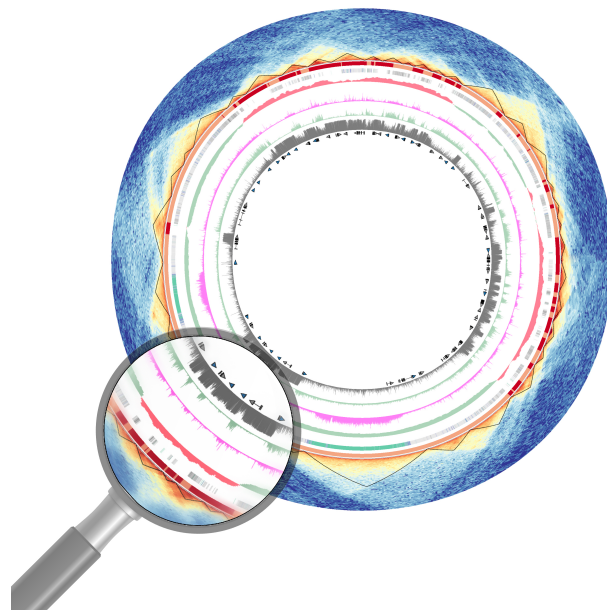


# **Genome-wide analysis of DNA methylation topology to understand cell fate**

Dissertation  
zur Erlangung des Grades  
des Doktors der Naturwissenschaften  
der Naturwissenschaftlich-Technischen Fakultät  
der Universität des Saarlandes

*von*  
ABDUL RAHMAN SALHAB



Saarbrücken  
2019

**Tag des Kolloquiums:** 29.05.2020

**Dekan:** Prof. Dr. Guido Kickelbick

**Vorsitzender:** Prof. Dr. Volkhard Helms

**Berichterstatter:** Prof. Dr. Jörn E. Walter  
Prof. Dr. Tobias Marschall

**Akad. Beisitzerin:** Dr. Nicole Ludwig

﴿وَقُلْ رَبِّ زِدْنِي عِلْمًا﴾

سورة طه، آية 114

وَمَنْ لَمْ يَدُقْ مَرُّ التَّعْلَمِ سَلَحًا ..... تَجَرَّعَ ذُلَّ الْجَهْلِ طُولَ حَيَاتِهِ

الإمام الشافعي ( 767 غزة - 820 القاهرة )

الْعِلْمُ يَرْفَعُ بَيْتًا لَا عِمَادَ لَهُ ..... وَالْجَهْلُ يَهْدِمُ بَيْتَ الْعِزِّ وَالشَّرَفِ

الشاعر أحمد شوقي ( 1868 - 1932 ، القاهرة )

﴿And pray: My lord, increase me in knowledge﴾

– Quran, Chapter 20, Verse 114

“Who do not slightly experience a hurdle of learning; will suffer from humiliation of ignorance throughout his life”

– Al-imam Al-Shafi'i (767, Gaza – 820, Cairo)

“Knowledge builds up a house that has no mainstays; ignorance demolishes the house of honor and glory”

– The poet, Ahmed Shawqi (1868 – 1932, Cairo)

# Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor Prof. Dr. Jörn Walter for the continuous support of my PhD study and research, his guidance and encouragement. His support has never stopped since he accepted me as master student in his lab, and later as PhD candidate. I am very much grateful to him for introducing me into epigenetics field and for giving me the opportunity to work closely and collaborate with great scientist from DEEP project. I should also thank him for his support in different occasions outside academia.

I would like to thank the committee's members for investing their time and effort in reviewing my thesis and giving me the opportunity to present my work in front of their groups.

My gratitude and deep appreciation to my mentor, Karl Nordström, who introduced me to NGS data processing and analysis, and bash scripting. He helped me a lot to shape my future career for many years ahead.

I would like to thank all co-authors and collaborators whom I worked with during my PhD period. Special thanks to Dr. Julia K. Polansky-Biskup from whom I learnt a lot about immune system, and a collaborative work with her represents one pillar of this thesis. Also, a great thanks to Peter Ebert, from whom I learnt about Chip-seq data processing and reproducibility in computational research.

I am very much grateful to Gilles Gasparoni and Pascal Giehr for reading my thesis and providing me critical comments and feedback. I learnt a lot about molecular biology from both of you, many thanks. I am also thankful to Gilles for his kind help in correcting the German version of the abstract.

Big thanks to Epigenetic group at Saarland university for the great time, the nice environment, and the continuous support. I should not forget the former member and a friend, Pavlo Lutsik, with whom I shared the working space for almost three years, getting direct support from him for methylation analysis with RnBeads.

I would like to thank Gilles, Annemarie, Kathrin, Anna and Pascal for asking me to process internal and external data related to their projects. This helped me to develop my skills in data processing and data integration, and enriched my knowledge in biology.

I would like to thank Karl, Gilles and Annemarie with whom I shared the working

space and exciting discussions in different topics outside science. I would like to thank everybody, who helped and supported me to improve my German language. Particularly, Annemarie, Nicole, Pascal, Judith, Karl, Gilles and Nina.

I would like to thank my brother, Mohammad Salhab, who helped me to prepare the cover picture.

I owe my deepest gratitude to my family inside and outside Syria, particularly my parents, my sisters, my brother, my brothers in law and my mother in law, from whom I derived my determination to pursue my PhD. Without their unconditional support and love, this work would not have been possible.

Last but not least, I would like to thank my lovely wife, Heba, for her love, support, passion to my work and her patience, especially during the writing phase of the thesis. She has always been supportive and provided the best conditions to finish my PhD.

*Dedicated to:*

*My family*

*My uncle, Mohammad Salhab (†2017)*

# Abstract

DNA methylation is an epigenetic modification associated with gene regulation. It has extensively been studied in the context of small regulatory regions. Yet, not so much is known about large domains characterized by fuzzy methylation patterns, termed Partially Methylated Domains (PMDs). The present thesis comprises PMD analyses in various contexts and provides several new aspects to study DNA methylation.

First, a comprehensive analysis of PMDs across a large cohort of WGBS samples was performed, to identify structural and functional features associated with PMDs. A newly developed approach, ChromH3M, was proposed for the analysis and integration of a large spectrum of WGBS data sets. Second, PMDs were found to be indicators of the cellular proliferation history and segmented loss of DNA methylation in PMDs supports the sequential linear differentiation model of memory T-cells. Third, assessment of genome-wide methylation changes in PMDs of Multiple Sclerosis-discordant monozygotic co-twins did not show significant differences, but local changes (DMRs) were identified.

Taken together, the outcomes of the presented studies shed light on a so far neglected aspect of DNA methylation, that is PMDs, in different contexts; lineage specialization, differentiation, replication, disease, chromatin organization and gene expression.

# Kurzfassung

Die DNA-Methylierung ist eine epigenetische Modifikation, die funktionell mit der Genregulation verbunden ist. Sie wurde bereits ausführlich im Kontext kleiner regulatorischer Regionen untersucht. Es ist jedoch noch nicht sehr viel bekannt über große Domänen, welche erstmals in WGBS-Daten beschrieben wurden. Sie werden als partiell methylierte Regionen (PMDs) bezeichnet und sind durch das Vorhandensein variabler Methylierungsmuster charakterisiert. Die vorliegende Arbeit umfasst PMD-Analysen in unterschiedlichen Kontexten und liefert verschiedene neue Aspekte zur Untersuchung der DNA-Methylierung.

Zuerst wurde eine umfassende Analyse von PMDs in einer großen Kohorte von WGBS-Proben durchgeführt, um strukturelle und funktionelle Merkmale zu identifizieren, die mit PMDs assoziiert sind. Ein neu entwickelter Ansatz, ChromH3M, wurde für die Analyse und Integration einer großen Kohorte von WGBS Datensätzen angewandt. Zweitens wurde festgestellt, dass PMDs Indikatoren für die Zellproliferationshistorie sind, und der zu beobachtende graduelle Verlust der globalen DNA-Methylierung bei der Differenzierung von T-Gedächtniszellen unterstützt die Hypothese der sequenziellen linearen Differenzierung. Drittens zeigte die Bewertung der genomweiten Methylierungsänderungen in PMDs von Multiple Sklerose-diskordanten monozygoten Zwillingen keine signifikanten Unterschiede, jedoch wurden lokale Änderungen (DMRs) identifiziert.

Insgesamt geben die Ergebnisse der vorgestellten Studien Aufschluss über einen bislang eher vernachlässigten Aspekt der DNA-Methylierung, d.h. PMDs, in verschiedenen Zusammenhängen: der Festlegung der Zell-entwicklungsbahnen, der Zelldifferenzierung, der Replikation, der Krankheit, der Organisation des Chromatins, sowie der Regulation der Genexpression.



# Contents

<b>Acknowledgments</b>	<b>i</b>
<b>Abstract</b>	<b>iv</b>
<b>Kurzfassung</b>	<b>v</b>
<b>List of Abbreviations</b>	<b>x</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Chromatin Organization . . . . .	1
1.1.1 Methods to study the 3D genome . . . . .	3
1.1.2 Nucleosome . . . . .	5
1.2 Histone Modifications . . . . .	5
1.2.1 Genome-wide profiling of protein–DNA interactions . . . . .	7
1.2.2 ChIP-Seq data analysis . . . . .	7
1.3 DNA Methylation . . . . .	8
1.3.1 Establishment, maintenance and erasing of DNA methylation . . .	10
1.3.2 DNA methylation in different genomic contexts . . . . .	12
1.3.3 DNA methylation detection . . . . .	13
1.3.4 WGBS data analysis . . . . .	15
1.4 International effort for profiling reference epigenome maps . . . . .	20
1.5 Aim of the thesis . . . . .	20
1.6 Results outline . . . . .	21
<b>2 A comprehensive analysis of 195 DNA methylomes reveals shared and cell-specific features of partially methylated domains</b>	<b>24</b>
2.1 Main Text . . . . .	25
2.2 Supplementary Material . . . . .	38
<b>3 Epigenomic Profiling of Human CD4<sup>+</sup> T Cells Supports a Linear Differentiation Model and Highlights Molecular Regulators of Memory Development</b>	<b>60</b>
3.1 Main Text . . . . .	61
3.2 Supplementary Material . . . . .	76

## CONTENTS

<b>4 DNA methylation signatures of monozygotic twins clinically discordant for multiple sclerosis</b>	<b>93</b>
4.1 Main Text . . . . .	94
4.2 Supplementary Material . . . . .	106
<b>5 Discussion</b>	<b>149</b>
5.1 DNA methylation and chromatin organization . . . . .	149
5.2 PMDs as discriminators for cellular origin . . . . .	151
5.3 PMDs as indicators of proliferation history of cells . . . . .	152
5.4 PMDs in cancerous cells and immortalized cell lines . . . . .	153
5.5 Heterochromatic signatures of PMDs and replication domains . . . . .	154
5.6 PMDs in MS-discordant monozygotic twins . . . . .	155
5.7 Conclusion . . . . .	156
5.8 Outlook . . . . .	156
<b>A Appendix: List Of Publications</b>	<b>159</b>
<b>B Appendix: License and Copyright Information</b>	<b>160</b>
B.1 Manuscripts . . . . .	160
B.1.1 Integrative Analysis of PMDs . . . . .	160
B.1.2 Memory T-cells Differentiation . . . . .	160
B.1.3 DNA Methylation of MZ Twins . . . . .	160
B.2 Figure Reprints . . . . .	161
<b>Bibliography</b>	<b>162</b>

## List of Figures

1.1	Chromosome structure at different genomic scales . . . . .	2
1.2	Hi-C protocol . . . . .	4
1.3	Histone tail modification sites . . . . .	6
1.4	ChromHMM segmentation . . . . .	9
1.5	DNA demethylation mechanisms . . . . .	11
1.6	Partially Methylated Domains . . . . .	14
1.7	General scheme of bisulfite sequencing . . . . .	16
1.8	MethylSeekR segmentation . . . . .	18
5.1	Compaction of different classes of PMDs . . . . .	155
5.2	PMDs in different contexts . . . . .	158

## List of Tables

B.1 Licensing information for figure reuse . . . . .	161
--	-----

# List of Abbreviations

**5caC** 5-carboxycytosine 10

**5fC** 5-formylcytosine 10

**5hmC** 5-hydroxymethylcytosine 10, 11, 15, 153

**5mC** 5-methylcytosine 8, 10, 11, 13, 15, 17, 153

**ChIP-Seq** Chromatin Immunoprecipitation Sequencing vi, 7, 8, 20

**ChromH3M** ChromHMM-meta 22, 151

**DEEP** German Epigenome Program 8, 19–22

**DMRs** Differential Methylated Regions iv, v, 12, 14, 17, 22, 149, 150, 156

**DNA** Deoxyribonucleic Acid 1

**FDR** False Discovery Rate 19

**FMRs** Fully Methylated Regions 19

**GAM** Genome Architecture Mapping 3

**HMDs** Highly Methylated Domains 17–19, 149–151, 153, 155–157

**HMM** Hidden Markov Model 8, 17, 151, 157

**IHEC** International Human Epigenome Consortium 20, 21, 152, 156, 157

**LMRs** Low Methylated Regions 8, 12, 17–19, 152

**MS** Multiple Sclerosis 22, 155

**MZ** Monozygotic 22

**NDRs** Nucleosome depleted Regions 12

## List of Abbreviations

**PBMCs** Peripheral Blood Mononuclear Cells 22

**PETs** Paired-End Tags 4

**PHH** Primary Human Hepatocytes 154

**PMDs** Partially Methylated Domains iv, v, vii, 13–15, 17–19, 21–23, 60, 93, 149–157

**PTM** Post Translational Modifications 5

**RRBS** Reduced Representation Bisulfite Sequencing 14, 15, 157

**TADs** Topologically Associated Domains 2, 3, 149

**Tcm** Central memory T-cells 152

**Tem** Effector memory cells 152

**Temra** Terminally differentiated cells 152

**Tmem** Memory T-cells 9, 14, 18, 19, 21, 22, 152, 156

**Tn** Naive T-cells 152

**UMRs** Unmethylated Regions 17–19, 152

**WGBS** Whole Genome Bisulfite Sequencing iv–vi, 13–18, 20–22, 93, 149, 151, 152, 156

# Chapter 1

## Introduction

### Preface

The definition of epigenetics has been redefined multiple times since the first time it was coined by Conrad Waddington in early 1940s (reviewed in [Cavalli and Heard, 2019]). According to Riggs and Holliday, epigenetics refers to the study of the heritable phenotype changes that do not entail alterations in the DNA sequence [Holliday, 1994]. These changes encompass DNA methylation and histone modifications which control gene expression. Moreover, the higher order structure of the chromatin plays a role in gene regulation by bringing loci far away from each other, in the linear DNA, into proximity in space. Hence, joint efforts from different biological fields are needed to better understand how epigenetic changes impact and regulate gene expression.

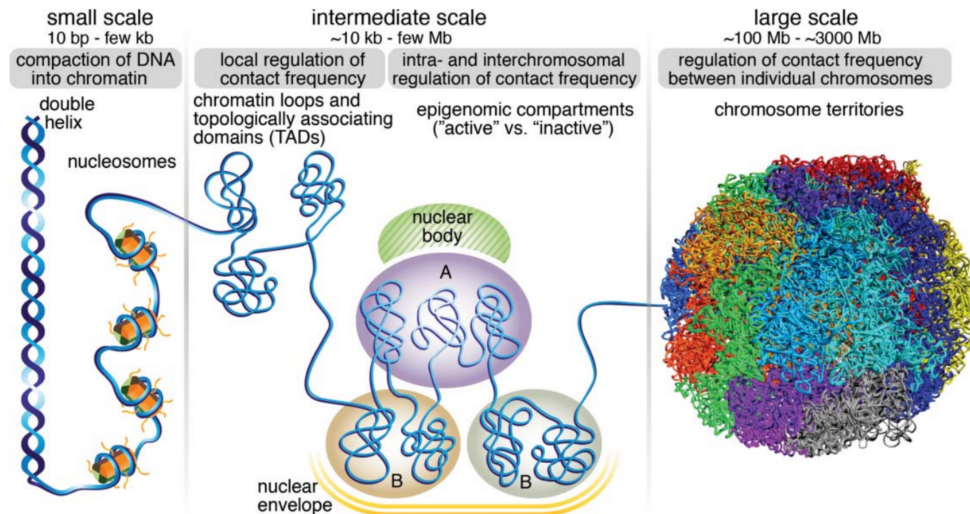
This chapter will guide you through a biological background of epigenetics and chromatin organization. Section 1.1 provides knowledge about the 3D chromatin architecture and the most used techniques to map chromatin contacts (associated to the results of Chapter 2). Section 1.2 and Section 1.3 outline the two epigenetic marks; DNA methylation and histone marks, and the current methods for charting these modifications (associated to Chapter 2, Chapter 3 and Chapter 4). Section 1.4 introduces an international effort to decipher high quality reference epigenomes of different primary human cell types. Section 1.5 and Section 1.6 introduce the goal of this thesis and outline the results of three studies.

### 1.1 Chromatin Organization

Chromatin is a complex of DNA and proteins in eukaryotic cells and it is compacted in such a way that 2m of DNA fits in a 10 $\mu$ m-sized nucleus [Nuebler et al., 2018]. This is guaranteed through the hierarchy of the 3D genome (reviewed in [Gibcus and Dekker, 2013]), where the genome is organized at different scales (Figure 1.1); (i) small scale where the DNA is wrapped around a histone octamer to form a nucle-

osome (see Section 1.1.2), (ii) intermediate scale<sup>‡</sup> which includes (a) loops that bring two (or more) loci far away from each other into proximity to modulate gene transcription, (b) Topologically Associating Domains (TADs) [Dixon et al., 2012, Nora et al., 2012, Sexton et al., 2012] which are domains, highly enriched by intra-chromosomal interactions, and (c) transcriptionally active compartments (A compartments) and transcriptionally inactive compartments (B compartments) as defined by Hi-C experiments [Lieberman-Aiden et al., 2009]. These domains were found to coincide with “metaTAD” tree (hierarchy of domains-within-domains [Fraser et al., 2015], (iii) finally, the largest scale comprises chromosomal territories [Cremer and Cremer, 2001, Manuelidis, 1985], which are made up by A- and B- compartments where neighboring chromosomes interweave. The stability of these different hierarchical structures is different within cells; the larger the structures, the more stable they are [Gibcus and Dekker, 2013].

Therefore, a better understanding how the genome operates and how genes are regulated requires to study the 3D genome organization and integration of this knowledge with information encoded in the linear genome. An overview of such technologies to study the higher-order chromatin structure is presented in the next section.



**Figure 1.1: DNA packaging into chromatin.** DNA is compacted in a hierarchical manner forming nucleosomes by wrapping DNA around ‘core’ histones. On a larger scale, chromatin loops are formed to bring loci into proximity. TADs are then formed with lengths  $\sim 10\text{kb}$  - few Mb. Multiple TADs make up active (A) and inactive (B) compartments which compose chromosome territories at a larger scale ( $\sim$  hundreds of Mb). (From [Hansen et al., 2018], see Table B.1 for the license.)

<sup>‡</sup>This thesis focuses on studying DNA methylation at this scale. However, the nomenclature “large scale” is used within this work as it contrasts studying DNA methylation at smaller scale, e.g., regulatory regions



### 1.1.1 Methods to study the 3D genome

#### Experimental methods

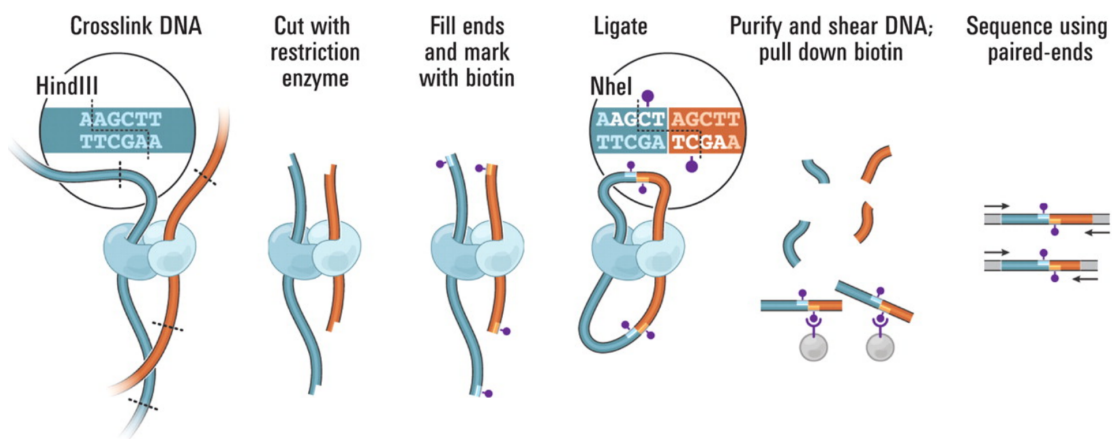
During the past decade new technologies (3C-based methods [Dekker et al., 2002]) have emerged to study the nuclear structure and three-dimensional organization of the genome, allowing high-throughput mapping of DNA to DNA contacts across different genomic regions and chromosomes (reviewed in [Pombo and Dillon, 2015]). These technologies comprise different flavours of 3C technology; Capturing Chromosome Conformation (3C) [Dekker et al., 2002], chromosome conformation capture-on-chip (4C) [Simonis et al., 2006], carbon-copy chromosome conformation capture (5C) [Dostie et al., 2006], Hi-C [Dixon et al., 2012, Lieberman-Aiden et al., 2009, Sexton et al., 2012] (which is relevant to this work, see Chapter 2), genome conformation capture (GCC) [Rodley et al., 2009], tethered conformation capture (TCC) [Kalhor et al., 2012], multiplexed 3C sequencing (3C-seq) [Stadhouders et al., 2013], capture-C [Hughes et al., 2014] and targeted chromatin capture (T2C) [Kolovos et al., 2014]. All these methods engage crosslinking and proximity based ligation of close-by regions, DNA fragmentation by restriction enzyme, followed by quantification of ligated products that are explained as chromatin contacts [Pombo and Dillon, 2015]. Despite the wealth of information that 3C-based methods has revealed (loops, TADs, A and B compartments), these methodologies have some limitations due to digestion and ligation principle; like GC-biases. They are inadequate to quantify simultaneous contacts between multiple chromatin regions, and they do not provide information about chromatin associations with the nuclear periphery. A new genome-wide ligation-free method, Genome Architecture Mapping (GAM), was introduced [Beagrie et al., 2017], which combines ultra-thin cryosectioning with laser microdissection and DNA sequencing. TSA-Seq is an immunocytochemistry method to study the cytological distance of genomic loci relative to a particular nuclear compartment [Chen et al., 2018]. Other researchers approached the same problem using imaging technologies utilizing super-resolution microscopes to visualize loci and sub-nuclear structures inside (living) cells [Chen et al., 2013, Hess et al., 2006, Nir et al., 2018].

As it is mentioned before there are many methods to study the higher-order chromatin structure and each has its own advantages and limitations. To harmonize the work and foster the efforts of the experts in the 3D genome field, the 4D Nucleome project was emerged [Dekker et al., 2017], which aims to better understand the three-dimensional organization of the nucleus in space and time (the 4<sup>th</sup> dimension) through developing standardized experimental and computational methods.

### Hi-C for mapping chromatin interactions

Hi-C is a genome-wide method to study long-range interactions by pairing proximity-based ligation with massively parallel sequencing [Lieberman-Aiden et al., 2009]. The major steps of this method can be summarized as follow (Figure 1.2):

1. cells are cross-linked with formaldehyde
2. DNA is digested with restriction enzyme like DpnII
3. ends are filled with biotin and ligated
4. DNA is purified and sheared
5. fragments with biotin are pulled down by streptavidin beads and followed by high-throughput sequencing



**Figure 1.2: Hi-C protocol.** DNA cross-linking; DNA digestion with a restriction enzyme; filling the ends with biotin; ends ligation; DNA purification and shearing; biotinylated junctions (which represent fragments that were originally in close spatial proximity) are sequenced. (From [Lieberman-Aiden et al., 2009]. Reprinted with permission from AAAS, see Table B.1 for license)

### Hi-C data analysis

Although there are many different bioinformatic tools available to process Hi-C data, most of them share the main steps: mapping paired-end reads to the reference genome (separately) to get the distal interacting tags, pairing to get the paired-end tags (PETs), filtering invalid PETs (singletons or randomly pulled down fragments), binning the whole genome (with multiple resolutions) to count interactions frequency between genomic regions, normalization to account for biases in Hi-C library, downstream analysis such as chromatin loop detection, TADs and A/B compartments calling. The interaction frequencies are usually visualized as heatmaps [Forcato et al.,

2017, Han and Wei, 2017].

