from lineage-negative bone marrow cells. Monocytes, B cells, and T cells were sorted from total bone marrow cells. Lineage-negative bone marrow cells were enriched using the following antibody cocktail: CD5, CD45R, CD11b, CD8a, Ly-6G, Ly-6C, and Ter119. Index sort information was recorded for all LSK and LSK CD150$^+$ single cells and used for in silico gating to annotate each cell as either HSC or MPP1-5 as described in section 4.1.1. The in silico gating was performed by Sina Stäble. Single-cell bisulfite sequencing (scBS-seq) was applied as described in [74, 76]. Paired-end sequencing was performed on the Illumina HiSeq X platform with a read length of 150 bp. Sequencing was performed at and with the support of the Genomics and Proteomics Core Facility at the DKFZ. Data management for the NGS data was provided by the Omics IT and Data Management Core Facility (ODCF) at the DKFZ.

4.1.4 Alignment, methylation calling, and quality control for the scBS-seq samples

I was responsible for the entire bioinformatical processing of the generated NGS data. To perform alignments, methylation calling, and quality control for the scBS-seq samples, I developed a comprehensive snakemake workflow, which I have published as an open-source package [SOFT5]. Adapter and quality trimming was performed with Cutadapt [212] using the following options: `-u 6 -U 6 --minimum-length 30 --max-n 0.3 -q 10` Read alignment was performed with Bismark [170]. To deal with chimeric reads, read pairs were first aligned through paired-end alignment, and unmapped read pairs were then subjected to single-end alignments to rescue the mappable read portions from chimeric reads. Methylation calling was subsequently performed with MethylDackel [171], using the following options -ignoreFlags 3840 -requireFlags 0 -q 30 -p 20 . PCR duplicates were marked using the MarkDuplicates algorithm from the Picard toolbox [210]. Read quality control was performed with FASTQC [213]. Alignment quality control was performed based on mapping rates (computed using samtools stats and samtools flagstats [211]), PCR duplicate rates, and insert size distributions.

Single-cell quality control was performed based on methylation calling and genomic coverage statistics computed with custom scripts. The scBS-seq protocol is plate-based, and the achieved coverage varied significantly between plates. Therefore, single-cell coverage filtering was performed for each plate individually. To exclude empty wells, a minimum number of C and CG motif methylation calls was defined for each plate based on the histogram of the C and CG coverage distribution. An upper threshold for the allowed C and CG coverage was also defined for each plate based on the overall coverage distribution to exclude multi-cell wells. The average methylation levels across all cytosines in a CG or CHH motif context were sufficiently comparable between plates to allow for consistent filtering thresholds. Cells with a CHH methylation level above 2% were excluded to exclude under-converted cells. Cells with a CHH methylation level below 60% were excluded to remove over-converted cells. A few cytosines exhibited methylation levels between 0 and 1 in each cell. These cytosines were excluded from analysis as proposed previously [75].

4.2 Construction of a genome-wide dual-layer DMR/DMCpG atlas

4.2.1 Integrated DMCpG and DMR calling with FDR control at the DMCpG level

Differentially methylated CpG (DMCpG) and differentially methylated region (DMR) calling was performed using R 3.4.1, and the Bioconductor DSS (version 2.26.0) and BSSeq (version 1.14) packages [82].

Pairwise DMCpG tests were performed with the `callDML` function, with argument `delta=0.1` to perform the tests using the posterior probability that the difference of the group means was greater than 10%. For each CpG, the set of null hypotheses $H_{hsc\_vs\_pop1}, \dots, H_{hsc\_vs\_popN}$ for equal group means between the HSC population and all other populations were tested. For each CpG, the resulting p-values were then combined to test the global null hypothesis that all null hypotheses $H_{hsc\_vs\_pop1}, \dots, H_{hsc\_vs\_popN}$ were true, versus the alternative that at least one of individual null hypotheses was false. To deal with the complex dependency structure of the individual p-values conservatively, the Bonferroni test was chosen for this purpose. I used the two-stage step-up method of Benjamini, Krieger, and Yekutieli [136] (BKY method) to control the FDR for the global null hypothesis tests at the level of 1%. CpGs for which the global null hypothesis could be rejected are referred to as hematopoietic DMCpGs.

DMR regions were called using the `callDMR` function from the DSS package, again in pairwise comparisons of the HSC population against all other populations. The DSS DMCpG test results were passed to `callDMR` without any multiple testing correction, as proposed by the DSS authors [82]. DMRs were called with a minimum size of 50 bp, a minimum number of three CpGs with a p-value below 0.01, and at least 50% CpGs in a DMR with a p-value below 0.01. Smoothing was not used, because it appeared to blur the generally sharp DMR boundaries observed in our data, leading to artificially extended DMR intervals. Next, the DMR intervals from all pairwise comparisons were merged to obtain candidate hematopoietic DMR regions.

To filter for DMCpGs with high methylation level shifts relative to the HSC population, only methylation level shifts observed in significant pairwise HSC vs. other comparisons were considered. To identify for each DMCpG the significant pairwise HSC vs. other comparisons, I performed multiple testing correction on the pairwise DMCpG calls from the individual HSC versus other comparisons, using the BKY two-stage procedure. When the FDR for the autosome-wide pairwise DMCpG calls was controlled at 1% for each HSC-versus-other comparison, few pairwise DMCpGs were found to be significant for the early progenitor populations, for which the expected number of true DMCpGs is small compared to the number of CpG sites [14, 122], while the extent of differential CpG methylation is expected to progressively increase from progenitor to mature populations [93, 118]. I, therefore, performed the pairwise DMCpG multiple testing correction while controlling the FDR at i) 5% for MPP1, MPP5, and MPP2 to trade off precision for sensitivity; ii) 1% for MPP3, MPP4, CMP CD55$^+$, preMegE, MkP and CMP CD55$^-$; and iii) 0.5% for all other populations, to limit the expected absolute number of false discoveries in the more differentiated populations, for which many true DMCpGs are expected. This resulted in a strong increase in detected DMCpGs for the early progenitor populations while providing high precision for the populations exhibiting large-scale methylome changes.

DMRs were filtered for methylation level shifts of at least 30% between the HSC population and at least one downstream population. Only statistically significant methylation level shifts were considered. Because the DMR calling strategy was heuristic and did not provide FDR control, the significance of a DMR methylation level shift in a given population was determined based on the presence of (BKY method-corrected, as detailed above) pairwise DMCpGs. A DMR was considered to be differentially methylated in a given population P compared to the HSC population if there was at least one pairwise HSC-versus-P DMCpG in the DMR. The methylation level shift was then computed based on the average DNAme levels across all DMCpGs within the DMR.

4.2.2 Calculation of DMCpG and DMR methylation levels

Methylation calls from all replicates of the same population were pooled. Methylation calls for the two cytosines in each CpG motif were merged. DMR methylation levels were computed as the mean methylation level across all DMCpGs within the DMR, excluding other CpGs within the DMR which were not identified as DMCpGs, because they showed no significant differential methylation and/or only small methylation level shifts across the hematopoietic system.

4.2.3 Annotation of genomic regions and potential target genes

Gene and genomic region annotation for the hematopoietic DMRs was performed with the gtfanno Python package (version 0.2.0) developed by me in the context of my thesis [SOFT4]. For this study, the gtfanno algorithm was parametrized such that for each DMR, residence in the following genomic location classes was queried in this order of precedence: i) promoter (5000 bp upstream to 1000 bp downstream of a transcription start site (TSS)); ii) 5'-untranslated region (5'-UTR) or 3'-UTR; iii) intron or exon; iv) candidate CREs (cCRE), within ±50 kb of a TSS but outside the promoter or gene body regions; and v) intergenic, if no annotation based on the preceding genomic location classes was possible. If a DMR was associated with multiple genes, only genes for which the DMR was located in the genomic region class with the highest precedence were considered. Within that class, all possible gene annotations were kept. For example, a DMR may have resided in the promoter region of two genes and the intron of a third gene. In this case, both promoter-based gene annotations were kept, but the intron-based gene annotation was discarded. In scenarios where a single gene annotation was required per DMR, DMR-to-gene annotations were further ranked based on the distance of the DMR center to the TSSs of the annotated genes.

4.2.4 Control of replicate homogeneity

To assess the similarity of replicates from the same population, the average methylation levels of the multi-cell tracks from the Ensembl Regulatory build (version 20161111, published in Ensembl release 91) [133] were used. Specifically, multi-cell enhancer, promoter, open chromatin, promoter flanking, and TFBS tracks were used. These tracks are enriched for genomic loci where robust methylation differences between different cell types and homogeneous methylation levels between cells of the same cell types can be expected. A GFF file mapping these genomic regions was obtained from `ftp://ftp.ensembl.org/pub/release-91/regulation/mus_musculus/mus_musculus.GRCm38.Regulatory_Build.regulatory_features.20161111.gff.gz`. Each replicate was characterized by a vector containing the average methylation levels across all of the individual regions gathered across all tracks. Distances between replicates were calculated as the Euclidean distance between their vectors.

4.3 Clustering analysis and annotation of DMR and DMCpG programming patterns

4.3.1 DMR and DMCpG clustering analysis

The aim of the DMR and DMCpG clustering analyses was to identify sets of co-regulated DMRs or DMCpGs with a well-defined relationship to distinct hematopoietic cell types. The DMR and DMCpG clusters were intended to serve as reference region sets for single-cell analyses of DNAme changes during hematopoietic differentiation, with a focus on early progenitor cells. To ensure that the clusters were not biased by the surface marker-defined progenitor populations in our data set, they were constructed using only data from the mature populations. These mature populations represent relatively homogeneous endpoints of differentiation in the hematopoietic system. Thus, differing methylation levels between these populations are unambiguously associated with functionally different cell types. In contrast, methylation level differences between potentially heterogeneous progenitor populations may be caused by convoluted shifts in cell type composition, making functional annotation challenging.

Clustering was performed using the Leiden community detection algorithm [143] on a weighted kNN graph based on the DMR or DMCpG methylation levels (section 4.2.2) of the nine mature populations in the data set (the CFU-E, monocyte, eosinophil, neutrophil, B cell, T Cell, cDC1, cDC2 and pDC populations). Specifically, I first computed a kNN graph using the `neighbors` preprocessing function from the ScanPy package [145], using the correlation distance and 15 neighbors. This function uses the nearest neighbor search algorithm provided by the UMAP reference implementation [214]. The scale invariance of the correlation distance was unproblematic for this data set, because all DMRs and all DMCpGs exhibited large methylation level shifts during hematopoiesis (all DMRs and all DMCpGs exhibited a methylation level shift compared to the HSC population of at least 30%). Next, I computed edge weights for the kNN graph based on the fuzzy simplicial set associated with the data, using the `fuzzy_simplicial_set` function from the UMAP package. Finally, I computed Leiden clustering using the `find_partition` function from the leidenalg Python package, with the `RBConfigurationVertexPartition` quality function, which implements Reichardt and Bornholdt's Potts model with a configuration null model [215].

The clustering resolution was selected by screening resolutions from 0.6 to 2.1 in steps of 0.1 for the DMR clustering analysis and from 0.8 to 1.5 in steps of 0.05 for the DMCpG clustering analysis, visually inspecting cluster homogeneity, and choosing the resolution that provided a good tradeoff between a comprehensible number of clusters and relatively high homogeneity of the programming patterns within the clusters. For the DMR clustering analysis, a resolution of 1.4 was chosen, which resulted in a partitioning with 28 DMR clusters. For the DMCpG clustering analysis, a resolution of 1.35 was chosen, which resulted in the identification of 30 DMCpG clusters.

The DMCpG clustering analysis consistently identified a small number of outlier DMCpGs across a range of resolution scores. At a resolution of 1.35, these DMCpGs were grouped in two outlier clusters, containing 114 and 97 DMCpGs, respectively. The other 30 DMCpG clusters exhibited cluster sizes ranging continuously from 4624 DMCpGs to 43207 DMCpGs. The outlier clusters were thus interpreted as tiny groupings of DMCpGs with non-recurring programming patterns and excluded from further analysis.

To improve the display of the DMR or DMCpG clustering in heatmap visualizations, the DMRs or DMCpGs within each of the clusters found by the Leiden clustering were ordered using hierarchical clustering with Ward's method, and z-score normalized DMR or DMCpG methylation levels, which effectively carries out hierarchical clustering based on a correlation distance, mirroring the use of the correlation distance in the Leiden clustering.

4.3.2 Annotation of marked and regulated populations for the loss of methylation clusters

The LOM DMR and DMCpG clusters were characterized by a range of multi-lineage-, lineage- and population-specific programming patterns. As a high-level characterization of these patterns, each of these DMR and DMCpG clusters was annotated with a set of "marked" populations, defined as the set of populations which exhibited markedly strong hypomethylation in the DMRs/DMCpGs of the cluster. Additionally, as a supplementary high-level characterization, each DMCpG cluster was annotated with "regulated" populations, defined as the set of all populations exhibiting a considerable fraction of unmethylated reads in the DMCpGs of the clusters. The set of regulated populations is thus a superset of the set of marked populations, which contains both populations with intermediate hypomethylation and populations with markedly strong hypomethylation in the DMCpGs of a cluster. A detailed definition is given in the following.

Marked populations: The set of marked populations for each LOM DMR cluster was defined to contain the population with the minimum mean DMR level across all DMRs in the cluster, as well as all populations with a mean DMR methylation level within 15% of this minimum. The definition was analogous for all LOM DMCpG clusters.

Regulated populations: To compute the set of regulated populations for each DMCpG cluster, the following steps were performed (separately for each DMCpG cluster): i) for each

DMCpG in the cluster, compute the methylation level shift compared to the HSC population; ii) classify the regulation state of each DMCpG in each population as either regulated or unregulated as follows: the DMCpG is considered regulated if the methylation level shift against the HSC population is at least 30%; this threshold is lowered to 20% for the MPP1-5 populations, to account for expected high heterogeneity; iii) for each population, determine the percentage of regulated DMCpGs; iv) classify each population as regulated or unregulated in the DMCpG cluster at hand as follows: the population is considered regulated if at least 70% of the DMCpGs are regulated in that population; otherwise it is considered unregulated. The threshold is lowered to 25% for the MPP1-5 populations to increase sensitivity in these expectedly heterogeneous populations.

4.3.3 Grouping and naming of DMR and DMCpG clusters based on the population-specificity of their regulatory profiles

The nomenclature of the DMR and DMCpG clusters was used to encode the population-specificity of their programming patterns, as well as the extent to which DNAme changes were observed across the mature hematopoietic system beyond the marked populations of each DMR or DMCpG cluster.

Nomenclature of the DMR clusters

Two DMR clusters showed a distinct regulatory pattern from all other DMR clusters, with the lowest DMR methylation levels observed in the HSC population. This indicated that methylation was primarily gained in the DMRs of these DMR clusters. These DMR clusters are therefore referred to as “gain of methylation” clusters. The names of these DMR clusters start with the prefix “H” (for “HSC”).

In contrast, for all other DMR clusters, the HSC population exhibited either the highest average methylation level (22 DMR clusters) or one of the highest average methylation levels surpassed minimally (<= 0.6%) only by one or more MPP populations (three DMR clusters). In these 25 DMR clusters, DMRs predominantly lost methylation compared to the HSC population, and these clusters are therefore referred to as “loss of methylation DMR clusters”. For each of these 25 loss of methylation DMR clusters, a set of marked populations was determined, defined to contain the population with the minimum average DMR methylation level, as well as all populations within 15% of this minimum (see also section 4.3.2). Two of the DMR clusters exhibited a very broad pattern of methylation loss across the hematopoietic system, with more than six marked populations across three lineages. The names of these clusters were prefixed with “P” to indicate their nearly “pan-hematopoietic” regulatory pattern. The remaining 23 loss of methylation DMR clusters were grouped based on their lineage-specificity. Sixteen DMR clusters exhibited regulatory patterns exclusively marking populations from a single lineage. The names of these DMR clusters were prefixed according to this lineage (i.e., E for erythroid, M for myeloid, D for dendritic, and L for lymphoid). The remaining 7 DMR clusters showed regulatory patterns marking populations across multiple lineages, and therefore their names were prefixed with “C” for “cross-lineage”.

The DMR clusters within each of these seven DMR cluster groups (H, P, E, M, D, L, C) were ordered based on how broadly methylation changes were observed across the mature hematopoietic system. To capture the broadness of the methylation changes in a DMR cluster, its mean DMR methylation level in each mature population was computed, and then the mean of these values was calculated, with each population mean weighted by the number of populations in our data set from the same lineage, to normalize for differences in coverage of the different lineages. This statistic quantified the average DMR methylation level across the mature hematopoietic system for each DMR cluster. The DMR clusters within each group were then ordered according to this statistic, either ascendingly (loss of methylation clusters) or descendingly (gain of methylation clusters). Thus within each group, the DMR clusters were ordered in the order of how broadly hypomethylation (for LOM clusters) or hypermethylation (for GOM clusters) occurred throughout the mature hematopoietic system. This ordering of the DMR clusters was encoded in the cluster names by the cluster number following the (H, P, E, M, D, L, or C) prefix. For example, the M1 and M5 DMR clusters both exhibited a myeloid lineage-specific regulatory pattern, but the M1 DMR cluster showed the broadest occurrence of hypomethylation across all mature populations (marking all myeloid populations and exhibiting partial hypomethylation in the dendritic cell populations), while the M5 DMR cluster had the most population-specific regulatory pattern of all myeloid clusters (marking only the eosinophil population and exhibiting partial hypomethylation only in the neutrophil population).

Nomenclature of the DMCpG clusters

The individual DMCpG clusters exhibited programming patterns that were largely analogous to the programming patterns of the DMR clusters. Therefore, a nomenclature analogous to that used for the DMR clusters was applied. To distinguish the DMR clusters from the CpG clusters, uppercase names were used for DMR clusters (e.g., H1), while lowercase names in italic were used for the CpG clusters (e.g., *l1*).

Two of the DMCpG clusters (*h1* and *h2*) were characterized by DMCpG programming patterns associated with gain of methylation compared to the HSC population.

The remaining 28 DMCpG clusters captured DMCpG programming associated with loss of methylation compared to the HSC population. As for the DMR cluster nomenclature, the nomenclature for these LOM clusters was based on the marked populations of each DMCpG cluster. In total, 18 DMCpG clusters exclusively marked populations from a single lineage. Their cluster names were prefixed to indicate their lineage-specificity, including three DMCpG clusters specifically marking the CFU-E population (*e1*-*e3*), five DMCpG clusters specifically marking myeloid populations (*m1*-*m5*), four clusters specifically marking dendritic cell populations (*d1*-*d4*) and five clusters specifically marking lymphoid populations (*l1*-*l5*). Nine DMCpG clusters marked populations across multiple lineages (cross-lineage clusters *c1*-*c9*). Finally, two DMCpG clusters marked more than seven populations across at least three lineages (pan-hematopoietic clusters *p1* and *p2*). A hypomethylation-specificity score for each DMCpG cluster was computed analogously to the same characterization of the DMR clusters. The specificity of the regulatory patterns was then denoted in the cluster names: the order of the DMCpG clusters within each group reflects the order of their hypomethylation-specificity scores, either in ascendingly sorted order for the loss of methylation clusters or in descendingly sorted order for the gain of methylation clusters.

4.3.4 Compilation of DMR cluster target gene sets

The GENCODE [216] release M25 for GRCm38.p6/mm10 was used for gene and transcript annotations as well as for gene track visualizations. Only a filtered subset of GENCODE was considered for gene annotations, which was restricted to splice variants for protein-coding genes with strong experimental support (referred to as the "GENCODE top protein-coding transcripts set" in this thesis). To select transcripts with strong experimental support, the following transcripts were excluded: i) incomplete transcripts (missing information at the 3'- or 5'-end); ii) transcripts with transcript support level 4 or 5; and iii) transcripts tagged as "NMD_exception", "NMD_likely_if_extended", "non_canonical_TEC", "non_submitted_evidence", or "not_organism_supported". For each gene, the longest transcript with an APPRIS principal tag [217] was considered to be its principal transcript.

The target gene set for each DMR cluster was computed using the proximity-based DMR-to-gene annotations (Methods, section 4.2.3, Results section 2.2.3). The target gene set for each DMR cluster was defined as the set of all genes that had at least one DMR from the DMR cluster annotated to them. DMR-to-gene annotations with large distances between the DMR and the putative target genes were not considered, because such long-range regulatory associations can only be made with low confidence when using proximity-based DMR-to-gene annotations [51, 142]. Specifically, only DMRs located within ±15 kb of the TSS of the gene they were annotated to were considered.

4.3.5 Compilation of hierarchical DMCpG sets

.

The hierarchical DMCpG sets were compiled as follows. First, for each DMR cluster, all DMCpGs residing in the DMRs of the DMR cluster were pooled. Second, within each such pool, DMCpGs were subgrouped according to the DMCpG cluster to which they belonged.

For each DMR cluster, only the (up to six) most frequent DMCpG clusters were considered, i.e., only the DMCpG clusters to which many of the DMCpGs within the DMRs of the DMR cluster belonged. This resulted in up to six distinct DMCpG sets per DMR cluster.

4.3.6 Single-cell clustering analysis

To compare the DNA methylome states of single cells within the HSPC compartment, each cell was characterized by its average methylation levels across the hierarchical DMCpG sets described in section 4.3.5. In addition to the general coverage quality control described above, cells were further filtered at this stage for a total coverage of at least 10,000 DMCpGs across all hierarchical DMCpG sets. Only hierarchical DMCpG sets with at least one methylation call in each cell were considered. Of note, among the remaining hierarchical DMCpG sets, the coverage across the selected cells was relatively high, so no further filtering of the hierarchical DMCpG sets was necessary (Figure 33). To order and partition the single cells, hierarchical clustering with Ward's method in combination with the cutreeHybrid partitioning algorithm [172] was applied. Hierarchical clustering was performed with scipy [218], and partitioning was performed with the Python implementation of the dynamicTreeCut R package [219]. The `cutreeHybrid` function was called with the following parameters `minClusterSize=1, deepSplit=3, pamStage=True`.

4.4 Profiling of a three-tier single-cell RNA-seq data set covering the hematopoietic system

4.4.1 Generation of 10x Genomics single-cell RNA-seq data for LSK, LK, and total bone marrow cells

All wet lab experiments were performed by collaboration partners. The experimental methods are detailed in the doctoral thesis of Maximilian Schönung [OWN2], who used and analyzed the data independently for a different, separate purpose in his doctoral project.

Briefly, mouse preparation and FACS-based cell sorting of 3984 total bone marrow cells, 3441 LK cells, and 1070 LSK cells (8495 cells in total) were performed by members of the Section Translational Cancer Epigenomics (DKFZ, Division Translational Medical Oncology) led by Daniel Lipka, with major contributions from Maximilian Schönung and Mark Hartmann. Bone marrow cells were isolated from the femora, tibiae, hips, and spines of sacrificed mice. Lineage-negative bone marrow cells were enriched using the following antibody cocktail: CD5, CD45R, CD11b, CD8a, Ly-6G, Ly-6C, and Ter119.

Library preparation was performed by Katharina Bauer at the single-cell Open Lab (scOpenLab) at the DKFZ, with support from Mark Hartmann. Library preparation was performed according to the manufacturer's instructions using the Single Cell 3' Reagent Kits v2 (10x Genomics) and the Chromium Controller (10x Genomics) to generate "Gel Bead-In-Emulsions" (GEMs).

All libraries were sequenced using paired-end (PE 26/96 bp) sequencing on the NovaSeq 6000 platform (Illumina). Sequencing was performed at the Genomics and Proteomics Core Facility at the DKFZ. Read alignment and data management with the Cell Ranger analysis pipeline system (10x Genomics) was provided by the Omics IT and Data Management Core Facility (ODCF).