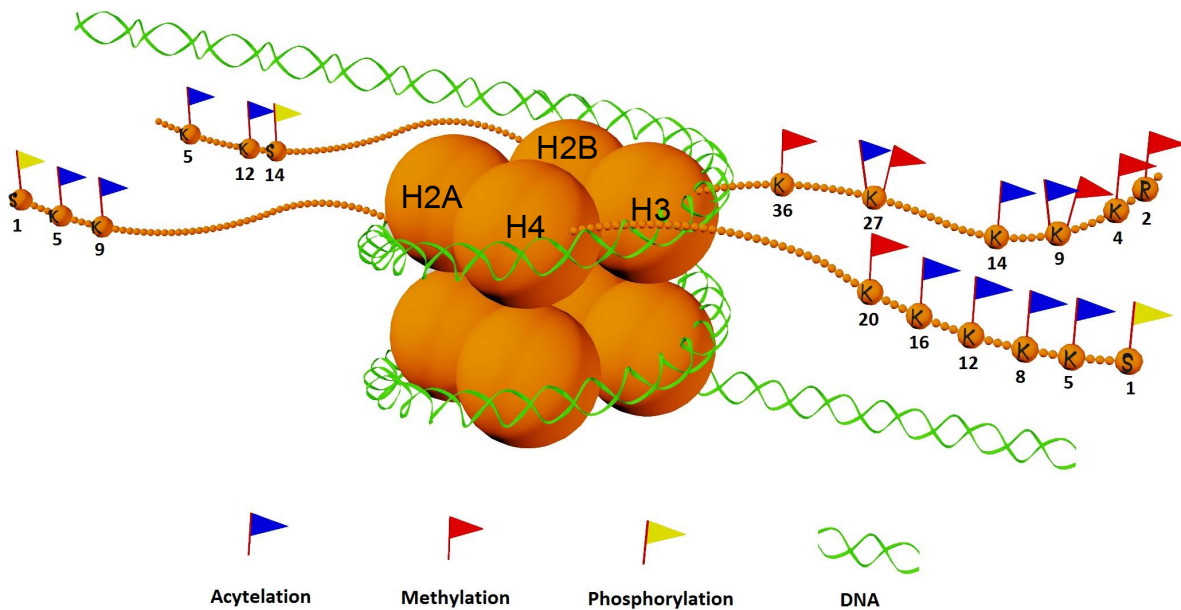
The Hi-C data used in Chapter 2 was processed and visualized using HiCExplorer [Ramírez et al., 2018, Wolff et al., 2018].

### 1.1.2 Nucleosome

Nucleosome is the basic building unit of chromatin and consists of 146 base pair of DNA wrapped around an octamer of histone proteins (two molecules of each ‘core’ histone H2A, H2B, H3 and H4) [Wolffe, 1998] (Figure 1.3). H1 histone enhances this structure and binds to the “linker DNA” region between nucleosomes, helping to stabilize the chromatin higher-order structure [Berezney and Jeon, 1995]. Histones have tails which can be subject to post translational modifications (PTM) (or histone marks) like methylation, acetylation, phosphorylation and ubiquitination (Figure 1.3) which can influence chromatin compaction, gene transcription and expression. More details about PTM will be given in the next section.

## 1.2 Histone Modifications

Histone can be modified, removed or reprogrammed by different groups of enzymes which leads to opening or closing the chromatin. While ‘writers’ (establishing enzymes) deposit histone marks, ‘erasers’ (de-modifying enzymes) remove them. For instance, acetylation of lysines is dynamic and governed by two enzyme families working in antagonistic manner. The ‘writer’ histone acetyltransferases (HATs) transfer acetyl-groups from Acetyl-Co-A to lysine facilitating the opening of the chromatin by reducing the DNA-histone contacts, but the ‘eraser’ histone deacetylases (HDACs) reverse the lysine acetylation. ‘readers’ which are usually part of a large complex bind to histone modifications to apply downstream functions [Strahl and Allis, 2000]. More than one hundred histone modifications have been reported with their functional role [Khare et al., 2011, Kouzarides, 2007] (Figure 1.3 shows some of them). The most two studied marks are methylation and acetylation of lysines. H3K4me3 is found to be associated with promoters of active genes [Barski et al., 2007, Bernstein et al., 2005, Santos-Rosa et al., 2002, Strahl et al., 1999] while H3K36me3 is found to be enriched in the gene bodies and associated with active transcription [Bannister et al., 2005]. H3K4me1 is considered as an enhancer mark, while H3K27ac as an active enhancer [Creyghton et al., 2010] and active promoter mark [Wang et al., 2008]. Heterochromatic regions are usually enriched by H3K9me3 [Barski et al., 2007, Hall et al., 2002, Lippman et al., 2004, Martens et al., 2005] and H3K27me3 [Barski et al., 2007, Litt et al., 2001, Ringrose and Paro, 2004]. However, it is not always easy and trivial to assign a functional role to a certain histone mark to describe the chromatin



**Figure 1.3: Histone tail modification sites.** Depiction of a nucleosome, the basic building unit of the chromatin. DNA (in green) is wrapped around an octamer of histone proteins. The tails of these histones may carry different post translational modifications. Active marks usually include lysine (K) acetylation (blue flag, such as H3K27ac), arginine (R) methylation (red flag) and some lysine methylation, such as H3K4me1, H3K4me3 and H3K36me3. Repressive marks include H3K9me3 and H3K27me3 (From [Salhab, 2014]).

state of a genomic region. In some cases, a chromatin state is characterized by co-occurrence of two (or more) antagonistic marks, such as “bivalent/poised” promoter or enhancer. The term “bivalent” domains was introduced by Bernstein et al. to describe the poised state of promoters of important developmental genes in ES cells that keep them ‘on-hold’ until they receive a suitable stimulus and become active rapidly [Bernstein et al., 2006]. The TSSs of these genes were occupied simultaneously by active (H3K4me3) and inactive (H3K27me3) marks. Upon differentiation, some of the bivalent genes became silent and lost H3K4me3 but preserved H3K27me3, while the expressed genes lost H3K27me3. Interestingly, the different histone marks exhibit different localization in the genome. The three active marks H3K4me1, H3K4me3 and H3K27ac present in the genome as sharp or ‘narrow’ marks, whereas the heterochromatic marks H3K27me3 and H3K9me3 distribute across ‘broad’ domains. Another active mark, H3K36me3 can also be seen as a ‘broad’ mark. Different techniques for genome-wide profiling for histone marks are discussed in the next section.

### 1.2.1 Genome-wide profiling of protein–DNA interactions

Chromatin Immunoprecipitation Sequencing (ChIP-Seq) is the most commonly utilized method for genome-wide profiling of histone marks [Barski et al., 2007, Johnson et al., 2007]. First, proteins are cross-linked with their bound DNA using formaldehyde, then the chromatin is sheared to fragment the DNA. Next, a specific antibody is used to capture the protein of interest (histone modification) and the associated DNA is isolated. Finally, after reverse cross-links, DNA fragments are purified, amplified (PCR) and sequenced. Although this method has some limitations, it is still widely accepted and used by the community (see Chapter 1.4). New methods have been developed to cope with the drawbacks of ChIP-Seq. For instance, DamID [van Steensel et al., 2001] and ChEC-seq [Zentner et al., 2015] do not include immunoprecipitation or crosslinking, that can lead to epitop masking, but rather they are dependent on DNA-modifying enzymes. CUT&RUN [Skene et al., 2018] allows to work with low cell-numbers (100 cells) making it suitable to study rare cell types. Additionally, it has lower background compared to ChIP-Seq and could be done in one day. CUT&TAG is an improvement of CUT&RUN to allow single cell application [Kaya-Okur et al., 2019]. The resolution of ChIP-Seq is limited by sonication, but methods like ChIP-exo [Rhee and Pugh, 2011], high-resolution X-ChIP [Skene et al., 2014] and ChIP-nexus [He et al., 2015] can improve its resolution by adding nuclease digestion steps. To enhance the scalability and universality of ChIP-Seq, a barcoding system RELACS was implemented allowing to profile hundreds of ChIP-Seq samples in three days [Arrigoni et al., 2018].

### 1.2.2 ChIP-Seq data analysis

The usual workflow for processing ChIP-Seq starts with filtering step for low quality reads and sequencing adaptors, followed by mapping the reads to a reference genome and finally peak calling to predict protein-DNA interaction sites. There are some technical issues should be taken into account when processing ChIP-Seq data:

- peak calling is a very hot topic and more attention should be brought to this issue. As it is mentioned earlier, histone marks have different localization and distribution across the genome. Hence, the proposed algorithms to call ‘narrow’ peaks may not be appropriate to call ‘broad’ domains. MACS is widely used tool for calling peaks from ChIP-Seq data [Zhang et al., 2008]. However, other tools have been implemented to call domains of broad histone marks, like histoneHMM [Heinig et al., 2015] and RSEG [Song and Smith, 2011].
- It was found that some genomic regions show artificially high signal often found at centromeres, telomeres and satellite repeats. Such regions are important to

be removed when computing Pearson correlation between genome-wide tracks. These regions were termed ‘Blacklist’ and was generated by ENCODE consortium [ENCODE et al., 2012].

ChIP-Seq samples used within this cumulative work (Chapter 2 and Chapter 3) were processed with DEEP pipeline (<http://doi.org/10.17617/1.2W>) [Ebert et al., 2015]. The above mentioned methods for peaks/domains calling are capable of analyzing each sample separately. To get a comprehensive picture of chromatin states when profiling multiple, histone marks segmentation tools are available to tackle this issue by integrating the different samples in a multivariate hidden markov model (HMM) (e.g., ChromHMM [Ernst and Kellis, 2012] and EpiCSeq [Mammana and Chung, 2015]). The output of such tools will be different states representing distinct signatures of histone marks (ChromHMM 18-state track in Figure 1.4a and the left panel in Figure 1.4b). The next step will be to interpret these states and assign meaningful biological labels for each state. Usually, this is done through computing overlap and neighborhood enrichment of each state across functional annotations (Figure 1.4b).

### 1.3 DNA Methylation

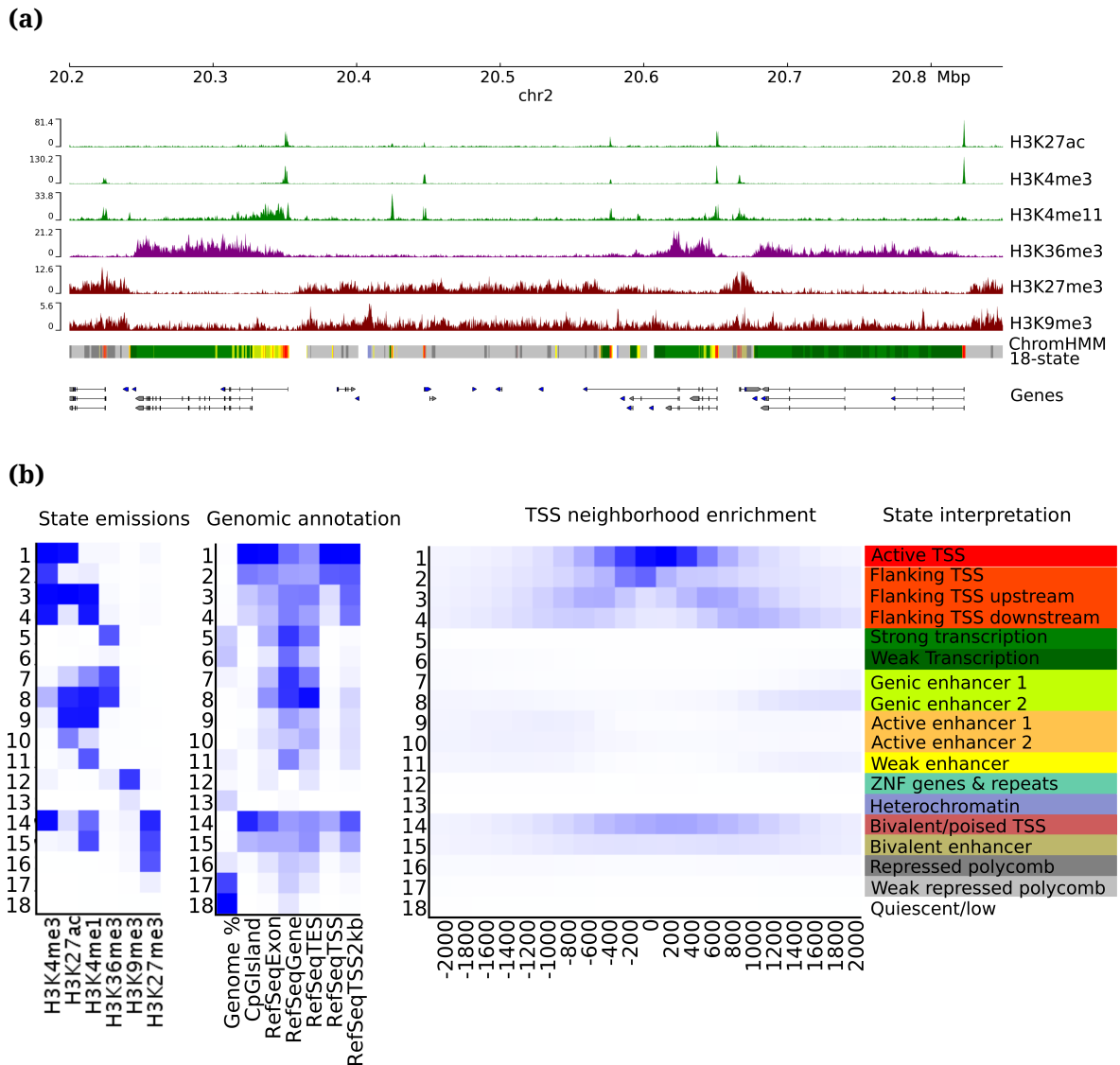
The process of adding a methyl group to DNA molecule is called DNA methylation. It can happen at the cytosine or adenine residues. The most common form of DNA methylation is methylation of cytosine at the fifth carbon atom of the pyrimidine ring which is termed ‘5-methylcytosine (5mC)<sup>‡</sup>. In 1975, 5mC in the context of CpG dinucleotides has been proposed as an epigenetic mark in vertebrates [Holliday and Pugh, 1975, Riggs, 1975] and it is common in eukaryotes and prokaryotes, although its rate is different across species. It is very abundant in mammalian genomes and occurs predominantly in the CpG context where the methylation, generally, are high except for short CpG dense regions (~ 1 kb length), so called CpG islands (CGI), which often are devoid for methylation [Feng et al., 2010, Zemach et al., 2010]. The majority of CGI are located at gene promoters. Additionally, CpG-poor regions were found to be lowly methylated (termed as LMRs; less than 50% methylated and contain less than 30 CpGs) and associated with distal regulatory elements and occupied by cell-type-specific transcription factors [Burger et al., 2013, Stadler et al., 2011]. Numerous studies clarified the impact of 5mC on gene control in different regulatory contexts (reviewed in [Jones, 2012]). For instance, the methylation of a promoter region is mainly associated to the silencing of the respective gene, but the high methylation in gene body may trigger the transcription elongation if the respective promoter is

---

<sup>‡</sup>There are different derivatives of 5mC (see 1.3.1, Active demethylation). However, the focus of this thesis is only on 5mC in CpG context.

## CHAPTER 1. INTRODUCTION

unmethylated and it may even influence splicing [Jones, 2012]. Moreover, methylation in repetitive regions like centromeres and telomeres is important for genome stability and integrity through the suppression of the expression of transposable elements [Moarefi and Chédin, 2011]. More details about the role of DNA methylation in



**Figure 1.4: ChromHMM output example.** **(a)** Tracks of the 6 core histone marks of DEEP effector memory T cell (Tmem) with chromatin states underneath generated by 18-state ChromHMM model (trained on 98 epigenome from ROADMAP consortium). The green tracks show three active 'narrow' marks, the 'broad' active histone mark is shown in purple and the two repressive marks are shown in dark red. Each color in ChromHMM 18-state segmentation represent distinct state named in **(b)**. **(b)** The left panel represent a heatmap of the emission parameters where the rows correspond to different states and the columns correspond to different marks. The darker the blue the more the propability of observing the mark in the state. Next to the left panel is a heatmap of overlap fold enrichment in TEM across different genomic annotations. The next panel shows the fold enrichment for each state for each 200 bp bin within 4kb around the TSSs. The right panel shows biological labels assigned to each state. Panel (a) was prepared using *pyGenomeTracks* package [Ramírez et al., 2018]

different functional annotations are presented in section 1.3.2

### 1.3.1 Establishment, maintenance and erasing of DNA methylation

In mammals, DNA methylation patterns are established during the early embryo development by *de novo* methylating enzymes called Dnmt3a and Dnmt3b and can be maintained during cell division by Dnmt1 allowing these patterns to be transmitted through cell generations [Allis et al., 2015]. Nevertheless, DNA methylation patterns are not permanent and changes can occur as a physiological response to environmental changes or can be associated with pathological processes. One mechanism that can remove DNA methylation marks is “active demethylation” in which DNA hydroxylases (TET proteins) are involved. Another mechanism “passive demethylation” is dependent on inhibition/loss of maintenance methyltransferase, Dnmt1, during DNA replication (see below and Figure 1.5) [Allis et al., 2015, p. 424].

DNA demethylation can happen genome wide or locally. After fertilization both maternal and paternal genomes undergo a massive genome wide loss of methylation throughout different proposed mechanisms. In the paternal genome, methylation is directly lost after fertilization and before the start of DNA replication, suggesting an active mechanism [Mayer et al., 2000, Oswald et al., 2000], whereas in the maternal genome the methylation is lost after consecutive cell divisions, suggesting a passive replication-dependent mechanism [Mayer et al., 2000]. Global DNA demethylation can also happen in primordial germ cells (PGCs) [Hajkova et al., 2002, Morgan et al., 2005]. On the other hand, gene-specific demethylation occurs during lineage-specific differentiation [Ji et al., 2010].

#### Passive demethylation<sup>‡</sup>

Passive demethylation (Figure 1.5) is triggered by cell division after successive rounds of DNA replication due to lack of Dnmt1 which is known to be responsible for maintenance methylation by recognizing hemimethylated CpG sites (CpG that is only methylated on one strand) and methylating the cytosine on the newly synthesized strand. In addition, it has been reported that Dnmt3a and Dnmt3b contribute to the maintenance of DNA methylation patterns [Chen et al., 2018, Liang et al., 2002].

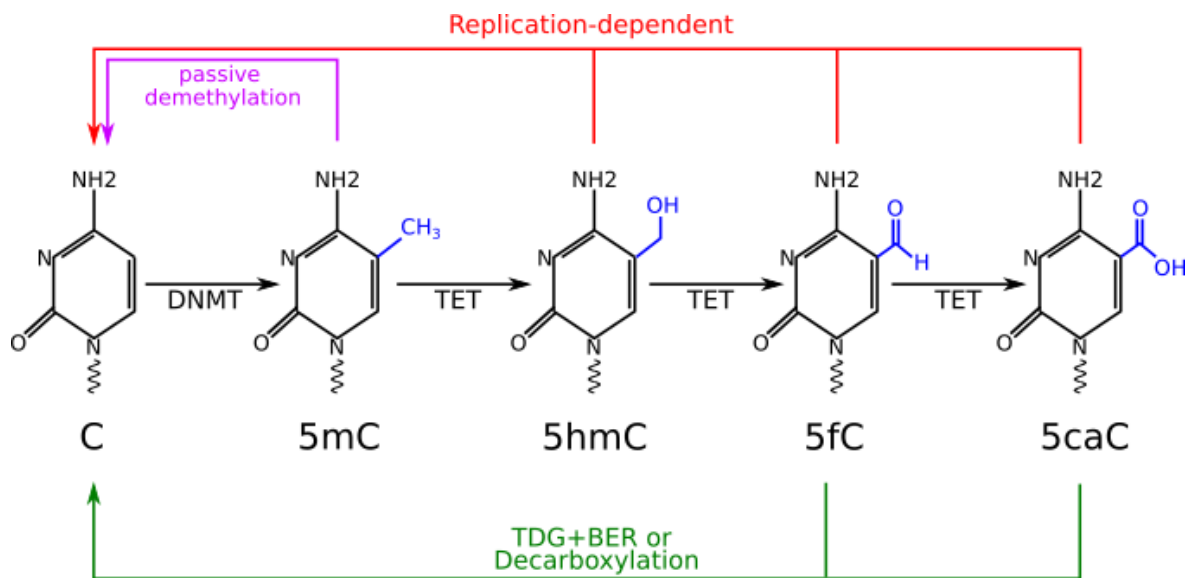
#### Active demethylation<sup>‡</sup>

Active demethylation is mediated by TET proteins that can oxidize iteratively the 5mC into 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxycytosine (5caC). These 5mC derivatives can be reversed into unmodified C by different mech-

---

<sup>‡</sup>This section is based on the review by [Wu and Zhang, 2017]





**Figure 1.5: DNA demethylation pathways.** Passive demethylation of 5mC (in purple) occurs as a result of DNA replication coupled with absence/block of Dnmt1. Active demethylation is mediated by TET enzymes that can oxidize 5mC into 5hmC, 5fC and 5caC. These oxidized forms can be reverted into unmodified C either through replication-dependent pathway (in red) or TDG+BER pathway (in green).

anisms (reviewed in [Bochtler et al., 2017, Wu and Zhang, 2014, 2010]). But the most accepted models are (Figure 1.5):

- **TDG-BER pathway:** through excision of 5fC/5caC by glycosylases such as TDG coupled with base excision repair (BER) to generate C [He et al., 2011, Maiti and Drohat, 2011, Weber et al., 2016]. This model is DNA replication-independent and is known as active modification-active removal (AM-AR) [Kohli and Zhang, 2013].
- **Replication-dependent dilution of oxidized 5mC:** this model is known as active modification-passive dilution (AM-PD) [Kohli and Zhang, 2013]. During DNA replication hemi-modified CpG dyads are created by incorporation of unmodified cytosine in the newly synthesized strand. UHRF1 recognizes 5mC:C dyad and recruits Dnmt1 to the hemi-methylated cytosines. Several *in vitro* studies showed that Dnmt1 is much less efficient at 5hmC:C, 5fC:C and 5caC:C dyads than at a 5mC:C dyad [Hashimoto et al., 2012, Ji et al., 2014, Otani et al., 2013] and hence after successive cell divisions, the oxidative forms of 5mC become demethylated.

### 1.3.2 DNA methylation in different genomic contexts

#### Methylation at transcription start sites

CGIs in somatic cells remain unmethylated and usually promoters of active genes are characterized by nucleosome depleted regions (NDRs) at the TSS, and these NDRs are usually enriched by H3K4me3 [Kelly et al., 2010]. Repression of CGI promoters can happen in different ways like i) repression mediated by polycomb proteins (H3K27me3) or ii) repression mediated by methylation of CGIs. So far it is not clear whether the silencing or the methylation comes first. [Lock et al., 1987] showed that methylation acts as a 'lock' to keep the previously silenced state of X-linked genes. Moreover, CGI promoters of mediated polycomb silenced genes are more likely to gain methylation in cancer cells. This suggests that methylation follows silencing, however the current data are not enough to make a strong statement (reviewed in [Jones, 2012]).

#### Methylation at enhancers

Enhancers are key regulatory elements that control gene expression in a tissue-specific manner. However, there is no clear definition for these elements, they are usually defined as genomic regions demarcated by H3K4me1 histone mark. In DNA methylation context they are linked to 'low methylated regions (LMRs)' which are CpG-poor regions (less than 30 CpGs) that have average methylation levels below 50% [Burger et al., 2013]. Enhancers have been linked to differentially methylated regions (DMRs) of differentiation specific genes when studying two closely related T cell populations [Schmidl et al., 2009].

#### Methylation at gene bodies

Long time ago gene body methylation was linked to gene transcription [Wolf et al., 1984] and a positive relationship between gene body methylation and active transcription has been confirmed on the active X chromosome [Hellman and Chess, 2007]. Moreover, gene body CGIs are highly methylated and do not prevent transcription elongation. One can conclude that the presence of 5mC regulates the transcription, and it is very much dependent on the genomic context of this mark. It was shown that exons are highly methylated compared to introns and there are transitions in DNA methylation level at the exon-intron boundaries, which suggests a role for DNA methylation in splice variance regulations [Laurent et al., 2010].

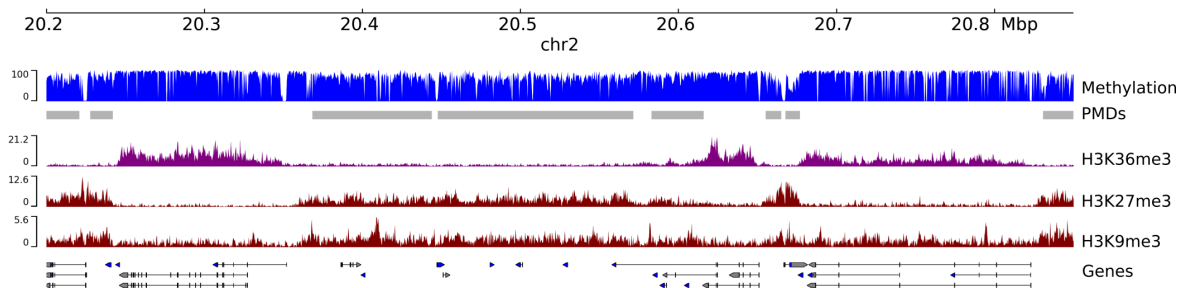
### Partially methylated domains

While DNA methylation has been extensively studied in the context of small regulatory regions like CpG-islands [Meissner et al., 2008], CpG shores [Irizarry et al., 2009], insulators [Bell and Felsenfeld, 2000, Shukla et al., 2011], promoters and enhancers [Burger et al., 2013, Stadler et al., 2011], not so much is known about large scale DNA methylation landscape and its impact on gene regulation and genome organization. With the first published whole genome bisulfite sequencing (WGBS) data set in 2009, a new term has been introduced, partially methylated domains (PMDs), referring into long genomic regions (mean length  $\sim$  150kb) characterized by fuzzy methylation patterns [Lister et al., 2009] (Figure 1.6). They were initially found in fibroblast cell line IMR90 but not in human embryonic stem cells H1. Since then PMDs have not gotten a great attention except for few isolated studies limited to cultured (cancer) cell lines such as medulloblastoma [Hovestadt et al., 2014], adipocyte tissue [Lister et al., 2011], SH-SY5Y neuronal cells [Schroeder et al., 2011], and human cancers [Berman et al., 2012, Hansen et al., 2011, Hon et al., 2012, Timp et al., 2014]. Moreover, some of these studies are based on 450K arrays data which only cover around 2% of CpGs in human. Schroeder et al. reported for the first time PMDs in a non-cancer primary human placenta tissue [Schroeder et al., 2013].

**PMD characterizations and functions** PMDs have been reported as long genomic regions that can cover up to 40% of the genome [Lister et al., 2009, 2011] and were characterized with highly disordered methylation levels. Later, it has been shown that PMDs are enriched by repressive histone marks H3K27me and H3K9me3, and are gene-poor regions and encompass lowly expressed and silenced genes (Figure 1.6) [Hon et al., 2012, Hovestadt et al., 2014], suggesting PMDs as marks of transcriptional repression. Moreover, they tend to be correlated with late replication timing [Aran et al., 2010] and nuclear lamina-associated regions [Berman et al., 2012].

### 1.3.3 DNA methylation detection

Several technologies have been developed to map 5mC levels on a genome-wide scale or in a locus-specific manner. These methods can be classified into three main groups (reviewed in [Yong et al., 2016]); **i) Restriction enzyme-based methods** in which a restriction enzyme has a preference to cut at certain sequence but it is sensitive to the DNA methylation state. e.g., the comprehensive high-throughput arrays for relative methylation (CHARM) method utilizes McrBC enzyme to fragment DNA and then uses array hybridization [Irizarry et al., 2008]. The method is quantitative and it has low cost allowing to profile large number of samples; **ii) Affinity enrichment-based methods** which use either methyl-CpG-binding domain (MBD) proteins or an-



**Figure 1.6: Partially methylated domains (PMDs).** A snapshot of  $\sim 600$ Kb of chr2 showing tracks of methylation and detected PMDs with corresponding three broad histone marks of Tmem cells (same sample of Figure 1.4a). The methylation values represented as blue bars which have values between 0 (unmethylated) and 100 (fully methylated). Below the methylation track is the detected PMDs by MethylSeekR tool [Burger et al., 2013]. PMDs enriched by heterochromatic marks H3K27me3 and H3K9me3 but depleted for the active mark H3K36me3. This figure was prepared using *pyGenomeTracks* package [Ramírez et al., 2018]

tibodies specific for 5mC to immunoprecipitate DNA with methylated CpG sites (like MeDIP). For the latter, DNA fractions can be then evaluated by arrays (MeDIP-chip) [Weng et al., 2009] or high-throughput sequencing (MeDIP-seq) [Zhao et al., 2014]; **iii) Bisulfite conversion-based methods** which is the most reliable method to test all cytosines in the genome. It involves sodium bisulfite treatment to modify the unmethylated cytosine into uracil (and during PCR, it is translated into thymine), whereas the methylated cytosines remain protected [Frommer et al., 1992]. Such methods provide single-base resolution and can be combined with array or high-throughput sequencing (WGBS) to study DNA methylation genome wide.

A benchmarking study showed that different technologies (MeDIP-seq, MethylCap-seq, RRBS and Infinium HumanMethylation27 array) measure DNA methylation accurately but they vary in the number of covered CpGs, sequencing depth and hence the cost [Bock et al., 2010]. Additionally, they differ in the number of detected DMRs. This can be related to the assay specific limitations, e.g., CpG-poor regions are difficult to be caught by MeDIP-seq and RRBS.

### 1.3.3.1 Array methods

These methods are based on hybridization of the converted DNA (through bisulfite treatment) to an array which have probes to discriminate between the methylated and unmethylated Cs. The most popular array is Illumina’s Infinium HumanMethylation450 BeadChip (HM450K) which has almost 450k probes covering mostly CpG islands, shores and shelves. The most recent version of this array is the Infinium MethylationEPIC BeadChip which covers  $>90\%$  of 450k sites in addition to CpG sites in enhancers defined by ENCODE [ENCODE et al., 2012] and FANTOM consortia [Forrest et al., 2014]. Many cancer studies use this method to study DNA methylation on

a genome wide level.

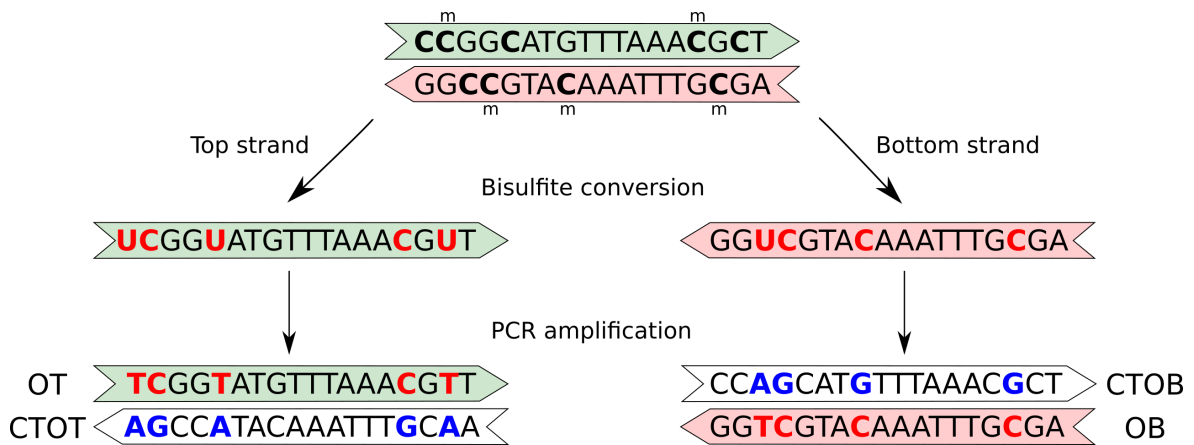
### 1.3.3.2 Whole genome bisulfite sequencing (WGBS)

The general WGBS protocol consists of the following main steps: shearing the genomic DNA into fragments which are then end-repaired, followed by adding an adenine overhang (A-tailing) to the 3' end, and ligation of the methylated adaptors into the fragments, which are then size-selected and bisulfite treated. Products are purified and amplified by PCR (Figure 1.7) and finally the resulting library is sequenced [Lister et al., 2009]. It should be noted that bisulfite treatment is a harsh treatment and lots of the starting material is lost, so it is recommended to start with a high amount of DNA to avoid over amplification. Additionally, bisulfite treatment can not distinguish between 5mC and 5hmC. On the other hand, WGBS measures almost all Cs in the genome allowing to study DNA methylation not only in CpG context but also in non CpG context. Moreover, WGBS allows to measure DNA methylation at CpG-poor regions which are hardly detectable by other methods. WGBS of IMR90 revealed long regions that compose one third of the genome, with average methylation levels less than 70%, termed PMDs (section 1.3.2) [Lister et al., 2009]. Although this method became a gold standard to study DNA methylation in many epigenomic consortia such as Roadmap, ENCODE, IHEC, Blueprint and DEEP (section 1.4), it remains the most expensive technique. To reduce the cost of WGBS, Meissner et al. developed a method called reduced representation bisulfite sequencing (RRBS) [Meissner et al., 2008]. This method is cost-effective in comparison to WGBS because only a fraction of the genome is sequenced (1-3% of the mammalian genome) which cover CpG-rich regions in close proximity to the restriction enzyme's recognition sequence (in this case MspI enzyme). A combination of different restriction enzymes (like Alu and HaeIII) can be used to increase the number of covered CGs but this will increase the sequencing cost as well.

### 1.3.4 WGBS data analysis

As it is mentioned in section 1.3.3 the bisulfite treatment involves C to T conversion. Hence, the quantification of methylated Cs will be done by identifying C-to-T conversions in the aligned reads and then calculating number of Cs divided by the sum of number of Cs and Ts for each cytosine in the genome.

The general workflow for processing WGBS data starts by trimming sequencing adaptors followed by the mapping the sequencing reads into the genome, then methylation calling for each CpG, and finally ends up with quality control using different measurements. The alignment step is not so trivial for many reasons; **i**) depletion of cytosines, due to the bisulfite treatment, reduces the sequence complexity of the



**Figure 1.7: General scheme of bisulfite treatment outcome.** Bisulfite treatment of DNA converts cytosine residues into uracil and then into thymine during PCR amplification, whereas methylated cytosine residues remain unmodified. PCR amplification of bisulfite converted DNA during WGBS library preparation will give rise to four different DNA sequences. OT, original top strand; CTOT, complement of the original top strand; OB, original bottom strand; CTOB, complement of the original bottom strand. This figure is adapted from [Krueger et al., 2012], see Table B.1 for the license.

resulting reads which leads to imperfect reads mapping or wrong alignment to the reference genome; **ii**) the bisulfite conversion rate is not always 100% efficient, meaning that some unmethylated cytosines will not be converted and then it will appear as methylated cytosines; **iii**) problems related to short-reads aligners including repetitive elements mapping and sequencing errors. An overview and recommendations for bisulfite alignment tools are available from [Bock, 2012, Krueger et al., 2012].

Two approaches have been developed to tackle these challenges; **i**) Wild-card aligners where Cs, in the genomic DNA sequence, are replaced by letter Y which match both Cs and Ts in the read, or the aligner does not penalize C to T mismatches. Examples of such aligners are BSMAP [Xi and Li, 2009], GSNAP [Wu and Nacu, 2010], Last [Frith et al., 2012], Pash [Coarfa et al., 2010], RMAP [Smith et al., 2009], RRBSMAP [Xi et al., 2011] and segemehl [Otto et al., 2012]; **ii**) three-letter aligners such as Bismark [Krueger and Andrews, 2011], BRAT [Harris et al., 2012, 2009], BS-Seeker [Chen et al., 2010] and MethylCoder [Pedersen et al., 2011] convert all Cs in the sequencing reads and in both strands of the reference genome into Ts and then apply standard mapping methods on this reduced base-space.

After the alignment step, DNA methylation levels are inferred by calculating the ratio between the observed Cs and the total number of Cs and Ts at each assayed CpG. However, the accuracy can be improved by including additional steps as Bis-SNP caller does [Liu et al., 2012]; first: local realignments in loci with known variation (could be taken from public database dbSNP or can be sample specific SNPs) and then marking duplicates and clipping overlapping ends in case the fragment size is too short; sec-

ond: base quality recalibration.

Methylation calling is followed usually by post quality control measurements such as number of called sites (usually  $\sim 26\text{M}$  CpGs), average CpG coverage, conversion rate (usually a sample should have at least 99.5% conversion rate), unique alignment rates and duplication rate.

Downstream analysis is the next logical step to be done and it is very dependent on the study design and the addressed biological questions. It might involve finding differential methylated regions DMRs, or genome-wide segmentation to define different classes of methylated regions/domains like UMRs, LMRs, PMDs and HMDs.

The focus of this thesis was analyzing WGBS data sets across a large spectrum of cell types and tissues. WGBS samples from DEEP project were uniformly processed according to DEEP pipeline (<http://doi.org/10.17617/1.2W>) [Ebert et al., 2015].

### Methylation segmentation

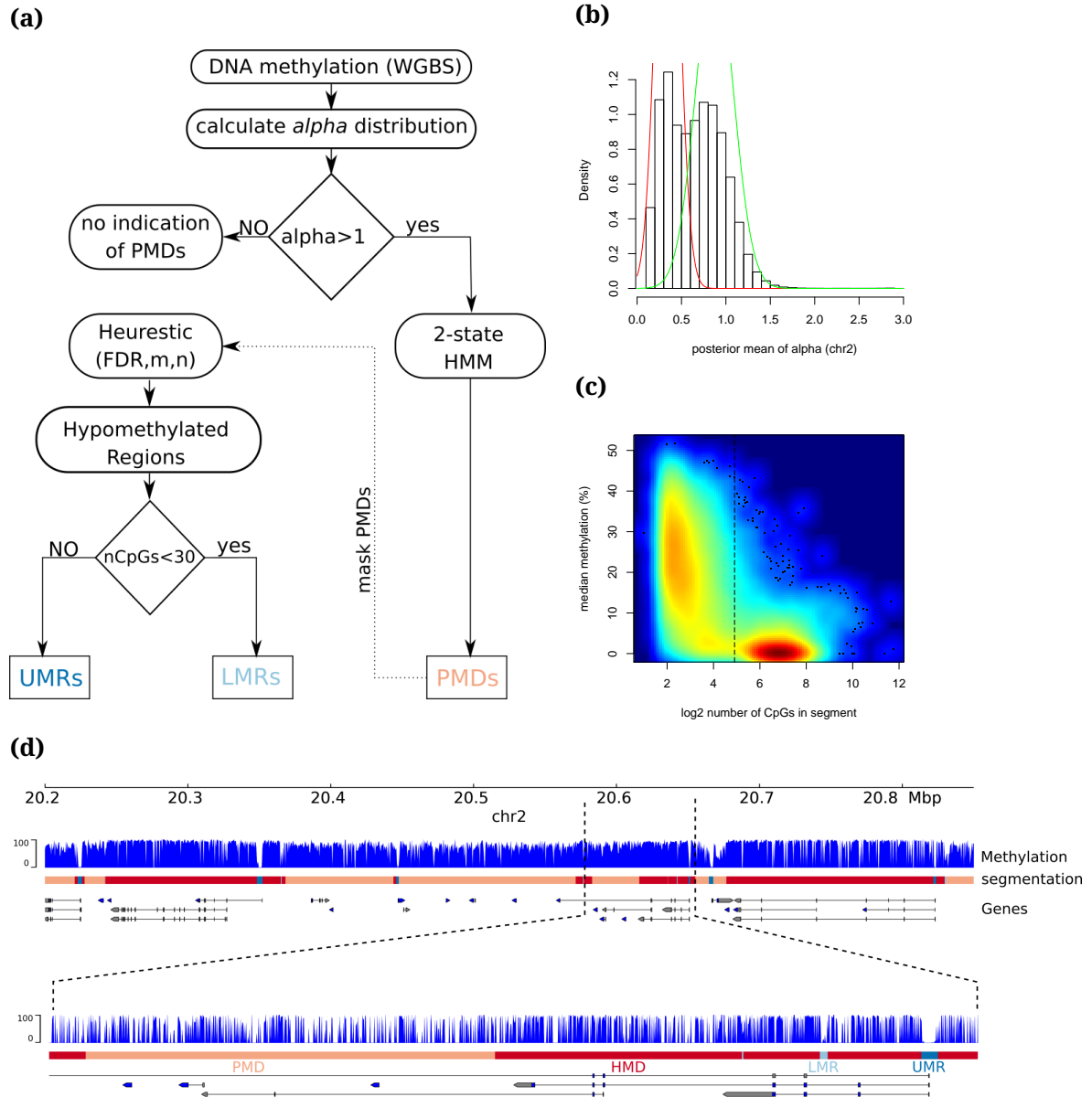
WGBS provides methylation information at base-pair resolution ( $\sim 26\text{M}$  CpGs). A visual inspection of such generated methylation profiles in a genome browser, one can directly notice a depletion of 5mC at some regions, mostly but not exclusively at CpG-rich regions. Such regions often represent regulatory elements like promoters. Hence, dividing the genome into different classes according to methylation levels is of great interest to define regulatory elements genome-wide. Stadler et.al implemented a 3-state HMM to define three regions; unmethylated state with  $\sim 0\%$  methylation; low methylated state with  $\sim 30\%$  methylation and fully methylated state with  $\sim 80\%$  methylation [Stadler et al., 2011]. Based on this idea, Burger et.al implemented an R package, called MethylSeek<sup>‡</sup>, to detect PMDs and discriminate them from the small unmethylated (UMRs) and low methylated regions (LMRs) [Burger et al., 2013] (the main workflow is shown in Figure 1.8a). PMDs detection using this tool is based on HMM method instead of defining them based on average methylation levels (less than 70% [Lister et al., 2009]). Instead of modeling PMDs at each CpG, summary statistics in 101 CpGs sliding windows (one CpG at a time) are used to characterize PMDs. More precisely, reads that cover each CpG are modeled as being generated from beta binomial distribution characterized with  $\alpha$  parameter. The distribution of  $\alpha$  values will tell if the sample contains PMDs or not. Distributions with  $\alpha < 1$  means that the methylome has a unimodal distribution, i.e., methylation levels are skewed toward 0 or 100%;  $\alpha = 1$  corresponds to uniform distribution;  $\alpha > 1$  means that the methylome has a bimodal distribution and it is an indicator of the presence of PMDs (Figure 1.8b). In such case, a two-state Hidden Markov Model (HMM) is trained with Gaussian emission on the  $\alpha$  values via standard expectation maximization algorithm and then

---

<sup>‡</sup>This tool was frequently used in this thesis

## CHAPTER 1. INTRODUCTION

PMDs are predicted using Viterbi algorithm. After detecting PMDs, they are masked and hypomethylated regions are detected as stretches of CpGs with average methylation levels below a certain cutoff (default 50%).



**Figure 1.8: MethylSeekR segmentation.** (a) MethylSeekR workflow for PMDs and short regulatory elements detection. (b) Distribution of posterior  $\alpha$  distribution for Tmem sample. The bimodal distribution and the tail where  $\alpha > 1$  suggest a presence of PMDs. (c) Two clouds represent LMRs as CpG-poor regions with methylation levels between 10 and 50%, and UMRs as CpG-rich regions with methylation levels between 0 and 10%. The vertical dashed line represents the used cut-off value (number of CG=30) of number of CGs per region for LMRs/UMRs classification. (d) Zoomed-in and zoomed-out snapshots of genome-wide segmentation of WGBS data. The segmentation track shows different genomic regions/domains. While PMDs (light brown) and HMDs (red) are observed as domains UMRs (dark blue) and LMRs (light blue) stretch over short regions which coincide with CpG-rich and CpG-poor regions, respectively.



These regions are further classified into low methylated regions LMRs and UMRs based on CG contents and minimum number of CpG within a region (controlled by FDR) (Figure 1.8c). In principle these are the main three segments that MethylSeekR defines, but as a notion of [Stadler et al., 2011] and [Schroeder et al., 2013, 2011] the rest of the genome, other than the three mentioned regions, is called either highly methylated domains (HMDs) or fully methylated regions (FMRs). Throughout this thesis the two terms were used interchangeably. Figure 1.8d shows snapshots of MethylSeekR segmentation for DEEP Tmem sample in zoomed-in and zoomed-out view.

In the context of this thesis and to automatize the segmentation process of hundreds of samples, a wrapper around MethylSeekR was implemented to generate a segmentation file, which can be used to explore the segmentation results in a genome browser. Additionally, a friendly HTML report is generated including figures related to segmentation results. e.g., distribution of average methylation of each segment type, segment length distributions and the percentage of each segment type. The input of this wrapper is a bed file that contains the following information for each CpG: chromosome, position of the CpG, methylation level, total number of reads covering this CpG position and the strand. The methylation levels from both strands are aggregated and then a file compatible with MethylSeekR input format is generated. PMDs, LMRs, UMRs are then generated by MethylSeekR. HMDs are calculated as the complementary of the genome excluding genomic gaps as annotated by UCSC [Rosenbloom et al., 2014] using bedtools [Quinlan and Hall, 2010]. The average methylation per segment are calculated and a bed file with color code is generated, which can be visualized in any genome browser. Finally, the above mentioned HTML report is generated.

Moreover, to explore PMDs across large number of samples, ChromH3M was implemented as a meta segmentation of MethylSeekR results. This workflow is presented in Chapter 2 (Figure S3). Briefly, it uses the bed files generated from the previously mentioned segmentation wrapper as input. 1kb windows across the genome are annotated with 0/1 according to absence/presence of PMDs for each sample (another window size is applied for LMRs and UMRs). ChromHMM [Ernst and Kellis, 2012] is then applied to the binarized matrix with different number of states provided by the user. The emission probabilities are then hierarchically clustered and annotations are added to the heatmap based on a provided sample sheet. This workflow and the segmentation wrapper script is available at <https://github.com/asalhab/ChromH3M>.

## 1.4 International effort for profiling reference epigenome maps

The International Human Epigenome Consortium (IHEC) coordinates the efforts of world-wide epigenomic consortia to map and decipher at least 1000 high resolution reference epigenomes of different primary human cell types to understand how the genome interacts with environment during development, and how the epigenome influences the diseases and health, aiming to accelerate applying of this knowledge to improve human health [Stunnenberg et al., 2016]. All consortia contribute to IHEC goals, but each consortium has also its own focus on specific cell types and diseases. Therefore, IHEC develops standards for data generation, software and methods to minimize the variance between the different consortia and harmonize data processing and data sharing which are important issues for integrative analysis and data interpretation later on. To this end, “EpiMap”<sup>‡</sup> project was evoked under IHEC umbrella aiming to reprocess samples from different consortia with the same pipelines and integrating them for later biological interpretation.

The current full members of IHEC (till the time of writing this thesis, September 2019) includes: CIHR and CEEHRC (Canada), BLUEPRINT Project (EU), DEEP (Germany), HKUST (Hong Kong), CREST (Japan), GIS (Singapore), KNIH (South Korea), NIH ROADMAP Epigenome Program (USA), NHGRI ENCODE Project (USA), 4D Nucleome Program (USA). Within the scope of IHEC and the context of the German Epigenome Program DEEP two projects, presented in Chapter 2 and Chapter 3, provide examples for large-scale data integration and data interpretation.