4.4.2 Clustering and cell type annotation

Single-cell transcriptome analysis was performed using the standard Scanpy workflow [145, 148, 220] with targeted modifications to account for the vast difference in cell type frequencies in the data. Cells with a fraction of mitochondrial reads above 5% were discarded. Cells from the LSK, LK, and total bone marrow surface marker-defined populations showed significantly different read count distributions. Therefore, the total bone marrow cells were filtered for a minimum read count of 1300 reads, while LSK and LK cells were filtered for a minimum read count of 2000 reads. Gene expression normalization was performed using the sctransform algorithm [146]. This approach is based on predicting gene expression levels in individual cells using a regularized negative binomial regression model where the cellular sequencing depth is utilized as the independent variable. The Pearson residuals from this model have been shown to represent normalized expression values with favorable properties [146, 147]. A challenge during the normalization of scRNA-seq data covering a range of cell types from progenitor to mature cells is that some of these cell types exhibit transcriptomes dominated by few strongly expressed genes. This effect was strongly pronounced in the given data set. This can skew standard gene expression normalization approaches [148, 149]. To address this challenge, I have exchanged the standard independent variable used for sctransform normalization (total sequencing depth) with an adjusted sequencing depth. This adjusted sequencing depth was computed while excluding i) all genes that in at least one cell possessed more than 5% of all the counts observed within that cell, and ii) all mitochondrial genes. This is likely to provide an improved size factor, following ideas initially proposed by Weinreb et al. [149] and recently promoted by an influential review of best practices in scRNA-seq data analysis [148]. The pearson residuals were clipped to the interval $[-\sqrt{n_{cells}/30}, +\sqrt{n_{cells}/30}]$, the default clipping range used in Seurat [168].

Clustering was performed with the PARC algorithm [150] for community detection. This algorithm extends the standard Leiden clustering algorithm with several preprocessing steps pruning the k-nearest neighbor graph. This has been shown to improve the clustering in the presence of strongly differing cluster sizes and between-cluster similarities [150]. PARC clustering was intentionally performed with a high resolution parameter of 1 to achieve sufficient partitioning to separate cell clusters for rare cell types, yielding 35 cell clusters that were partially redundant. These clusters were annotated with cell type labels using literature-based cell type markers [29, 45, 151, 152]. Clusters representing sub-types of a particular mature hematopoietic cell type were merged, because such high resolution among the mature hematopoietic cell types was not required. For example, there were three different clusters labeled as erythroblasts1-3 due to their high expression of *Hba-a2* and other erythroid markers. The cells from these clusters were merged to obtain a single erythroblast cluster for downstream analysis. The final clustering result comprised 18 distinct, cell type-annotated, single-cell clusters.

I acknowledge that several other doctoral students have done independent work on the same scRNA-seq data set. I have discussed some aspects of the analysis of these data with several of these colleagues, including Maximilian Schönung, Sina Stäble, Mariam Hakobyan, and Abdelrahman Mahmoud. I have carried out an earlier version of the data analysis in cooperation with Sina Stäble. Sina Stäble has shown parts of this collaborative work in her thesis [131], in combination with independent work performed by her and other collaboration partners. Later, Maximilian Schönung carried out an independent, similar clustering analysis as the one presented in this thesis and provided it to me for reference. Nevertheless, the analysis in this thesis stands out as an independent and original analysis, with distinct goals, scope, and methodological complexity compared to other analyses of the data. The analysis presented in this thesis was conceptualized and coded by me and differed in various key aspects from the parallel efforts of my colleagues. My analysis uses a different analysis framework (Python/ScanPy instead of R). I use different computational strategies, for example with regard to expression normalization and clustering (which I have provided in part to Sina Stäble for reproduction in her thesis). My analysis also differs in its focus on the identification of clean clusters of rare cell populations in the data set, such as the eosinophil population, which was of particular interest to this project.

4.4.3 Differential expression testing and computation of DMR cluster target gene set expression scores

Differential gene expression testing was performed using the `rank_genes_groups` function from Scanpy. Testing was performed using the Wilcoxon rank-sum test based on the sctransform-normalized expression levels. Testing with CPM-normalized log+1 expression values was considered but performed worse. This was likely due to the problem that some cell types exhibited skewed transcriptomes with dominant expression of some transcripts, as detailed above. Enrichment of highly expressed genes within each single-cell cluster was tested against the background of all other clusters, and multiple testing correction was performed with the Benjamini-Hochberg (BH) method (FDR < 0.01%). Only enrichments with a $\log_2$ fold change > 1.25, and only genes that were expressed in more than 25 cells were considered. Log-fold changes were computed based on CPM-normalized log+1 expression values, because log-fold changes computed on Pearson residuals are not suited in this context. For each single-cell cluster, I collected the 50 most enriched genes.

To compute the average expression of the target gene sets for each DMR cluster (section 4.3.4), the gene expression vectors were first z-score normalized and clipped to the interval $[-3, +3] to equally weigh all genes independent of their expression level. The gene expression vectors were then min-max normalized, and the mean of all gene expression vectors for each geneset was computed.

4.5 Enrichment analysis

4.5.1 Clustering of hematopoietic enhancer regions

Average H3K4me1 read counts and genomic positions for the 48,415 hematopoietic enhancer regions described by Lara-Astiaso *et al.* [107] were retrieved from Supplementary Table 2 of the publication. In this study, the enhancers were clustered based on the H3K4me1 read counts, using K-means clustering (K=9). This led to the identification of nine different enhancer activation programs during hematopoiesis. The clustering information was not made available with the publication. Therefore, the K-means clustering analysis was repeated based on the published data, yielding clusters matching the published clusters in both size and the characteristic pattern of enhancer activity.

I would like to acknowledge that Sina Stäble has independently generated an alternative re-analysis of the enhancer clustering for her own doctoral thesis. Her analysis differed from the published enhancer clustering by Lara-Astiaso et al. with regard to the applied normalization of the H3K4me1 counts. My independent analysis uses the same normalization as Lara-Astiaso et al.} and thus reproduces the published enhancer clusters.

4.5.2 DMR cluster annotation through gene set and region set enrichment analysis

Enrichments of gene sets or region sets in individual DMR clusters were computed by comparing the numbers of the DMRs in the foreground cluster which possessed membership in the gene set or region set against the number of all other DMRs which possessed this membership, using Fisher's exact test (two-sided). P-value adjustment into q-values was performed using the BH method [154, 155]. Thus, the enrichment analysis allowed multiple

DMRs annotated against the same gene within a DMR cluster to contribute separately to the enrichment, as opposed to classic overrepresentation analysis, in which the set of genes associated with foreground DMRs is contrasted with the set of genes associated with the background DMRs.

For gene set enrichment analyses, only DMRs that were annotated to a gene (i.e., which were not intergenic) and whose center lay within ±50 kb of the TSS of their annotated target gene were considered. Gene set membership for each DMR was determined based on its gene annotations (section 4.2.3). If a DMR had multiple gene annotations, the DMR was considered to be a member of a given geneset if any of its annotated genes belonged to this geneset. To test for the enrichment of hematopoietic cell type expression markers, I used cell type expression marker gene sets computed based on an in-house scRNA-seq data set of murine hematopoietic cells (section 4.4.3). Enrichments against the hematopoietic enhancer clusters published by Lara-Astiaso et al. [107] were performed based on a reconstruction of the published enhancer clusters using the original tag counts and methods made accessible with the publication (section 4.5.1).

4.5.3 DMR subregion-resolved transcription factor binding motif enrichment analysis

Archetype TFBM positions were generated by Vierstra et al. [169]. A BED file containing the genomic positions of all archetype TFBMs (version 1.0) was downloaded from `https://resources.altius.org/~jvierstra/projects/motif-clustering/releases/v1.0/mm10.archetype_motifs.v1.0.bed.gz`. Annotations for each archetype TFBM were downloaded from `https://resources.altius.org/~jvierstra/projects/motif-clustering/releases/v1.0/motif_annotations.xlsx`.

Enrichments were performed using a novel paradigm for TF enrichment testing: testing was neither performed at the DMR nor at the DMCpG level. Instead, DMCpGs which resided in DMRs of the same DMR cluster and belonged to the same DMCpG cluster were grouped together. Then testing was performed over these groups. This enabled testing for TFBM enrichments within systematically defined subregions of DMRs. The construction of these hierarchical DMCpG sets is described in section 4.3.5. As an additional filtering step for the enrichment analysis, for each such DMCpG set, I only allowed one randomly chosen DMCpG from each individual DMR. I reasoned that DMCpGs within a single DMR were more likely to introduce dependency structures into the enrichment tests, which would skew the testing results. To make statistical testing across all DMCpG sets comparable, I considered only DMCpG sets with at least 650 DMCpGs. DMCpGs sets containing more than 650 DMCpGs were downsampled to contain exactly 650 DMCpGs. I then screened each group of DMCpGs for associations with the archetype TFBMs, using Fisher's exact test to test each DMCpG group against the background of all other DMCpG groups. P-value adjustment into q-values was performed using the BH method [154, 155].

4.6 Quantification and visualization of DMR seeding and expansion during hematopoietic differentiation

4.6.1 Classification of DMR expansion states

This study introduces two novel methods for the classification of DMR expansion states. The first method can be applied to any WGBS data set using basic data analysis operations. The second method requires the construction of a dual-layer DMR/DMCpG atlas (section 4.2) as a prerequisite step. The first method is thus easier to interpret and implement, while the second method provides robust information sharing across samples in multi-group comparisons. Both methods were applied as described below across all DMRs in the hematopoietic DMR/DMCpG atlas generated in this study. These methods could also be used to estimate DMR expansion across single-cell clusters in scBS-seq data analysis if sufficiently high cell numbers are available. To apply these methods to other data sets, minor modifications would be necessary. For example, the HSC population was used as a reference root population in this study. Other data sets would use other root populations or root single-cell clusters.

Method 1: DMCpG methylation level-based DMR expansion state classification

The DMR expansion state of each DMR in each population was classified as either unregulated, seeded, intermediate, or completed. For each DMR in each population, I first determined the number of regulated DMCpGs. DMCpGs were considered regulated if they exhibited an absolute methylation level shift of at least 30% compared to the HSC methylation level; for the MPP1-5 populations, this threshold was lowered to a shift of 20%. The threshold was reduced for the MPP1-5 populations because, due to their high heterogeneity, a significant shift of the frequency of DNAme at a given CpG may only occur in a population subset and still be biologically meaningful. For each DMR, I noted the maximal number of regulated DMCpGs observed in any population. The DMR state for each population was determined based on the percentage of regulated DMCpGs relative to the maximum observed count: DMRs were thus classified as unregulated (0% regulated CpGs), seeded ($< 45\%$), intermediate ($< 81\%$), or completed ($\geq 81\%$).

Method 2: DMCpG cluster-based DMR expansion state classification

This method is largely identical to method 1, except for the first analysis steps. The method requires the construction of a dual-layer DMR/DMCpG atlas (section 4.2) with cluster annotations for both layers (section 4.3). First, each DMCpG cluster was annotated with a set of regulated populations, as detailed in section 4.3.2. Briefly, the set of regulated populations for a particular DMCpG cluster comprised all populations exhibiting a large percentage of at least partially hypomethylated DMCpGs among the DMCpGs of the cluster. Next, the number of regulated DMCpGs in a particular DMR in a particular population could be determined by counting the number of DMCpGs in the DMR which belonged to DMCpG clusters for which the population was listed as a regulated population. Using this approach, the number of regulated DMCpGs was determined for each DMR in each population. The next steps were identical to method 1, i.e., for each DMR, the maximal number of regulated DMCpGs observed in any population was noted. The DMR state for each population was determined based on the percentage of regulated DMCpGs relative to the maximum observed count: DMRs were thus classified as unregulated (0% regulated CpGs), seeded ($< 45\%$), intermediate ($< 81\%$), or completed ($\geq 81\%$). DMRs with at least five regulated DMCpGs were considered to be in an intermediate expansion state, even if these five DMCpGs represented less than 45% of all DMCpGs in the DMR. In summary, for the DMCpG cluster-based approach, only information about the different DMCpG clusters to which the DMCpGs in a particular DMR belonged was used, and not the concrete methylation level of each DMCpG measured for that DMR (which could be subject to potentially substantial sampling noise, depending on the coverage).

4.6.2 Asymmetry-aware visualization of DMR profiles and DMR expansion

To enable comparisons across DMRs of varying sizes, each individual DMR was segmented into seven bins with an average size of 39 bp. For visual displays, the flanking regions around the DMRs were also segmented into bins using this average size. However, for the alignment computation described in the following, only the information from within the DMR regions was used. For each DMR, a methylation level vector was computed, representing the average CpG methylation levels for each of these bins along the plus strand.

Initial analysis revealed that many DMRs expanded asymmetrically from a hypomethylated seed region that emerged in progenitor populations: within these DMRs, the DMR expansion occurred predominantly along the plus or the minus strand (each case was observed in about half of the DMRs). To align all methylation level vectors by their DMR expansion direction, the methylation level vectors of all DMRs where the major direction of DMR expansion occurred along the minus strand were reversed. To determine the direction of major expansion for each DMR, I exploited the asymmetry of the methylation levels within DMRs where a seeded or intermediate DMR expansion state is followed by asymmetric DMR expansion. For each DMR cluster, I first selected the population which most prominently exhibited asymmetric DMR methylation level states in that cluster, defined as the population for which the average absolute methylation level difference between the second and first half of the methylation level vector DMR was maximal. These populations were most clearly and most often in a seeded or intermediate state, followed by asymmetric expansion in their corresponding DMR cluster. For each DMR within each DMR cluster, I computed the absolute methylation level difference between the right and left half of the methylation level vector using the selected populations. If this difference was greater than 0, I could assume that the seed region was mostly located in the first half of the vector and that the DMR expanded from there towards the second half, corresponding to DMR expansion mainly along the plus strand. Conversely, for a negative difference, I could assume that the main direction of DMR expansion was along the minus strand. DMRs with perfectly symmetric DMR expansion (difference of 0) were not observed, although DMRs with near-symmetric expansion did exist (difference close to 0).

For displaying these data in a heatmap representation, the methylation levels for bins with missing values were interpolated with a Gaussian kernel (kernel width = 7 bins, s.d. = 1) smoothing, using the astropy.convolve algorithm, which can perform convolution in the presence of missing values.

For displaying the average methylation levels across the DMRs of the D1 DMR cluster, I first computed the average methylation levels for each of these populations using unsmoothed and uninterpolated methylation level vectors, with DMR size stratification. Specifically, all DMR methylation level vectors for each population were grouped by the size of their corresponding DMRs, and the average methylation level vector within each group was computed. Grouping was performed by these DMR size intervals: 0 to 100 bp, 100 to 200 bp, 200 to 350 bp, 350 to 500 bp, and finally 500 bp and all DMR sizes above. All intervals before the last were defined as left-closed, right-open intervals. For display, the resulting size-stratified average DMR methylation levels were individually smoothed using Gaussian kernel smoothing. For this purpose, the discrete DMR methylation level vectors (with one value per bin) were interpolated into curve data with one value per bp using linear interpolation. Then these curves were smoothed with a Gaussian kernel (width = 40 bp, s.d. = 20 bp).

4.7 Data and code

4.7.1 Code availability

The complete analysis code for this thesis is documented in Jupyter Notebooks and deposited on GitHub (`https://github.com/stephenkraemer/mouse_hema_meth`). Access to the repository is given upon request. All software packages developed for the analyses in this thesis are available as open source Python packages on GitHub, including bistro [SOFT1], codaplot [SOFT3], gtfanno [SOFT4], and methlevels [SOFT2].

4.7.2 Data availability

The complete, uniformly processed alignments and methylation calls for all bulk samples as well as the alignments and methylation calls for the scBS-seq samples are available upon request. FASTQ files and methylation calls will be made publically available on GEO when the manuscript describing the main findings of this thesis is published (the manuscript is currently in preparation).

The original FASTQ files and methylation calls for the T-WGBS data initially published in [14, 122] are available under accession number GSE52709 (containing the HSC, MPP1, MPP2 populations, as well as a combined MPP3/4 population which was not considered for this thesis). These methylation calls were generated by Qi Wang [14, 122]; they are different from the methylation calls used for this thesis.

A subset of the uniformly processed data set generated by me was published in parallel to my work on this thesis and is available under accession number GSE164124 (containing the MDP, CDP, cMoP, cDC1, cDC2, pDC populations).

4.7.3 Programming languages and software packages

All custom analysis modules and open source software packages developed for this project were written in Python. The bistro methylation caller was written in Cython. Various software packages from the Python data science ecosystem were repeatedly utilized, including the core Python data science libraries NumPy [221], pandas [222], SciPy [218], and statsmodels [223] as well as the core Python bioinformatics libraries PyRanges [224] and ScanPy [145]. Specialized Python packages used for specific tasks are cited in the corresponding sections of the Methods and Results chapters. In addition to the Python ecosystem, a few tasks were carried out using Bioconductor/R packages [225]. These packages are also cited in the corresponding Methods and Results chapters.

Chapter 5

Supplementary Materials

5.1 Supplementary figures

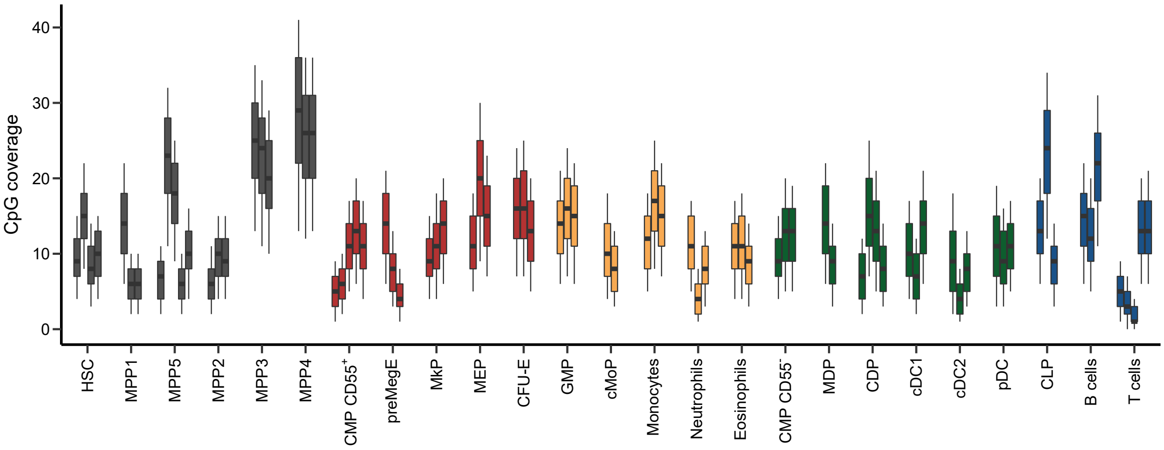

Figure S1: Boxplots showing the distribution of the autosomal CpG coverage (number of methylation calls per CpG dinucleotide) in each replicate. Whiskers represent the 10th and 90th percentiles. Tagmentation-based whole-genome bisulfite sequencing (T-WGBS) was performed on three or more replicates for all populations, except for the cMoP and MDP populations (two replicates). A uniform alignment and methylation calling pipeline was applied to the T–WGBS data of each replicate. Methylation calls from all replicates within a population were combined for population-level analyses, the resulting population-level CpG coverage is shown in Figure 6. Alignments were performed by the Omics IT and Data Management Core Facility (ODCF) at the German Cancer Research Center, using an updated version of the T–WGBS alignment workflow described by Wang et al. [68]. Methylation calling was performed using the *bistro* software package [SOFT1], which offers automatic detection and filtering of methylation calls affected by gap repair nucleotides or M-bias.

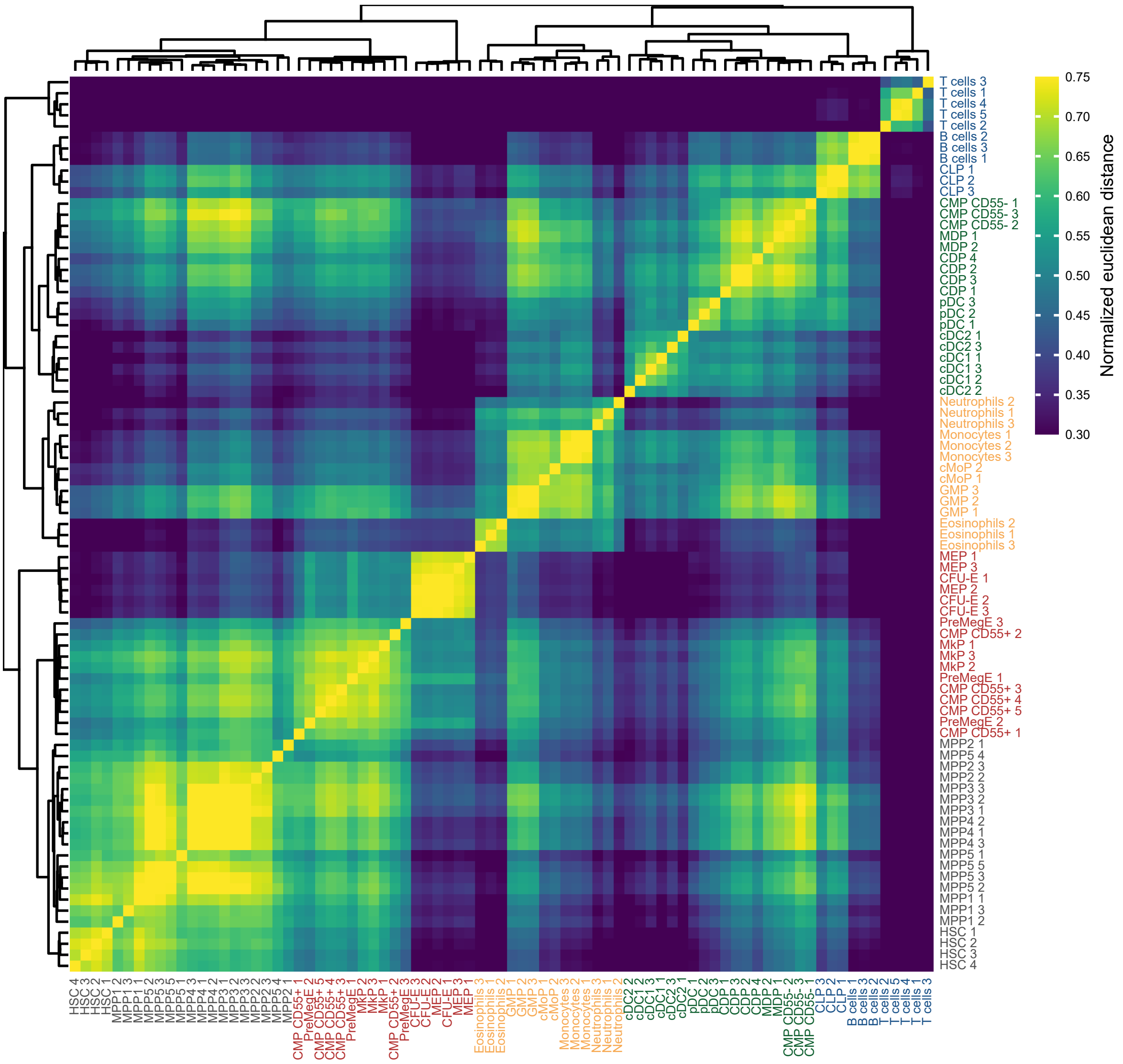

Figure S2: Methylation levels in the Ensembl Regulatory Build intervals are highly similar between replicates of the same population. Average methylation levels in all murine candidate cis-regulatory regions identified in the Ensembl Regulatory Build were computed. To account for global shifts of the CpG methylation levels between the hematopoietic cell types, each methylation level vector was corrected by subtracting its mean methylation level. The heatmap shows min-max-normalized pairwise Euclidean distances between all replicates. Replicates were clustered using unsupervised hierarchical clustering with Ward's method.

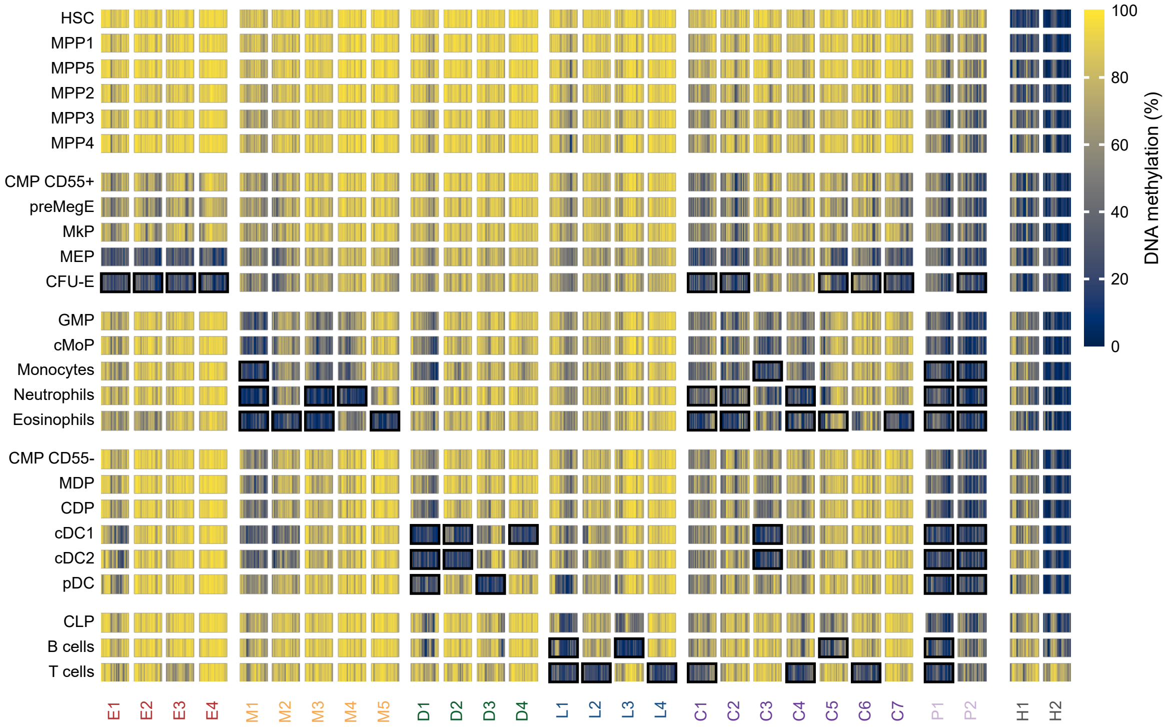

Figure S3: DMR cluster compactness and separation is maintained when DMR methylation levels are considered directly. For each cluster, methylation levels for 200 randomly selected DMRs are shown. This figure complements the analogous display of the z-score transformed DMR methylation levels for the same DMRs in Figure 13. DMR clustering was performed using Leiden clustering with the correlation distance. DMRs were thus grouped in a scale and location invariant approach, such that high homogeneity of the individual, z-score transformed DMR methylation level profiles within each DMR cluster could be achieved. This figure demonstrates that the DMR clusters remain clearly separated when the DMR methylation levels are directly considered.

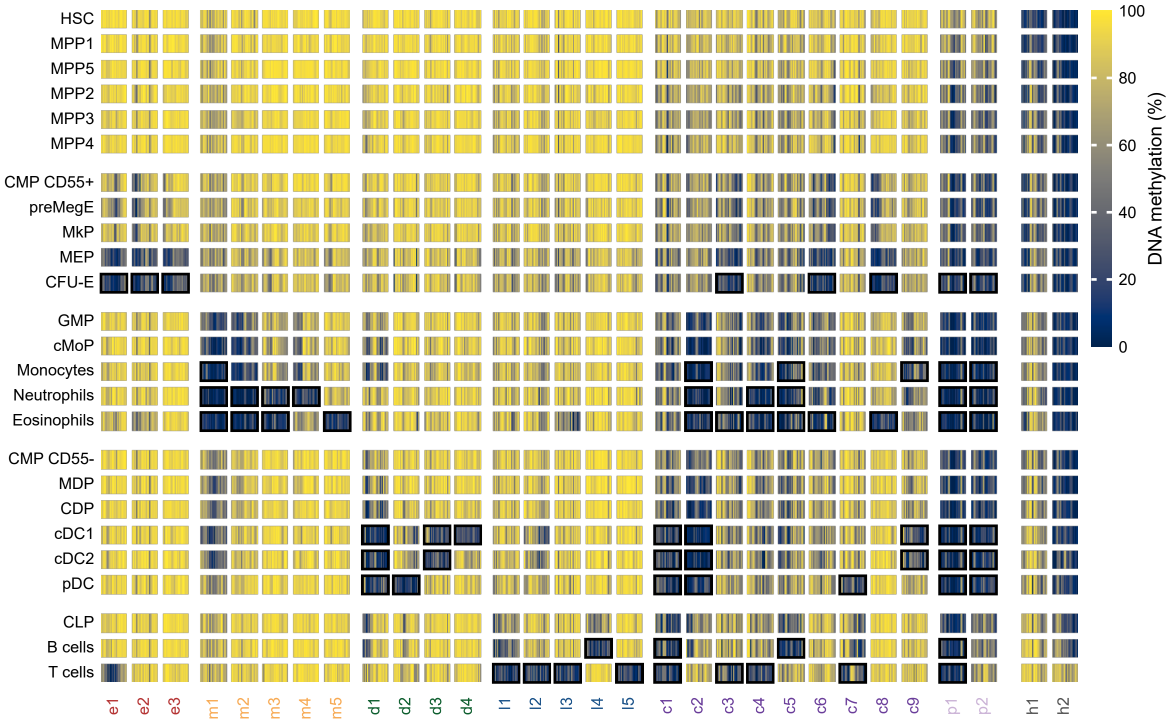

Figure S4: DMCpG cluster compactness and separation is maintained when DMCpG methylation levels are considered directly. For each cluster, methylation levels for 200 randomly selected DMCpGs are shown. This figure complements the analogous display of the z-score transformed DMCpG methylation levels for the same DMCpGs in Figure 21. DMCpG clustering was performed using Leiden clustering with the correlation distance. DMCpGs were thus grouped in a scale and location invariant approach, such that high homogeneity of the individual, z-score transformed DMCpG methylation level profiles within each DMCpG cluster could be achieved. This figure demonstrates that the DMCpG clusters remain clearly separated when the DMCpG methylation levels are directly considered.

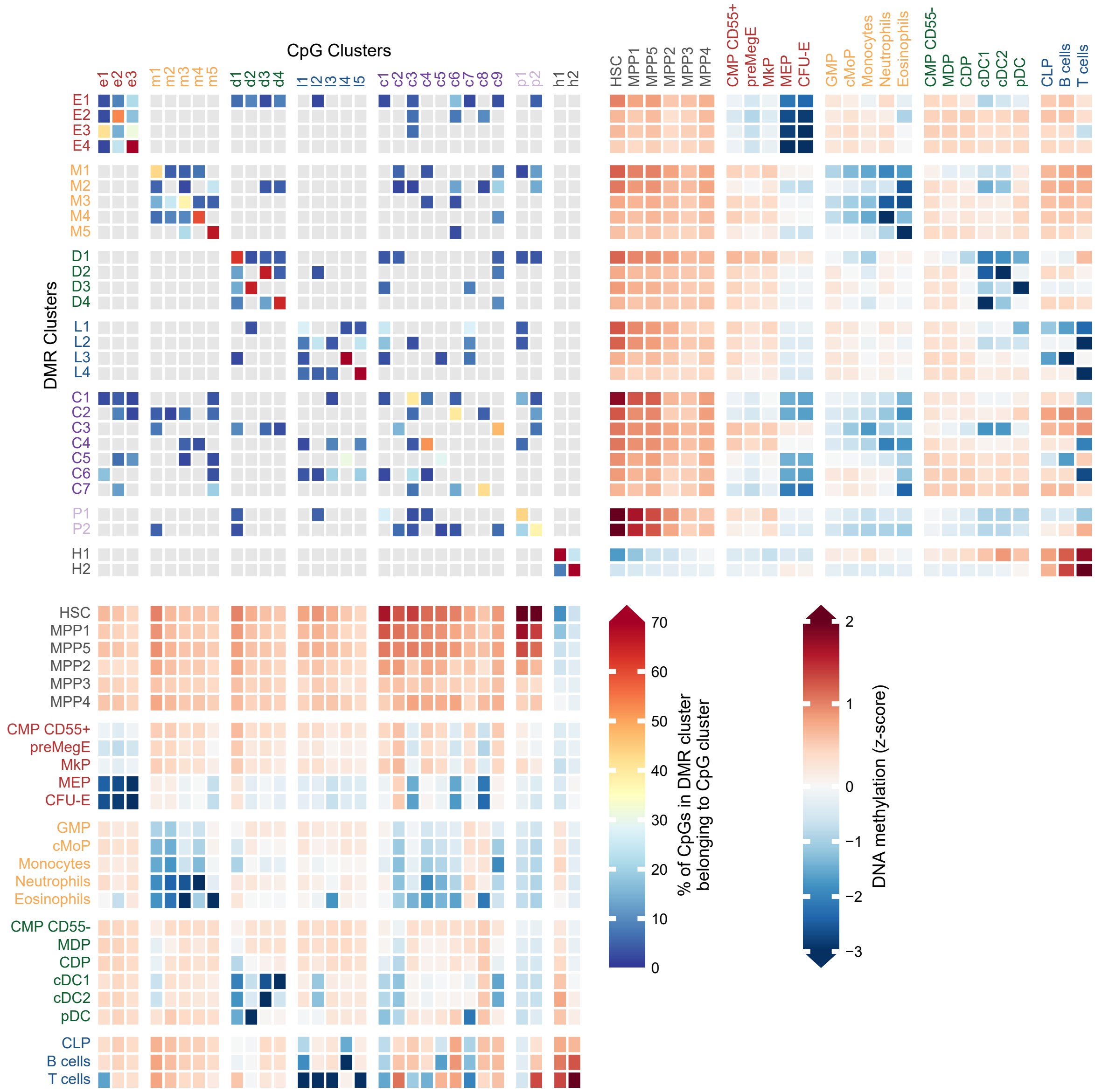

Figure S5: Within most DMR clusters, the predominant majority of all contained DMCpGs is characterized by one or two characteristic DMCpG programming patterns which closely match the overall programming pattern of the DMR cluster. The complex heatmap juxtaposes i) the z-score transformed mean DMR methylation levels across all populations for each DMR cluster; ii) the z-score transformed mean DMCpG methylation levels across all populations for each DMCpG cluster; and iii) for each DMR cluster, the percentage of all DMCpGs contained in the DMRs of the cluster exhibiting each DMCpG programming pattern. This visualization complements Figure 24, which shows the percentage of the DMRs within each DMR cluster exhibiting each DMCpG programming pattern at least once.

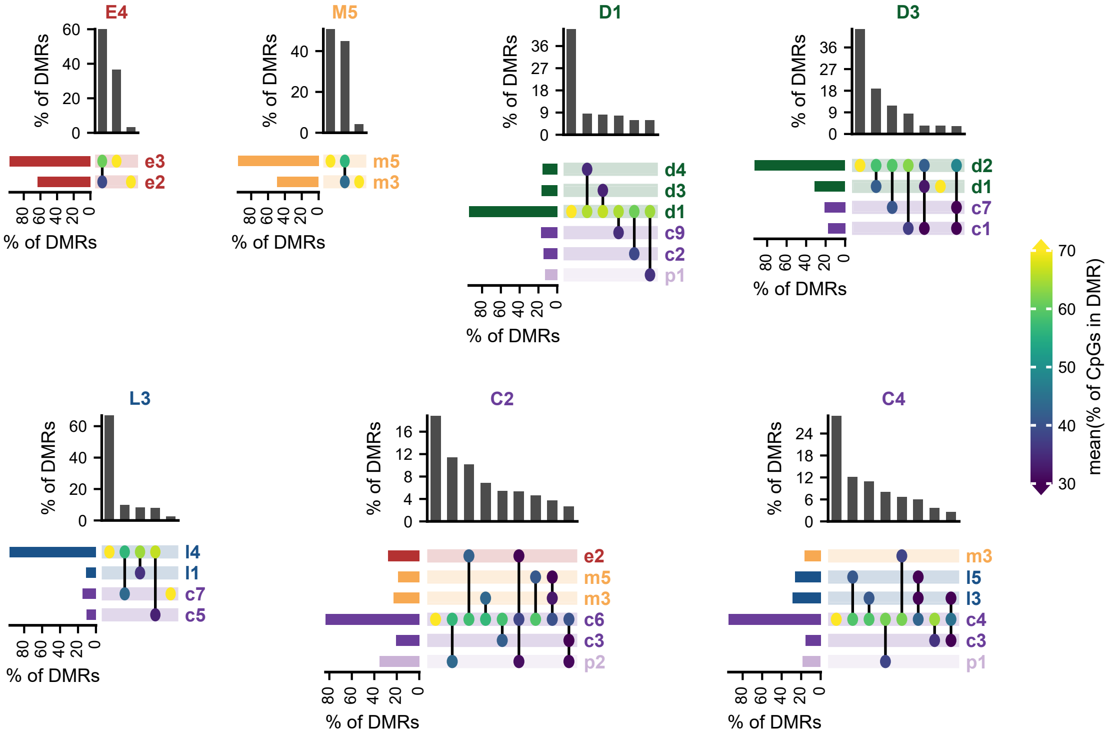

Figure S6: Programming within DMRs typically involves a predominant DMCpG–PP shaping the population-specificity of the DMR, in combination with less frequent, less population-specific DMCpG-PPs which indicate preceding DMR seeding and expansion steps during differentiation. UpSet plots detailing properties of the most frequent DMCpG-PPs for representative DMR clusters. This figure provides supplementary characterization of additional DMR clusters to extend the analysis introduced in Figure 29.

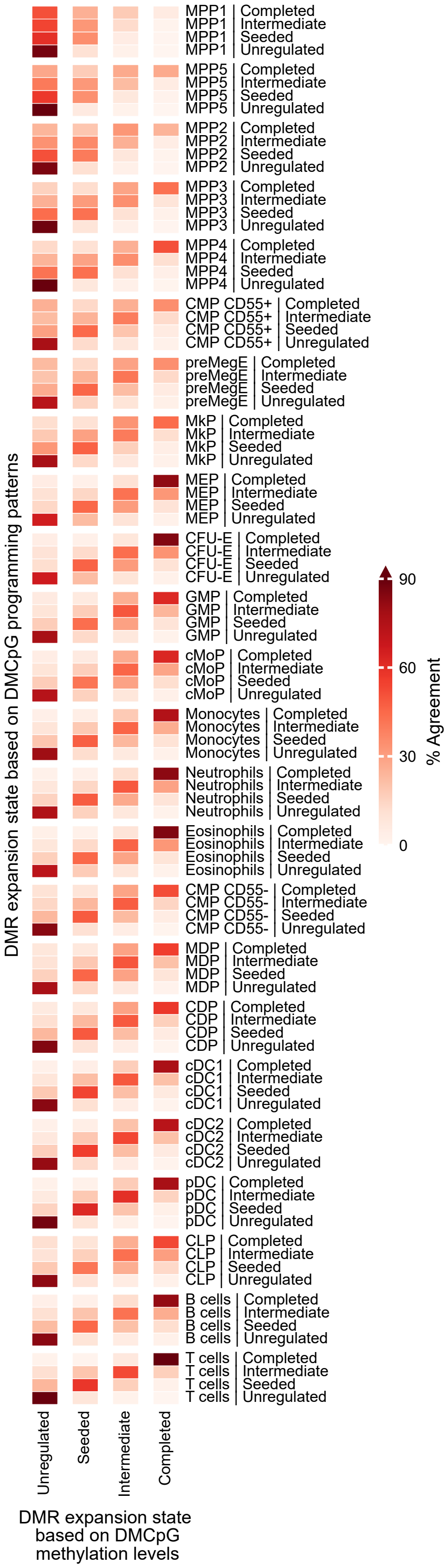

Figure S7: Predominant agreement between DMR expansion state classifications based on DMCpG programming patterns or based on individual DMCpG methylation levels. The heatmap visualizes a confusion matrix showing the percentual overlap between the DMR state classifications from both methods. Both methods were used to classify the expansion state of each DMR as unregulated, seeded, intermediate or completed. The DMR expansion state map obtained through the methylation level-based method was introduced in Figure 17, and the alternative DMR expansion state map obtained through the DMCpG programming pattern-based method was introduced in Figure 27.

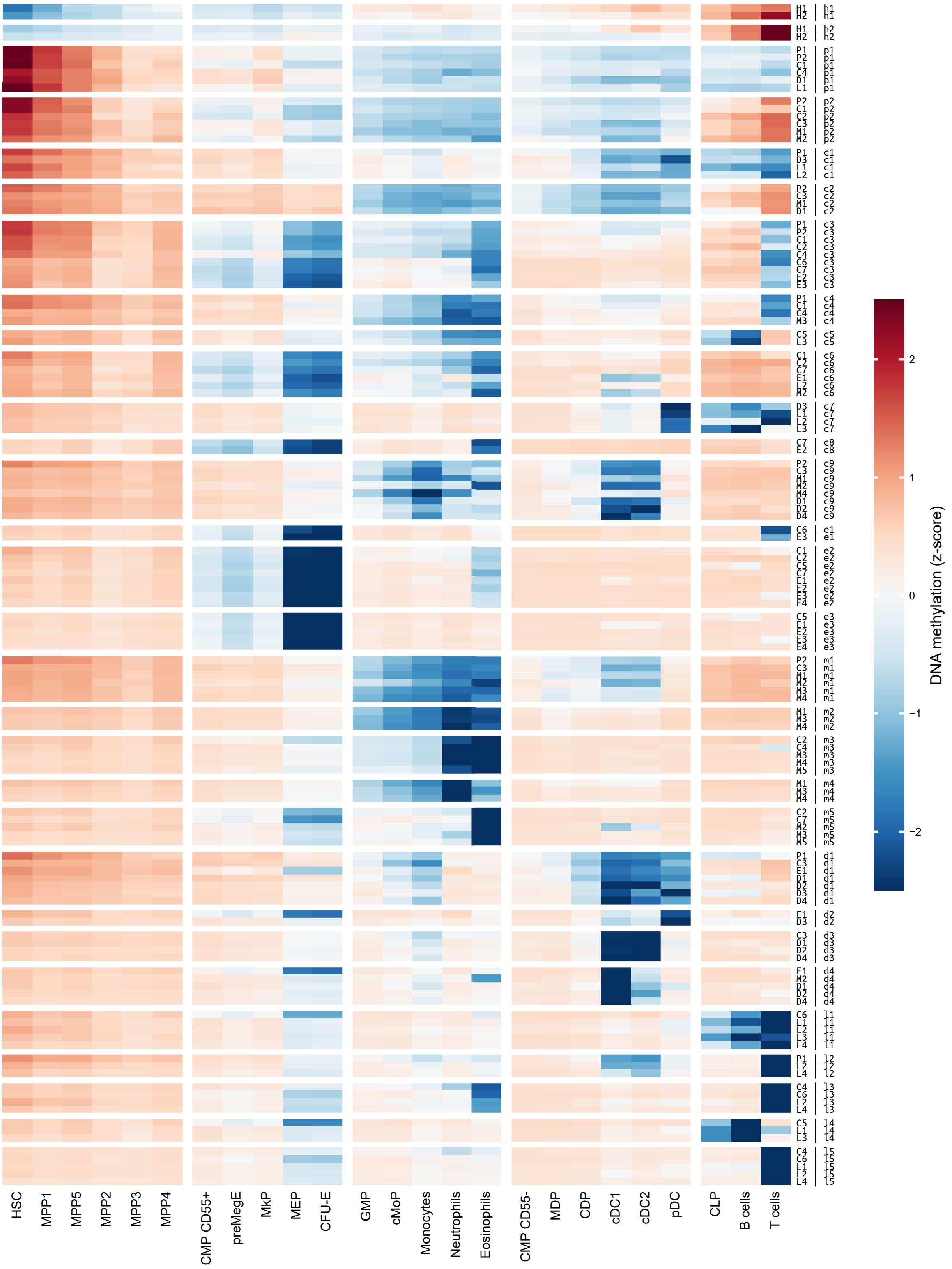

Figure S8: DMCpGs from the same DMCpG cluster share highly related programming patterns independently of the DMR cluster membership of the DMRs they reside in. DMCpGs were grouped hierarchically, first by the DMR cluster in which they occurred, then by the DMCpG cluster to which they belonged. Only the (up to six) most frequent DMCpG–PPs for each DMR cluster were considered. The heatmap shows the average z-score transformed DNA methylation levels for each hierarchical DMCpG set.

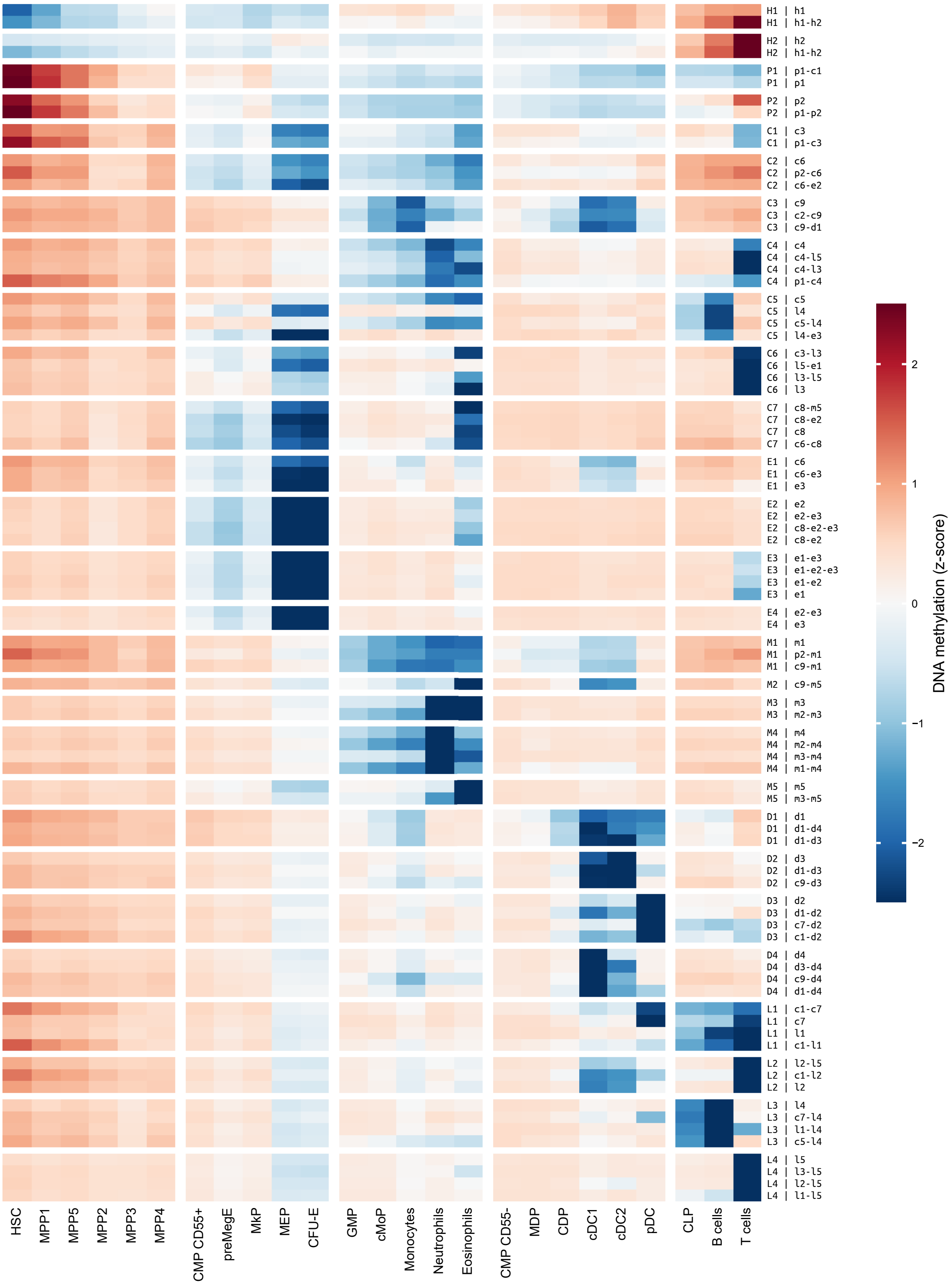

Figure S9: DMRs from the same DMR cluster share highly related programming patterns, independently of which DMCpG clusters the individual DMRs contain. For each DMR cluster, the most frequent combinations of contained DMCpG clusters were determined. Only combinations arising in at least 8% of all DMRs in the cluster were considered. Then DMRs within each DMR cluster were grouped by these combinations. The heatmap shows the average z-score transformed DMR methylation levels for each such group of DMRs.