The reference epigenomes belonging to IHEC should include the following assays for each sample: DNA methylation (WGBS), ChIP-Seq of six core histone modifications plus input (H3K27ac, H3K4me3, H3K4me1, H3K36me3, H3K27me3 and H3K9me3), gene expression (RNA-Seq) and optional open chromatin assay (ATAC-, DNaseI- or NOMe-Seq). All samples should meet IHEC QC standards<sup>‡‡</sup>, defined by Assay Standards Working Group, in order to be considered as IHEC samples.

## 1.5 Aim of the thesis

As it is mentioned in Section 1.4 a large epigenome cohort for broad spectrum of primary human cell types has been generated by IHEC and integrating such data is a big computational challenge, specially when considering the different assays generated by IHEC.

---

<sup>‡</sup>Working title defined by IHEC Integrative Analysis Work Group

<sup>‡‡</sup><https://github.com/IHEC/ihec-assay-standards>

This thesis presents the results of three projects; two of them (Chapter 2 and Chapter 3) have been conducted within DEEP consortium and toward the overarching IHEC vision. The third one (Chapter 4) is outside DEEP's context yet related to the results of the first two projects. The focus of these projects is investigating DNA methylation landscape genome-wide across a comprehensive spectrum of WGBS samples generated by IHEC members and others to define common and cell type specific features of Partially Methylated Domains (PMDs) as very abundant topological units of genomes, and focusing on PMDs in differentiated CD4<sup>+</sup> T memory cells (Tmem) and CD4<sup>+</sup> cells of four Multiple Sclerosis-discordant female Monozygotic twin pairs. More specifically, understanding the developmental relationship of Tmem cell subsets to identify the most likely differentiation model, i.e., either the sequential or the parallel model. Moreover, DNA methylation was integrated in the context of PMDs with other (epi)genetic data like histone modifications, gene expression, replication timing and 3D chromatin conformation data, aiming to understand the interplay between DNA methylation, heterochromatin and the higher-order chromatin structure.

### 1.6 Results outline

This section will guide you throughout the outcomes of this cumulative work, structured in three chapters and give a summary of the aim and the results of each project.

**Chapter 2** presents a comprehensive comparative analysis of 195 DNA methylome in primary cells and tissues to get insight into PMDs, their genome-wide organization, structural and functional features associated with them. This data set includes samples that have been produced by different members of IHEC in addition to publicly available data. This work spotlights a different aspect for studying DNA methylation beyond its role in small regulatory elements, that is, focusing on the large-scale facet. Additionally, it delivers an important message about using cell lines as *in-vitro* model to study DNA methylation changes in cancer. The major results can be summarized as follow:

- PMDs are strong cell-type specific discriminators and cover up to 75% of the genome regardless of their tissues or cell origin. Each cell type has distinct average methylation levels in PMDs. Myeloid cells have homogeneous average PMD methylation levels compared to lymphoid cells (proliferating cells).
- A decreased DNA methylation at PMDs in immortalized cells is linked to an increase in heterochromatinization and to a decrease of gene expression.

## CHAPTER 1. INTRODUCTION

- Distinct heterochromatic signatures at PMDs demarcate distinct domains of late DNA replication.
- PMDs are epigenological features beside their role in gene regulation.
- The implemented ChromH3M workflow is a straightforward framework for integration of large-scale WGBS data.

This work was originally published in a peer-reviewed journal [Salhab et al., 2018] (see Chapter 2 for the detailed contribution), and it was conducted in the context of DEEP project. The main text is at page 25 and the supplementary material is at page 38.

**Chapter 3** focuses on another DEEP-related project, and its aim was to study the influence of epigenetics on the differentiation of memory T-cells (Tmem) and to identify the key molecular regulators. Specifically, understanding whether different Tmem subtypes arise from a common progenitor or whether they represent stages of a sequential differentiation process. Accordingly, a comprehensive epigenetic data was generated for naive, central-, effector-, and terminally differentiated CD4<sup>+</sup> Tmem human cells from blood and CD69<sup>+</sup> Tmem human cells from bone marrow.

A major findings of this study is observing a progressive genome-wide loss of DNA methylation during Tmem differentiation which found to be more prominent in PMDs. Similar observations are found during the memory differentiation of B cells, but not during the differentiation of monocytes into macrophages. These findings support the linear sequential differentiation model of T- and B-cells, and suggest PMDs-associated loss of DNA methylation as an indicator of the proliferation history of the cells. Moreover, PMDs were used in this work as an adjusting tool for detecting ‘functional’ differential methylated regions (DMRs) in highly proliferative cells.

This work was originally published in a peer-reviewed journal [Durek et al., 2016] (see Chapter 3 for the detailed contribution). The main text is at page 61 and the supplementary material is at page 76.

**Chapter 4** deals with studying DNA methylation signatures of CD4<sup>+</sup> memory T cells of monozygotic twins (MZ) clinically discordant for Multiple Sclerosis (MS), an autoimmune disease affecting the central nervous system. The aim was to identify changes in DNA methylation related to MS in peripheral blood mononuclear cells (PBMCs), and to check how MS treatment affects the methylome. The major part of this paper was based on analyzing EPIC arrays, and in a small part WGBS data was used to assess the global DNA methylation differences between the MS-discordant co-twins. The results showed that there was no significant genome-wide changes in

## CHAPTER 1. INTRODUCTION

DNA methylation in PMDs between the MS-discordant co-twins. However, a local prominent MS-DMR in *FIRRE* (located on X-chromosome) was identified.

This work was originally published in a peer-reviewed journal [Souren et al., 2019] (see Chapter 4 for the detailed contribution). The main text is at page 94 and the supplementary material is at page 106.

## Chapter 2

# **A comprehensive analysis of 195 DNA methylomes reveals shared and cell-specific features of partially methylated domains**

The full text of this chapter was originally published as:

Salhab, A., Nordström, K., Gasparoni, G., Kattler, K., Ebert, P., Ramirez, F., Arigoni, L., Müller, F., Polansky, J. K., Cadenas, C., G.Hengstler, J., Lengauer, T., Manke, T., DEEP Consortium, and Walter, J. (2018). A comprehensive analysis of 195 dna methylomes reveals shared and cell-specific features of partially methylated domains. *Genome Biology*, 19(1):150.

under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>).


The author of this thesis contributed to the study design, integrative analysis, Hi-C data analysis and ChromH3M implementation. He generated all the main and the supplementary figures (except Figure S9, the left panel). Together with J.W. and K.N. and contribution from other authors he wrote the manuscript. All authors read and accepted the final version of the manuscript.

RESEARCH

Open Access



# A comprehensive analysis of 195 DNA methylomes reveals shared and cell-specific features of partially methylated domains

Abdulrahman Salhab<sup>1</sup>, Karl Nordström<sup>1</sup>, Gilles Gasparoni<sup>1</sup>, Kathrin Kattler<sup>1</sup>, Peter Ebert<sup>2</sup>, Fidel Ramirez<sup>3</sup>, Laura Arrigoni<sup>3</sup>, Fabian Müller<sup>2</sup>, Julia K. Polansky<sup>4,5</sup>, Cristina Cadenas<sup>6</sup>, Jan G.Hengstler<sup>6</sup>, Thomas Lengauer<sup>2</sup>, Thomas Manke<sup>3</sup>, DEEP Consortium and Jörn Walter<sup>1\*</sup> 

## Abstract

**Background:** Partially methylated domains are extended regions in the genome exhibiting a reduced average DNA methylation level. They cover gene-poor and transcriptionally inactive regions and tend to be heterochromatic. We present a comprehensive comparative analysis of partially methylated domains in human and mouse cells, to identify structural and functional features associated with them.

**Results:** Partially methylated domains are present in up to 75% of the genome in human and mouse cells irrespective of their tissue or cell origin. Each cell type has a distinct set of partially methylated domains, and genes expressed in such domains show a strong cell type effect. The methylation level varies between cell types with a more pronounced effect in differentiating and replicating cells. The lowest level of methylation is observed in highly proliferating and immortal cancer cell lines. A decrease of DNA methylation within partially methylated domains tends to be linked to an increase in heterochromatic histone marks and a decrease of gene expression. Characteristic combinations of heterochromatic signatures in partially methylated domains are linked to domains of early and middle S-phase and late S-G2 phases of DNA replication.

**Conclusions:** Partially methylated domains are prominent signatures of long-range epigenomic organization. Integrative analysis identifies them as important general, lineage- and cell type-specific topological features. Changes in partially methylated domains are hallmarks of cell differentiation, with decreased methylation levels and increased heterochromatic marks being linked to enhanced cell proliferation. In combination with broad histone marks, partially methylated domains demarcate distinct domains of late DNA replication.

**Keywords:** Partially methylated domains, Heterochromatin, Replication timing, Proliferation

## Background

DNA methylation is an epigenetic hallmark with an important role in gene and genome regulation. Changes in the genome-wide landscape of DNA methylation are extensively studied in the context of small regulatory regions like CpG islands [1], CpG shores [2], and proximal and distal regulatory regions [3]. With the first genome-wide bisulfite-based DNA methylation analyses, a new

term, partially methylated domains (PMDs), was introduced by Lister et al. [4] referring to long genomic regions in the range of hundreds of kilo-basepairs (kb) characterized by highly disordered methylation levels. They were initially discovered in the fibroblast cell line IMR90 but cannot be observed in human embryonic stem cells H1.

It has been shown later that PMDs are enriched with heterochromatic histone modifications such as H3K27me3 and that they are gene-poor and less active [5, 6] than other genomic regions. Several studies have since reported PMDs in various cell types: medulloblastoma [6], adipocyte tissue [7], SH-SY5Y neuronal cells [8], and human cancers [5, 9–11]. PMDs in cancer cells are

\*Correspondence: [j.walter@mx.uni-saarland.de](mailto:j.walter@mx.uni-saarland.de)

<sup>1</sup>Department of Genetics, Saarland University, Campus Saarbrücken, 66123 Saarbrücken, Germany

Full list of author information is available at the end of the article



© The Author(s). 2018 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

linked to late replication and nuclear lamina-associated regions [10]. The first non-cancer primary human tissue type with PMDs has been reported in placenta [12] and were defined using a hidden Markov model (HMM) rather than applying a threshold of methylation level. Recently, we, as part of the DEEP consortium <http://www.deutsches-epigenom-programm.de/>, published the first primary human cells, CD4+ T cells, with PMDs [13]. We showed that progressive loss of DNA methylation correlates with T cell memory differentiation and happens predominantly in PMDs. Burger et al. [3] implemented an HMM-based detection method called MethylSeekR to define PMDs and separate them from highly methylated domains (HMDs) and short (regulatory) regions that come in two types (methylation below 50%): lowly methylated regions (LMRs, CpG poor regions, less than 30 CpGs) and unmethylated regions (UMRs, mostly CpG islands, more than 30 CpGs). LMRs and UMRs are relatively short (a few hundred to a few thousand basepairs) and correspond to distal and proximal regulatory elements, respectively [14]. Tools such as MethylSeekR are very useful for exploring the methylome landscape on a large scale and help to discriminate the large domains, from the small regulatory regions.

In collaboration with our colleagues in the international human epigenome consortium IHEC <http://ihec-epigenomes.org/>, we contributed to generating a large epigenome cohort for numerous primary cell types from human and mouse. WGBS data serve as an invaluable resource for studying PMDs in primary cells. PMDs represent a new aspect for studying the DNA methylation landscape on a genome-wide level apart from the context of regulatory regions that have been studied extensively and pose the question whether DNA methylation has an impact on the genome organization. At the same time, it has become quite clear that cells in vitro behave differently from primary cells, for instance regarding methylation levels. Thus, it is important to compare the methylome of primary cells and cell lines in order to validate in vitro systems and afford an appropriate interpretation of the data.

Here, we investigate the genome-wide organization of PMDs across a comprehensive spectrum of available WGBS data generated by IHEC members, DEEP <http://www.deutsches-epigenom-programm.de/>, Blueprint [www.blueprint-epigenome.eu](http://www.blueprint-epigenome.eu), and Roadmap <http://www.roadmappigenomics.org/>, together with other public data in order to gain insights into PMDs. In addition, we integrated WGBS data with other epigenetic data, ChIP-seq, RNA-seq, Hi-C and Repli-seq, in an attempt to describe the interaction between DNA methylation and chromatin formation in order to understand how they impact cell division, differentiation, and the higher order chromatin structure. Moreover, we propose a

new integrative approach to exploring and interpreting methylome topologies using WGBS data, an approach very much needed as the amount of such data is growing rapidly.

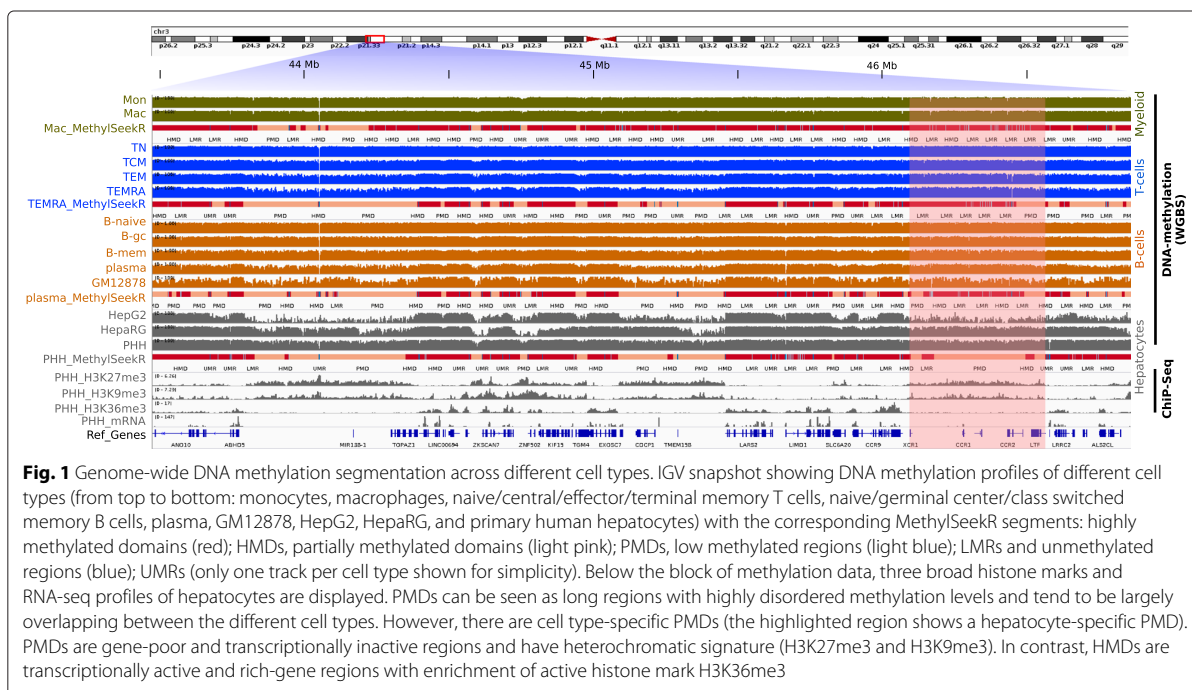
## Results

### Partially methylated domains are cell type discriminators

We collected and surveyed 171 public human WGBS datasets of different primary cell types (hepatocytes, T cells, B cells, monocytes, macrophages, eosinophils, neutrophils, dendritic cells, natural killer cells, endothelial cells, and thymocytes) and tissues (liver, intestine, spleen, esophagus, stomach gastric, colon sigmoid, colon mucosa, heart, and pancreas) for which we identified PMDs with MethylSeekR (see Additional file 1 for the complete list of samples). Figure 1 shows methylomes of different cell types with the corresponding segmentation tracks. The lengths of PMDs vary broadly, ranging from 100 kb up to 20 Mb (Additional file 2: Figure S1). PMDs cover a large portion of the genome (50–75%). The average and individual levels of PMD methylation vary between different cell types (boxplots in Fig. 2a). While PMD positions in the genomes are highly conserved across cell types, in general, only roughly 26% of the genome is annotated as completely shared PMDs across all cell types (Fig. 2b). Overall, PMDs are enriched for the broad heterochromatic marks H3K27me3 and H3K9me3 and depleted for the broad euchromatic mark H3K36me3. The latter is also reflected in the low appearance of annotated transcriptional units within PMDs and an overall low average transcription of genes located in PMDs (Fig. 1 and Additional file 2: Figure S2).

To gain a deeper insight into the cell-specific and genome-wide distribution of PMD methylation profiles, we generated and applied a modified ChromHMM [15] approach, “ChromH3M,” as an abbreviation for ChromHMMmeta segmentation (see details in the “Methods” section and Additional file 2: Figure S3). In brief, we bin the genome into 1 kb tiled windows, labeled as 1 or 0 according to the presence/absence of PMDs for each sample. This binarized signal is then processed with ChromHMM to generate a 15-state model. The emission probabilities are displayed after hierarchical clustering. This approach generates PMD clusters discriminating cell type origin and/or cell-related subgroups (Fig. 2a). Only five out of 171 samples did not cluster together with samples of similar origin. This approach is surprisingly stable even across cells which differ strongly in their overall methylation level (shown as box plots in Fig. 2a). We also used shorter LMR and UMR regions for such a ChromH3M meta-segmentation and roughly obtained the main subgroups in hierarchical clusters using 10,000 bootstraps and an “au” threshold of 97 (see Additional file 2: Figure S4 and the “Methods” section for details).





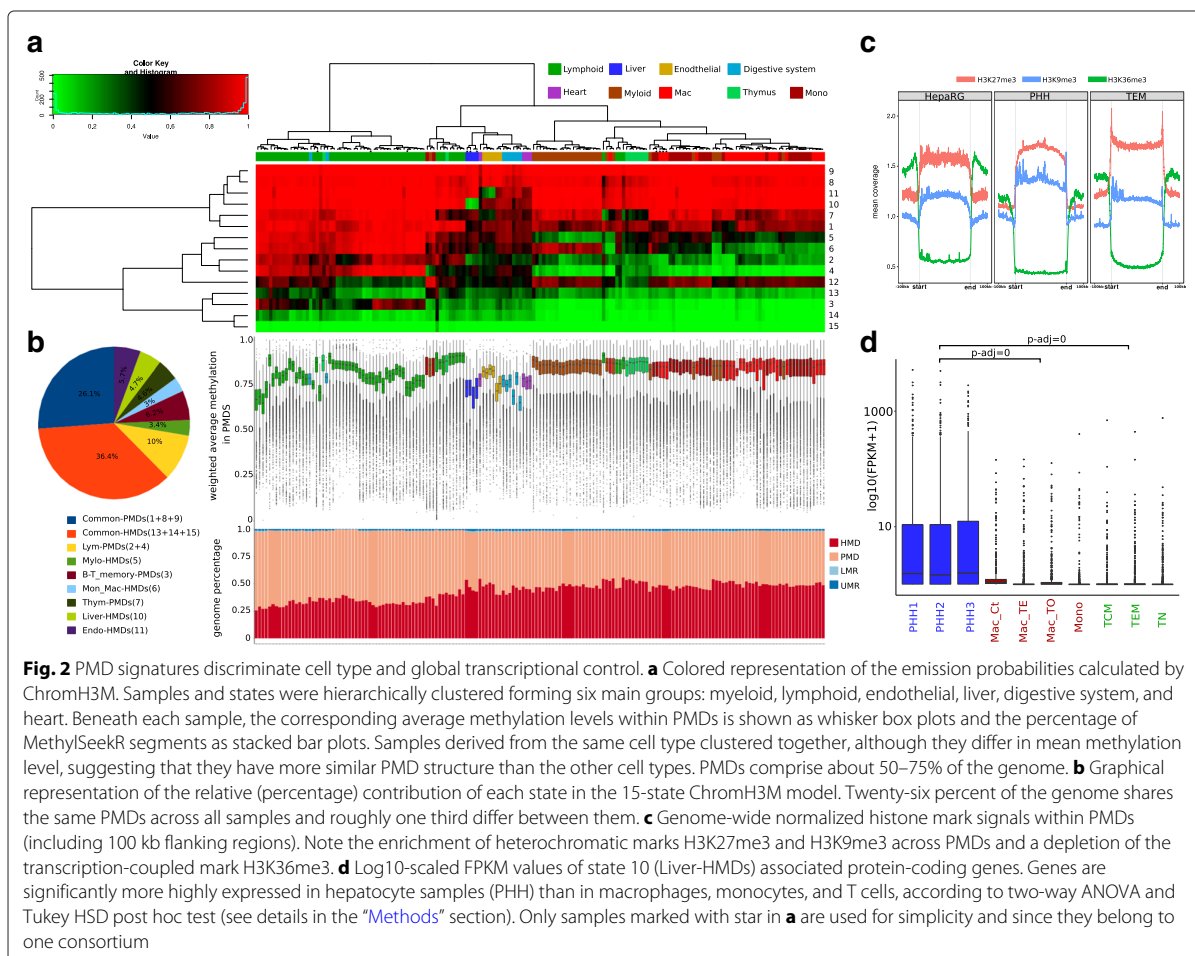
We conclude that PMDs are strong cell-type-specific discriminators comparable with regulatory changes in short UMRs/LMRs.

For 171 available human methylomes of tissues and primary cells, ChromH3M generates a tree with six main branches separating myeloid from lymphoid cells, endothelial tissues, liver tissue, and tissues of the digestive system and heart. Myeloid cells split into two subclusters: the granulocytes (neutrophils and eosinophils) and agranulocytes (monocytes, macrophages, and dendritic cells). Members of both subclusters have similar average PMD methylation. B cells and T cells form a lymphocyte cluster which branches off into subgroups of memory T and B cells, i.e., central and effector memory T cells, germinal center and memory B cells, respectively. This indicates that cell types not only display a distinct overall PMD topology but also acquire distinct PMD substructures upon proliferation and differentiation [13].

We furthermore observe that, in general, PMDs have extended heterochromatic signatures in both primary cells and permanent cell lines (Fig. 2c). PMDs cover relatively gene-poor regions with mostly lowly/unexpressed genes (Additional file 2: Figure S2). The ChromH3M analysis reveals a couple of distinct features of cell-type-specific PMDs (Fig. 2a). For instance, states 10 and 11 comprise regions that only are HMDs in liver and endothelial cell types, respectively. State 4 discriminates myeloid HMDs from PMDs in other cell types. State 3 defines B and T cell-specific PMDs (Fig. 2a, b). The shared

PMDs are defined by states 1, 8, and 9, while states 14 and 15 define shared HMDs.

To explore the biological functions of genes present within cell-type-specific HMDs/PMDs, we performed a functional annotation analysis with DAVID [16, 17] for genes in state 10 and state 3. For the former, liver-specific HMDs, the GO terms liver tissue expression, Rotor syndrome disease (lack of hepatocyte pigment deposits), and the KEGG pathway for drug metabolism through cytochrome P450 were obtained. These genes exhibit significantly higher expression in liver tissue/hepatocytes than in other cell types (two-way ANOVA and Tukey HSD post hoc test,  $p$  adj = 0) (Fig. 2d). Furthermore, these HMDs are largely devoid of heterochromatic marks and enriched for the transcriptional elongation mark H3K36me3 across gene bodies (Additional file 2: Figure S5, left panel). This is exemplified by two hepatocyte-specific gene loci CYP2B6 and FMO6P (Additional file 2: Figure S6). The latter state, number 3, marks B and T cell-specific PMDs. Hence, these regions in B and T cells are enriched with the repressive mark H3K27me3 and, to a lower degree, with H3K36me3. Further, the functional analysis provides cell-type-associated terms, cell differentiation, inflammatory response, adaptive immune response and specific surface antigen MHC class I, in addition to the KEGG pathway for the hematopoietic cell lineage. Interestingly, the expression levels of these genes are downregulated in accordance with their PMD annotation. However, regarding only the DNA methylation signal, there is



a trend to split the B and T cells into naive versus memory cells. This discrimination can neither be confirmed by ChIP-seq nor by RNA-seq (see Additional file 2: Figure S5, right panel). This could be due to the limitation in detecting the precise boundaries of shallow PMDs in naive cells.

In summary, the ChromH3M results indicate a domain-wide transition of cell-type-specific PMDs into HMDs and vice versa along with transcriptional regulation. The direction of this transition couples with specific changes in heterochromatic states.

A ChromH3M analysis on 24 WGBS mouse samples (Additional file 2: Figure S7) shows a similar classification and distribution of PMD states, confirming that our findings not only hold for human but describe a feature apparently conserved among mammals. In mouse, we identify cell-type/tissue-specific PMDs for neuron, intestine, colon, and mammary epithelial cells. Furthermore, the epithelial cells group into cells of the luminal and the basal compartment. We conclude that in human

and mouse, PMDs are excellent epigenome classifiers of cell-type-specific topologies.

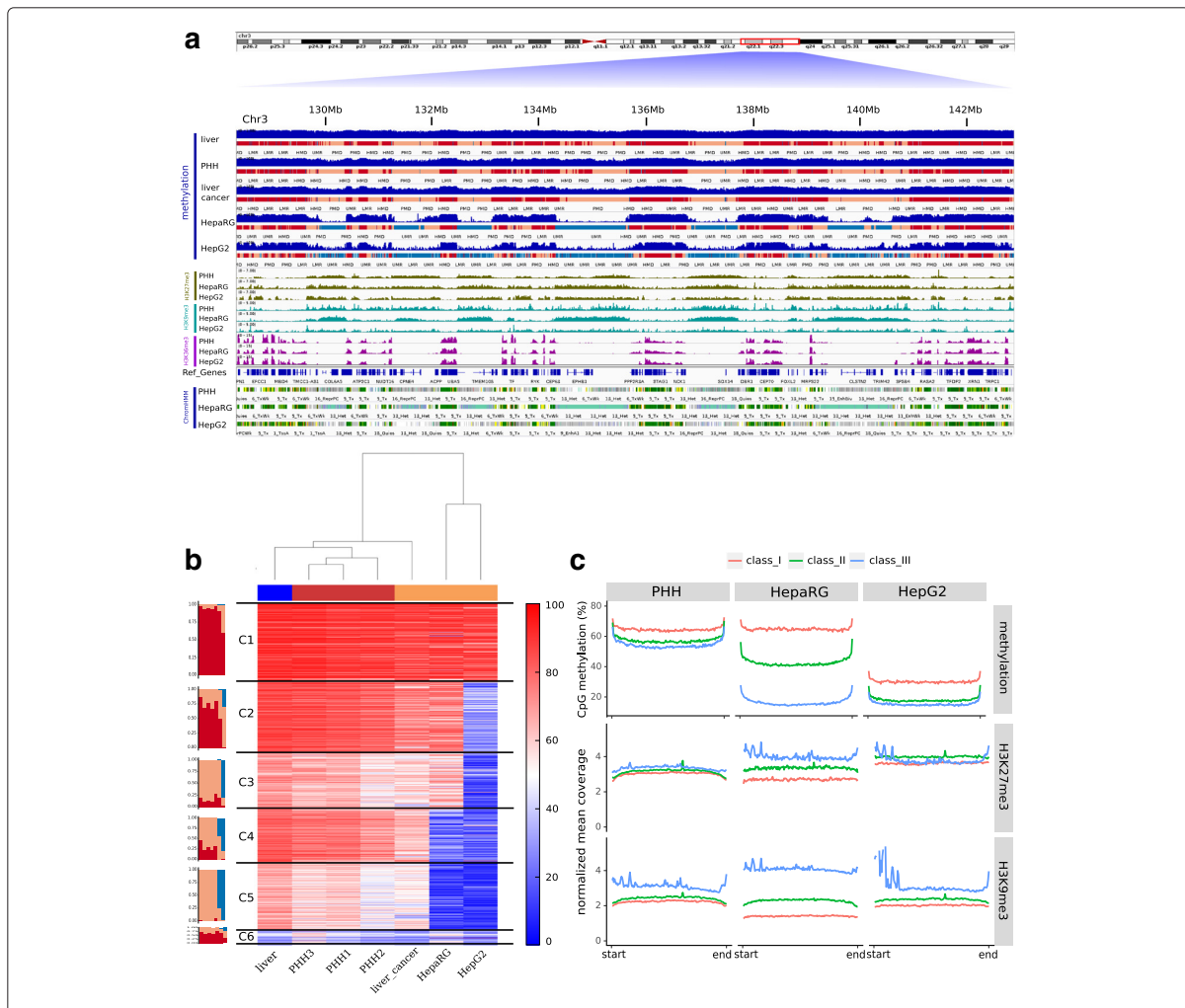
### Chromatin compaction increases with DNA methylation erosion at PMDs in immortalized cells

Immortalized cell lines are widely used for studying cellular mechanisms including the influence of epigenetic control. However, it is known that cells in culture undergo drastic epigenetic alterations linked to passaging and cell replication numbers [18]. To investigate the epigenome-wide changes occurring between primary cells and immortal cell lines, we compared the methylomes of primary cells and cell lines of the same origin. With this comparison, we wanted to monitor the impact of cultivation and cancer-specific changes on PMD formation. We generated epigenome data for isolated primary hepatocytes (PHH) and two hepatic cancer cell lines: the hepatic progenitor cell line (HepaRG) and the liver hepatocellular carcinoma cell line (HepG2). We also include in our comparison results on publicly

available liver cancer cells and noncancerous liver tissues (Fig. 3a).

First, we calculated the average methylation across the samples in 10 Kb bins. We then performed *k*-means clustering forming six disjoint clusters which were subsequently annotated by the MethySeekR segmentation

(Fig. 3b). Cluster 1 defines highly methylated bins across all samples while the other clusters show progressive loss of methylation in the order: liver tissue >PHH >liver cancer >HepaRG >HepG2. Interestingly, primary liver cancer was more similar to the non-cancerous liver and PHH than to the cancer cell lines HepaRG and HepG2. Both



**Fig. 3** Heterochromatinization accompanied by DNA methylation erosion at PMDs in cancers. **a** A snapshot of 14 Mb of chr3 showing the relevant epigenetic marks. Top: distinct DNA methylation tracks and the MethySeekR segmentation of liver tissue, isolated hepatocytes (PHH), liver cancer tissue, HepaRG, and HepG2 cell lines, respectively. PMDs of primary cells and normal and cancerous tissues are extensively and selectively less methylated in cancer cell lines (largely converted into unmethylated regions). Middle: histone marks H3K27me3, H3K9me3, and H3K36me3 in the same samples. Bottom: ChromHMM segmentation based on these three histone modifications in addition to H3K4me3, H3K4me1, H3K27ac, and Input (see details in the “Methods” section). **b** *K*-means clustering ( $k = 6$ ) based on the averaged methylation in 10 Kb bins. Cluster 1 represents the most (almost fully) methylated bins across all samples, while the other clusters are ordered according to the progressive erosion of methylation in PMDs. Bar plots (left) beside the heatmap show the percentage of the annotated bins as HMD, PMDs, and UMR for each sample in each cluster. **c** Progressive change of DNA methylation in PMDs across cancer cell lines. The top of the figure shows classified and grouped PMDs (three classes) based on the average PMD methylation levels in PHH and their corresponding overall levels in HepaRG and HepG2, respectively. Note the intermediate status of HepaRG, e.g., with a higher similarity to PHH in class\_I (most highly methylated), an intermediate status in class\_II and a higher similarity to HepG2 in class\_III (lowest methylation level). The bottom shows the PMD wide changes in heterochromatic marks across the clusters defined by DNA methylation. The inverse correlation to DNA methylation is most obvious for HepaRG (class\_I and class\_III)

cancer cell lines have lower methylation levels compared to the primary cells, as seen in clusters 4 and 5. This indicates a different epigenomic pattern in cultivated cancer cells in comparison to primary cancer cells. To gain deeper understanding of the features governing the development and changes in cell-type-specific PMDs, we focused on the analysis of liver PMDs that exhibit changes in the average methylation level in the cancer cell lines. We first extracted PMDs of PHH cells, which exhibit large overlap with PMDs of liver tissue and liver cancer tissue (data not shown). Such primary liver cell PMDs split into three subclasses, with respect to changes in DNA methylation in HepaRG and HepG2 (Fig. 3c). In the first subclass (class\_I, red), PHH and HepaRG exhibit the same average degree of methylation (65%), but show a very low methylation state in HepG2. In the second class (class\_II, green), HepaRG methylation levels are intermediate between PHH and HepG2, while in the third class (class\_III, blue) both HepaRG and HepG2 show the same low average methylation as compared to the primary cells.

Along with the progressive loss of DNA methylation in these three subclasses, we observe a distinct gain of heterochromatic marks (Fig. 3c), suggesting a compensatory effect. The effect is most obvious in the HepaRG cell line which shows an intermediate level of PMD methylation. Moreover, H3K36me3 is positively correlated with DNA methylation across the gene body in the three subclasses (Additional file 2: Figure S8). We confirmed this observation by calculating the average methylation across ChromHMM segments of HepaRG (Additional file 2: Figures S9 and S10). PMDs associated with stronger transcription are higher methylated, on average, and marked by lower levels of heterochromatic marks.

We conclude that in immortalized cells, a progressive erosion of DNA methylation mainly in PMDs is linked to a substantial gain of heterochromatic marks. This is likely to be accompanied by differences in chromatin compaction and regulation in the immortalized cells with a prolonged proliferation. The conversion of PMDs and sometimes of HMDs, found in cancer tissues, into low methylated domains as seen for HepG2 indicate that epigenetic changes found in model cell lines should be interpreted with great care, as they may reflect the properties more of the cell's proliferation history and less of the cancer state or cell-specific origin.

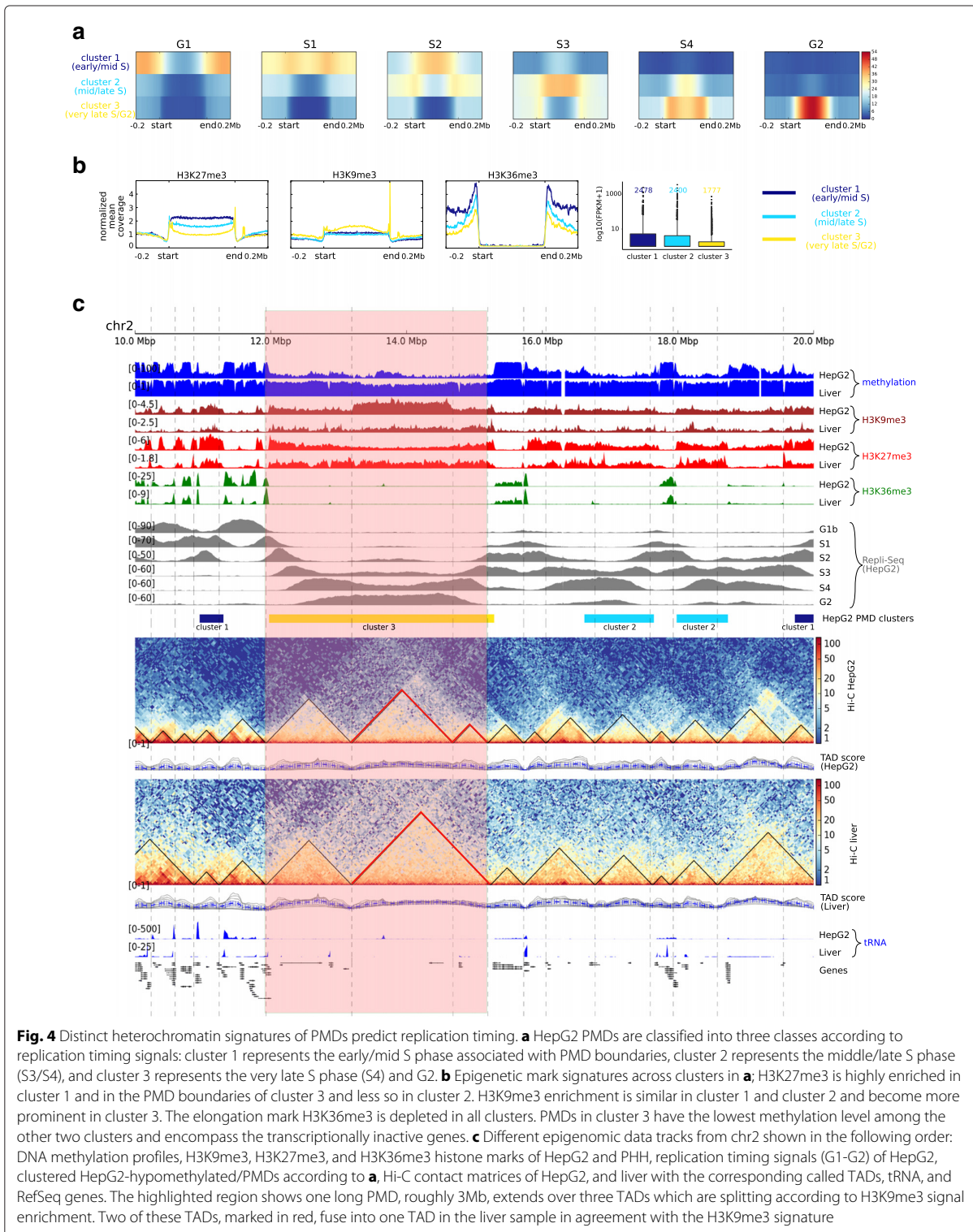
#### Distinct heterochromatic signatures of PMDs predict replication timing

It has been shown that during cell division, late-replicating regions can become gradually demethylated [19] and long PMDs show widespread H3K9me3 marks bordered by H3K27me3, whereas shorter PMDs are enriched by H3K27me3 only [6]. So far, these features have not been

deeply investigated and analyzed in an integrated fashion, i.e., combining DNA methylation and chromatin marks. Using replication timing data for HepG2 from the ENCODE project [20, 21], we clustered HepG2 hypomethylated/PMDs regions (longer than 300 kb), by the *k*-means algorithm, into three clusters (see the “Methods” section and Fig. 4a). We observe that these clusters display distinct combinations of histone modification and DNA methylation (Fig. 4b and Additional file 2: Figure S11). Cluster of early/mid S phase (dark blue) is associated with shortest PMDs, and the chromatin is enriched for the two repressive marks H3K27me3 and H3K9me3. Mid/late S phase cluster (light blue) comprises longer PMDs which are less highly enriched for H3K27me3 compared to the previous cluster. In very late S/G2 phase cluster (yellow), PMDs extend over very long regions making up roughly 50% of the total PMDs/hypomethylated regions (Additional file 2: Figures S12 and S13). These PMDs are strongly enriched for H3K9me3, bordered by H3K27me3. We confirm our clustering by using the three broad histone modification signals in PMDs as predictors and observe a high average prediction accuracy of 0.77 for HepG2 (0.81 for IMR90) (notice that this is a three-class prediction, details in the “Methods” section). The very late replicating regions S/G2 phase have the highest prediction accuracy, suggesting a distinct chromatin signature in this phase. These findings extend previous results [19] indicating that combinations of heterochromatic marks and DNA methylation define early and middle S-phase and late S-G2 phase (Fig. 4b).

#### PMDs organization and topologically associated domains (TADs)

To analyze the overall relationship between PMDs and TADs (topologically associated domains), we generated Hi-C data for HepG2 and used available liver Hi-C data [22]. Using HiCExplorer tool [23], we identified 3217 and 4021 TADs in liver and HepG2, respectively. As a consequence, TADs of HepG2 are shorter than those in liver (Additional file 2: Figure S14). This finding is in agreement with Taberlay et al. [24], who showed that cancer cells in general form smaller TADs and establish new boundaries. We find that TAD borders are significantly closer to PMD borders as compared to randomized test borders ( $p$  value  $< 2.2e-16$ , Wilcoxon test) (Additional file 2: Figure S15). Moreover, 94% of base pairs within PMDs, in HepG2 and liver, are also annotated as heterochromatic TADs (Additional file 2: Figure S16) (details in “Methods” section). The light red box in Fig. 4c highlights a typical example of a region in which several TADs exist in HepG2 and primary liver that are belonging to one PMD. In HepG2, we observe the formation of an extra TAD (marked in red) but not the formation of an additional PMD boundary. This extra TAD shows a strong enrichment of H3K9me3.



**Fig. 4** Distinct heterochromatin signatures of PMDs predict replication timing. **a** HepG2 PMDs are classified into three classes according to replication timing signals: cluster 1 represents the early/mid S phase associated with PMD boundaries, cluster 2 represents the middle/late S phase (S3/S4), and cluster 3 represents the very late S phase (S4) and G2. **b** Epigenetic mark signatures across clusters in **a**; H3K27me3 is highly enriched in cluster 1 and in the PMD boundaries of cluster 3 and less so in cluster 2. H3K9me3 enrichment is similar in cluster 1 and cluster 2 and become more prominent in cluster 3. The elongation mark H3K36me3 is depleted in all clusters. PMDs in cluster 3 have the lowest methylation level among the other two clusters and encompass the transcriptionally inactive genes. PMDs in cluster 3 have the lowest methylation level among the other two clusters and encompass the transcriptionally inactive genes. **c** Different epigenomic data tracks from chr2 shown in the following order: DNA methylation profiles, H3K9me3, H3K27me3, and H3K36me3 histone marks of HepG2 and PHH, replication timing signals (G1–G2) of HepG2, clustered HepG2-hypomethylated/PMDs according to **a**, Hi-C contact matrices of HepG2, and liver with the corresponding called TADs, tRNA, and RefSeq genes. The highlighted region shows one long PMD, roughly 3Mb, extends over three TADs which are splitting according to H3K9me3 signal enrichment. Two of these TADs, marked in red, fuse into one TAD in the liver sample in agreement with the H3K9me3 signature

## Discussion

Our comprehensive integrated analysis of primary human cells adds valuable novel insights into the structure and function of PMDs in human and mouse epigenomes. Building on our first report of PMD changes in primary T cells [13], we here integrated publicly available WGBS datasets fulfilling high-quality standards to systematically analyze the features of PMDs in primary cells, primary tissues, and immortalized cells from human and mouse. We apply a new integrative “ChromH3M” approach which combines existing tools and represents an easy and straightforward method for analyzing and integrating a large cohort of WGBS datasets. This allowed us to define and compare PMDs across hundreds of WGBS samples, revealing a couple of intriguing new DNA methylation properties with respect to genome organization, timing of DNA replication and cell-type-specific gene regulation.

We find that PMDs comprise up to 75% of all epigenomes. However, only roughly 26% of the genome consists of PMDs that are shared across all investigated cells (shared PMDs). These common/shared PMDs have also been a focus of a recent analysis by Zhou et al. [25] confirming some of our findings. As a major difference, we also find that PMDs serve as excellent cell type classifiers as cells with functional similarities show a more similar PMD arrangement and topology, arguing for a shared developmental origin of lineage-specific PMDs. Finally, we observe that PMD methylation changes in some cells when they proliferate and show strong methylation decreases in immortal cell lines.

Analyzing the PMD topology in more detail, we observe that the epigenomes of cells are partitioned into long regions of PMDs interspersed with HMDs. These two classes of epi-domains show contrasting chromatin signatures. While PMDs are more heterochromatic and gene-poor regions, HMDs show strong transcriptional activity and enrichment of genes. This finding generalizes previous isolated observations reported by [4, 8, 10, 13, 26] to a number of different cancer types and cell lines. We also find cell-type-specific changes from PMD to HMD, and vice versa, occurring in genomic regions that contain genes functionally enriched for cell-type-specific properties. This finding points towards a developmental control of PMD and HMD formation. The complete understanding which partitioning of HMDs and PMDs defines a precursor ground state of a cell type needs more investigation. Such knowledge will help understanding the role of epigenetic domains in cell differentiation.

We find that long PMDs have a lower density of protein-coding genes, lincRNA, and pseudogenes relative to the shorter PMDs. In general, protein-coding genes are less highly expressed in long PMDs than in shorter PMDs and HMDs (Additional file 2: Figures S17 and S18).

We hypothesize that this is also reflected in the more pronounced constitutive heterochromatic nature of long PMDs as compared to the more facultative heterochromatic nature of shorter PMDs [27]. Shorter PMDs retain more epigenomic plasticity with more pronounced cell-type-specific features.

PMDs can also be divided into different subclasses which are observed in different stages of DNA replication. A hallmark of the late stages of replication (S4 and G2) is their length and the presence of the constitutive heterochromatic mark H3K9me3 together with a characteristic enrichment of H3K27me3 at the boundaries. On the other hand, the early/middle S phases (S1-3) PMDs are shorter and exhibit a higher overall proportion of H3K27me3. The length-dependent histone modification pattern is in agreement with previous findings in medulloblastoma [6]. The differences are strong enough to be used as predictors of the replication phases. Our results are consistent with a previous report [19] and extend its results by providing a detailed characterization of chromatin state and DNA methylation at PMDs in relation to cell cycle. Moreover, the long PMDs associated with the late S4-G2 phases overlap with 56% of the bases within shared PMDs. The shorter PMDs retain a greater variability, confirming our hypothesis that shorter PMDs possess epigenetically more less rigid heterochromatic structures than longer ones. This characteristic could be relevant for differentiation, cell-fate determination, and/or cell maturation processes.

To deeper understand the cell-type-specific changes occurring at PMDs (and HMDs), in cancer and cancer cell lines, we compared the DNA methylation landscape of primary human hepatocytes (PHH) to liver cancer tissue and hepatocellular carcinoma cell lines (HepaRG and HepG2). Notably, the methylome of primary liver cancer retained a PMD structure highly similar to primary cells. PMDs in cancer tissue show a mild but clearly reduced level of methylation. In cancer cell lines, however, the DNA methylation in PHH-specific PMDs strongly decreases. The regions with lower methylation still retain the typical PMD histone marks even if they are completely unmethylated. So far, we have no explanation to how this aligns with models suggesting that global demethylation is caused by a global loss of heterochromatic marks such as H3K9me2/3 and consequently a lack of UHRF1 activity during replication by deregulation of DNMT1 [28]. When counting the unmethylated regions as PMDs, the overall PMD structure of cell lines is hepatocyte-like. It is likely that the strong erosion of DNA methylation is the consequence of extensive cultivation leading to a proliferation-dependent loss of methylation while maintaining or even enforcing heterochromatic marks such as H3K9me2/3. An alternative hypothesis is that the loss of PMD methylation is caused by the selection/expansion

of cell subpopulations with lower methylation. To better understand the generation of increased demethylation in PMDs as a consequence of cell proliferation (cell division), we performed an experiment outlined in Additional file 2: Figure S19A. Human T memory cells were obtained from three different donors, and from each, such “bulk” sample single cells were isolated and clonally expanded following TCR stimulation. After proliferative expansion, single clonal cultures were analyzed and compared to the starting “bulk” samples (mixed T cells) by RRBS. We observe a preferential loss of methylation in PMDs (Additional file 2: Figure S19B, confirming our results reported in Durek et al. [13]). We calculated the percentage of fully methylated, fully unmethylated, and mixed patterns of four consecutive CpGs within single read (see Additional file 2: Supplementary methods for details). Interestingly, the fraction of mixed patterns within PMDs remains constant in all three single cell clones while the fraction of fully unmethylated patterns expands and the fraction of fully methylated patterns diminishes (Additional file 2: Figure S19C). Moreover, 20–35% of CpGs in the fully methylated fraction loses methylation in all three clonally expanded populations. This strongly argues for a gradual loss of methylation coupled to cell division rather than a clonal selection considering that the analyzed cell populations arise from three independent single cells/donor. Overall, our findings are in agreement with result of a recent paper [29] which suggested that at least in cancers, hypomethylation is unlikely to be the result of a “population level effect” only and the extent of hypomethylation is proportional to the cell division rate of the tissue.

A genome-wide decrease of methylation is also seen in early human and mouse embryos [30–34]. Schroeder et al. [35] reported that PMDs can be detected in the oocyte and early embryos of several species but that they are not detectable in placenta, a tissue that shows a low level of overall DNA methylation. Upon differentiation, the genome-wide DNA methylation levels (also in PMDs) increase in somatic cells probably to prevent genomic and transcriptional instability that is observed in fast proliferating cancer cells that usually show a pronounced erosion of PMDs [11]. These findings are in line with our analysis suggesting that while PMDs are general features of (adult) somatic cells, proliferation, differentiation, and development have an impact on PMD topology and genome-wide epigenetic memory.