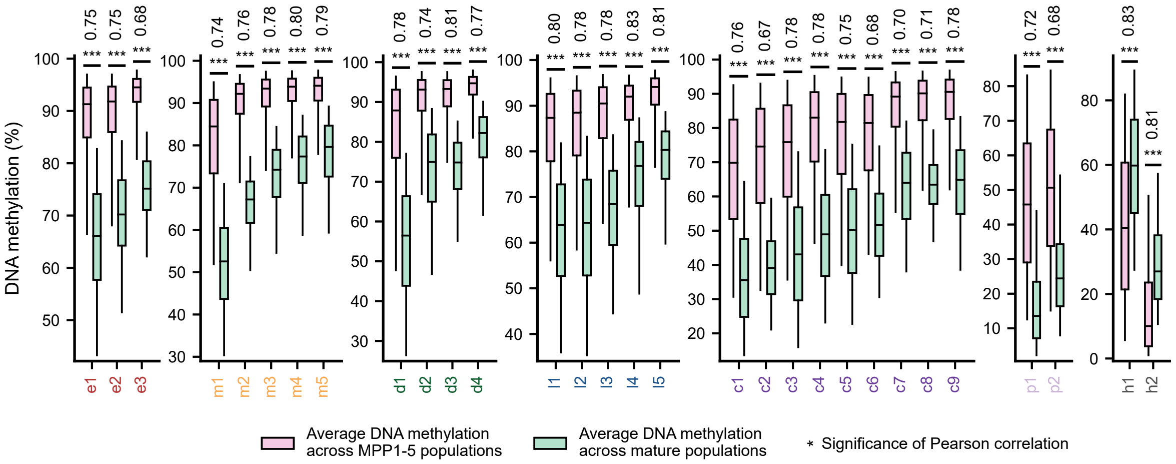

Figure S10: Strong correlation betweeen the extent of early DMCpG programming and the breadth of hypomethylation across the mature populations within each DMCpG cluster. Box plots show the range of the mean DNAme levels across the mature populations and of the mean DNAme levels across the MPP1-5 populations across all individual DMCpGs within each DMCpG cluster. Whiskers indicate the first and 99th percentile. For each DMCpG cluster, the correlation between these two statistics was computed across all DMCpGs within the cluster. The Pearson correlation coefficient for each comparison is indicated above the boxplots. The significance of the pearson correlation for each comparison was assessed with a permutation test, using 1000 randomly sampled DMCpGs and 10,000 permutations. All significance tests showed the minimal possible p-value of 1e-4 after Benjamini-Hochberg correction, indicated by the three stars above each pairwise comparison. This figure supplements the global correlation analysis presented in Figure 30.

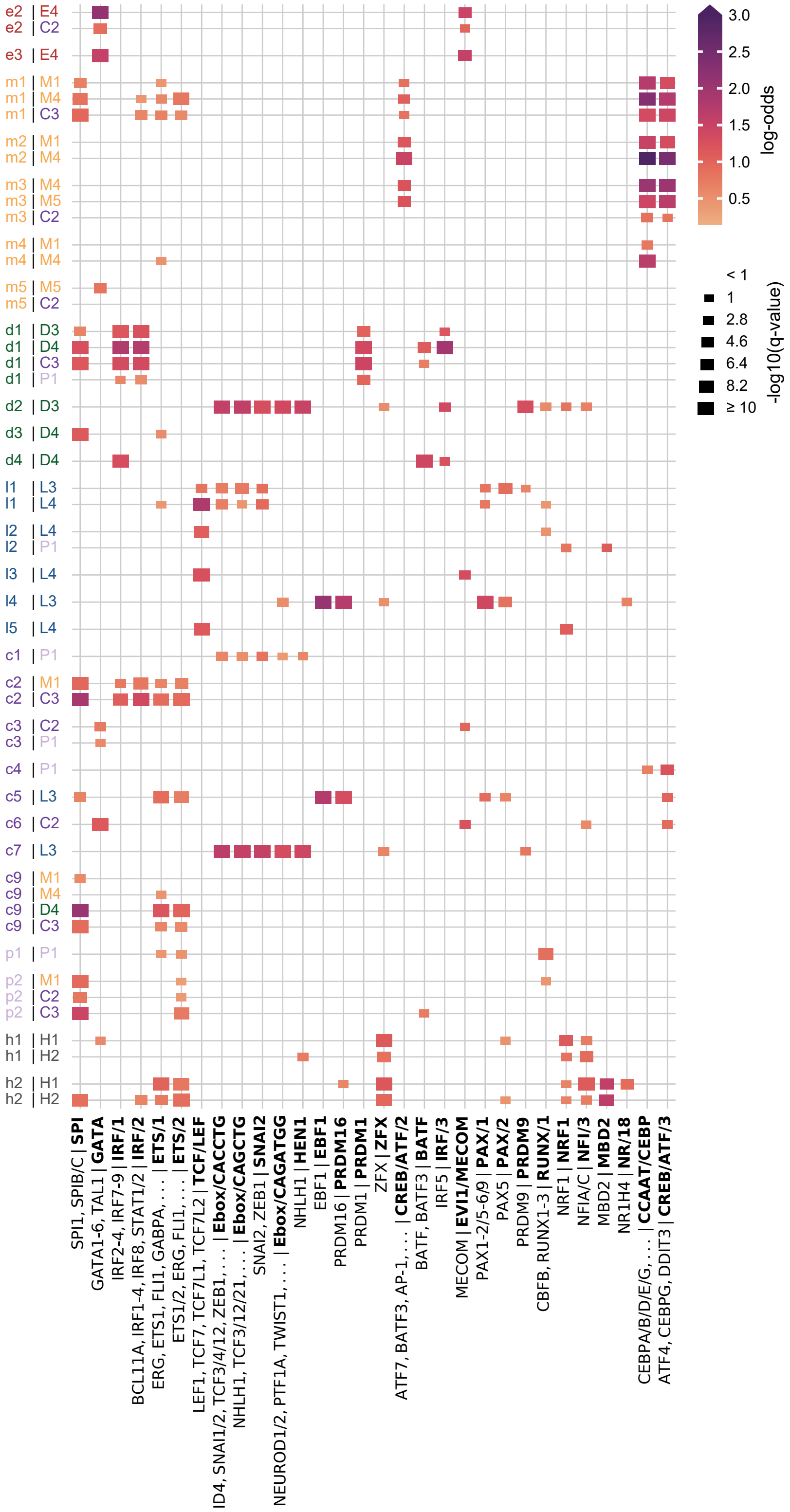

Figure S11: Programming with different DMCpG programming patterns is associated with the activity of distinct transcription factors. This figure supplements Figure 31. It shows the same data through the same heatmap encoding, with a different sorting order: the DMCpG sets are sorted by their associated DMCpG cluster, instead of by their associated DMR cluster. This facilitates the comparison of the TFBM enrichments for the same CpG cluster across multiple DMR clusters.

5.2 Supplementary tables

Table S1: Overview of the tagmentation-based whole-genome bisulfite sequencing (T-WGBS) dataset. All sequencing data were uniformly processed for this study. T-WGBS sequencing data for nine of the populations were previously published (HSC, MPP1 and MPP2 in [14, 122]; MDP, CDP, cMoP, cDC1, cDC2 and pDC in [OWN1]). For the remaining populations, new sequencing data were generated for this study, in at least three replicates per population. Moreover, an additional replicate for the HSC population was generated, complementing the three HSC population replicates published in earlier studies. The experimental sample generation was performed by collaborators, as detailed in the doctoral thesis of Sina Stäble [131], who performed a significant part of this experimental work as part of her doctoral project. The table lists the original publication and the GEO accession number of the published data. The table shows the coverage per replicate and the coverage achieved for each population by aggregating the information across all replicates. The available coverage is measured as the average number of methylation calls across the autosomal CpGs. Uniformly processed methylation calls for all replicates from all datasets are available on request and will be made available publicly when the results from this study are published. Lipka2014, data published in [14, 122]; Czeh2022, data published in [OWN1].

Population	Replicate	Replicate coverage	Population coverage	Dataset	Accession number
HSC	1	9.4	42.4	Lipka2014	GSE52709
	2	14.7			
	3	8.5			
	4	9.7		unpublished	on request
MPP1	1	14.1	26.0	Lipka2014	GSE52709
	2	6.0			
	3	5.8			
MPP5	1	6.7	62.9	unpublished	on request
	2	22.1			
	3	17.6			
	4	6.0			
	5	10.3			
MPP2	1	6.2	25.2	Lipka2014	GSE52709
	2	9.5			
	3	9.2			
MPP3	1	24.4	67.1	unpublished	on request
	2	22.7			
	3	19.8			
MPP4	1	28.2	78.0		
	2	24.7			
	3	24.9			
CMP $CD55^+$	1	5.1	46.3		
	2	6.1			
	3	10.7			
	4	13.2			
	5	11.0			
preMegE	1	13.7	26.0		
	2	7.8			
	3	4.4			
MkP	1	9.5	34.0		
	2	11.0			
	3	13.3			
MEP	1	11.4	46.1		
	2	19.6			
	3	14.9			

continued on next page

Table S1: continued

Population	Replicate	Replicate coverage	Population coverage	Dataset	Accession number
CFU-E	1	15.5	44.8		
	2	16.4			
	3	12.8			
GMP	1	13.7	44.4		
	2	16.1			
	3	14.6			
cMoP	1	10.8	18.7	Czeh2022	GSE164124
	2	7.9			
Monocytes	1	11.6	43.3	unpublished	on request
	2	16.6			
	3	14.9			
Neutrophils	1	11.3	24.0		
	2	4.4			
	3	8.3			
Eosinophils	1	10.6	30.7		
	2	11.3			
	3	8.7			
CMP CD55$^-$	1	9.5	34.8		
	2	12.6			
	3	12.5			
MDP	1	14.2	23.0	Czeh2022	GSE164124
	2	8.7			
CDP	1	7.2	44.3		
	2	15.5			
	3	13.1			
	4	8.4			
cDC1	1	10.5	31.3		
	2	7.1			
	3	13.6			
cDC2	1	9.5	21.8		
	2	4.3			
	3	7.9			
pDC	1	11.6	32.5		
	2	9.7			
	3	11.1			
CLP	1	13.1	45.1	unpublished	on request
	2	23.0			
	3	8.7			
B cells	1	14.3	48.1		
	2	12.6			
	3	21.1			
T cells	1	5.0	36.9		
	2	3.7			
	3	1.7			
	4	13.0			
	5	13.4			

Table S2: References for the surface marker definitions of the 25 hematopoietic populations analyzed by tagmentation-based whole-genome bisulfite sequencing. This table is based on information from the doctoral thesis of Sina Stäble [131] and on personal communication with Sina Stäble and Daniel Lipka.

Population	References
LSK	[17, 226]
HSC	[12–14, 226]
MPP1	[12–14, 226]
MPP2	[12–14, 226]
MPP3	[12–14, 226]
MPP4	[12–14, 226]
MPP5	[12–14, 226]
CMP CD55$^-$	[7, 22, 226]
CMP CD55$^+$	[7, 22, 226]
GMP	[7, 226]
MEP	[7, 226]
CLP	[21, 226]
preMegE	[23]
MkP	[23]
CFU-E	[23]
MDP	[24, 227, 228]
CDP	[24, 26, 227, 228]
cMoP	[144]
Mono	[229, 230]
Neutro	[230]
Eosino	[230]
cDC1	[230, 231]
cDC2	[230, 231]
pDC	[232, 233]
B cell	[234]
T cell	[234]

Table S3: Surface marker definitions for the 25 hematopoietic populations analyzed by tagmentation-based whole-genome bisulfite sequencing. This table is based on information from the doctoral thesis of Sina Stäble [131] and on personal communication with Sina Stäble and Daniel Lipka.

Abbreviation	Full name	Source	Surface markers
Lin-	lineage negative (Lin-) cells	bone marrow	CD5(-) CD8(-) B220(-) Ter-119(-) CD11b(-) Gr-1(-)
LSK	LSK cells	bone marrow	Lin(-) c-Kit(+) Sca-1(+)
HSC	hematopoietic stem cell	bone marrow	Lin(-) c-Kit(+) Sca-1(+) CD150(+) CD48(-) CD34(-)
MPP1	multipotent progenitor 1	bone marrow	Lin(-) c-Kit(+) Sca-1(+) CD150(+) CD48(-) CD34(+)
MPP2	multipotent progenitor 2	bone marrow	Lin(-) c-Kit(+) Sca-1(+) CD150(+) CD48(+)
MPP3	multipotent progenitor 3	bone marrow	Lin(-) c-Kit(+) Sca-1(+) CD150(-) CD48(+) CD135(-)
MPP4	multipotent progenitor 4	bone marrow	Lin(-) c-Kit(+) Sca-1(+) CD150(-) CD48(+) CD135(+)
MPP5	multipotent progenitor 5	bone marrow	Lin(-) c-Kit(+) Sca-1(+) CD150(-) CD48(-)
CMP CD55$^-$	common myeloid progenitor CD55 negative	bone marrow	Lin(-) c-Kit(+) Sca-1(-) CD16/32(low) CD34(+) CD55(-)
CMP CD55$^+$	common myeloid progenitor CD55 positive	bone marrow	Lin(-) c-Kit(+) Sca-1(-) CD16/32(low) CD34(+) CD55(+)
GMP	granulocyte/macrophage progenitor	bone marrow	Lin(-) c-Kit(+) Sca-1(-) CD16/32(hi) CD34(+)
MEP	megakaryocyte/erythrocyte progenitor	bone marrow	Lin(-) c-Kit(+) Sca-1(-) CD16/32(low) CD34(-)
CLP	common lymphoid progenitor	bone marrow	Lin(-) c-Kit(mid) Sca-1(mid) CD127(+)
preMegE	pre-megakaryocyte/erythrocyte progenitor	bone marrow	Lin(-) c-Kit(+) Sca-1(-) CD150(+) CD41(+)
MkP	megakaryocyte progenitor	bone marrow	Lin(-) c-Kit(+) Sca-1(-) CD150(+) CD41(-) CD105(-)
CFU-E	colony-forming unit-erythroid cells	bone marrow	Lin(-) c-Kit(+) Sca-1(-) CD150(-) CD41(+) CD105(+)
MDP	monocyte-dendritic cell progenitor	bone marrow	Lin(-) c-Kit(+) CD135(+) CD115(+)
CDP	common dendritic cell progenitor	bone marrow	Lin(-) c-Kit(low/int) CD135(+) CD115(+)
cMoP	common monocyte progenitor	bone marrow	Lin(-) c-Kit(high) CD135(-) CD115(+) CD11b(-) Ly6C(+)
monocyte	monocyte	bone marrow	CD11b(+) Ly6C(hi)
neutrophil	neutrophil	bone marrow	CD11b(+) Ly6G(+)
eosinophil	eosinophil	bone marrow	CD11b(+) Ly6G(-) SiglecF(+)
cDC1	conventional type 1 dendritic cells	spleen	CD11c(hi) MHCII(+) CD11b(-) CD8a(+)
cDC2	conventional type 2 dendritic cells	spleen	CD11c(hi) MHCII(+) CD11b(+) CD8a(-)
pDC	plasmacytoid dendritic cell	spleen	CD11c(mid) PDCA(+)
B cell	B cell	bone marrow	B220(+)
T cell	T cell	bone marrow	CD4(+) CD8(+)

Table S4: Alignment statistics for the uniformly processed tagmentation-based whole-genome bisulfite sequencing data. Sequencing reads were aligned and non-properly paired reads as well as duplicate reads were discarded. The table details alignment statistics for each replicate of the 25 hematopoietic populations. Shown are the number of sequenced reads, the percentages of properly paired and duplicate reads, as well as the number and percentage of reads which were used for methylation calling (i.e., which passed alignment and read quality filtering).

Population	Replicate	Total number of reads	Number of used reads	Proper pairs (%)	Duplicates (%)	Used reads (%)
B cells	1	831423192	671040372	98.5	17.79	80.71
	2	761121148	502059090	97.42	31.46	65.96
	3	1808562416	913386973	98.28	47.78	50.5
CDP	1	1195292564	267522509	96.84	74.46	22.38
	2	1278536984	536583585	96.95	54.98	41.97
	3	1242720486	457981037	97.06	60.21	36.85
	4	1261803882	318097765	96.96	71.75	25.21
CFU-E	1	950480552	677304798	98.39	27.13	71.26
	2	1046551228	693388398	98.27	32.01	66.25
	3	917527402	616246197	98.59	31.42	67.16
CLP	1	973955604	564688584	98.16	40.18	57.98
	2	2215493248	1199017528	98.33	44.21	54.12
	3	913864712	397113914	98.14	54.68	43.45
CMP CD55$^+$	1	1039727722	420544842	97.94	57.49	40.45
	2	1179200110	613064855	96.44	44.45	51.99
	4	1276305762	646181720	96.04	45.41	50.63
	5	1288023820	758791759	95.44	36.53	58.91
	6	1378574370	671078536	94.93	46.25	48.68
CMP CD55$^-$	1	1082037830	519242750	97.85	49.86	47.99
	3	1142956790	750994860	97.99	32.29	65.71
	4	1294469752	683774473	96.09	43.27	52.82
Eosinophils	1	1400627838	694665201	96.04	46.44	49.6
	2	1371859344	586282756	96.53	53.79	42.74
	3	1230096396	470758403	96.47	58.2	38.27
GMP	1	1959483690	902573260	98.04	51.98	46.06
	2	1029242956	719214643	98.41	28.53	69.88
	4	1491196424	866545922	98.38	40.27	58.11
HSC	1	652419082	424610362	98.21	33.13	65.08
	2	1865615966	665561526	97.9	62.22	35.68
	3	1562987486	733603387	96.04	49.1	46.94
	4	685556166	409105014	96.19	36.52	59.67
MDP	1	1238731344	525348077	96.62	54.21	42.41
	2	1181609490	324335298	96.77	69.32	27.45
MEP	1	974337654	522924402	98.11	44.44	53.67
	2	1065330108	774886001	98.36	25.62	72.74
	3	2242955400	1425790876	98.06	34.5	63.57
MPP1	1	801210388	600523159	98.42	23.47	74.95
	2	1219334550	332683964	97.96	70.68	27.28
	3	1393529748	537257173	96.38	57.82	38.55
MPP2	1	1340694118	353562742	97.9	71.53	26.37
	2	1570802106	852400639	97.6	43.33	54.27
	3	1623055560	874086746	97.74	43.89	53.85
MPP3	1	1487478786	1114025931	98.21	23.32	74.89
	2	1516043846	1080071077	98.19	26.94	71.24
	3	1400549226	942100632	97.97	30.7	67.27
MPP4	1	1439293448	1149934148	98.5	18.6	79.9
	2	1454308136	1031533406	98.49	27.56	70.93

continued on next page

Table S4: continued

Population	Replicate	Total number of reads	Number of used reads	Proper pairs (%)	Duplicates (%)	Used reads (%)
	3	1367926332	1049108764	98.19	21.49	76.69
MPP5	1	825465982	458450977	97.45	41.91	55.54
	2	1413395096	1014666401	98.15	26.36	71.79
	3	1460232116	890965119	98.5	37.49	61.02
	4	1107946010	286591697	98.36	72.5	25.87
	5	701393666	567698725	98.39	17.46	80.94
MkP	1	899902444	486697447	97.98	43.9	54.08
	2	1038472438	672717274	97.89	33.11	64.78
	4	713973264	574214077	98.56	18.14	80.43
Monocytes	1	754485832	531389320	98.37	27.94	70.43
	2	885887984	687708083	98.51	20.88	77.63
	3	1001826598	732049809	98.53	25.46	73.07
Neutrophils	1	1388507998	496881878	96.49	60.7	35.79
	2	1039518370	218090226	96.17	75.19	20.98
	3	1115693220	367495202	96.16	63.22	32.94
T cells	1	912497756	508679011	97.39	41.65	55.75
	2	531534340	336776696	97.79	34.43	63.36
	3	324199408	139320350	97.39	54.42	42.97
	4	1199960356	664662881	98.52	43.13	55.39
	5	1162218390	651855229	98.57	42.48	56.09
cDC2	1	946764708	427846718	96.7	51.51	45.19
	2	961484750	166590296	97.12	79.79	17.33
	3	1364938782	358734450	96.48	70.2	26.28
cDC1	1	1005231122	440756526	96.61	52.76	43.85
	2	883139862	273514920	96.64	65.67	30.97
	3	1359692196	562766270	96.58	55.19	41.39
cMoP	1	1106899496	360033249	97.12	64.59	32.53
	2	1116251064	284288965	97.25	71.78	25.47
pDC	1	1072177778	461316855	97.03	54.0	43.03
	2	914564414	362784440	96.95	57.28	39.67
	3	1345969226	493063461	96.27	59.64	36.63
preMegE	1	986958544	620417397	98.34	35.48	62.86
	2	1147615282	621256854	98.17	44.03	54.13
	3	1049547834	340738257	98.4	65.93	32.47

Table S5: CpG methylation levels and CHH conversion rates for all T-WGBS replicates. The table lists the average autosomal CpG methylation levels for each replicate and the mean and standard deviation of the average replicate CpG methylation levels per population. The standard deviations of the replicate methylation levels within a population were between 0.04% and 0.79%. The table also lists the conversion rate of the autosomal CHHs. The mean and standard deviation of the CHH conversion rate across all replicates was 99.43 ± 0.21% (mean ± s.d.). Repl., replicate; Pop., population.