In general, levels of PMD methylation should be considered when comparing local epigenetic states *in vitro* particularly when comparing healthy and cancerous tissues to immortal cell lines. In our recent study [13], we suggested a way to consider such global demethylation effects for the detection of differentially methylated region (DMR). Here, we screened for DMRs based on their deviation from the global methylation change rather than

applying a fixed cutoff (for more details, see [13]). DMRs were stratified over PMDs and HMDs such that many DMRs simply following the global change of methylation could be excluded. In B cells, this procedure reduced the number of DMRs within PMDs tremendously (from 28,014 to 8338 using adaptive filtering when comparing naive B cells with plasma cells). On the other hand, DMRs, within PMDs, that gain DNA methylation upon differentiation are increased (2811 DMRs in comparison to 95 retrieved by basic thresholding method) (Additional file 2: Figure S20). Genomic region enrichment analysis for such DMRs using GREAT [36] provides cell differentiation and development relevant as major terms (Additional file 2: Figure S20). These findings demonstrate the advantage of stratifying DMRs according to increasing and decreasing of DNA methylation in HMDs and PMDs, affording more insight into the biological role of the genes associated with these DMRs.

A very important observation is that PMD and HMD prediction can be used as a proxy for and/or support Hi-C data when detecting and classifying TADs. When overlaying TAD and PMD predictions, we observe that they largely co-localize and often share the same boundaries. Specifically, PMDs almost completely overlap with heterochromatic TADs. However, we also observe that multiple TADs can overlap with one single PMD. This suggests that either PMDs cover domains larger than TADs or indicates that Hi-C data provide a more fine-grained resolution for domain boundaries. Overall, we observe that there are commonalities as well as differences when aligning TADs and PMDs, and their topological organization and functional relation will have to be further investigated to better understand their dependencies. A recent study by Nothjunge et al. [37] showed that the establishment of heterochromatic (B) compartments precedes PMD formation. As this study only focused on DNA methylation, it remains an open question if B compartments are indeed established prior to a heterochromatic domain formation which we see as one feature of cell-type-specific PMDs.

## Conclusions

We provide a comprehensive analysis of PMDs for 195 human and mouse methylomes including more than 157 primary cell samples. Our analysis adds a new dimension to studying DNA methylation on a large scale extending beyond the context of cis-regulatory elements that has been studied extensively. Our results show that PMDs are an excellent classifier of cellular origin and confirm that they are indicators of the cellular proliferation history. In addition, PMD heterochromatic histone mark signatures serve as an effective classifier for distinguishing early from middle and late replication domains. ChromH3M is an easy and straightforward framework for integrated analysis of large-scale WGBS data and can highlight specific

combinatorial patterns of PMDs across large number of samples. PMDs are also a useful adjusting tool for detecting functional DMRs in highly proliferative cells. We believe that PMDs are a crucial epitopological signature beside their role in gene regulation. Our analysis reveals an important limitation in using cultivated cells for disease-associated epigenetic studies as they undergo strong changes in their epigenetic topology.

## Methods

### WGBS

Coverage and methylation fraction of human samples were downloaded from the Roadmap Epigenomic Project <http://egg2.wustl.edu/roadmap/data/byDataType/dnamethylation/WGBS>. Blueprint data was downloaded from [ftp://ftp.ebi.ac.uk/pub/databases/blueprint/data/homo\\_sapiens/GRCh38/](ftp://ftp.ebi.ac.uk/pub/databases/blueprint/data/homo_sapiens/GRCh38/) and then mapped to hg19 using liftOver from UCSC [38]. DEEP data was taken from previous studies [13, 39]. Bed files containing the coverage and methylation levels at CpG resolution from [40] were directly used in the analysis. We list all samples with the relevant sources in the Additional file 1.

### MethylSeekR segmentation

All samples were segmented into partially methylated domain (PMDs), lowly methylated regions (LMRs), and unmethylated regions (UMRs) using the MethylSeekR tool [3]. The rest of the genome, excluding gaps as annotated by UCSC [38], was denoted as highly methylated domains (HMDs). We ran MethylSeekR with default parameters: a coverage cutoff at five reads per CpG, methylation level threshold at 0.5, and maximum FDR of 0.05 for detection of hypomethylated regions, resulting in a threshold of at least four CpGs per LMR, 101 CpGs per sliding window  $nCGbin = 101$ , and smoothing over 3 CpGs. Methylation levels of both strands were aggregated and weighted average methylation levels were plotted as box plots across PMDs.

### ChromH3M segmentation

In order to explore PMDs and find combinatorial patterns across samples, we binned the genome into 1 kb windows and annotated each of them with 1 if the bin overlaps with a PMD and 0 otherwise across all samples. We used ChromHMM [15] to train this binarized signal with a 15-state HMM. We termed this method “ChromH3M.” The emission probabilities and states were hierarchically clustered using Euclidean distance and ward.D2 as an agglomeration method in the R environment [41]. The very same analysis was performed for LMRs and UMRs, respectively. To assess the uncertainty in the hierarchical clustering, we calculated an unbiased  $p$  value (AU  $p$  value) via multiscale bootstrap resampling ( $n = 10,000$ ). The two cell line samples

HepaRG and HepG2 were not included in ChromH3M analysis.

The normalized mean coverage of three broad histone marks (H3K27me3, H3K36me3, and H3K9me3), generated by the DEEP pipeline <http://doi.org/10.17617/1.2W> [42], were plotted genome-wide across the PMDs with proper flanking regions using deepTools [43]. The number of protein-coding genes falling within PMDs was calculated, demanding a minimum of 80% of the gene length to be overlapping with the segment. A pseudocount of 1 was added to FPKM to avoid zeros in the box plots.

The heatmap in Fig. 3b was generated by binning the genome into 1 kb windows and averaging the methylation levels across all samples resulting in roughly 280,000 windows which then were clustered by  $k$ -means into six clusters and annotated with methylSeekR segments. Samples were hierarchically clustered with ward.D2 and Euclidean distance. Sex chromosomes were excluded from the aforementioned analyses.

### Clustering of PHH PMDs and cancer cell lines

PHH PMDs shorter than 20 kb were filtered, and a matrix of methylation levels in 1 kb windows across PHH, HepaRG, and HepG2 was calculated after normalizing all PMDs to the same length of 150 kb using deepTools [43]. The windows were clustered with  $k$ -means method into three clusters. H3K27me3, H3K9me3, and DNA methylation signals were plotted along PMDs of each cluster using deepTools [43].

### Analysis of replication domains

Replication timing signals were downloaded from ENCODE project and used directly (details about this data are available from <https://www.encodeproject.org/documents/50ccff70-1305-4312-8b09-0311f7681881/@@download/attachment/wgEncodeUwRepliSeq.html.pdf>). A two-state HMM was used to segment the HepG2 methylation profile into highly methylated and PMDs/hypomethylated regions using the “Hidden-Markov” R package [44], assuming that each CpG may have one of the two states: foreground state with high methylation level and background state with low methylation level. Regions shorter than 300 kb were filtered. The mean coverage of replication signals (G1, S1-S4, and G2) was calculated in 1 kb bins across normalized (to 500 kb length) and flanked PMDs (250 kb up and downstream) using deepTools [43]. PMDs were then clustered using  $k$ -means into three classes: early/middle S phase, middle/late S phase, and late S/G2 phase. The mean coverage signals of H3K27me3, H3K9me3, and H3K36me3, and DNA methylation levels were plotted across the PMDs of each class using deepTools. The number of protein-coding genes falling into each class was calculated, demanding 80% of the gene length to be within the



PMD. FPKM values were plotted as box plots in log scale with pseudocount of 1 to avoid zeros. For the prediction of replication domains, we built a multiclass classification model using the counts of reads of each histone mark in 1 kb bins as predictors and the three aforementioned clusters as response at each PMD. We split the data into 75% training set and 25% test set. We trained the model with a random forest classifier and selected the model using 10-fold CV repeated five times. The prediction accuracy was calculated based on the confusion matrix between the predicted and the reference values. One-versus-all accuracy was calculated and then the average accuracy was calculated. This analysis was performed using the caret package <https://github.com/topepo/caret/> in the R environment. The analysis of genomic regions regarding DMRs, with and without adjusting for the global DNA demethylation, was carried out using the GREAT tool [36]. GO analysis was done using DAVID [16, 17].

#### Chromatin state segmentation

All Chip-Seq samples, listed in Additional file 2, were pre-processed starting from raw BAM files as follows: duplicate reads were removed using samtools version 1.3 with the filter “-F 1024.” Regions of known artifacts (“blacklist regions”) taken from the ENCODE project <https://www.encodeproject.org/> [20], which we adapted to account for differences between ENCODE’s hg19 and DEEP’s hs37d5 assembly, were filtered out using bedtools version 2.20.1 with the subcommand “pairtobed” and the option “-type neither.” After preprocessing, the filtered BAM files for all six histone marks plus Input were used as input for the chromatin state segmentation using ChromHMM version 1.11 (Java 1.7) with default parameters. We did not train a dedicated ChromHMM model for our dataset, but used the available ROADMAP 18-state model [45] to benefit from its biologically meaningful state labeling, which enabled us to immediately interpret the chromatin state maps in the context of this work.

#### HepG2 Hi-C

HepG2 cells have been fixed for 10 min using 1% formaldehyde in D-MEM and quenched for 5 min in 125 mM glycine. After two PBS washes, cells have been collected by scraping them off the plate and snap-frozen in liquid nitrogen. Hi-C experiments have been conducted as previously described [23], with the following modifications. Nuclei from cell pellets containing about four million of cells have been extracted by sonication [46] using the following parameters: 75 W peak power, 2% duty factor, 200 cycles/burst, and 180 s, using Covaris milliTubes and Covaris E220 sonicator. After nuclei permeabilization, chromatin has been digested overnight at 37 °C using HindIII high fidelity (80 units per million cells; R3104S, NEB). Biotin incorporation has been carried out at 37 °C for 1 h in 300  $\mu$ l volume using these

reaction conditions: 50 mM of each nucleotide (dATP, dTTP, dGTP, biotin-14-dCTP, from Life Technologies, 19518-018) and 8 U of Klenow (NEB, M0210L). Ligase mix has been added to each sample followed by 4 h of incubation at room temperature under rotation. After nuclei lysis, protein digestion and overnight de-crosslink, DNA has been precipitated and sonicated to 100–600 bp. Biotinylated DNA has been pulled down as previously described. One hundred nanograms of DNA bound to beads have been used for library preparation using a modification of the NEBNext Ultra DNA library preparation workflow (NEB, E7370). DNA bound to beads has been end-repaired, A-tailed, adaptor-ligated, and USER-treated following manufacturer’s instruction. After a bead wash, DNA has been eluted from the beads by incubating at 98 °C for 10 min. Adaptor-ligated DNA has been PCR amplified using 7 PCR cycles. Libraries have been sequenced paired-end, with a read length of 75 bp, on the Illumina NextSeq 500 instrument.

#### Hi-C data processing

Reads were mapped to the human reference genome hg19 (37d5) using bowtie2 [47], and then samtools [48] was used to convert the reads to BAM format. A matrix of read counts over the bins in the genome, considering the sites around the restriction site AAGCTT was built using the hicBuildMatrix function from HiCEXplorer [23]. Ten bins were merged with hicMergeMatrixBins and then the matrix was corrected for GC bias and very low/high contact regions. To compute the TADs we first calculated the TAD scores by “hicFindTADs TAD\_score” command with the following parameters “-minDepth 300000 -maxDepth 2000000 -step 70000” and then TADs were identified by “hicFindTADs find\_TADs” command. The interaction matrix and other signal tracks were also visualized using HiCEXplorer.

#### Comparison between TADs and PMDs

To test the consistency between TAD borders and PMD borders, we generated an equally sized set of randomized borders and calculated the shortest distance between TAD borders and (i) PMD borders and (ii) the randomized borders. A Wilcoxon test was carried out between the two distance distributions. To calculate the overlap between PMDs and heterochromatic TADs (generated as described above), we classified TADs using histone marks into two classes by *k*-means from deepTools. One class is enriched by heterochromatic marks and the other by euchromatic mark. We counted the number of overlapping base-pairs between PMDs and the heterochromatic TADs and then plotted the results as venn diagrams. The comparison was done for liver and HepG2. In this analysis, PMDs within a distance of 50 kb were fused and only those longer than 300 kb were included. This was done to exclude intersecting LMRs and UMRs.

## Additional files

**Additional file 1:** List of samples. XLSX sheet with list of samples and the corresponding URLs and accession numbers of the raw data. (XLSX 36 kb)

**Additional file 2:** Supplementary materials. PDF document with supplementary figures and supplementary methods. (PDF 11,526 kb)

**Additional file 3:** Review history. (DOCX 58 kb)

## Abbreviations

HMDs: Highly methylated domains; HMM: Hidden Markov model; LMRs: Lowly methylated regions; PMDs: Partially methylated domains; PHH: Primary human hepatocytes; TADs: Topologically associated domains; UMRs: Unmethylated regions

## Acknowledgements

We are grateful to all colleagues from the IHEC consortium for making their data available. A full list of the investigators who contributed to the generation of the epigenomic data used in our study can be found under the respective homepages of the “NIH Epigenomics Roadmap”, “ENCODE”, the German epigenome consortium DEEP, and BLUEPRINT. We would like to thank Jasmin Kirch for technical support in NGS experiments. We would like to thank the Flow Cytometry Core Facility at the DRFZ for cell sorting. The full list of DEEP investigators is available at DEEP website <http://www.deutsches-epigenom-programm.de/project/groups/>.

## Funding

This work was mainly supported by the German Epigenome Programme (DEEP) of the German Federal Ministry of Education and Research (BMBF) (01KU1216F). JW and KN received partial support by the European Union’s Seventh Framework Programme (FP7/2007-2013) under grant agreement no 282510 – BLUEPRINT.

## Availability of data materials

Methylation calls and coverage files of WGBS are available from different sources: IHEC portal [49] (for DEEP samples), Blueprint consortium [50], NIH ROADMAP consortium [51], Ziller et al. [40], and Heyn et al. [52]. Sample details and raw data accession numbers are listed in Additional file 1. WGBS mouse data (coverage and methylation calls) were downloaded from MethBase [53]. The original studies from where the data was obtained are listed in Additional file 1. Chip-Seq and RNA-Seq DEEP data are available from the IHEC portal [54]. Samples are listed in Additional file 1. Liver Hi-C data was taken from [22] and can be accessed under the GEO accession number GSM1419086. HepG2 Hi-C data and PMD tracks of 195 samples generated in this study are available under GEO accession number GSM3105137. Replication timing signals of HepG2 and IMR90 were downloaded from the ENCODE project data portal [20, 21], and they are accessible under GEO accession numbers GSM923446 and GSM923447. Clonal T memory sequencing data has been deposited at the European Genome-phenome Archive (EGA), which is hosted by the EBI and the CRG, under accession number EGAS00001003157. All DEEP samples were uniformly processed as described in [13] according to DEEP pipelines [42, 55]. ChromH3M code has been deposited at Zenodo [56] and is available from GitHub [57].

## Review history

The review history is included as Additional file 3.

## Authors’ contributions

AS, KN, and JW contributed to the study design. AS contributed to the integrative analysis. PE contributed to the chromatin state segmentation analysis. LA and TM contributed to the HepG2 Hi-C data generation. AS and FR contributed to the Hi-C data analysis. GG contributed to the hepatocytes, HepaRG, HepG2, T cells WGBS data generation. KK contributed to the HepaRG and HepG2 Chip-Seq data generation. FM contributed to the preparation of Blueprint WGBS data in hg19 assembly. CC provided the hepatocyte samples. JP contributed to the clonal cultures of T memory cells experiment. AS, KN, and JW wrote the manuscript with contributions from other authors. All authors read and accepted the final version of the manuscript.

## Ethics approval and consent to participate

Ethical approvals for human and mouse samples were present following the standards outlined in DEEP, Blueprint, and Roadmap guidelines. In our study we adhere to the rules defined by the data access committees.

## Consent for publication

Not applicable.

## Competing interests

The authors declare that they have no competing interests.

## Publisher’s Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Author details

<sup>1</sup>Department of Genetics, Saarland University, Campus Saarbrücken, 66123 Saarbrücken, Germany. <sup>2</sup>Department of Computational Biology and Applied Algorithmics, Max Planck Institute for Informatics, 66123 Saarbrücken, Germany. <sup>3</sup>Max Planck Institute of Immunobiology and Epigenetics, 79108 Freiburg, Germany. <sup>4</sup>Berlin-Brandenburg Center for Regenerative Therapies at the Charité, Berlin, Germany. <sup>5</sup>University Medicine Berlin and German Rheumatism Research Centre, Berlin, Germany. <sup>6</sup>Leibniz Research Center for working Environment and Human Factors IfADo, 44139 Dortmund, Germany.

Received: 19 January 2018 Accepted: 20 August 2018

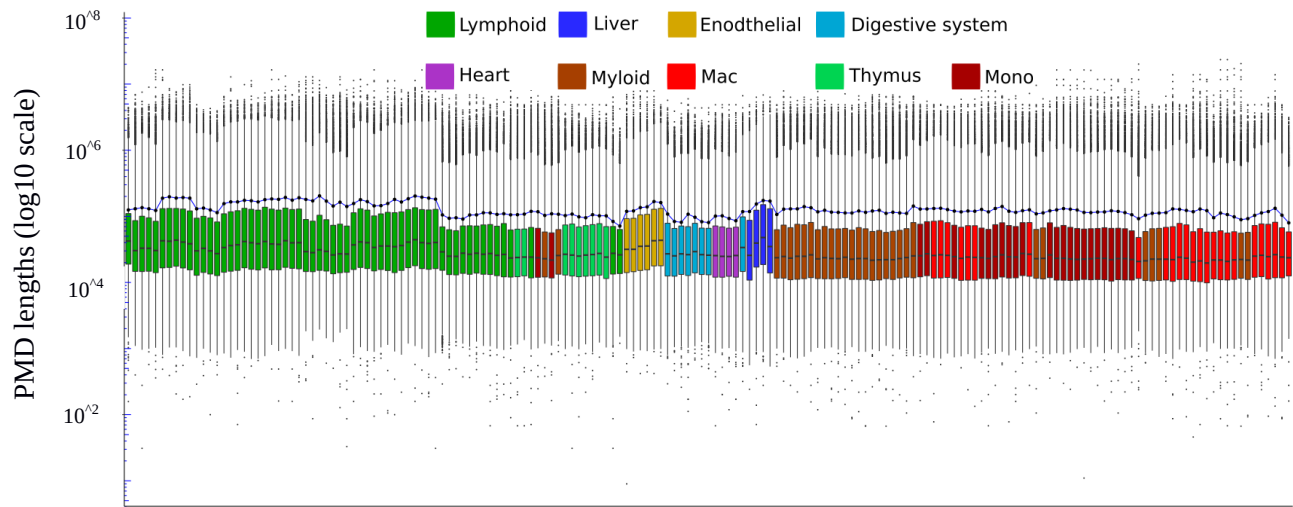
Published online: 28 September 2018

## References

- Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*. 2008;454(7205):766–70.
- Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, Cui H, Gabo K, Rongione M, Webster M, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CPG island shores. *Nat Genet*. 2009;41(2):178–86.
- Burger L, Gaidatzis D, Schübeler D, Stadler MB. Identification of active regulatory regions from DNA methylation data. *Nucleic Acids Res*. 2013;41(16):e155. <https://doi.org/10.1093/nar/gkt599>.
- Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, et al. Human dna methylomes at base resolution show widespread epigenomic differences. *Nature*. 2009;462(7271):315–22.
- Hon GC, Hawkins RD, Caballero OL, Lo C, Lister R, Pelizzola M, Valsesia A, Ye Z, Kuan S, Edsall LE, et al. Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Res*. 2012;22(2):246–58.
- Hovestadt V, Jones DT, Picelli S, Wang W, Kool M, Northcott PA, Sultan M, Stachurski K, Ryzhova M, Warnatz HJ, et al. Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. *Nature*. 2014;510(7506):537–41.
- Lister R, Pelizzola M, Kida YS, Hawkins RD, Nery JR, Hon G, Antosiewicz-Bourget J, O’Malley R, Castanon R, Klugman S, et al. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature*. 2011;471(7336):68–73.
- Schroeder DI, Lott P, Korf I, LaSalle JM. Large-scale methylation domains mark a functional subset of neuronally expressed genes. *Genome Res*. 2011;21(10):1583–91.
- Hansen KD, Timp W, Bravo HC, Sabuncyan S, Langmead B, McDonald OG, Wen B, Wu H, Liu Y, Diep D, et al. Increased methylation variation in epigenetic domains across cancer types. *Nat Genet*. 2011;43(8):768–75.
- Berman BP, Weisenberger DJ, Aman JF, Hinoue T, Ramjan Z, Liu Y, Noushmehr H, Lange CP, van Dijk CM, Tollenaar RA, et al. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat Genet*. 2012;44(1):40–6.
- Timp W, Bravo HC, McDonald OG, Goggins M, Umbricht C, Zeiger M, Feinberg AP, Irizarry RA. Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors. *Genome Med*. 2014;6(8):61.