Population	Replicate	Repl. mean CpG meth. (%)	Pop. CpG meth. (%) (mean ± s.d.)	CHH conversion rate (%)
HSC	1	81.89	82.0 ± 0.1%	99.49
	2	82.08		99.42
	3	81.86		99.17
	4	82.01		99.56
MPP1	1	81.6	81.1 ± 0.6%	99.5
	2	80.45		99.16
	3	81.3		99.33
MPP5	1	81.67	81.9 ± 0.3%	99.7
	2	82.13		99.5
	3	81.89		99.55
	4	82.22		99.11
	5	81.57		99.46
MPP2	1	81.31	81.1 ± 0.2%	98.56
	2	80.88		99.29
	3	81.03		99.23
MPP3	1	81.22	81.20 ± 0.04%	99.67
	2	81.23		99.6
	3	81.3		99.55
MPP4	1	81.68	81.8 ± 0.1%	99.69
	2	81.74		99.67
	3	81.84		99.66
CMP CD55^{+}	1	79.33	79.2 ± 0.5%	99.58
	2	78.28		99.54
	3	79.41		99.23
	4	79.46		99.29
	5	79.4		99.24
preMegE	1	77.54	77.4 ± 0.1%	99.58
	2	77.44		99.59
	3	77.3		99.6
MkP	1	77.59	78.2 ± 0.5%	99.54
	2	78.21		99.59
	3	78.64		99.52
MEP	1	71.74	71.4 ± 0.6%	99.43
	2	70.66		99.65
	3	71.7		99.44
CFU-E	1	70.98	71.00 ± 0.06%	99.6
	2	70.97		99.58
	3	71.07		99.64
GMP	1	78.21	79.1 ± 0.8%	99.12
	2	79.7		99.59
	3	79.41		99.29
cMoP	1	78.38	78.8 ± 0.6%	99.32
	2	79.23		99.16
Monocytes	1	76.03	76.2 ± 0.2%	99.55
	2	76.41		99.61
	3	76.33		99.67
Neutrophils	1	76.96	77.2 ± 0.3%	99.64

continued on next page

Table S5: continued

Population	Replicate	Repl. mean CpG meth. (%)	Pop. CpG meth. (%) (mean ± s.d.)	CHH conversion rate (%)
	2	77.56		99.42
	3	77.16		99.57
Eosinophils	1	74.86	75.0 ± 0.2%	99.27
	2	75.05		99.29
	3	75.2		99.36
CMP CD55$^-$	1	80.48	80.6 ± 0.2%	99.05
	2	80.43		99.56
	3	80.88		99.2
MDP	1	80.7	80.9 ± 0.2%	99.17
	2	81.03		99.11
CDP	1	80.44	80.4 ± 0.1%	98.77
	2	80.45		99.35
	3	80.47		99.34
	4	80.2		99.14
cDC1	1	78.55	78.4 ± 0.2%	99.56
	2			99.5
	3	78.23		99.62
cDC2	1	78.99	78.5 ± 0.4%	99.58
	2	78.16		99.46
	3	78.33		99.59
pDC	1	79.54	79.1 ± 0.4%	99.61
	2	78.92		99.53
	3	78.74		99.61
CLP	1	80.53	80.4 ± 0.3%	99.53
	2	80.49		99.26
	3	80.04		99.44
B cells	1	78.84	79.1 ± 0.3%	99.61
	2	79.05		99.43
	3	79.41		99.25
T cells	1	78.8	78.7 ± 0.2%	99.5
	2	78.47		99.43
	3	79.04		99.44
	4	78.61		99.24
	5	78.59		99.27

Chapter 6

Bibliography

Publications

[OWN1] Melinda Czeh, Sina Stäble, **Stephen Krämer**, Lena Tepe, Sweta Talyan, Joana Carrelha, Yiran Meng, Barbara Heitplatz, Marius Schwabenland, Michael D. Milsom, Christoph Plass, Marco Prinz, Matthias Schlesner, Miguel A. Andrade-Navarro, Claus Nerlov, Sten Eirik W. Jacobsen, Daniel B. Lipka, and Frank Rosenbauer. “DNMT1 Deficiency Impacts on Plasmacytoid Dendritic Cells in Homeostasis and Autoimmune Disease.” In: *The Journal of Immunology* 208.2 (Jan. 2022), pp. 358–370. ISSN: 0022-1767. DOI: `10.4049/jimmunol.2100624`. URL: `https://doi.org/10.4049/jimmunol.2100624` (visited on 03/15/2023).

[OWN2] Maximilian Schönung, Mark Hartmann, **Stephen Krämer**, Sina Stäble, Mariam Hakobyan, Emely Kleinert, Theo Aurich, Defne Cobanoglu, Florian H. Heidel, Stefan Fröhling, Michael D. Milsom, Matthias Schlesner, Pavlo Lutsik, and Daniel B. Lipka. “Dynamic DNA Methylation Reveals Novel Cis-Regulatory Elements in Mouse Hematopoiesis.” In: *Experimental Hematology* 117 (Jan. 2023), 24–42.e7. ISSN: 0301-472X. DOI: `10.1016/j.exphem.2022.11.001`. URL: `https://www.sciencedirect.com/science/article/pii/S0301472X22008050` (visited on 03/15/2023).

[OWN3] Marina Scheller, Anne Kathrin Ludwig, Stefanie Göllner, Christian Rohde, **Stephen Krämer**, Sina Stäble, Maike Janssen, James-Arne Müller, Lixiazi He, Nicole Bäumer, Christian Arnold, Joachim Gerß, Maximilian Schönung, Christian Thiede, Christian Niederwieser, Dietger Niederwieser, Hubert Serve, Wolfgang E. Berdel, Ulrich Thiem, Inga Hemmerling, Florian Leuschner, Christoph Plass, Matthias Schlesner, Judith Zaugg, Michael D. Milsom, Andreas Trumpp, Caroline Pabst, Daniel B. Lipka, and Carsten Müller-Tidow. “Hotspot DNMT3A Mutations in Clonal Hematopoiesis and Acute Myeloid Leukemia Sensitize Cells to Azacytidine via Viral Mimicry Response.” In: *Nature Cancer* 2.5 (May 2021), pp. 527–544. ISSN: 2662-1347. DOI: `10.1038/s43018-021-00213-9`. URL: `https://www.nature.com/articles/s43018-021-00213-9` (visited on 04/22/2022).

[OWN4] Daniel Hüebschmann, Nils Kurzawa, Sebastian Steinhauser, Philipp Rentzsch, **Stephen Krämer**, Carolin Andresen, Jeongbin Park, Roland Eils, Matthias Schlesner, and Carl Herrmann. "Deciphering Programs of Transcriptional Regulation by Combined Deconvolution of Multiple Omics Layers." In: *bioRxiv* (Oct. 2017), p. 199547. DOI: `10.1101/199547`. URL: `https://www.biorxiv.org/content/10.1101/199547v1` (visited on 05/23/2019).

[OWN5] Andres Quintero, Daniel Hübschmann, Nils Kurzawa, Sebastian Steinhauser, Philipp Rentzsch, **Stephen Krämer**, Carolin Andresen, Jeongbin Park, Roland Eils, Matthias Schlesner, and Carl Herrmann. "ShinyButchR: Interactive NMF-based Decomposition Workflow of Genome-Scale Datasets." In: *Biology Methods and Protocols* 5.bpaa022 (Jan. 2020). ISSN: 2396-8923. DOI: `10.1093/biomethods/bpaa022`. URL: `https://doi.org/10.1093/biomethods/bpaa022` (visited on 04/22/2021).

[OWN6] Daniel Hübschmann, Lea Jopp-Saile, Carolin Andresen, **Stephen Krämer**, Zuguang Gu, Christoph E. Heilig, Simon Kreutzfeldt, Veronica Teleanu, Stefan Fröhling, Roland Eils, and Matthias Schlesner. "Analysis of Mutational Signatures with yet Another Package for Signature Analysis." In: *Genes, Chromosomes and Cancer* 60.5 (2021), pp. 314–331. ISSN: 1098-2264. DOI: `10.1002/gcc.22918`. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1002/gcc.22918` (visited on 05/15/2023).

[OWN7] Ani Grigoryan, Johannes Pospiech, **Stephen Krämer**, Daniel Lipka, Thomas Liehr, Hartmut Geiger, Hiroshi Kimura, Medhanie A. Mulaw, and Maria Carolina Florian. "Attrition of X Chromosome Inactivation in Aged Hematopoietic Stem Cells." In: *Stem Cell Reports* 16.4 (Apr. 2021), pp. 708–716. ISSN: 2213-6711. DOI: `10.1016/j.stemcr.2021.03.007`. URL: `https://www.sciencedirect.com/science/article/pii/S2213671121001375` (visited on 04/22/2021).

[OWN8] Ruzhica Bogeska, Ana-Matea Mikecin, Paul Kaschutnig, Malak Fawaz, Marleen Büchler-Schäff, Duy Le, Miguel Ganuza, Angelika Vollmer, Stella V. Paffenholz, Noboru Asada, Esther Rodriguez-Correa, Felix Frauhammer, Florian Buettner, Melanie Ball, Julia Knoch, Sina Stäble, Dagmar Walter, Amelie Petri, Martha J. Carreño-Gonzalez, Vinona Wagner, Benedikt Brors, Simon Haas, Daniel B. Lipka, Marieke A. G. Essers, Vivienn Weru, Tim Holland-Letz, Jan-Philipp Mallm, Karsten Rippe, **Stephan Krämer**, Matthias Schlesner, Shannon McKinney Freeman, Maria Carolina Florian, Katherine Y. King, Paul S. Frenette, Michael A. Rieger, and Michael D. Milsom. "Inflammatory Exposure Drives Long-Lived Impairment of Hematopoietic Stem Cell Self-Renewal Activity and Accelerated Aging." In: *Cell*

Stem Cell 29.8 (Aug. 2022), 1273–1284.e8. ISSN: 1934-5909. DOI: 10.1016/j.stem.2022.06.012. URL: https://www.sciencedirect.com/science/article/pii/S1934590922002612 (visited on 03/15/2023).

[OWN9] Felicitas Bossler, Bianca J. Kuhn, Thomas Günther, **Stephen J. Kraemer**, Prajakta Khalkar, Svenja Adrian, Claudia Lohrey, Angela Holzer, Mitsugu Shimobayashi, Matthias Dürst, Arnulf Mayer, Frank Rösl, Adam Grundhoff, Jeroen Krijgsveld, Karin Hoppe-Seyler, and Felix Hoppe-Seyler. "Repression of Human Papillomavirus Oncogene Expression under Hypoxia Is Mediated by PI3K/mTORC2/AKT Signaling." In: *mBio* 10.1 (Feb. 2019), e02323–18. ISSN: 2150-7511. DOI: 10.1128/mBio.02323-18. URL: https://mbio.asm.org/content/10/1/e02323-18 (visited on 05/23/2019).

[OWN10] Stefan Gröschel, Daniel Hübschmann, Francesco Raimondi, Peter Horak, Gregor Warsow, Martina Fröhlich, Barbara Klink, Laura Gieldon, Barbara Hutter, Kortine Kleinheinz, David Bonekamp, Oliver Marschal, Priya Chudasama, Jagoda Mika, Marie Groth, Sebastian Uhrig, **Stephen Krämer**, Christoph Heining, Christoph E. Heilig, Daniela Richter, Eva Reisinger, Katrin Pfütze, Roland Eils, Stephan Wolf, Christof von Kalle, Christian Brandts, Claudia Scholl, Wilko Weichert, Stephan Richter, Sebastian Bauer, Roland Penzel, Evelin Schröck, Albrecht Stenzinger, Richard F. Schlenk, Benedikt Brors, Robert B. Russell, Hanno Glimm, Matthias Schlesner, and Stefan Fröhling. "Defective Homologous Recombination DNA Repair as Therapeutic Target in Advanced Chordoma." In: *Nature Communications* 10.1 (Apr. 2019), p. 1635. ISSN: 2041-1723. DOI: 10.1038/s41467-019-09633-9. URL: https://www.nature.com/articles/s41467-019-09633-9 (visited on 05/23/2019).

Manuscripts

[PLANNED1] **Stephen Krämer**, Sina Stäble, Maximilian Schönung, Mark Hartmann, Jens Langstein, Ruzhica Bogeska, Melinda Czeh, Julia Knoch, Philipp Rentzsch, Charles Imbusch, Qi Wang, Matthias Bieg, Natasha Anstee, Julius Graesel, Lars Feuerbach, Weichenhan Dieter, Benedikt Brors, Karsten Rippe, Simon Haas, Jan-Philipp Mallm, Frank Rosenbauer, Daniel Hübschmann, Roland Eils, Christoph Plass, Matthias Schlesner, Michael D. Milsom, and Daniel B. Lipka. "Hierarchical DNA methylation programming during hematopoietic differentiation at single-CpG and single-cell resolution." (In preparation).

[PLANNED2] **Stephen Krämer**, Michael D. Milsom, Eils Roland, Daniel B. Lipka, and Matthias Schlesner. "Dissecting the rich but complex information content of the DNA methylome at CpG-resolution." (Planned).

[PLANNED3] **Stephen Krämer**, Eils Roland, and Matthias Schlesner. "Codaplot - flexible, multi-layered and modular complex heatmaps within the Python ecosystem." (In preparation).

[PLANNED4] **Stephen Krämer**, Eils Roland, and Matthias Schlesner. "locplot - genomic region plotting with tight matplotlib integration." (In preparation).

[PLANNED5] **Stephen Krämer**, Eils Roland, Daniel B. Lipka, and Matthias Schlesner. "Automated, multi-dimensional M-bias filtering with bistro." (Planned).

Software packages

[SOFT1] Stephen Krämer, Daniel B. Lipka, Roland Eils, and Matthias Schlesner. *Bistro*. URL: `https://github.com/stephenkraemer/bistro`.

[SOFT2] Stephen Krämer, Roland Eils, and Matthias Schlesner. *Methlevels*. URL: `https://github.com/stephenkraemer/methlevels`.

[SOFT3] Stephen Krämer, Roland Eils, and Matthias Schlesner. *Codaplot*. URL: `https://github.com/stephenkraemer/codaplot`.

[SOFT4] Stephen Krämer, Daniel B. Lipka, Roland Eils, and Matthias Schlesner. *Gtfanno*. URL: `https://github.com/stephenkraemer/gtfanno`.

[SOFT5] Stephen Krämer, Roland Eils, and Matthias Schlesner. *Smk_wgbs*. URL: `https://github.com/stephenkraemer/smk_wgbs`.

References

[1] Stephen J. Loughran et al. “Lineage Commitment of Hematopoietic Stem Cells and Progenitors: Insights from Recent Single Cell and Lineage Tracing Technologies.” In: *Experimental Hematology* 88 (Aug. 2020), pp. 1–6. ISSN: 0301-472X. DOI: 10.1016/j.exphem.2020.07.002. URL: https://www.sciencedirect.com/science/article/pii/S0301472X20302678 (visited on 03/08/2023).

[2] Ron Sender and Ron Milo. “The Distribution of Cellular Turnover in the Human Body.” In: *Nature Medicine* 27.1 (Jan. 2021), pp. 45–48. ISSN: 1546-170X. DOI: 10.1038/s41591-020-01182-9. URL: https://www.nature.com/articles/s41591-020-01182-9 (visited on 03/08/2023).

[3] Stuart H. Orkin and Leonard I. Zon. “Hematopoiesis: An Evolving Paradigm for Stem Cell Biology.” In: *Cell* 132.4 (Feb. 2008), pp. 631–644. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2008.01.025. URL: https://www.cell.com/cell/abstract/S0092-8674(08)00125-6 (visited on 03/08/2023).

[4] Gerald J. Spangrude, Shelly Heimfeld, and Irving L. Weissman. “Purification and Characterization of Mouse Hematopoietic Stem Cells.” In: *Science* 241.4861 (July 1988), pp. 58–62. DOI: 10.1126/science.2898810. URL: https://www.science.org/doi/10.1126/science.2898810 (visited on 03/10/2023).

[5] Jun Seita and Irving L. Weissman. “Hematopoietic Stem Cell: Self-Renewal versus Differentiation.” In: *Wiley Interdisciplinary Reviews. Systems Biology and Medicine* 2.6 (2010 Nov-Dec), pp. 640–653. ISSN: 1939-005X. DOI: 10.1002/wsbm.86.

[6] Motonari Kondo, Irving L. Weissman, and Koichi Akashi. “Identification of Clonogenic Common Lymphoid Progenitors in Mouse Bone Marrow.” In: *Cell* 91.5 (Nov. 1997), pp. 661–672. ISSN: 0092-8674. DOI: 10.1016/S0092-8674(00)80453-5. URL: http://www.sciencedirect.com/science/article/pii/S0092867400804535 (visited on 11/27/2019).

[7] K. Akashi et al. “A Clonogenic Common Myeloid Progenitor That Gives Rise to All Myeloid Lineages.” In: *Nature* 404.6774 (2000), pp. 193–197. DOI: 10.1038/35004599.

[8] Simon Haas, Andreas Trumpp, and Michael D. Milsom. “Causes and Consequences of Hematopoietic Stem Cell Heterogeneity.” In: *Cell Stem Cell* 22.5 (May 2018), pp. 627–638. ISSN: 1934-5909. DOI: 10.1016/j.stem.2018.04.003. URL: http://www.sciencedirect.com/science/article/pii/S1934590918301656 (visited on 07/08/2019).

[9] Cornelis Murre. “Defining the Pathways of Early Adult Hematopoiesis.” In: *Cell Stem Cell* 1.4 (Oct. 2007), pp. 357–358. ISSN: 1934-5909. DOI: 10.1016/j.stem.2007.09.008. URL: http://www.sciencedirect.com/science/article/pii/S1934590907001774 (visited on 11/25/2019).

[10] Jörgen Adolfsson et al. “Identification of Flt3+ Lympho-Myeloid Stem Cells Lacking Erythro-Megakaryocytic Potential: A Revised Road Map for Adult Blood Lineage Commitment.” In: *Cell* 121.2 (Apr. 2005), pp. 295–306. ISSN: 0092-8674. DOI: 10.1016/j.cell.2005.02.013. URL: http://www.sciencedirect.com/science/article/pii/S0092867405001583 (visited on 11/25/2019).

[11] Yojiro Arinobu et al. “Reciprocal Activation of GATA-1 and PU.1 Marks Initial Specification of Hematopoietic Stem Cells into Myeloerythroid and Myelolymphoid Lineages.” In: *Cell Stem Cell* 1.4 (Oct. 2007), pp. 416–427. ISSN: 1875-9777. DOI: 10.1016/j.stem.2007.07.004.

[12] Anne Wilson et al. “Hematopoietic Stem Cells Reversibly Switch from Dormancy to Self-Renewal during Homeostasis and Repair.” In: *Cell* 135.6 (Dec. 2008), pp. 1118–1129. ISSN: 0092-8674. DOI: 10.1016/j.cell.2008.10.048. URL: http://www.sciencedirect.com/science/article/pii/S009286740801386X (visited on 05/15/2019).

[13] Eric M. Pietras et al. “Functionally Distinct Subsets of Lineage-Biased Multipotent Progenitors Control Blood Production in Normal and Regenerative Conditions.” In: *Cell Stem Cell* 17.1 (July 2015), pp. 35–46. ISSN: 1875-9777. DOI: 10.1016/j.stem.2015.05.003.

[14] Nina Cabezas-Wallscheid et al. “Identification of Regulatory Networks in HSCs and Their Immediate Progeny via Integrated Proteome, Transcriptome, and DNA Methylome Analysis.” In: *Cell Stem Cell* 15.4 (Oct. 2014), pp. 507–522. ISSN: 1934-5909. DOI: 10.1016/j.stem.2014.07.005. URL: http://www.sciencedirect.com/science/article/pii/S1934590914003014 (visited on 05/15/2019).

[15] Pia Sommerkamp et al. “Mouse Multipotent Progenitor 5 Cells Are Located at the Interphase between Hematopoietic Stem and Progenitor Cells.” In:

Blood 137.23 (June 2021), pp. 3218–3224. ISSN: 1528-0020. DOI: 10.1182/blood.2020007876.

[16] J. Adolfsson et al. “Upregulation of Flt3 Expression within the Bone Marrow Lin(-)Sca1(+)c-Kit(+) Stem Cell Compartment Is Accompanied by Loss of Self-Renewal Capacity.” In: *Immunity* 15.4 (Oct. 2001), pp. 659–669. ISSN: 1074-7613. DOI: 10.1016/s1074-7613(01)00220-5.

[17] S. Okada et al. “In Vivo and in Vitro Stem Cell Function of C-Kit- and Sca-1-positive Murine Hematopoietic Cells.” In: *Blood* 80.12 (Dec. 1992), pp. 3044–3050. ISSN: 0006-4971.

[18] K. Ikuta and I. L. Weissman. “Evidence That Hematopoietic Stem Cells Express Mouse C-Kit but Do Not Depend on Steel Factor for Their Generation.” In: *Proceedings of the National Academy of Sciences of the United States of America* 89.4 (Feb. 1992), pp. 1502–1506. ISSN: 0027-8424. DOI: 10.1073/pnas.89.4.1502.

[19] Eric M. Pietras. “Inflammation: A Key Regulator of Hematopoietic Stem Cell Fate in Health and Disease.” In: *Blood* 130.15 (Oct. 2017), pp. 1693–1698. ISSN: 0006-4971. DOI: 10.1182/blood-2017-06-780882. URL: https://doi.org/10.1182/blood-2017-06-780882 (visited on 04/02/2023).

[20] Alejo E. Rodriguez-Fraticelli et al. “Clonal Analysis of Lineage Fate in Native Haematopoiesis.” In: *Nature* 553.7687 (Jan. 2018), pp. 212–216. ISSN: 1476-4687. DOI: 10.1038/nature25168. URL: https://www.nature.com/articles/nature25168 (visited on 12/02/2019).

[21] Motonari Kondo, Irving L. Weissman, and Koichi Akashi. “Identification of Clonogenic Common Lymphoid Progenitors in Mouse Bone Marrow.” In: *Cell* 91.5 (Nov. 1997), pp. 661–672. ISSN: 0092-8674. DOI: 10.1016/S0092-8674(00)80453-5. URL: http://www.sciencedirect.com/science/article/pii/S0092867400804535 (visited on 11/25/2019).

[22] Guoji Guo et al. “Mapping Cellular Hierarchy by Single-Cell Analysis of the Cell Surface Repertoire.” In: *Cell Stem Cell* 13.4 (Oct. 2013), pp. 492–505. ISSN: 1934-5909. DOI: 10.1016/j.stem.2013.07.017. URL: http://www.sciencedirect.com/science/article/pii/S1934590913003627 (visited on 11/24/2019).

[23] Cornelis J. H. Pronk et al. “Elucidation of the Phenotypic, Functional, and Molecular Topography of a Myeloerythroid Progenitor Cell Hierarchy.” In: *Cell Stem Cell* 1.4 (Oct. 2007), pp. 428–442. ISSN: 1934-5909. DOI: 10.1016/j.stem.2007.07.005. URL: http://www.sciencedirect.com/science/article/pii/S1934590907000719 (visited on 01/04/2021).

[24] Kang Liu et al. "In Vivo Analysis of Dendritic Cell Development and Homeostasis." In: *Science (New York, N.Y.)* 324.5925 (Apr. 2009), pp. 392–397. ISSN: 0036-8075. DOI: 10.1126/science.1170540. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2803315/ (visited on 01/04/2021).

[25] Shalin H. Naik et al. "Development of Plasmacytoid and Conventional Dendritic Cell Subtypes from Single Precursor Cells Derived in Vitro and in Vivo." In: *Nature Immunology* 8.11 (Nov. 2007), pp. 1217–1226. ISSN: 1529-2916. DOI: 10.1038/ni1522. URL: https://www.nature.com/articles/ni1522 (visited on 04/04/2023).