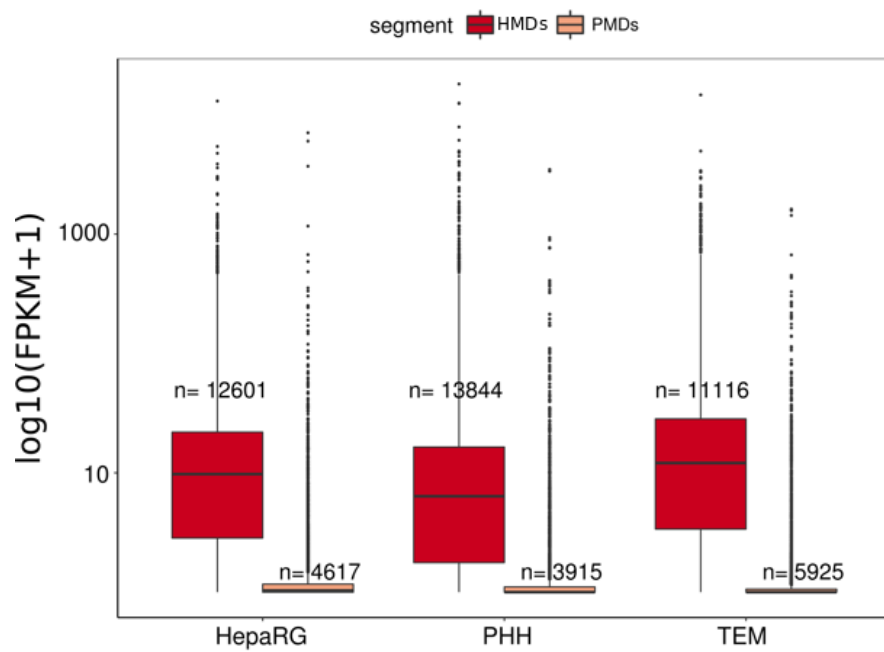
12. Schroeder DJ, Blair JD, Lott P, Yu HOK, Hong D, Cray F, Ashwood P, Walker C, Korf I, Robinson WP, et al. The human placenta methylome. *Proc Natl Acad Sci*. 2013;110(15):6037–42.
13. Durek P, Nordström K, Gasparoni G, Salhab A, Kressler C, De Almeida M, Bassler K, Ulas T, Schmidt F, Xiong J, et al. Epigenomic profiling of human CD4+ T cells supports a linear differentiation model and highlights molecular regulators of memory development. *Immunity*. 2016;45(5):1148–61.
14. Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, van Nimwegen E, Wirblich C, Oakeley EJ, Gaidatzis D, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*. 2011;480(7378):490–5.
15. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods*. 2012;9(3):215–6.
16. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat Protoc*. 2009;4(1):44–57.
17. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. 2009;37(1):1–13.
18. Shipony Z, Mukamel Z, Cohen NM, Landan G, Chomsky E, Zelig SR, Fried YC, Ainbinder E, Friedman N, Tanay A. Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. *Nature*. 2014;513(7516):115–9.
19. Aran D, Toperoff G, Rosenberg M, Hellman A. Replication timing-related and gene body-specific methylation of active human genes. *Hum Mol Genet*. 2011;20(4):670–80.
20. Consortium EP, et al. An integrated encyclopedia of dna elements in the human genome. *Nature*. 2012;489(7414):57–74.
21. Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, Hitz BC, Gabdank I, Narayanan AK, Ho M, Lee BT, et al. Encode data at the encode portal. *Nucleic Acids Res*. 2016;44(D1):726–32.
22. Leung D, Jung I, Rajagopal N, Schmitt A, Selvaraj S, Lee AY, Yen CA, Lin S, Lin Y, Qiu Y, et al. Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature*. 2015;518(7539):350–4.
23. Ramirez F, Bhardwaj V, Villaveces J, Arrigoni L, Gruening BA, Lam KC, Habermann B, Akhtar A, Manke T. High-resolution tads reveal DNA sequences underlying genome organization in flies. *bioRxiv*. 2017;115063.
24. Taberlay PC, Achinger-Kawecka J, Lun AT, Buske FA, Sabir K, Gould CM, Zotenko E, Bert SA, Giles KA, Bauer DC, et al. Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. *Genome Res*. 2016;26(6):719–31.
25. Zhou W, Dinh HQ, Ramjan Z, Weisenberger DJ, Nicolet CM, Shen H, Laird PW, Berman BP. Dna methylation loss in late-replicating domains is linked to mitotic cell division. *Nat Genet*. 2018;50(4):591.
26. Lin IH, Chen DT, Chang YF, Lee YL, Su CH, Cheng C, Tsai YC, Ng SC, Chen HT, Lee MC, et al. Hierarchical clustering of breast cancer methylomes revealed differentially methylated and expressed breast cancer genes. *PLoS ONE*. 2015;10(2):0118453.
27. Saksouk N, Simboeck E, Déjardin J. Constitutive heterochromatin formation and transcription in mammals. *Epigenetics Chromatin*. 2015;8(1):3.
28. von Meyenn F, Iurlaro M, Habibi E, Liu NQ, Salehzadeh-Yazdi A, Santos F, Petrini E, Milagre I, Yu M, Xie Z, et al. Impairment of dna methylation maintenance is the main cause of global demethylation in naive embryonic stem cells. *Mol Cell*. 2016;62(6):848–61.
29. Dmitrijeva M, Ossowski S, Serrano L, Schaefer MH. Tissue-specific dna methylation loss during ageing and carcinogenesis is linked to chromosome structure, replication timing and cell division rates. *Nucleic Acids Res*. 2018;46(14):7022–39.
30. Kobayashi H, Sakurai T, Imai M, Takahashi N, Fukuda A, Yayoi O, Sato S, Nakabayashi K, Hata K, Sotomaru Y, et al. Contribution of intragenic DNA methylation in mouse gametic DNA methylomes to establish oocyte-specific heritable marks. *PLoS Genet*. 2012;8(11):1002440.
31. Wang L, Zhang J, Duan J, Gao X, Zhu W, Lu X, Yang L, Zhang J, Li G, Ci W, et al. Programming and inheritance of parental DNA methylomes in mammals. *Cell*. 2014;157(4):979–91.
32. Guo H, Zhu P, Yan L, Li R, Hu B, Lian Y, Yan J, Ren X, Lin S, Li J, et al. The DNA methylation landscape of human early embryos. *Nature*. 2014;511(7511):606.
33. Lee HJ, Hore TA, Reik W. Reprogramming the methylome: erasing memory and creating diversity. *Cell Stem Cell*. 2014;14(6):710–9.
34. Okae H, Chiba H, Hiura H, Hamada H, Sato A, Utsunomiya T, Kikuchi H, Yoshida H, Tanaka A, Suyama M, et al. Genome-wide analysis of dna methylation dynamics during early human development. *PLoS Genet*. 2014;10(12):1004868.
35. Schroeder DJ, Jayashankar K, Douglas KC, Thirkill TL, York D, Dickinson PJ, Williams LE, Samollow PB, Ross PJ, Bannasch DL, et al. Early developmental and evolutionary origins of gene body DNA methylation patterns in mammalian placentas. *PLoS Genet*. 2015;11(8):1005442.
36. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. Great improves functional interpretation of cis-regulatory regions. *Nat Biotechnol*. 2010;28(5):495–501.
37. Nothjunge S, Nührenberg TG, Grüning BA, Doppler SA, Preissl S, Schwaderer M, Rommel C, Krane M, Hein L, Gilsbach R. Dna methylation signatures follow preformed chromatin compartments in cardiac myocytes. *Nat Commun*. 2017;8(1):1667.
38. Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, et al. The UCSC genome browser database: 2015 update. *Nucleic Acids Res*. 2015;43(D1):670–81.
39. Wallner S, Schröder C, Leitão E, Berulava T, Haak C, Beißer D, Rahmann S, Richter AS, Manke T, Bönisch U, et al. Epigenetic dynamics of monocyte-to-macrophage differentiation. *Epigenetics Chromatin*. 2016;9(1):33.
40. Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LT-Y, Kohlbacher O, De Jager PL, Rosen ED, Bennett DA, Bernstein BE, et al. Charting a dynamic dna methylation landscape of the human genome. *Nature*. 2013;500(7463):477.
41. R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2013. <http://www.R-project.org/>.
42. Ebert P, Müller F, Nordström K, Lengauer T, Schulz MH. A general concept for consistent documentation of computational analyses. *Database*. 2015;2015:bav050.
43. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res*. 2016;44:W160–57.
44. Harte D. HiddenMarkov: Hidden Markov Models. Wellington: Statistics Research Associates; 2017. Statistics Research Associates. R package version 1.8-11. <http://www.statsresearch.co.nz/dsh/sslub/>.
45. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, et al. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015;518(7539):317–30.
46. Arrigoni L, Richter AS, Betancourt E, Bruder K, Diehl S, Manke T, Bönisch U. Standardizing chromatin research: a simple and universal method for chip-seq. *Nucleic Acids Res*. 2016;44(7):67–7.
47. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nat Methods*. 2012;9(4):357–9.
48. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, et al. The sequence alignment/map format and samtools. *Bioinformatics*. 2009;25(16):2078–9.
49. Consortium D. WGBS of DEEP Data. <http://epigenomesportal.ca/ihec/download.html?b=2017-10&as=1&i=6&session=>.
50. Epigenome B. WGBS of Blueprint Data. [ftp://ftp.ebi.ac.uk/pub/databases/blueprint/data/homo\\_sapiens/GRCh38/](ftp://ftp.ebi.ac.uk/pub/databases/blueprint/data/homo_sapiens/GRCh38/). Accessed Nov 2016.
51. Roadmap N. WGBS of Roadmap Data. <https://egg2.wustl.edu/roadmap/data/byDataType/dnamethylation/WGBS/>. Accessed Mar 2016.
52. Heyn H, Vidal E, Ferreira HJ, Vizoso M, Sayols S, Gomez A, Moran S, Boque-Sastre R, Guil S, Martinez-Cardus A, et al. Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer. *Genome Biol*. 2016;17(1):11.
53. Song Q, Decato B, Hong EE, Zhou M, Fang F, Qu J, Garvin T, Kessler M, Zhou J, Smith AD. A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLoS ONE*. 2013;8(12):81148.
54. IHEC. International Human Epigenome Consortium. <http://epigenomesportal.ca/ihec/grid.html>.
55. Ebert P. DEEP Computational Metadata: Max Planck Inst Inform; 2016. <https://doi.org/10.17617/1.2w>.
56. Salhab A. asalhab/ChromH3M: Paper release version. 2018. <https://doi.org/10.5281/zenodo.1326417>.
57. Salhab A. asalhab/ChromH3M: GitHub repository. 2018. <https://github.com/asalhab/ChromH3M/releases/tag/v1.0>. Accessed 2 Aug 2018.

Figure S1



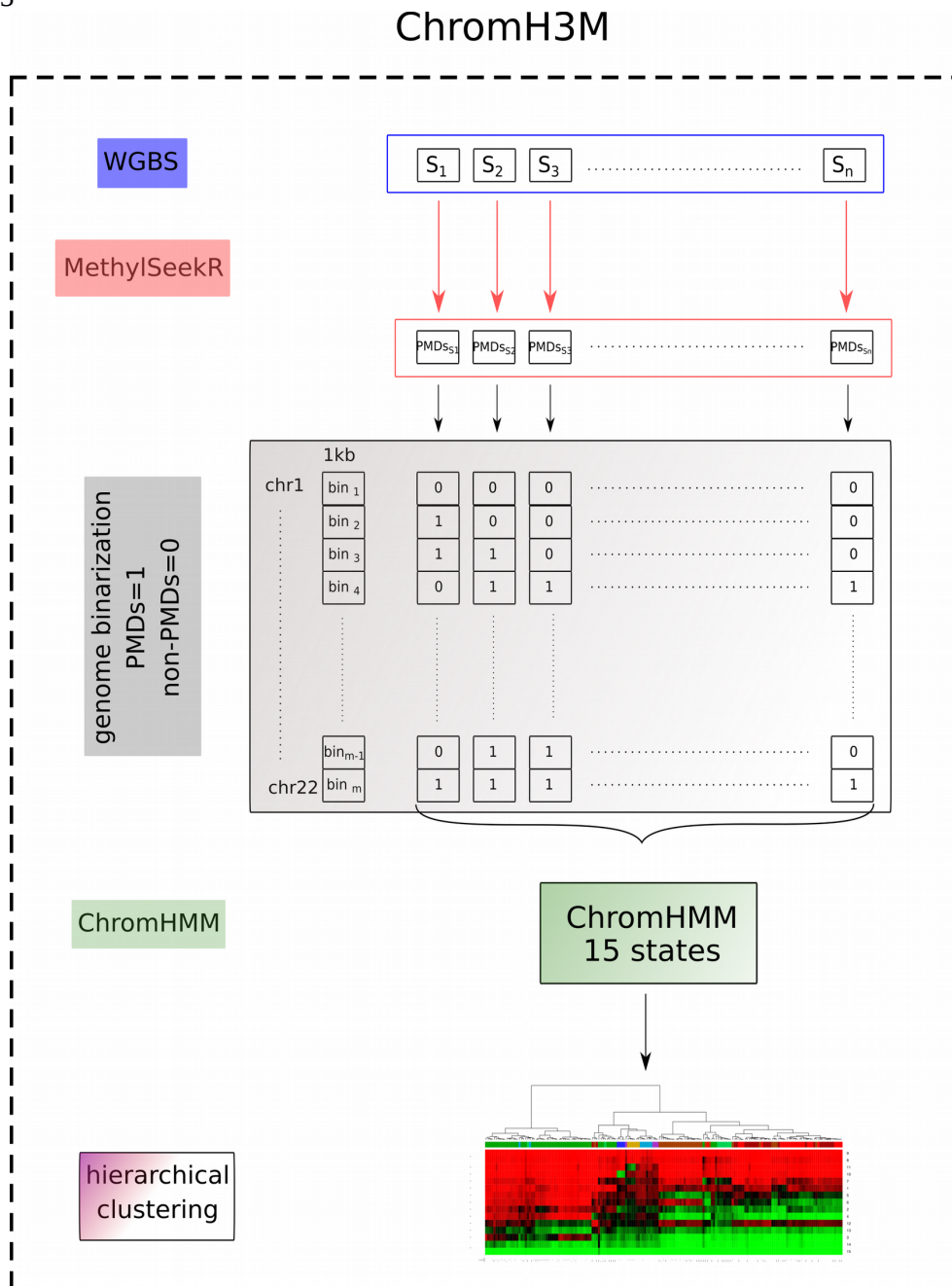
**Figure S1. PMDs length distribution.** The mean length distribution of PMDs is ~150 kb

Figure S2



**Figure S2. PMDs are gene-poor regions.** Number of genes and their FPKM values in PMDs/HMDs in one cell line (HepaRG) and two primary cells (hepatocytes and effector memory T-cells).

Figure S3



**Figure S3. ChromH3M workflow.** MethylSeekR is applied to each sample to identify PMDs. 1Kb bins across the genome are annotated with 1/0 according to presence/absence of PMD for each sample. The binarized signal is loaded into ChromHMM and a model with 15 states is trained. The emission probabilities and states are then hierarchically clustered.

Figure S4

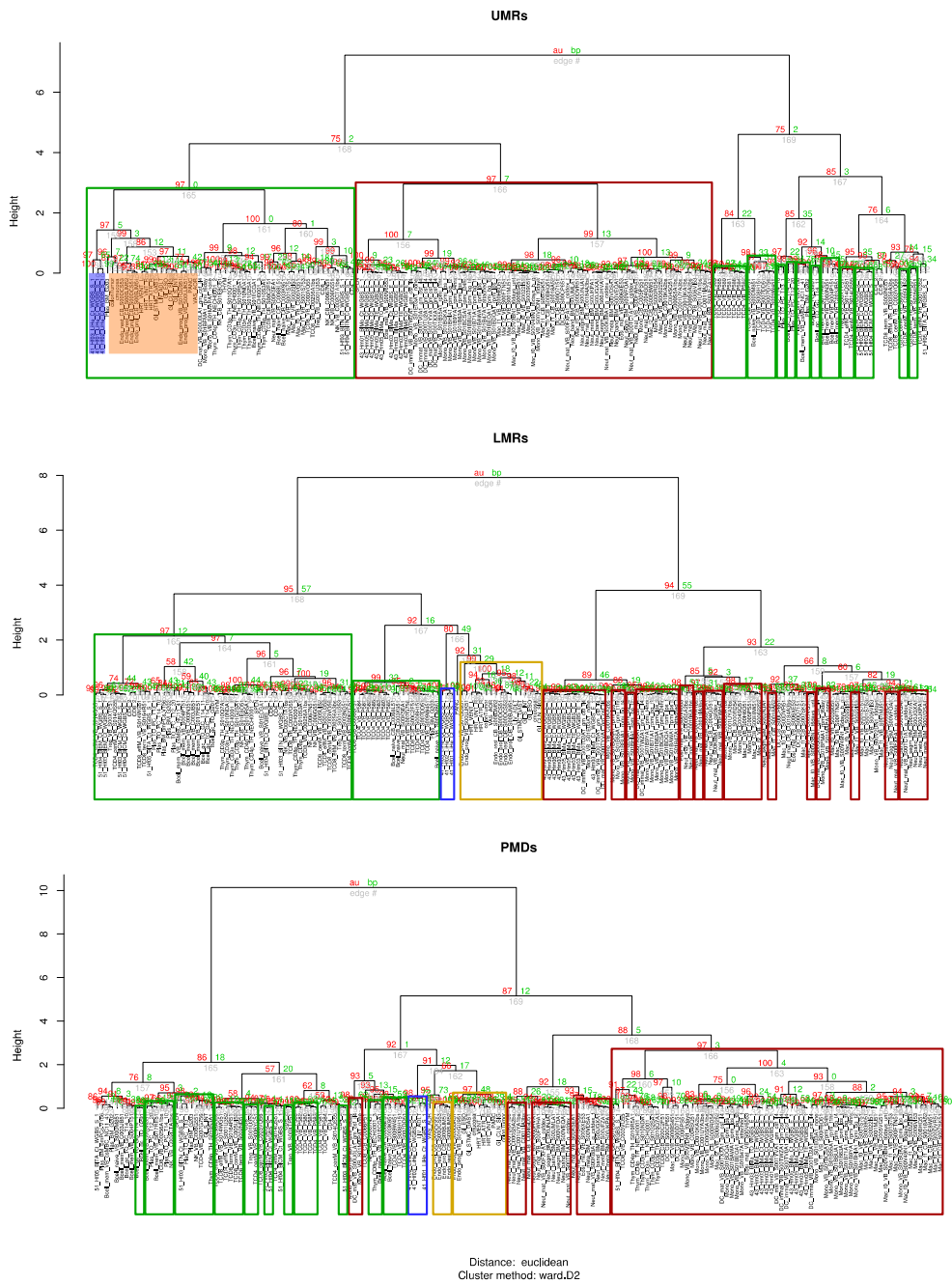
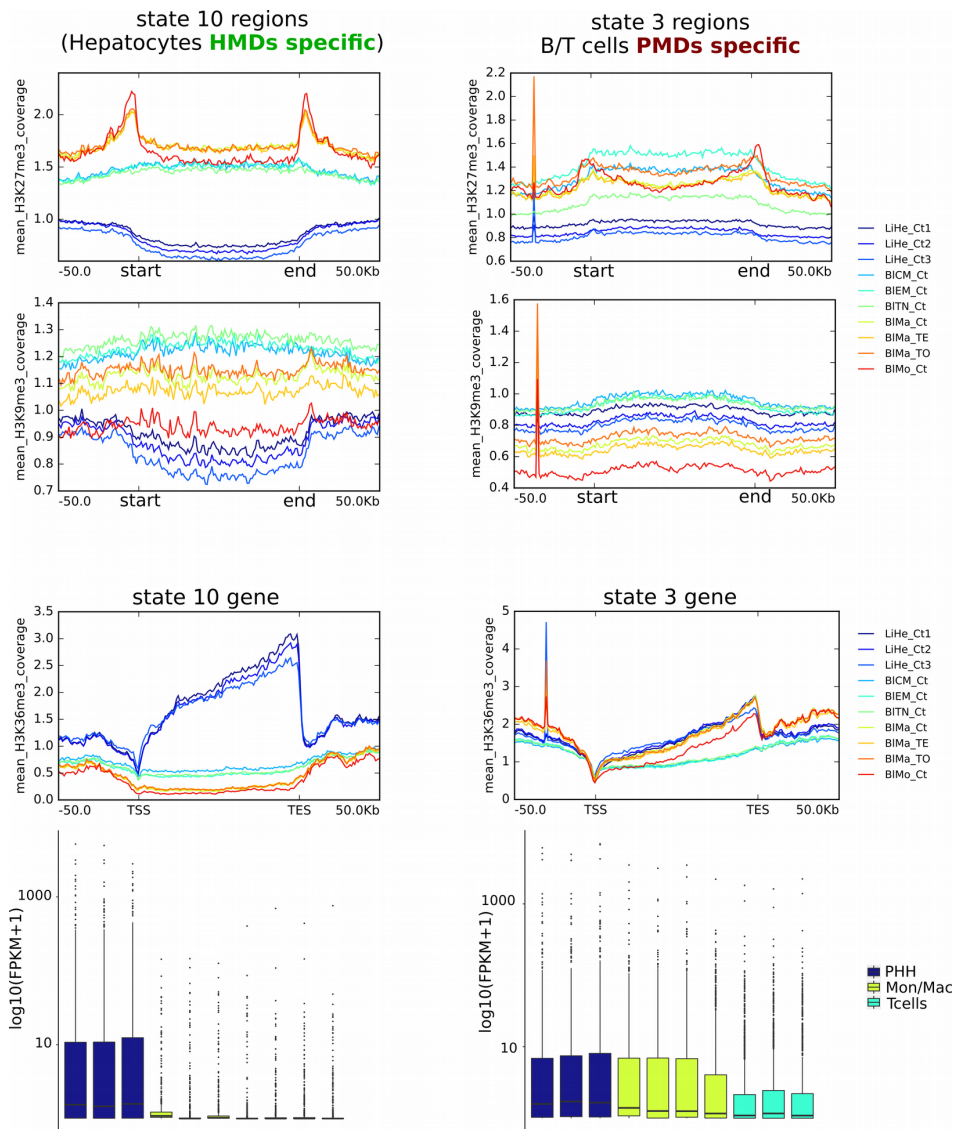


Figure S4. Dendrograms calculated from emission probabilities derived from ChromH3M model for PMDs, LMRs and UMRs analyzed with 10000 bootstrap replications. Red values are AU (Approximately Unbiased) p-value and green values are BP (Bootstrap Probability) value. Colored

boxes are the clusters with AU greater than 97% and they contain the same samples across the analyzed three segments. The two shaded boxes in UMRs correspond to the two same colored boxes (blue and brown) in LMRs and PMDs.

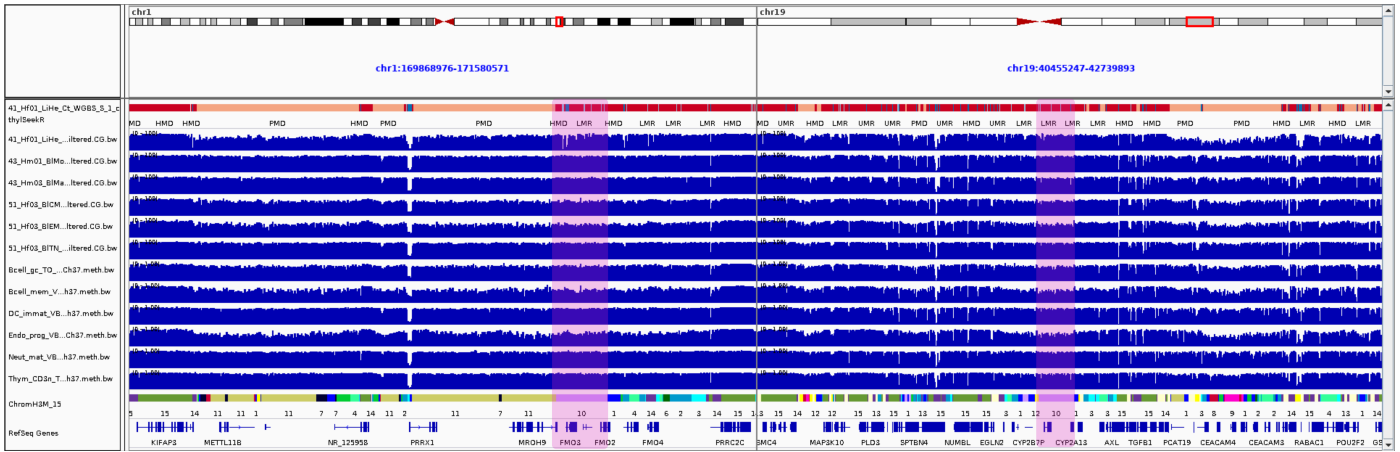


Figure S5



**Figure S5. Cell-type specific PMDs/FMRs and their heterochromatic signature.** The left panel represents the broad histone mark signals of four cell types; hepatocytes, monocytes, macrophages and T-cells in hepatocyte-specific HMDs together with FPKM values of the associated genes. The right panel is the same but for B/T cell-specific PMDs.

Figure S6



**Figure S6. hepatocyte gene specific locus.** Two hepatocyte gene families CYP and FMO which have been identified to be part of state10 displayed in Figure 2A.