[26] Nobuyuki Onai et al. "Identification of Clonogenic Common Flt3+M-CSFR+ Plasmacytoid and Conventional Dendritic Cell Progenitors in Mouse Bone Marrow." In: *Nature Immunology* 8.11 (Nov. 2007), pp. 1207–1216. ISSN: 1529-2908. DOI: 10.1038/ni1518.

[27] Nobuyuki Onai, Markus G. Manz, and Michael A. Schmid. "Isolation of Common Dendritic Cell Progenitors (CDP) from Mouse Bone Marrow." In: *Methods in Molecular Biology (Clifton, N.J.)* 595 (2010), pp. 195–203. ISSN: 1940-6029. DOI: 10.1007/978-1-60761-421-0_13.

[28] Boris Reizis. "Regulation of Plasmacytoid Dendritic Cell Development." In: *Current Opinion in Immunology*. Lymphocyte Development Tumour Immunology 22.2 (Apr. 2010), pp. 206–211. ISSN: 0952-7915. DOI: 10.1016/j.coi.2010.01.005. URL: http://www.sciencedirect.com/science/article/pii/S0952791510000063 (visited on 11/01/2018).

[29] Franziska Paul et al. "Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors." In: *Cell* 163.7 (Dec. 2015), pp. 1663–1677. ISSN: 0092-8674. DOI: 10.1016/j.cell.2015.11.013. URL: http://www.sciencedirect.com/science/article/pii/S0092867415014932 (visited on 11/27/2019).

[30] Franziska Paul et al. "Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors." In: *Cell* 163.7 (Dec. 2015), pp. 1663–1677. ISSN: 0092-8674. DOI: 10.1016/j.cell.2015.11.013. URL: https://www.sciencedirect.com/science/article/pii/S0092867415014932 (visited on 03/22/2023).

[31] Lars Velten et al. "Human Haematopoietic Stem Cell Lineage Commitment Is a Continuous Process." In: *Nature Cell Biology* 19.4 (Apr. 2017), p. 271. ISSN: 1476-4679. DOI: 10.1038/ncb3493. URL: https://www.nature.com/articles/ncb3493 (visited on 05/15/2019).

[32] Faiyaz Notta et al. "Distinct Routes of Lineage Development Reshape the Human Blood Hierarchy across Ontogeny." In: *Science* 351.6269 (Jan. 2016).

ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aab2116. URL: https://science.sciencemag.org/content/351/6269/aab2116 (visited on 11/27/2019).

[33] Leïla Perié et al. "The Branching Point in Erythro-Myeloid Differentiation." In: *Cell* 163.7 (Dec. 2015), pp. 1655–1662. ISSN: 0092-8674. DOI: 10.1016/j.cell.2015.11.059. URL: http://www.sciencedirect.com/science/article/pii/S0092867415016219 (visited on 11/27/2019).

[34] Louise E. Purton and David T. Scadden. "Limiting Factors in Murine Hematopoietic Stem Cell Assays." In: *Cell Stem Cell* 1.3 (Sept. 2007), pp. 263–270. ISSN: 1875-9777. DOI: 10.1016/j.stem.2007.08.016.

[35] Daniel E. Wagner and Allon M. Klein. "Lineage Tracing Meets Single-Cell Omics: Opportunities and Challenges." In: *Nature Reviews Genetics* 21.7 (July 2020), pp. 410–427. ISSN: 1471-0064. DOI: 10.1038/s41576-020-0223-2. URL: https://www.nature.com/articles/s41576-020-0223-2 (visited on 04/05/2023).

[36] Ryo Yamamoto et al. "Clonal Analysis Unveils Self-Renewing Lineage-Restricted Progenitors Generated Directly from Hematopoietic Stem Cells." In: *Cell* 154.5 (Aug. 2013), pp. 1112–1126. ISSN: 0092-8674. DOI: 10.1016/j.cell.2013.08.007. URL: http://www.sciencedirect.com/science/article/pii/S0092867413009641 (visited on 11/25/2019).

[37] Brad Dykstra et al. "Long-Term Propagation of Distinct Hematopoietic Differentiation Programs in Vivo." In: *Cell Stem Cell* 1.2 (Aug. 2007), pp. 218–229. ISSN: 1875-9777. DOI: 10.1016/j.stem.2007.05.015.

[38] Yohei Morita, Hideo Ema, and Hiromitsu Nakauchi. "Heterogeneity and Hierarchy within the Most Primitive Hematopoietic Stem Cell Compartment." In: *The Journal of Experimental Medicine* 207.6 (June 2010), pp. 1173–1182. ISSN: 1540-9538. DOI: 10.1084/jem.20091318.

[39] Christa E. Muller-Sieburg et al. "Myeloid-Biased Hematopoietic Stem Cells Have Extensive Self-Renewal Capacity but Generate Diminished Lymphoid Progeny with Impaired IL-7 Responsiveness." In: *Blood* 103.11 (June 2004), pp. 4111–4118. ISSN: 0006-4971. DOI: 10.1182/blood-2003-10-3448.

[40] Christa E. Müller-Sieburg et al. "Deterministic Regulation of Hematopoietic Stem Cell Self-Renewal and Differentiation." In: *Blood* 100.4 (Aug. 2002), pp. 1302–1309. ISSN: 0006-4971.

[41] Fuwei Shang and Hans-Reimer Rodewald. "Toward the Dissection of Hematopoietic Stem Cell Fates and Their Determinants." In: *Current Opinion in Genetics & Development* 75 (Aug. 2022), p. 101945. ISSN: 0959-437X. DOI: 10.1016/j.gde.2022.101945. URL: https://www.scienced

irect.com/science/article/pii/S0959437X22000545 (visited on 04/05/2023).

[42] Katrin Busch et al. “Fundamental Properties of Unperturbed Haematopoiesis from Stem Cells *in Vivo*.” In: *Nature* 518.7540 (Feb. 2015), pp. 542–546. ISSN: 1476-4687. DOI: 10.1038/nature14242. URL: https://www.nature.com/articles/nature14242 (visited on 05/15/2019).

[43] Elisa Laurenti and Berthold Göttgens. “From Haematopoietic Stem Cells to Complex Differentiation Landscapes.” In: *Nature* 553.7689 (Jan. 2018), pp. 418–426. ISSN: 1476-4687. DOI: 10.1038/nature25022. URL: https://www.nature.com/articles/nature25022 (visited on 04/05/2023).

[44] Sten Eirik W. Jacobsen and Claus Nerlov. “Haematopoiesis in the Era of Advanced Single-Cell Technologies.” In: *Nature Cell Biology* 21.1 (Jan. 2019), pp. 2–8. ISSN: 1476-4679. DOI: 10.1038/s41556-018-0227-8. URL: https://www.nature.com/articles/s41556-018-0227-8 (visited on 04/05/2023).

[45] Amir Giladi et al. “Single-Cell Characterization of Haematopoietic Progenitors and Their Trajectories in Homeostasis and Perturbed Haematopoiesis.” In: *Nature Cell Biology* 20.7 (July 2018), p. 836. ISSN: 1476-4679. DOI: 10.1038/s41556-018-0121-4. URL: https://www.nature.com/articles/s41556-018-0121-4 (visited on 05/12/2019).

[46] Caleb Weinreb et al. “Lineage Tracing on Transcriptional Landscapes Links State to Fate during Differentiation.” In: *Science* 367.6479 (Feb. 2020), eaaw3381. DOI: 10.1126/science.aaw3381. URL: https://www.science.org/doi/10.1126/science.aaw3381 (visited on 04/05/2023).

[47] Weike Pei et al. “Resolving Fates and Single-Cell Transcriptomes of Hematopoietic Stem Cell Clones by PolyloxExpress Barcoding.” In: *Cell Stem Cell* 27.3 (Sept. 2020), 383–395.e8. ISSN: 1934-5909. DOI: 10.1016/j.stem.2020.07.018. URL: https://www.sciencedirect.com/science/article/pii/S1934590920303568 (visited on 04/07/2023).

[48] Carrie Deans and Keith A. Maggert. “What Do You Mean, “Epigenetic”?” In: *Genetics* 199.4 (Apr. 2015), pp. 887–896. ISSN: 0016-6731. DOI: 10.1534/genetics.114.173492. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4391566/ (visited on 04/05/2023).

[49] Dirk Schübeler. “Function and Information Content of DNA Methylation.” In: *Nature* 517.7534 (Jan. 2015), pp. 321–326. ISSN: 1476-4687. DOI: 10.1038/nature14192. URL: https://www.nature.com/articles/nature14192 (visited on 07/05/2019).

[50] Daria Shlyueva, Gerald Stampfel, and Alexander Stark. “Transcriptional Enhancers: From Properties to Genome-Wide Predictions.” In: *Nature Reviews*

Genetics 15.4 (Apr. 2014), pp. 272–286. ISSN: 1471-0064. DOI: 10.1038/nrg3682. URL: https://www.nature.com/articles/nrg3682 (visited on 04/12/2023).

[51] Hideyuki Yoshida et al. "The Cis-Regulatory Atlas of the Mouse Immune System." In: *Cell* 176.4 (Feb. 2019), 897–912.e20. ISSN: 0092-8674. DOI: 10.1016/j.cell.2018.12.036. URL: https://www.sciencedirect.com/science/article/pii/S0092867418316507 (visited on 04/09/2023).

[52] Dario Nicetto and Kenneth S. Zaret. "Role of H3K9me3 Heterochromatin in Cell Identity Establishment and Maintenance." In: *Current Opinion in Genetics & Development*. Genome Architecture and Expression 55 (Apr. 2019), pp. 1–10. ISSN: 0959-437X. DOI: 10.1016/j.gde.2019.04.013. URL: https://www.sciencedirect.com/science/article/pii/S0959437X19300127 (visited on 04/12/2023).

[53] Christian Beisel and Renato Paro. "Silencing Chromatin: Comparing Modes and Mechanisms." In: *Nature Reviews Genetics* 12.2 (Feb. 2011), pp. 123–135. ISSN: 1471-0064. DOI: 10.1038/nrg2932. URL: https://www.nature.com/articles/nrg2932 (visited on 04/12/2023).

[54] Ana Pombo and Niall Dillon. "Three-Dimensional Genome Architecture: Players and Mechanisms." In: *Nature Reviews Molecular Cell Biology* 16.4 (Apr. 2015), pp. 245–257. ISSN: 1471-0080. DOI: 10.1038/nrm3965. URL: https://www.nature.com/articles/nrm3965 (visited on 04/12/2023).

[55] Shaohui Hu et al. "DNA Methylation Presents Distinct Binding Sites for Human Transcription Factors." In: *eLife* 2 (Sept. 2013). Ed. by Danny Reinberg, e00726. ISSN: 2050-084X. DOI: 10.7554/eLife.00726. URL: https://doi.org/10.7554/eLife.00726 (visited on 07/12/2019).

[56] Yimeng Yin et al. "Impact of Cytosine Methylation on DNA Binding Specificities of Human Transcription Factors." In: *Science* 356.6337 (May 2017), eaaj2239. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aaj2239. URL: https://science.sciencemag.org/content/356/6337/eaaj2239 (visited on 07/12/2019).

[57] X. Nan et al. "Transcriptional Repression by the Methyl-CpG-binding Protein MeCP2 Involves a Histone Deacetylase Complex." In: *Nature* 393.6683 (May 1998), pp. 386–389. ISSN: 0028-0836. DOI: 10.1038/30764.

[58] Aurélien A. Sérandour et al. "Epigenetic Switch Involved in Activation of Pioneer Factor FOXA1-dependent Enhancers." In: *Genome Research* 21.4 (Apr. 2011), pp. 555–565. ISSN: 1088-9051. DOI: 10.1101/gr.111534.110. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3065703/ (visited on 04/05/2023).

[59] Julie Dubois-Chevalier et al. “The Ubiquitous Transcription Factor CTCF Promotes Lineage-Specific Epigenomic Remodeling and Establishment of Transcriptional Networks Driving Cell Differentiation.” In: *Nucleus* 6.1 (Jan. 2015), pp. 15–18. ISSN: 1949-1034. DOI: 10.1080/19491034.2015.1004258. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4615151/ (visited on 04/05/2023).

[60] Makiko Iwafuchi-Doi and Kenneth S. Zaret. “Cell Fate Control by Pioneer Transcription Factors.” In: *Development (Cambridge, England)* 143.11 (June 2016), pp. 1833–1837. ISSN: 1477-9129. DOI: 10.1242/dev.133900.

[61] Peter A. Jones. “Functions of DNA Methylation: Islands, Start Sites, Gene Bodies and Beyond.” In: *Nature Reviews Genetics* 13.7 (July 2012), pp. 484–492. ISSN: 1471-0064. DOI: 10.1038/nrg3230. URL: https://www.nature.com/articles/nrg3230 (visited on 07/01/2019).

[62] Lisa D. Moore, Thuc Le, and Guoping Fan. “DNA Methylation and Its Basic Function.” In: *Neuropsychopharmacology* 38.1 (Jan. 2013), pp. 23–38. ISSN: 1740-634X. DOI: 10.1038/npp.2012.112. URL: https://www.nature.com/articles/npp2012112 (visited on 04/10/2023).

[63] Alika K. Maunakea et al. “Conserved Role of Intragenic DNA Methylation in Regulating Alternative Promoters.” In: *Nature* 466.7303 (July 2010), pp. 253–257. ISSN: 1476-4687. DOI: 10.1038/nature09165. URL: https://www.nature.com/articles/nature09165 (visited on 04/10/2023).

[64] Vionnie W. C. Yu et al. “Epigenetic Memory Underlies Cell-Autonomous Heterogeneous Behavior of Hematopoietic Stem Cells.” In: *Cell* 167.5 (Nov. 2016), 1310–1322.e17. ISSN: 0092-8674. DOI: 10.1016/j.cell.2016.10.045. URL: http://www.sciencedirect.com/science/article/pii/S0092867416314660 (visited on 05/15/2019).

[65] Mirang Kim and Joseph Costello. “DNA Methylation: An Epigenetic Mark of Cellular Memory.” In: *Experimental & Molecular Medicine* 49.4 (Apr. 2017), e322–e322. ISSN: 2092-6413. DOI: 10.1038/emm.2017.10. URL: https://www.nature.com/articles/emm201710 (visited on 04/10/2023).

[66] Bérengère de Laval et al. “C/EBPβ-Dependent Epigenetic Memory Induces Trained Immunity in Hematopoietic Stem Cells.” In: *Cell Stem Cell* 26.5 (May 2020), 657–674.e8. ISSN: 1934-5909. DOI: 10.1016/j.stem.2020.01.017. URL: https://www.sciencedirect.com/science/article/pii/S1934590920300175 (visited on 04/10/2023).

[67] Ryan Lister et al. “Human DNA Methylomes at Base Resolution Show Widespread Epigenomic Differences.” In: *Nature* 462.7271 (Nov. 2009), pp. 315–322. ISSN: 1476-4687. DOI: 10.1038/nature08514. URL: https://www.nature.com/articles/nature08514 (visited on 07/12/2019).

[68] Qi Wang et al. “Tagmentation-Based Whole-Genome Bisulfite Sequencing.” In: *Nature Protocols* 8.10 (Oct. 2013), pp. 2022–2032. ISSN: 1750-2799. DOI: 10.1038/nprot.2013.118.

[69] Andrew Adey and Jay Shendure. “Ultra-Low-Input, Tagmentation-Based Whole-Genome Bisulfite Sequencing.” In: *Genome Research* 22.6 (June 2012), pp. 1139–1143. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.136242.111. URL: https://genome.cshlp.org/content/22/6/1139 (visited on 04/10/2023).

[70] 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. ISSN: 1362-4962. DOI: 10.1093/nar/gki901.

[71] Fumihito Miura et al. “Amplification-Free Whole-Genome Bisulfite Sequencing by Post-Bisulfite Adaptor Tagging.” In: *Nucleic Acids Research* 40.17 (Sept. 2012), e136. ISSN: 1362-4962. DOI: 10.1093/nar/gks454.

[72] Sébastien A. Smallwood et al. “Single-Cell Genome-Wide Bisulfite Sequencing for Assessing Epigenetic Heterogeneity.” In: *Nature Methods* 11.8 (Aug. 2014), pp. 817–820. ISSN: 1548-7105. DOI: 10.1038/nmeth.3035. URL: https://www.nature.com/articles/nmeth.3035 (visited on 08/13/2019).

[73] M. Farlik et al. “Single-Cell DNA Methylome Sequencing and Bioinformatic Inference of Epigenomic Cell-State Dynamics.” In: *Cell Reports* 10.8 (2015), pp. 1386–1397. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2015.02.001.

[74] Stephen J. Clark et al. “Genome-Wide Base-Resolution Mapping of DNA Methylation in Single Cells Using Single-Cell Bisulfite Sequencing (scBS-seq).” In: *Nature Protocols* 12.3 (Mar. 2017), pp. 534–547. ISSN: 1750-2799. DOI: 10.1038/nprot.2016.187. URL: https://www.nature.com/articles/nprot.2016.187 (visited on 03/04/2019).

[75] Tony Hui et al. “High-Resolution Single-Cell DNA Methylation Measurements Reveal Epigenetically Distinct Hematopoietic Stem Cell Subpopulations.” In: *Stem Cell Reports* 11.2 (Aug. 2018), pp. 578–592. ISSN: 2213-6711. DOI: 10.1016/j.stemcr.2018.07.003. URL: http://www.sciencedirect.com/science/article/pii/S2213671118303084 (visited on 05/15/2019).

[76] Stephen J. Clark et al. “scNMT-seq Enables Joint Profiling of Chromatin Accessibility DNA Methylation and Transcription in Single Cells.” In: *Nature Communications* 9.1 (Feb. 2018), p. 781. ISSN: 2041-1723. DOI: 10.1038/s41467-018-03149-4. URL: https://www.nature.com/articles/s41467-018-03149-4 (visited on 07/22/2019).

[77] Simone Picelli et al. “Full-Length RNA-seq from Single Cells Using Smart-seq2.” In: *Nature Protocols* 9.1 (Jan. 2014), pp. 171–181. ISSN: 1750-2799. DOI: 10.1038/nprot.2014.006. URL: https://www.nature.com/articles/nprot.2014.006 (visited on 04/11/2023).

[78] Jose Ramon Hernandez Mora et al. “Single-Cell Multi-Omic Analysis Profiles Defective Genome Activation and Epigenetic Reprogramming Associated with Human Pre-Implantation Embryo Arrest.” In: *Cell Reports* 42.2 (Feb. 2023), p. 112100. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2023.112100. URL: https://www.sciencedirect.com/science/article/pii/S2211124723001110 (visited on 04/10/2023).

[79] Ricard Argelaguet et al. “Multi-Omics Profiling of Mouse Gastrulation at Single-Cell Resolution.” In: *Nature* 576.7787 (Dec. 2019), pp. 487–491. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1825-8. URL: https://www.nature.com/articles/s41586-019-1825-8 (visited on 04/10/2023).

[80] Agostina Bianchi et al. “scTAM-seq Enables Targeted High-Confidence Analysis of DNA Methylation in Single Cells.” In: *Genome Biology* 23.1 (Oct. 2022), p. 229. ISSN: 1474-760X. DOI: 10.1186/s13059-022-02796-7. URL: https://doi.org/10.1186/s13059-022-02796-7 (visited on 04/10/2023).

[81] Kasper D. Hansen, Benjamin Langmead, and Rafael A. Irizarry. “BSmooth: From Whole Genome Bisulfite Sequencing Reads to Differentially Methylated Regions.” In: *Genome Biology* 13.10 (Oct. 2012), R83. ISSN: 1474-760X. DOI: 10.1186/gb-2012-13-10-r83. URL: https://doi.org/10.1186/gb-2012-13-10-r83 (visited on 04/10/2022).

[82] Hao Feng, Karen N. Conneely, and Hao Wu. “A Bayesian Hierarchical Model to Detect Differentially Methylated Loci from Single Nucleotide Resolution Sequencing Data.” In: *Nucleic Acids Research* 42.8 (Apr. 2014), e69. ISSN: 0305-1048. DOI: 10.1093/nar/gku154. URL: https://doi.org/10.1093/nar/gku154 (visited on 03/09/2023).

[83] Yongseok Park and Hao Wu. “Differential Methylation Analysis for BS-seq Data under General Experimental Design.” In: *Bioinformatics* 32.10 (May 2016), pp. 1446–1453. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw026. URL: https://doi.org/10.1093/bioinformatics/btw026 (visited on 05/03/2022).

[84] Altuna Akalin et al. “methylKit: A Comprehensive R Package for the Analysis of Genome-Wide DNA Methylation Profiles.” In: *Genome Biology* 13.10 (2012), R87. ISSN: 1465-6906. DOI: 10.1186/gb-2012-13-10-r87. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3491415/ (visited on 04/06/2023).

[85] Deqiang Sun et al. "MOABS: Model Based Analysis of Bisulfite Sequencing Data." In: *Genome Biology* 15.2 (2014), R38. DOI: 10.1186/gb-2014-15-2-r38. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4054608/ (visited on 04/06/2023).

[86] Zachary D. Smith and Alexander Meissner. "DNA Methylation: Roles in Mammalian Development." In: *Nature Reviews Genetics* 14.3 (Mar. 2013), pp. 204–220. ISSN: 1471-0064. DOI: 10.1038/nrg3354. URL: https://www.nature.com/articles/nrg3354 (visited on 04/06/2023).

[87] Keegan Korthauer et al. "Detection and Accurate False Discovery Rate Control of Differentially Methylated Regions from Whole Genome Bisulfite Sequencing." In: *Biostatistics* 20.3 (July 2019), pp. 367–383. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxy007. URL: https://doi.org/10.1093/biostatistics/kxy007 (visited on 04/25/2022).

[88] Mark D. Robinson et al. "Statistical Methods for Detecting Differentially Methylated Loci and Regions." In: *Frontiers in Genetics* 5 (2014). ISSN: 1664-8021. URL: https://www.frontiersin.org/article/10.3389/fgene.2014.00324 (visited on 04/25/2022).

[89] Frank Jühling et al. "Metilene: Fast and Sensitive Calling of Differentially Methylated Regions from Bisulfite Sequencing Data." In: *Genome Research* 26.2 (Feb. 2016), pp. 256–262. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.196394.115. URL: https://genome.cshlp.org/content/26/2/256 (visited on 03/19/2023).

[90] Sébastien A. Smallwood et al. "Single-Cell Genome-Wide Bisulfite Sequencing for Assessing Epigenetic Heterogeneity." In: *Nature Methods* 11.8 (Aug. 2014), pp. 817–820. ISSN: 1548-7105. DOI: 10.1038/nmeth.3035. URL: https://www.nature.com/articles/nmeth.3035 (visited on 04/06/2023).

[91] Christof Angermueller et al. "Parallel Single-Cell Sequencing Links Transcriptional and Epigenetic Heterogeneity." In: *Nature Methods* 13.3 (Mar. 2016), pp. 229–232. ISSN: 1548-7105. DOI: 10.1038/nmeth.3728. URL: https://www.nature.com/articles/nmeth.3728 (visited on 04/06/2023).

[92] Chongyuan Luo et al. "Single-Cell Methylomes Identify Neuronal Subtypes and Regulatory Elements in Mammalian Cortex." In: *Science* 357.6351 (Aug. 2017), pp. 600–604. DOI: 10.1126/science.aan3351. URL: https://www.science.org/doi/10.1126/science.aan3351 (visited on 04/06/2023).

[93] Matthias Farlik et al. "DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation." In: *Cell Stem Cell* 19.6 (Dec. 2016), pp. 808–822. ISSN: 1875-9777. DOI: 10.1016/j.stem.2016.10.019.

[94] Youjin Hu et al. "Simultaneous Profiling of Transcriptome and DNA Methylome from a Single Cell." In: *Genome Biology* 17.1 (May 2016), p. 88. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0950-z. URL: https://doi.org/10.1186/s13059-016-0950-z (visited on 04/06/2023).

[95] Anna Danese et al. "EpiScanpy: Integrated Single-Cell Epigenomic Analysis." In: *Nature Communications* 12.1 (Sept. 2021), p. 5228. ISSN: 2041-1723. DOI: 10.1038/s41467-021-25131-3. URL: https://www.nature.com/articles/s41467-021-25131-3 (visited on 04/06/2023).

[96] Qi Tian et al. "scMelody: An Enhanced Consensus-Based Clustering Model for Single-Cell Methylation Data by Reconstructing Cell-to-Cell Similarity." In: *Frontiers in Bioengineering and Biotechnology* 10 (2022). ISSN: 2296-4185. URL: https://www.frontiersin.org/articles/10.3389/fbioe.2022.842019 (visited on 04/11/2023).

[97] Christof Angermueller et al. "DeepCpG: Accurate Prediction of Single-Cell DNA Methylation States Using Deep Learning." In: *Genome Biology* 18.1 (Apr. 2017), p. 67. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1189-z. URL: https://doi.org/10.1186/s13059-017-1189-z (visited on 04/11/2023).

[98] ChantrioInt-Andreas Kapourani and Guido Sanguinetti. "Melissa: Bayesian Clustering and Imputation of Single-Cell Methylomes." In: *Genome Biology* 20.1 (Mar. 2019), p. 61. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1665-8. URL: https://doi.org/10.1186/s13059-019-1665-8 (visited on 04/11/2023).

[99] ChantrioInt-Andreas Kapourani et al. "scMET: Bayesian Modeling of DNA Methylation Heterogeneity at Single-Cell Resolution." In: *Genome Biology* 22.1 (Apr. 2021), p. 114. ISSN: 1474-760X. DOI: 10.1186/s13059-021-02329-8. URL: https://doi.org/10.1186/s13059-021-02329-8 (visited on 04/10/2023).

[100] Camila P. E. de Souza et al. "Epiclomal: Probabilistic Clustering of Sparse Single-Cell DNA Methylation Data." In: *PLOS Computational Biology* 16.9 (Sept. 2020), e1008270. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1008270. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008270 (visited on 04/06/2023).

[101] Gaetan De Waele et al. "CpG Transformer for Imputation of Single-Cell Methylomes." In: *Bioinformatics* 38.3 (Jan. 2022), pp. 597–603. ISSN: 1367-

4803. DOI: 10.1093/bioinformatics/btab746. URL: https://doi.org/10.1093/bioinformatics/btab746 (visited on 04/11/2023).

[102] Jianxiong Tang et al. "CaMelia: Imputation in Single-Cell Methylomes Based on Local Similarities between Cells." In: *Bioinformatics* 37.13 (July 2021), pp. 1814–1820. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btab029. URL: https://doi.org/10.1093/bioinformatics/btab029 (visited on 04/11/2023).

[103] M. Ryan Corces et al. "Lineage-Specific and Single-Cell Chromatin Accessibility Charts Human Hematopoiesis and Leukemia Evolution." In: *Nature Genetics* 48.10 (Oct. 2016), pp. 1193–1203. ISSN: 1546-1718. DOI: 10.1038/ng.3646. URL: https://www.nature.com/articles/ng.3646 (visited on 07/07/2021).

[104] Jason D. Buenrostro et al. "Single-Cell Chromatin Accessibility Reveals Principles of Regulatory Variation." In: *Nature* 523.7561 (July 2015), pp. 486–490. ISSN: 1476-4687. DOI: 10.1038/nature14590. URL: https://www.nature.com/articles/nature14590 (visited on 08/13/2019).

[105] Jason D. Buenrostro et al. "Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation." In: *Cell* 173.6 (May 2018), 1535–1548.e16. ISSN: 0092-8674. DOI: 10.1016/j.cell.2018.03.074. URL: http://www.sciencedirect.com/science/article/pii/S009286741830446X (visited on 05/15/2019).

[106] A.M. Ranzoni et al. "Integrative Single-Cell RNA-Seq and ATAC-Seq Analysis of Human Developmental Hematopoiesis." In: *Cell Stem Cell* 28.3 (2021), 472–487.e7. ISSN: 1934-5909. DOI: 10.1016/j.stem.2020.11.015.

[107] David Lara-Astiaso et al. "Chromatin State Dynamics during Blood Formation." In: *Science* 345.6199 (Aug. 2014), pp. 943–949. ISSN: 0036-8075. DOI: 10.1126/science.1256271.

[108] P. Zeller et al. "Single-Cell sortChIC Identifies Hierarchical Chromatin Dynamics during Hematopoiesis." In: *Nature Genetics* 55.2 (2023), pp. 333–345. ISSN: 1061-4036. DOI: 10.1038/s41588-022-01260-3.

[109] 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 (Jan. 2009), pp. 80–93. ISSN: 1934-5909. DOI: 10.1016/j.stem.2008.11.011. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2785912/ (visited on 04/09/2023).

[110] Timothy J. Ley et al. "DNMT3A Mutations in Acute Myeloid Leukemia." In: *New England Journal of Medicine* 363.25 (Dec. 2010), pp. 2424–2433.

ISSN: 0028-4793. DOI: 10.1056/NEJMoa1005143. URL: https://doi.org/10.1056/NEJMoa1005143 (visited on 04/09/2023).

[111] François Delhommeau et al. "Mutation in TET2 in Myeloid Cancers." In: *New England Journal of Medicine* 360.22 (May 2009), pp. 2289–2301. ISSN: 0028-4793. DOI: 10.1056/NEJMoa0810069. URL: https://doi.org/10.1056/NEJMoa0810069 (visited on 04/09/2023).

[112] Lambert Busque et al. "Recurrent Somatic TET2 Mutations in Normal Elderly Individuals with Clonal Hematopoiesis." In: *Nature Genetics* 44.11 (Nov. 2012), pp. 1179–1181. ISSN: 1546-1718. DOI: 10.1038/ng.2413. URL: https://www.nature.com/articles/ng.2413 (visited on 04/09/2023).

[113] Sagi Abelson et al. "Prediction of Acute Myeloid Leukaemia Risk in Healthy Individuals." In: *Nature* 559.7714 (July 2018), pp. 400–404. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0317-6. URL: https://www.nature.com/articles/s41586-018-0317-6 (visited on 04/09/2023).

[114] G.A. Challen et al. "Dnmt3a Is Essential for Hematopoietic Stem Cell Differentiation." In: *Nature Genetics* 44.1 (2012), pp. 23–31. DOI: 10.1038/ng.1009.

[115] Ann-Marie Bröske et al. "DNA Methylation Protects Hematopoietic Stem Cell Multipotency from Myeloerythroid Restriction." In: *Nature Genetics* 41.11 (Nov. 2009), pp. 1207–1215. ISSN: 1546-1718. DOI: 10.1038/ng.463. URL: https://www.nature.com/articles/ng.463 (visited on 07/05/2019).

[116] Jennifer J. Trowbridge et al. "DNA Methyltransferase 1 Is Essential for and Uniquely Regulates Hematopoietic Stem and Progenitor Cells." In: *Cell Stem Cell* 5.4 (Oct. 2009), pp. 442–449. ISSN: 1934-5909. DOI: 10.1016/j.stem.2009.08.016. URL: https://www.sciencedirect.com/science/article/pii/S1934590909004007 (visited on 04/09/2023).

[117] Grant A. Challen et al. "Dnmt3a and Dnmt3b Have Overlapping and Distinct Functions in Hematopoietic Stem Cells." In: *Cell Stem Cell* 15.3 (Sept. 2014), pp. 350–364. ISSN: 1934-5909. DOI: 10.1016/j.stem.2014.06.018. URL: https://www.sciencedirect.com/science/article/pii/S1934590914002665 (visited on 04/09/2023).

[118] Hong Ji et al. "Comprehensive Methylome Map of Lineage Commitment from Haematopoietic Progenitors." In: *Nature* 467.7313 (Sept. 2010), pp. 338–342. ISSN: 1476-4687. DOI: 10.1038/nature09367. URL: https://www.nature.com/articles/nature09367 (visited on 05/16/2019).

[119] E. Hodges et al. "Directional DNA Methylation Changes and Complex Intermediate States Accompany Lineage Specificity in the Adult Hema-

topoietic Compartment.” In: *Molecular Cell* 44.1 (2011), pp. 17–28. DOI: 10.1016/j.molcel.2011.08.026.

[120] Michael T. Bocker et al. “Genome-Wide Promoter DNA Methylation Dynamics of Human Hematopoietic Progenitor Cells during Differentiation and Aging.” In: *Blood* 117.19 (May 2011), e182–189. ISSN: 1528-0020. DOI: 10.1182/blood-2011-01-331926.

[121] Christoph Bock et al. “DNA Methylation Dynamics during In Vivo Differentiation of Blood and Skin Stem Cells.” In: *Molecular Cell* 47.4 (Aug. 2012), pp. 633–647. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2012.06.019. URL: http://www.sciencedirect.com/science/article/pii/S1097276512005448 (visited on 10/30/2018).

[122] D.B. Lipka et al. “Identification of Dna Methylation Changes at Cis-Regulatory Elements during Early Steps of Hsc Differentiation Using Tagmentation-Based Whole Genome Bisulfite Sequencing.” In: *Cell Cycle* 13.22 (2014), pp. 3476–3487. DOI: 10.4161/15384101.2014.973334.

[123] Christopher C Oakes et al. “DNA Methylation Dynamics during B Cell Maturation Underlie a Continuum of Disease Phenotypes in Chronic Lymphocytic Leukemia.” In: *Nature genetics* 48.3 (Mar. 2016), pp. 253–264. ISSN: 1061-4036. DOI: 10.1038/ng.3488. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4963005/ (visited on 04/09/2023).

[124] Claudia Benz et al. “Hematopoietic Stem Cell Subtypes Expand Differentially during Development and Display Distinct Lymphopoietic Programs.” In: *Cell Stem Cell* 10.3 (Mar. 2012), pp. 273–283. ISSN: 1934-5909. DOI: 10.1016/j.stem.2012.02.007. URL: https://www.sciencedirect.com/science/article/pii/S1934590912000653 (visited on 04/10/2023).

[125] David G. Kent et al. “Prospective Isolation and Molecular Characterization of Hematopoietic Stem Cells with Durable Self-Renewal Potential.” In: *Blood* 113.25 (June 2009), pp. 6342–6350. ISSN: 0006-4971. DOI: 10.1182/blood-2008-12-192054. URL: https://www.sciencedirect.com/science/article/pii/S0006497120372530 (visited on 04/10/2023).

[126] Franco Izzo et al. “DNA Methylation Disruption Reshapes the Hematopoietic Differentiation Landscape.” In: *Nature Genetics* 52.4 (Apr. 2020), pp. 378–387. ISSN: 1546-1718. DOI: 10.1038/s41588-020-0595-4. URL: https://www.nature.com/articles/s41588-020-0595-4 (visited on 07/07/2021).

[127] Guanjue Xiang et al. “An Integrative View of the Regulatory and Transcriptional Landscapes in Mouse Hematopoiesis.” In: *Genome Research* 30.3 (Mar. 2020), pp. 472–484. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr

.255760.119. URL: https://genome.cshlp.org/content/30/3/472 (visited on 07/15/2021).

[128] Jill E. Moore et al. "Expanded Encyclopaedias of DNA Elements in the Human and Mouse Genomes." In: *Nature* 583.7818 (July 2020), pp. 699–710. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2493-4. URL: https://www.nature.com/articles/s41586-020-2493-4 (visited on 04/09/2023).

[129] Netanel Loyfer et al. "A DNA Methylation Atlas of Normal Human Cell Types." In: *Nature* 613.7943 (Jan. 2023), pp. 355–364. ISSN: 1476-4687. DOI: 10.1038/s41586-022-05580-6. URL: https://www.nature.com/articles/s41586-022-05580-6 (visited on 04/06/2023).

[130] Roshni Roy et al. "DNA Methylation Signatures Reveal That Distinct Combinations of Transcription Factors Specify Human Immune Cell Epigenetic Identity." In: *Immunity* 54.11 (Nov. 2021), 2465–2480.e5. ISSN: 1074-7613. DOI: 10.1016/j.immuni.2021.10.001. URL: https://www.sciencedirect.com/science/article/pii/S1074761321004076 (visited on 04/26/2023).

[131] Sina Stäble. "Deconvolution of Hematopoietic Commitment Decisions by Genome-wide Analysis of Progressive DNA Methylation Changes." PhD thesis. University of Heidelberg, Feb. 2019.

[132] Eva Reisinger et al. "OTP: An Automatized System for Managing and Processing NGS Data." In: *Journal of Biotechnology*. Bioinformatics Solutions for Big Data Analysis in Life Sciences Presented by the German Network for Bioinformatics Infrastructure 261 (Nov. 2017), pp. 53–62. ISSN: 0168-1656. DOI: 10.1016/j.jbiotec.2017.08.006. URL: https://www.sciencedirect.com/science/article/pii/S0168165617315924 (visited on 04/10/2022).

[133] Daniel R. Zerbino et al. "The Ensembl Regulatory Build." In: *Genome Biology* 16.1 (Mar. 2015), p. 56. ISSN: 1465-6906. DOI: 10.1186/s13059-015-0621-5. URL: https://doi.org/10.1186/s13059-015-0621-5 (visited on 04/14/2022).

[134] Melanie D. Mumau et al. "Identification of a Multipotent Progenitor Population in the Spleen That Is Regulated by NR4A1." In: *The Journal of Immunology* 200.3 (Feb. 2018), pp. 1078–1087. ISSN: 0022-1767. DOI: 10.4049/jimmunol.1701250. URL: https://doi.org/10.4049/jimmunol.1701250 (visited on 02/27/2023).

[135] Christopher G. Duncan et al. "Dosage Compensation and DNA Methylation Landscape of the X Chromosome in Mouse Liver." In: *Scientific Reports* 8.1 (July 2018), p. 10138. ISSN: 2045-2322. DOI: 10.1038/s41598-018-28356

-3. URL: https://www.nature.com/articles/s41598-018-28356-3 (visited on 03/07/2023).

[136] Yoav Benjamini, Abba M. Krieger, and Daniel Yekutieli. "Adaptive Linear Step-up Procedures That Control the False Discovery Rate." In: *Biometrika* 93.3 (Sept. 2006), pp. 491–507. ISSN: 0006-3444. DOI: 10.1093/biomet/93.3.491. URL: https://doi.org/10.1093/biomet/93.3.491 (visited on 01/06/2021).

[137] Bernd Schröder. "The Multifaceted Roles of the Invariant Chain CD74 — More than Just a Chaperone." In: *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 1863.6, Part A (June 2016), pp. 1269–1281. ISSN: 0167-4889. DOI: 10.1016/j.bbamcr.2016.03.026. URL: https://www.sciencedirect.com/science/article/pii/S0167488916300799 (visited on 03/10/2023).

[138] F Momburg et al. "Differential Expression of Ia and Ia-associated Invariant Chain in Mouse Tissues after in Vivo Treatment with IFN-gamma." In: *The Journal of Immunology* 136.3 (Feb. 1986), pp. 940–948. ISSN: 0022-1767. DOI: 10.4049/jimmunol.136.3.940. URL: https://doi.org/10.4049/jimmunol.136.3.940 (visited on 03/10/2023).

[139] Tomohiko Ishibashi et al. "ESAM Is a Novel Human Hematopoietic Stem Cell Marker Associated with a Subset of Human Leukemias." In: *Experimental Hematology* 44.4 (Apr. 2016), 269–281.e1. ISSN: 0301-472X. DOI: 10.1016/j.exphem.2015.12.010. URL: https://www.exphem.org/article/S0301-472X(16)00005-9/fulltext (visited on 03/10/2023).

[140] Christian Schmidl et al. "The Enhancer and Promoter Landscape of Human Regulatory and Conventional T-cell Subpopulations." In: *Blood* 123.17 (Apr. 2014), e68–e78. ISSN: 0006-4971. DOI: 10.1182/blood-2013-02-486944. URL: https://doi.org/10.1182/blood-2013-02-486944 (visited on 07/20/2022).

[141] Guillaume Devailly and Anagha Joshi. "Insights into Mammalian Transcription Control by Systematic Analysis of ChIP Sequencing Data." In: *BMC Bioinformatics* 19.14 (Nov. 2018), p. 409. ISSN: 1471-2105. DOI: 10.1186/s12859-018-2377-x. URL: https://doi.org/10.1186/s12859-018-2377-x (visited on 07/20/2022).

[142] Jill E. Moore et al. "A Curated Benchmark of Enhancer-Gene Interactions for Evaluating Enhancer-Target Gene Prediction Methods." In: *Genome Biology* 21.1 (Jan. 2020), p. 17. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1924-8. URL: https://doi.org/10.1186/s13059-019-1924-8 (visited on 07/20/2022).

[143] V. A. Traag, L. Waltman, and N. J. van Eck. “From Louvain to Leiden: Guaranteeing Well-Connected Communities.” In: *Scientific Reports* 9.1 (Mar. 2019), p. 5233. ISSN: 2045-2322. DOI: 10.1038/s41598-019-41695-z. URL: https://www.nature.com/articles/s41598-019-41695-z (visited on 03/05/2023).

[144] Jan Hettinger et al. “Origin of Monocytes and Macrophages in a Committed Progenitor.” In: *Nature Immunology* 14.8 (Aug. 2013), pp. 821–830. ISSN: 1529-2916. DOI: 10.1038/ni.2638.

[145] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. “SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis.” In: *Genome Biology* 19.1 (Feb. 2018), p. 15. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1382-0. URL: https://doi.org/10.1186/s13059-017-1382-0 (visited on 08/14/2019).

[146] Christoph Hafemeister and Rahul Satija. “Normalization and Variance Stabilization of Single-Cell RNA-seq Data Using Regularized Negative Binomial Regression.” In: *bioRxiv* (Mar. 2019), p. 576827. DOI: 10.1101/576827. URL: https://www.biorxiv.org/content/10.1101/576827v1 (visited on 06/08/2019).

[147] Jan Lause, Philipp Berens, and Dmitry Kobak. “Analytic Pearson Residuals for Normalization of Single-Cell RNA-seq UMI Data.” In: *Genome Biology* 22.1 (Sept. 2021), p. 258. ISSN: 1474-760X. DOI: 10.1186/s13059-021-02451-7. URL: https://doi.org/10.1186/s13059-021-02451-7 (visited on 03/22/2023).

[148] Malte D Luecken and Fabian J Theis. “Current Best Practices in Single-Cell RNA-seq Analysis: A Tutorial.” In: *Molecular Systems Biology* 15.6 (June 2019), e8746. ISSN: 1744-4292. DOI: 10.15252/msb.20188746. URL: https://www.embopress.org/doi/full/10.15252/msb.20188746 (visited on 04/10/2020).

[149] Caleb Weinreb, Samuel Wolock, and Allon M. Klein. “SPRING: A Kinetic Interface for Visualizing High Dimensional Single-Cell Expression Data.” In: *Bioinformatics (Oxford, England)* 34.7 (Apr. 2018), pp. 1246–1248. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btx792.

[150] Shobana V. Stassen et al. “PARC: Ultrafast and Accurate Clustering of Phenotypic Data of Millions of Single Cells.” In: *Bioinformatics* (2020). DOI: 10.1093/bioinformatics/btaa042. URL: https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btaa042/5714737 (visited on 04/10/2020).

[151] Chiara Baccin et al. “Combined Single-Cell and Spatial Transcriptomics Reveal the Molecular, Cellular and Spatial Bone Marrow Niche Organiza-

tion." In: *Nature Cell Biology* 22.1 (Jan. 2020), pp. 38–48. ISSN: 1476-4679. DOI: 10.1038/s41556-019-0439-6. URL: https://www.nature.com/articles/s41556-019-0439-6 (visited on 08/01/2021).

[152] Franco Izzo et al. "DNA Methylation Disruption Reshapes the Hematopoietic Differentiation Landscape." In: *Nature Genetics* 52.4 (Apr. 2020), pp. 378–387. ISSN: 1546-1718. DOI: 10.1038/s41588-020-0595-4. URL: https://www.nature.com/articles/s41588-020-0595-4 (visited on 02/26/2023).

[153] Yoav Benjamini and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1 (1995), pp. 289–300. ISSN: 2517-6161. DOI: 10.1111/j.2517-6161.1995.tb02031.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1995.tb02031.x (visited on 03/23/2023).

[154] Daniel Yekutieli and Yoav Benjamini. "Resampling-Based False Discovery Rate Controlling Multiple Test Procedures for Correlated Test Statistics." In: *Journal of Statistical Planning and Inference* 82.1 (Dec. 1999), pp. 171–196. ISSN: 0378-3758. DOI: 10.1016/S0378-3758(99)00041-5. URL: https://www.sciencedirect.com/science/article/pii/S0378375899000415 (visited on 03/23/2023).

[155] John D. Storey and Robert Tibshirani. "Statistical Significance for Genomewide Studies." In: *Proceedings of the National Academy of Sciences* 100.16 (Aug. 2003), pp. 9440–9445. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1530509100. URL: https://www.pnas.org/content/100/16/9440 (visited on 09/14/2020).

[156] Miroslawa Siatecka and James J. Bieker. "The Multifunctional Role of EKLF/KLF1 during Erythropoiesis." In: *Blood* 118.8 (Aug. 2011), pp. 2044–2054. ISSN: 0006-4971. DOI: 10.1182/blood-2011-03-331371. URL: https://doi.org/10.1182/blood-2011-03-331371 (visited on 03/22/2023).

[157] J. Edgeworth et al. "Identification of P8,14 as a Highly Abundant Heterodimeric Calcium Binding Protein Complex of Myeloid Cells." In: *The Journal of Biological Chemistry* 266.12 (Apr. 1991), pp. 7706–7713. ISSN: 0021-9258.

[158] David F. Stroncek, Lorraine Caruccio, and Maria Bettinotti. "CD177: A Member of the Ly-6 Gene Superfamily Involved with Neutrophil Proliferation and Polycythemia Vera." In: *Journal of Translational Medicine* 2.1 (Mar. 2004), p. 8. ISSN: 1479-5876. DOI: 10.1186/1479-5876-2-8. URL: https://doi.org/10.1186/1479-5876-2-8 (visited on 03/22/2023).

[159] Gabrielle Faure-André et al. "Regulation of Dendritic Cell Migration by CD74, the MHC Class II-associated Invariant Chain." In: *Science (New York, N.Y.)* 322.5908 (Dec. 2008), pp. 1705–1710. ISSN: 1095-9203. DOI: 10.1126/science.1159894.

[160] Jiquan Zhang et al. "Characterization of Siglec-H as a Novel Endocytic Receptor Expressed on Murine Plasmacytoid Dendritic Cell Precursors." In: *Blood* 107.9 (May 2006), pp. 3600–3608. ISSN: 0006-4971. DOI: 10.1182/blood-2005-09-3842. URL: https://doi.org/10.1182/blood-2005-09-3842 (visited on 03/22/2023).

[161] Joseph M. Dal Porto, Kathy Burke, and John C. Cambier. "Regulation of BCR Signal Transduction in B-1 Cells Requires the Expression of the Src Family Kinase Lck." In: *Immunity* 21.3 (Sept. 2004), pp. 443–453. ISSN: 1074-7613. DOI: 10.1016/j.immuni.2004.07.018.

[162] D. B. Straus and A. Weiss. "Genetic Evidence for the Involvement of the Lck Tyrosine Kinase in Signal Transduction through the T Cell Antigen Receptor." In: *Cell* 70.4 (Aug. 1992), pp. 585–593. ISSN: 0092-8674. DOI: 10.1016/0092-8674(92)90428-f.

[163] DY Mason et al. "CD79a: A Novel Marker for B-cell Neoplasms in Routinely Processed Tissue Samples." In: *Blood* 86.4 (Aug. 1995), pp. 1453–1459. ISSN: 0006-4971. DOI: 10.1182/blood.V86.4.1453.bloodjournal8641453. URL: https://doi.org/10.1182/blood.V86.4.1453.bloodjournal8641453 (visited on 03/22/2023).

[164] Claus Nerlov and Thomas Graf. "PU.1 Induces Myeloid Lineage Commitment in Multipotent Hematopoietic Progenitors." In: *Genes & Development* 12.15 (Aug. 1998), pp. 2403–2412. ISSN: 0890-9369. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC317050/ (visited on 03/22/2023).

[165] Zeenath Unnisa et al. "Meis1 Preserves Hematopoietic Stem Cells in Mice by Limiting Oxidative Stress." In: *Blood* 120.25 (Dec. 2012), p. 4973. DOI: 10.1182/blood-2012-06-435800. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3525022/ (visited on 03/22/2023).

[166] Vincenzo Calvanese et al. "MLLT3 Governs Human Haematopoietic Stem-Cell Self-Renewal and Engraftment." In: *Nature* 576.7786 (Dec. 2019), pp. 281–286. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1790-2.

[167] Simon Haas et al. "Inflammation-Induced Emergency Megakaryopoiesis Driven by Hematopoietic Stem Cell-like Megakaryocyte Progenitors." In: *Cell Stem Cell* 17.4 (Oct. 2015), pp. 422–434. ISSN: 1875-9777. DOI: 10.1016/j.stem.2015.07.007.

[168] Yuhan Hao et al. "Integrated Analysis of Multimodal Single-Cell Data." In: *Cell* 184.13 (June 2021), 3573–3587.e29. ISSN: 0092-8674. DOI: 10.1016

/j.cell.2021.04.048. URL: https://www.sciencedirect.com/science/article/pii/S0092867421005833 (visited on 03/22/2023).

[169] Jeff Vierstra et al. "Global Reference Mapping of Human Transcription Factor Footprints." In: *Nature* 583.7818 (July 2020), pp. 729–736. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2528-x. URL: https://www.nature.com/articles/s41586-020-2528-x (visited on 09/27/2021).

[170] Felix Krueger and Simon R. Andrews. "Bismark: A Flexible Aligner and Methylation Caller for Bisulfite-Seq Applications." In: *Bioinformatics* 27.11 (June 2011), pp. 1571–1572. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr167. URL: https://doi.org/10.1093/bioinformatics/btr167 (visited on 05/14/2023).

[171] Devon Ryan. *MethylDackel*. URL: https://github.com/dpryan79/MethylDackel.

[172] Peter Langfelder, Bin Zhang, and Steve Horvath. "Defining Clusters from a Hierarchical Cluster Tree: The Dynamic Tree Cut Package for R." In: *Bioinformatics* 24.5 (Mar. 2008), pp. 719–720. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btm563. URL: https://doi.org/10.1093/bioinformatics/btm563 (visited on 05/12/2023).

[173] Zuguang Gu, Roland Eils, and Matthias Schlesner. "Complex Heatmaps Reveal Patterns and Correlations in Multidimensional Genomic Data." In: *Bioinformatics* 32.18 (Sept. 2016), pp. 2847–2849. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw313. URL: https://doi.org/10.1093/bioinformatics/btw313 (visited on 05/16/2023).

[174] Florian Hahne and Robert Ivanek. "Visualizing Genomic Data Using Gviz and Bioconductor." In: *Statistical Genomics: Methods and Protocols*. Ed. by Ewy Mathé and Sean Davis. Methods in Molecular Biology. New York, NY: Springer, 2016, pp. 335–351. ISBN: 978-1-4939-3578-9. DOI: 10.1007/978-1-4939-3578-9_16. URL: https://doi.org/10.1007/978-1-4939-3578-9_16 (visited on 05/16/2023).

[175] Lucille Lopez-Delisle et al. "pyGenomeTracks: Reproducible Plots for Multivariate Genomic Datasets." In: *Bioinformatics* 37.3 (Apr. 2021), pp. 422–423. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa692. URL: https://doi.org/10.1093/bioinformatics/btaa692 (visited on 05/16/2023).

[176] M.J. Ziller et al. "Charting a Dynamic DNA Methylation Landscape of the Human Genome." In: *Nature* 500.7463 (2013), pp. 477–481. DOI: 10.1038/nature12433.

[177] Matthew D. Schultz et al. "Human Body Epigenome Maps Reveal Noncanonical DNA Methylation Variation." In: *Nature* 523.7559 (July 2015),

pp. 212–216. ISSN: 1476-4687. DOI: 10.1038/nature14465. URL: https://www.nature.com/articles/nature14465 (visited on 04/25/2023).

[178] Philippe Gascard et al. "Epigenetic and Transcriptional Determinants of the Human Breast." In: *Nature Communications* 6.1 (Feb. 2015), p. 6351. ISSN: 2041-1723. DOI: 10.1038/ncomms7351. URL: https://www.nature.com/articles/ncomms7351 (visited on 04/25/2023).

[179] Katherine E. Varley et al. "Dynamic DNA Methylation across Diverse Human Cell Lines and Tissues." In: *Genome Research* 23.3 (Mar. 2013), pp. 555–567. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.147942.112. URL: https://genome.cshlp.org/content/23/3/555 (visited on 04/25/2023).

[180] Bethan Psaila and Adam J. Mead. "Single-Cell Approaches Reveal Novel Cellular Pathways for Megakaryocyte and Erythroid Differentiation." In: *Blood* 133.13 (Mar. 2019), pp. 1427–1435. ISSN: 0006-4971. DOI: 10.1182/blood-2018-11-835371. URL: https://doi.org/10.1182/blood-2018-11-835371 (visited on 04/29/2023).

[181] Gilad Landan et al. "Epigenetic Polymorphism and the Stochastic Formation of Differentially Methylated Regions in Normal and Cancerous Tissues." In: *Nature Genetics* 44.11 (Nov. 2012), pp. 1207–1214. ISSN: 1546-1718. DOI: 10.1038/ng.2442. URL: https://www.nature.com/articles/ng.2442 (visited on 04/28/2023).

[182] C. Anthony Scott et al. "Identification of Cell Type-Specific Methylation Signals in Bulk Whole Genome Bisulfite Sequencing Data." In: *Genome Biology* 21.1 (July 2020), p. 156. ISSN: 1474-760X. DOI: 10.1186/s13059-020-02065-5. URL: https://doi.org/10.1186/s13059-020-02065-5 (visited on 04/28/2023).

[183] 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." In: *Nature Genetics* 49.4 (Apr. 2017), pp. 635–642. ISSN: 1546-1718. DOI: 10.1038/ng.3805. URL: https://www.nature.com/articles/ng.3805 (visited on 04/30/2023).

[184] Yupeng He et al. "Spatiotemporal DNA Methylome Dynamics of the Developing Mouse Fetus." In: *Nature* 583.7818 (July 2020), pp. 752–759. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2119-x. URL: https://www.nature.com/articles/s41586-020-2119-x (visited on 04/26/2023).

[185] Christopher E. Schlosberg, Nathan D. VanderKraats, and John R. Edwards. "Modeling Complex Patterns of Differential DNA Methylation That Associate with Gene Expression Changes." In: *Nucleic Acids Research* 45.9 (May 2017), pp. 5100–5111. ISSN: 0305-1048. DOI: 10.1093/nar/gkx078. URL:

https://academic.oup.com/nar/article/45/9/5100/2972665 (visited on 07/16/2019).

[186] Mira Jeong et al. “Large Conserved Domains of Low DNA Methylation Maintained by Dnmt3a.” In: *Nature Genetics* 46.1 (Jan. 2014), pp. 17–23. ISSN: 1546-1718. DOI: 10.1038/ng.2836.

[187] Laura Wiehle et al. “Tet1 and Tet2 Protect DNA Methylation Canyons against Hypermethylation.” In: *Molecular and Cellular Biology* 36.3 (Jan. 2016), pp. 452–461. ISSN: 0270-7306. DOI: 10.1128/MCB.00587-15. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4719427/ (visited on 04/27/2023).

[188] Elisabeth R. Wilson et al. “Focal Disruption of DNA Methylation Dynamics at Enhancers in IDH-mutant AML Cells.” In: *Leukemia* 36.4 (Apr. 2022), pp. 935–945. ISSN: 1476-5551. DOI: 10.1038/s41375-021-01476-y. URL: https://www.nature.com/articles/s41375-021-01476-y (visited on 05/01/2023).

[189] Shamika Ketkar et al. “Remethylation of Dnmt3a-/- Hematopoietic Cells Is Associated with Partial Correction of Gene Dysregulation and Reduced Myeloid Skewing.” In: *Proceedings of the National Academy of Sciences* 117.6 (Feb. 2020), pp. 3123–3134. DOI: 10.1073/pnas.1918611117. URL: https://www.pnas.org/doi/10.1073/pnas.1918611117 (visited on 05/01/2023).

[190] Christoph Bock. “Analysing and Interpreting DNA Methylation Data.” In: *Nature Reviews Genetics* 13.10 (Oct. 2012), pp. 705–719. ISSN: 1471-0056. DOI: 10.1038/nrg3273.

[191] Kirsty Minton. “Mapping the Minutiae of the Human Methylome.” In: *Nature Reviews Genetics* 24.3 (Mar. 2023), pp. 139–139. ISSN: 1471-0064. DOI: 10.1038/s41576-023-00576-y. URL: https://www.nature.com/articles/s41576-023-00576-y (visited on 04/06/2023).

[192] Florian Eckhardt et al. “DNA Methylation Profiling of Human Chromosomes 6, 20 and 22.” In: *Nature Genetics* 38.12 (Dec. 2006), pp. 1378–1385. ISSN: 1546-1718. DOI: 10.1038/ng1909. URL: https://www.nature.com/articles/ng1909 (visited on 04/28/2023).

[193] Weiwei Zhang et al. “Predicting Genome-Wide DNA Methylation Using Methylation Marks, Genomic Position, and DNA Regulatory Elements.” In: *Genome Biology* 16.1 (Jan. 2015), p. 14. ISSN: 1465-6906. DOI: 10.1186/s13059-015-0581-9. URL: https://doi.org/10.1186/s13059-015-0581-9 (visited on 04/28/2023).

[194] Arif Harmanci et al. “EpiSAFARI: Sensitive Detection of Valleys in Epigenetic Signals for Enhancing Annotations of Functional Elements.” In:

Bioinformatics 36.4 (Feb. 2020), pp. 1014–1021. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz702. URL: https://doi.org/10.1093/bioinformatics/btz702 (visited on 04/06/2023).

[195] Ko Hashimoto et al. "Regulated Transcription of Human Matrix Metalloproteinase 13 (MMP13) and Interleukin-1β (IL1B) Genes in Chondrocytes Depends on Methylation of Specific Proximal Promoter CpG Sites*." In: *Journal of Biological Chemistry* 288.14 (Apr. 2013), pp. 10061–10072. ISSN: 0021-9258. DOI: 10.1074/jbc.M112.421156. URL: https://www.sciencedirect.com/science/article/pii/S0021925820673653 (visited on 04/06/2023).

[196] Shimrat Mamrut et al. "DNA Methylation of Specific CpG Sites in the Promoter Region Regulates the Transcription of the Mouse Oxytocin Receptor." In: *PLOS ONE* 8.2 (Feb. 2013), e56869. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0056869. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0056869 (visited on 04/06/2023).

[197] Christoper J. Nile et al. "Methylation Status of a Single CpG Site in the IL6 Promoter Is Related to IL6 Messenger RNA Levels and Rheumatoid Arthritis." In: *Arthritis & Rheumatism* 58.9 (2008), pp. 2686–2693. ISSN: 1529-0131. DOI: 10.1002/art.23758. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/art.23758 (visited on 04/06/2023).

[198] Rainer W. Fürst et al. "A Differentially Methylated Single CpG-site Is Correlated with Estrogen Receptor Alpha Transcription." In: *The Journal of Steroid Biochemistry and Molecular Biology* 130.1 (May 2012), pp. 96–104. ISSN: 0960-0760. DOI: 10.1016/j.jsbmb.2012.01.009. URL: https://www.sciencedirect.com/science/article/pii/S0960076012000313 (visited on 04/06/2023).

[199] Kouki Tsuboi et al. "Single CpG Site Methylation Controls Estrogen Receptor Gene Transcription and Correlates with Hormone Therapy Resistance." In: *The Journal of Steroid Biochemistry and Molecular Biology* 171 (July 2017), pp. 209–217. ISSN: 0960-0760. DOI: 10.1016/j.jsbmb.2017.04.001. URL: https://www.sciencedirect.com/science/article/pii/S0960076017301000 (visited on 04/06/2023).

[200] Sebastian Kaluscha et al. "Evidence That Direct Inhibition of Transcription Factor Binding Is the Prevailing Mode of Gene and Repeat Repression by DNA Methylation." In: *Nature Genetics* 54.12 (Dec. 2022), pp. 1895–1906. ISSN: 1546-1718. DOI: 10.1038/s41588-022-01241-6. URL: https://www.nature.com/articles/s41588-022-01241-6 (visited on 04/28/2023).

[201] Takahiro Suzuki et al. “RUNX1 Regulates Site Specificity of DNA Demethylation by Recruitment of DNA Demethylation Machineries in Hematopoietic Cells.” In: *Blood Advances* 1.20 (Sept. 2017), pp. 1699–1711. ISSN: 2473-9529. DOI: 10.1182/bloodadvances.2017005710. URL: https://doi.org/10.1182/bloodadvances.2017005710 (visited on 04/27/2023).

[202] François Spitz and Eileen E. M. Furlong. “Transcription Factors: From Enhancer Binding to Developmental Control.” In: *Nature Reviews Genetics* 13.9 (Sept. 2012), pp. 613–626. ISSN: 1471-0064. DOI: 10.1038/nrg3207. URL: https://www.nature.com/articles/nrg3207 (visited on 04/28/2023).

[203] C. Anthony Scott et al. “Identification of Cell Type-Specific Methylation Signals in Bulk Whole Genome Bisulfite Sequencing Data.” In: *Genome Biology* 21.1 (July 2020), p. 156. ISSN: 1474-760X. DOI: 10.1186/s13059-020-02065-5. URL: https://doi.org/10.1186/s13059-020-02065-5 (visited on 04/24/2023).

[204] Nathan D. VanderKraats et al. “Discovering High-Resolution Patterns of Differential DNA Methylation That Correlate with Gene Expression Changes.” In: *Nucleic Acids Research* 41.14 (Aug. 2013), pp. 6816–6827. ISSN: 0305-1048. DOI: 10.1093/nar/gkt482. URL: https://doi.org/10.1093/nar/gkt482 (visited on 04/28/2023).

[205] Christopher E Schlosberg et al. “ME-Class2 Reveals Context Dependent Regulatory Roles for 5-Hydroxymethylcytosine.” In: *Nucleic Acids Research* 47.5 (Mar. 2019), e28. ISSN: 0305-1048. DOI: 10.1093/nar/gkz001. URL: https://doi.org/10.1093/nar/gkz001 (visited on 04/28/2023).

[206] Zohar Shipony et al. “Dynamic and Static Maintenance of Epigenetic Memory in Pluripotent and Somatic Cells.” In: *Nature* 513.7516 (Sept. 2014), pp. 115–119. ISSN: 1476-4687. DOI: 10.1038/nature13458.

[207] 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.” In: *Nature Genetics* 49.4 (Apr. 2017), p. 635. ISSN: 1546-1718. DOI: 10.1038/ng.3805. URL: https://www.nature.com/articles/ng.3805 (visited on 05/16/2019).

[208] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” In: *Bioinformatics* 30.15 (Aug. 2014), pp. 2114–2120. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu170. URL: https://doi.org/10.1093/bioinformatics/btu170 (visited on 04/14/2023).

[209] Heng Li. *Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM*. May 2013. DOI: 10.48550/arXiv.1303.3997. arXiv:

1303.3997 [q-bio]. URL: http://arxiv.org/abs/1303.3997 (visited on 04/14/2023).

[210] *Picard Toolkit*. 2019. URL: https://broadinstitute.github.io/picard/.

[211] Petr Danecek et al. "Twelve Years of SAMtools and BCFtools." In: *GigaScience* 10.2 (Feb. 2021), giab008. ISSN: 2047-217X. DOI: 10.1093/gigascience/giab008. URL: https://doi.org/10.1093/gigascience/giab008 (visited on 04/14/2023).

[212] Marcel Martin. "Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads." In: *EMBnet.journal* 17.1 (May 2011), pp. 10–12. ISSN: 2226-6089. DOI: 10.14806/ej.17.1.200. URL: https://journal.embnet.org/index.php/embnetjournal/article/view/200 (visited on 05/14/2023).

[213] Simon Andrews. *FASTQC*. URL: https://github.com/s-andrews/FastQC.

[214] Leland McInnes, John Healy, and James Melville. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." In: *arXiv:1802.03426 [cs, stat]* (Feb. 2018). arXiv: 1802.03426 [cs, stat]. URL: http://arxiv.org/abs/1802.03426 (visited on 05/12/2019).

[215] Joerg Reichardt and Stefan Bornholdt. "Statistical Mechanics of Community Detection." In: *Physical Review E* 74.1 (July 2006), p. 016110. ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.74.016110. arXiv: cond-mat/0603718. URL: http://arxiv.org/abs/cond-mat/0603718 (visited on 04/10/2020).

[216] Adam Frankish et al. "GENCODE Reference Annotation for the Human and Mouse Genomes." In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D766–D773. ISSN: 0305-1048. DOI: 10.1093/nar/gky955. URL: https://doi.org/10.1093/nar/gky955 (visited on 04/14/2023).

[217] Jose Manuel Rodriguez et al. "APPRIS: Annotation of Principal and Alternative Splice Isoforms." In: *Nucleic Acids Research* 41.D1 (Jan. 2013), pp. D110–D117. ISSN: 0305-1048. DOI: 10.1093/nar/gks1058. URL: https://doi.org/10.1093/nar/gks1058 (visited on 03/19/2023).

[218] Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python." In: *Nature Methods* 17.3 (Mar. 2020), pp. 261–272. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0686-2. URL: https://www.nature.com/articles/s41592-019-0686-2 (visited on 05/05/2023).

[219] Kyle Smith. *dynamicTreeCut (Python)*. URL: https://github.com/kylessmith/dynamicTreeCut.

[220] Isaac Virshup et al. “The Scverse Project Provides a Computational Ecosystem for Single-Cell Omics Data Analysis.” In: *Nature Biotechnology* (Apr. 2023), pp. 1–3. ISSN: 1546-1696. DOI: 10.1038/s41587-023-01733-8. URL: https://www.nature.com/articles/s41587-023-01733-8 (visited on 04/15/2023).

[221] Charles R. Harris et al. “Array Programming with NumPy.” In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2649-2. URL: https://www.nature.com/articles/s41586-020-2649-2 (visited on 05/05/2023).

[222] Wes McKinney. “Data Structures for Statistical Computing in Python.” In: *Python in Science Conference*. Austin, Texas, 2010, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a. URL: https://conference.scipy.org/proceedings/scipy2010/mckinney.html (visited on 05/05/2023).

[223] Skipper Seabold and Josef Perktold. “Statsmodels: Econometric and Statistical Modeling with Python.” In: *Python in Science Conference*. Austin, Texas, 2010, pp. 92–96. DOI: 10.25080/Majora-92bf1922-011. URL: https://conference.scipy.org/proceedings/scipy2010/seabold.html (visited on 05/05/2023).

[224] Endre Bakken Stovner and Pål Sætrom. “PyRanges: Efficient Comparison of Genomic Intervals in Python.” In: *Bioinformatics* 36.3 (Feb. 2020), pp. 918–919. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz615. URL: https://doi.org/10.1093/bioinformatics/btz615 (visited on 05/05/2023).

[225] Wolfgang Huber et al. “Orchestrating High-Throughput Genomic Analysis with Bioconductor.” In: *Nature Methods* 12.2 (Feb. 2015), pp. 115–121. ISSN: 1548-7105. DOI: 10.1038/nmeth.3252. URL: https://www.nature.com/articles/nmeth.3252 (visited on 05/05/2023).

[226] Marcus Eich, Andreas Trumpp, and Steffen Schmitt. “OMIP-059: Identification of Mouse Hematopoietic Stem and Progenitor Cells with Simultaneous Detection of CD45.1/2 and Controllable Green Fluorescent Protein Expression by a Single Staining Panel.” In: *Cytometry Part A* 95.10 (2019), pp. 1049–1052. ISSN: 1552-4930. DOI: 10.1002/cyto.a.23845. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cyto.a.23845 (visited on 01/04/2021).

[227] Simon Yona et al. “Fate Mapping Reveals Origins and Dynamics of Monocytes and Tissue Macrophages under Homeostasis.” In: *Immunity* 38.1 (Jan. 2013), pp. 79–91. ISSN: 1074-7613. DOI: 10.1016/j.immuni.2012.12.001. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3908543/ (visited on 01/04/2021).

[228] Andreas Schlitzer et al. “Identification of cDC1- and cDC2-committed DC Progenitors Reveals Early Lineage Priming at the Common DC Progenitor Stage in the Bone Marrow.” In: *Nature Immunology* 16.7 (July 2015), pp. 718–728. ISSN: 1529-2916. DOI: 10.1038/ni.3200. URL: https://www.nature.com/articles/ni.3200 (visited on 01/04/2021).

[229] Chao Shi et al. “Bone Marrow Mesenchymal Stem and Progenitor Cells Induce Monocyte Emigration in Response to Circulating Toll-like Receptor Ligands.” In: *Immunity* 34.4 (Apr. 2011), pp. 590–601. ISSN: 1097-4180. DOI: 10.1016/j.immuni.2011.02.016.

[230] Ying-Ying Hey, Jonathan K. H. Tan, and Helen C. O’Neill. “Redefining Myeloid Cell Subsets in Murine Spleen.” In: *Frontiers in Immunology* 6 (2016), p. 652. ISSN: 1664-3224. DOI: 10.3389/fimmu.2015.00652.

[231] Dalia Pakalniškytė and Barbara U. Schraml. “Tissue-Specific Diversity and Functions of Conventional Dendritic Cells.” In: *Advances in Immunology* 134 (2017), pp. 89–135. ISSN: 1557-8445. DOI: 10.1016/bs.ai.2017.01.003.

[232] Amanda L. Blasius et al. “Siglec-H Is an IPC-specific Receptor That Modulates Type I IFN Secretion through DAP12.” In: *Blood* 107.6 (Mar. 2006), pp. 2474–2476. ISSN: 0006-4971. DOI: 10.1182/blood-2005-09-3746. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1895736/ (visited on 01/04/2021).

[233] Luisa Cervantes-Barragan et al. “Plasmacytoid Dendritic Cells Control T-cell Response to Chronic Viral Infection.” In: *Proceedings of the National Academy of Sciences of the United States of America* 109.8 (Feb. 2012), pp. 3012–3017. ISSN: 1091-6490. DOI: 10.1073/pnas.1117359109.

[234] BD Biosciences. *Human and Mouse CD Marker Handbook*. 2010.