Figure S7

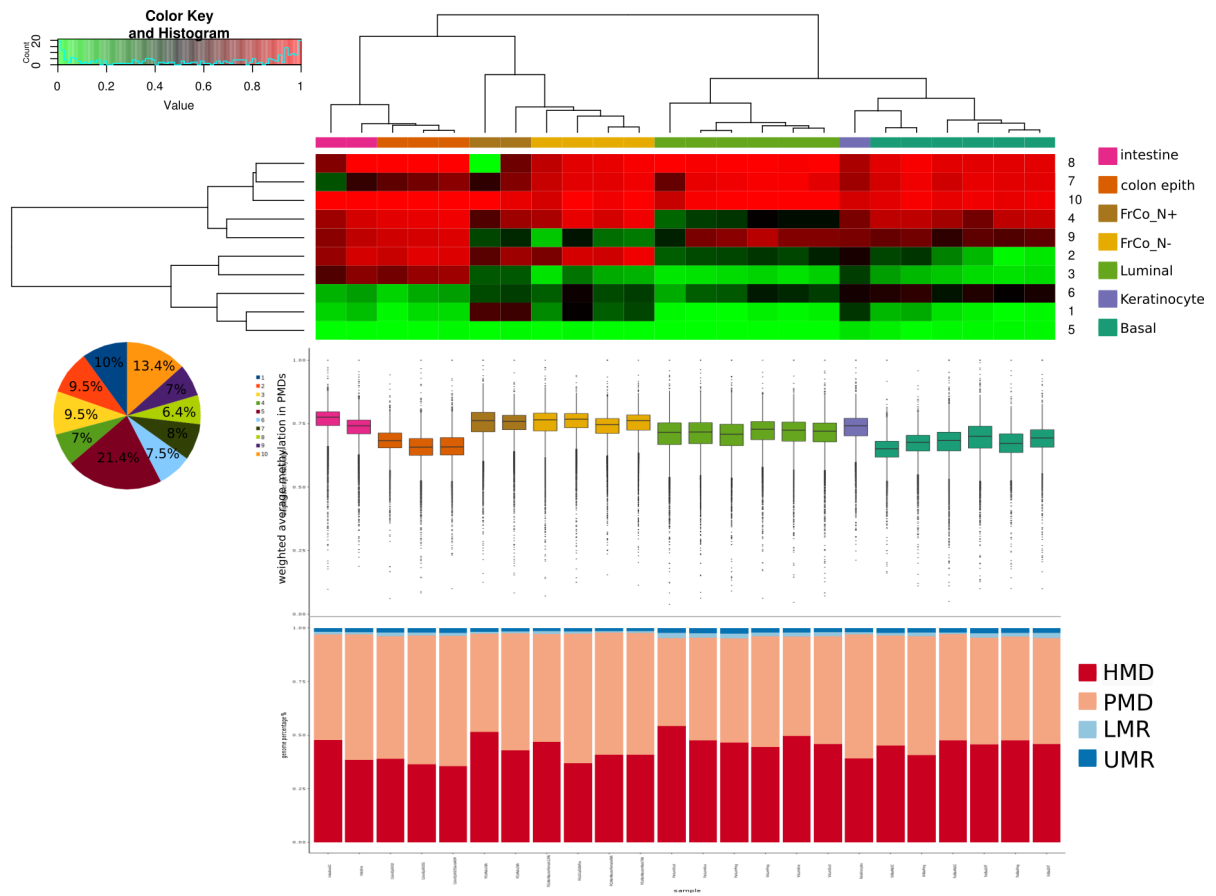
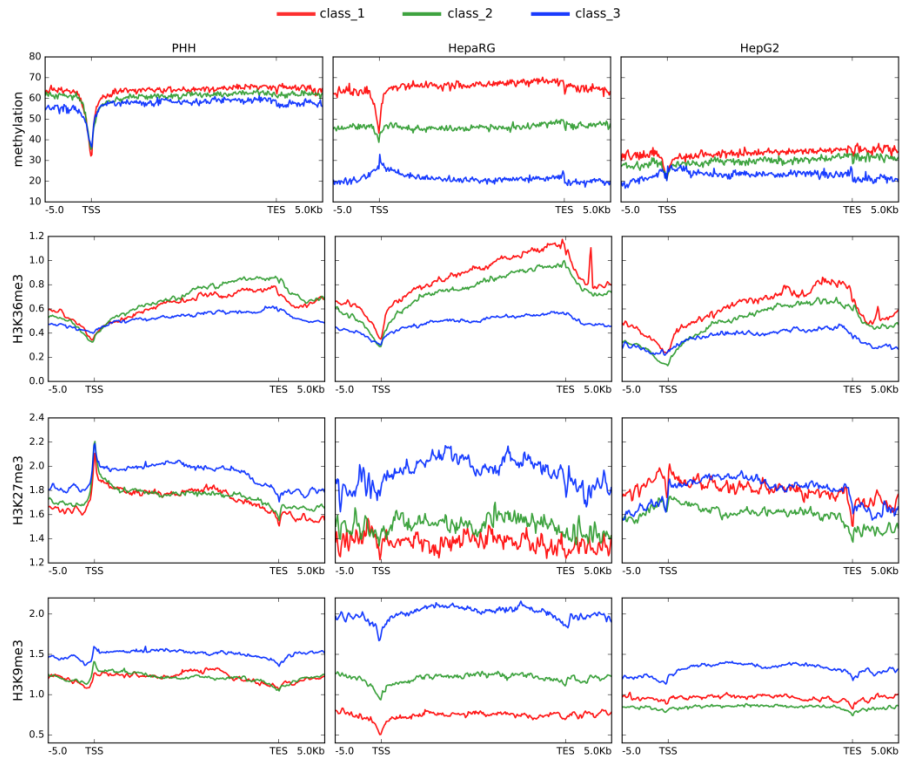


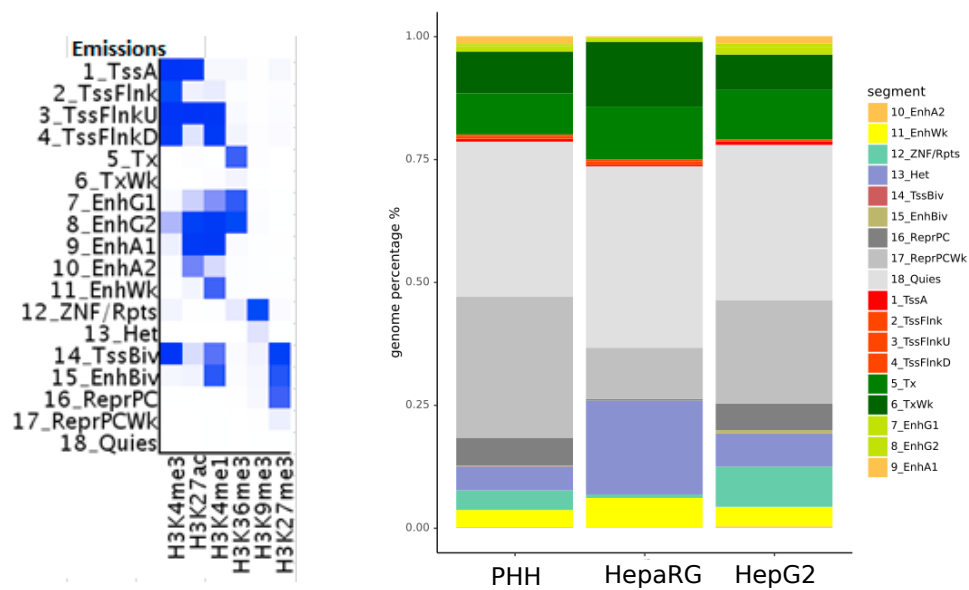
Figure S7. Analogue to Figure 2A for mouse data

Figure S8



**Figure S8. Epigenetic modification signatures in PHH<sub>PMDs</sub>.** DNA-methylation, H3K36me3, H3K27me3 and H3K9me3 signal across the gene bodies in three classes according to Figure 3C.

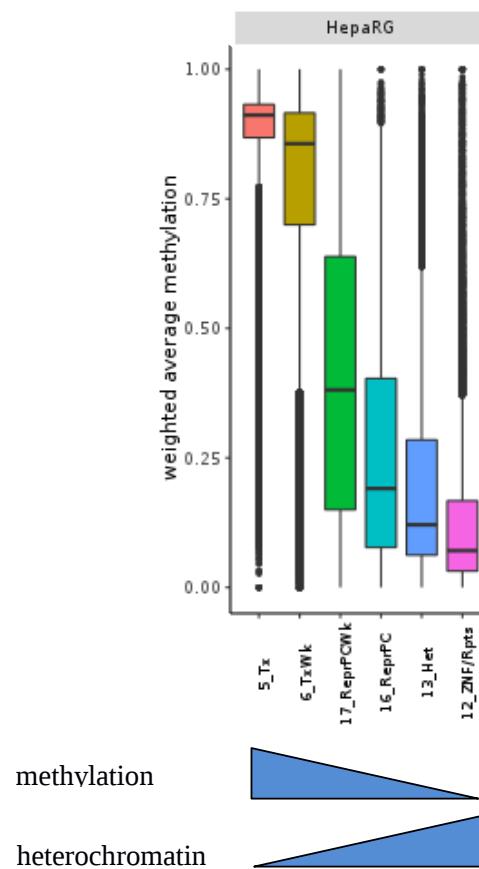
Figure S9



**Figure S9. chromatin state segmentation.** Emission probabilities of ChromHMM model and ChromHMM segments fraction related to Figure 3A. labels are the following:

1\_TssA: Active TSS, 2\_TssAFlank: Flanking active TSS, 3\_TxFlankU: Flanking TSS Upstream, 3\_TxFlankD: Flanking TSS Downstream, 5\_Tx: Strong transcription, 6\_TxWk: Weak transcription, 7\_EnhG1: Genic enhancer1, 7\_EnhG2: Genic enhancer2, 9\_EnhA1: Active Enhancer 1, 10\_EnhA2: Active Enhancer 2, 11\_EnhWk: Weak Enhancer, 12\_ZNF/Rpts: ZNF genes & repeats, 13\_Het: Heterochromatin, 14\_TssBiv: Bivalent/Poised TSS, 15\_EnhBiv: Bivalent Enhancer, 16\_ReprPC: Repressed PolyComb, 17\_ReprPCWk: Weak Repressed PolyComb, 18\_Quies: Quiescent/Low

Figure S10



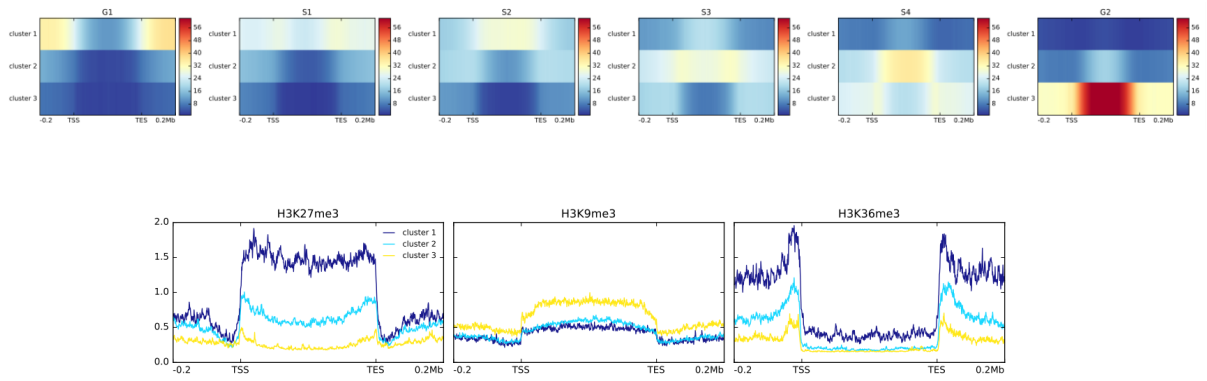
**Figure S10. DNA methylation erosion is accompanied by increasing of heterochromatic marks.**

Methylation levels across selected ChromHMM segments showing that DNA-demethylation is accompanied by heterochromatinization in HepaRG. labels are the following:

5\_Tx: Strong transcription, 6\_TxWk: Weak transcription, 12\_ZNF/Rpts: ZNF genes & repeats,

13\_Het: Heterochromatin, 16\_ReprPC: Repressed PolyComb, 17\_ReprPCWk: Weak Repressed PolyComb

Figure S11



**Figure S11. PMDs and heterochromatic marks demarcate distinct domains of late DNA-replication.** Heterochromatization at PMDs during cell cycle in IMR90 (related to Figure 4A, B of HepG2)

Figure S12

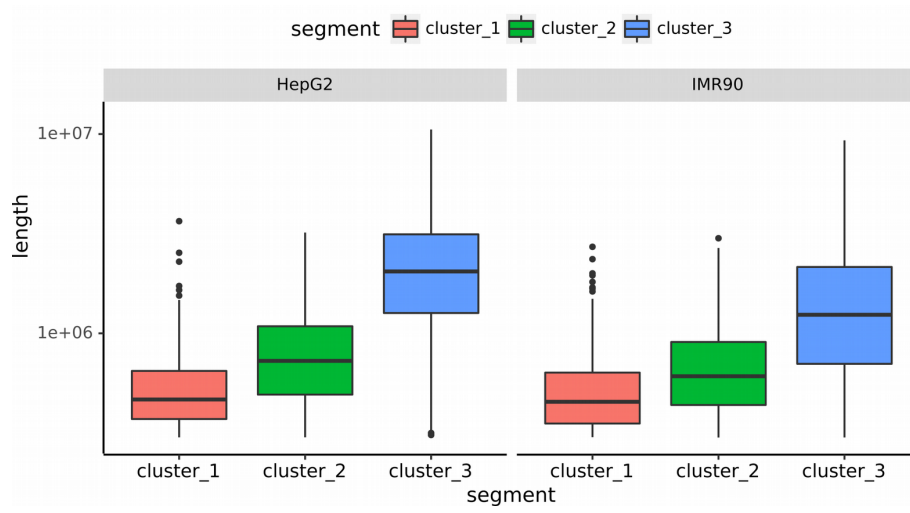


Figure S12. Cluster lengths



Figure S13

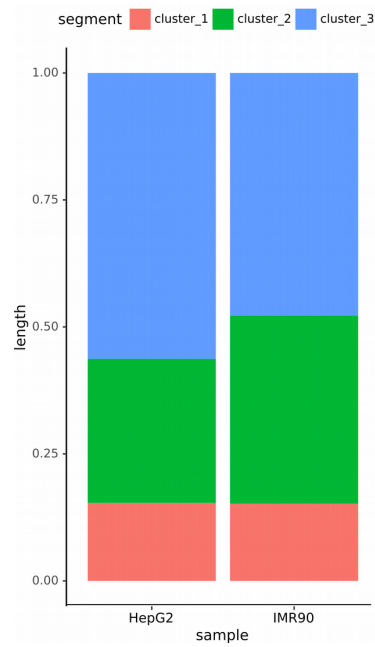
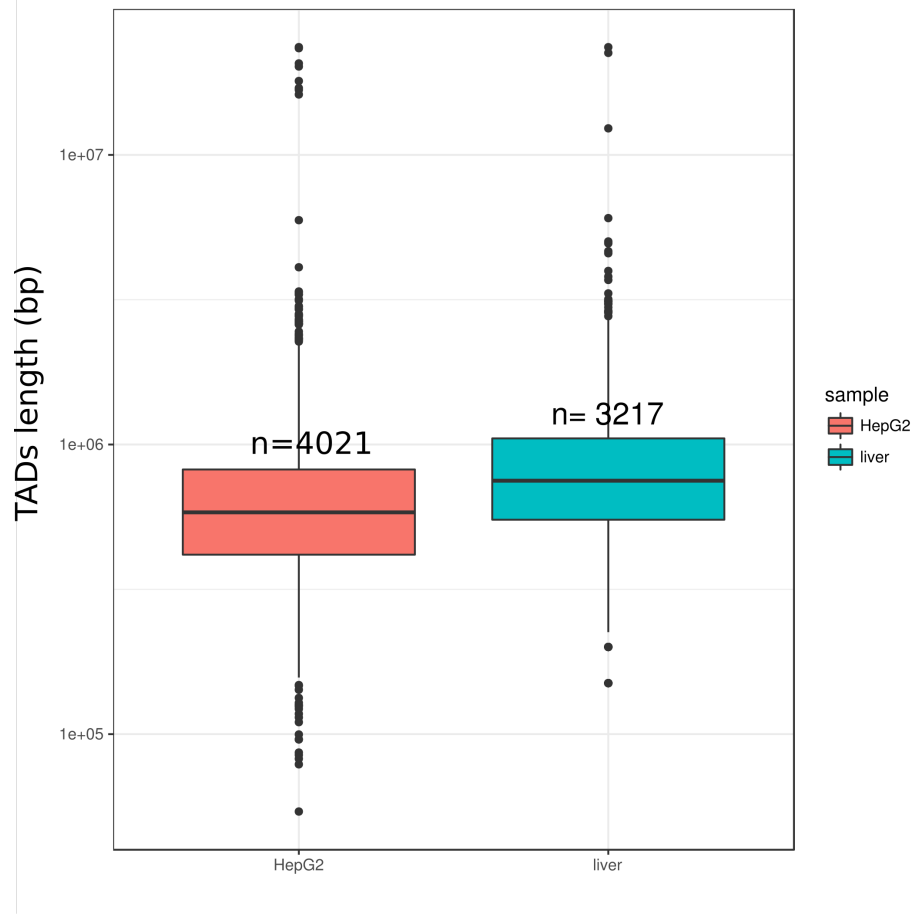


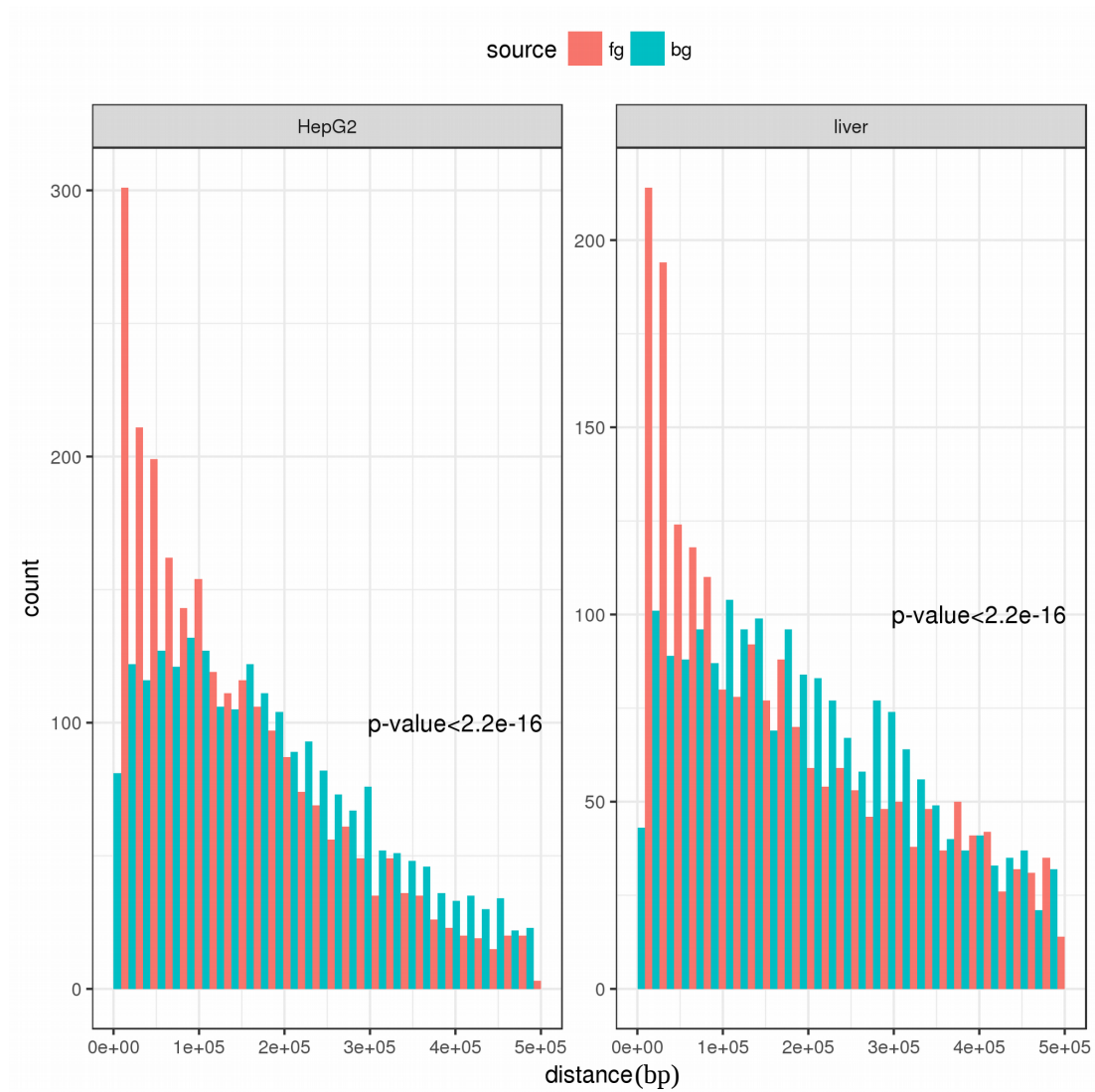
Figure S13. Clusters percentage of the genome

Figure S14



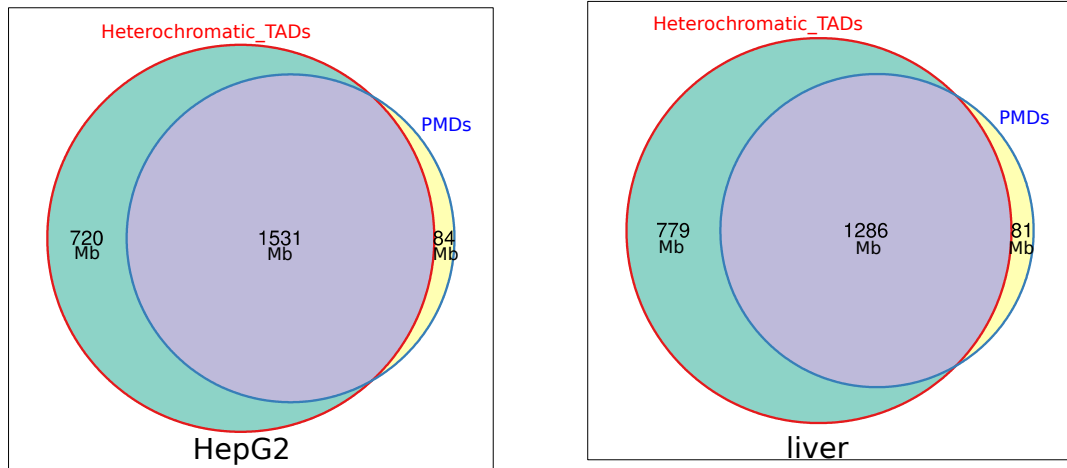
**Figure S14. TAD lengths.** TAD lengths in log10 scale for HepG2 and liver. The number above each boxplot represent the number of TADs for the corresponding sample.

Figure S15



**Figure S15. Distance distribution between TAD borders and PMD borders.** TAD borders are closer to PMD borders than randomized set of borders according to Wilcoxon test ( $p\text{-value} < 2.2e-16$ ). Fg (in red) represent the distance distribution between TAD borders and PMD borders. Bg (in green) represent the distance distribution between TAD borders and randomized test borders.

Figure S16



**Figure S16. PMDs and heterochromatic TADs overlap.** 94% of PMDs overlap with heterochromatic TADs in HepG2 and liver. Heterochromatic TADs in HepG2 form ~ 2.25 Gb of the genome while they are less in liver ~ 2.0 Gb.

Figure S17

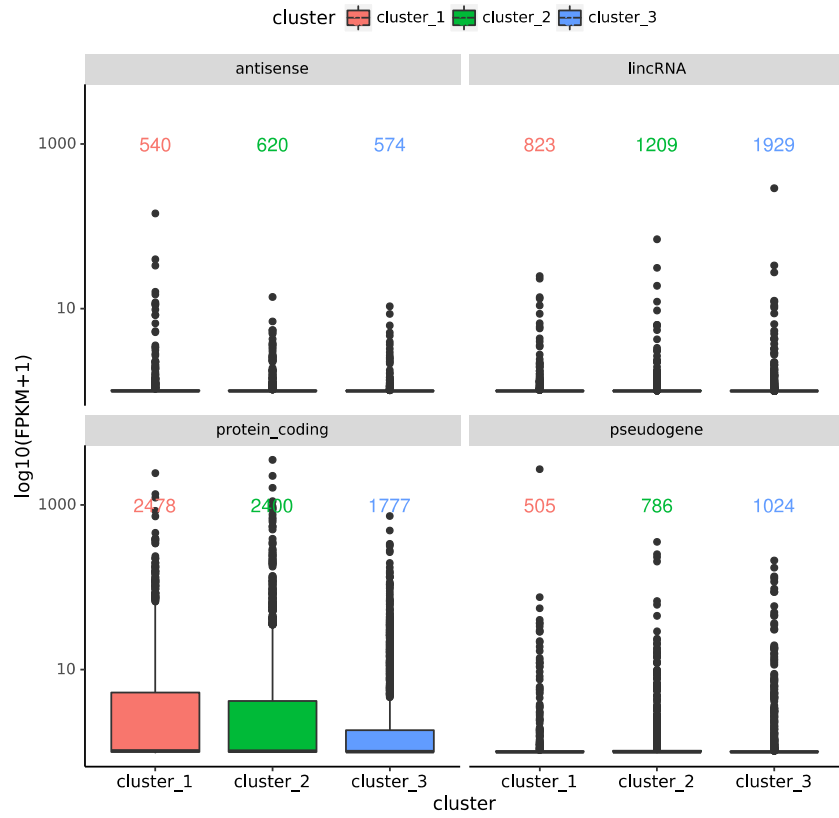


Figure S17. Number of the genes and FPKM values in some gene classes in each cluster

Figure S18

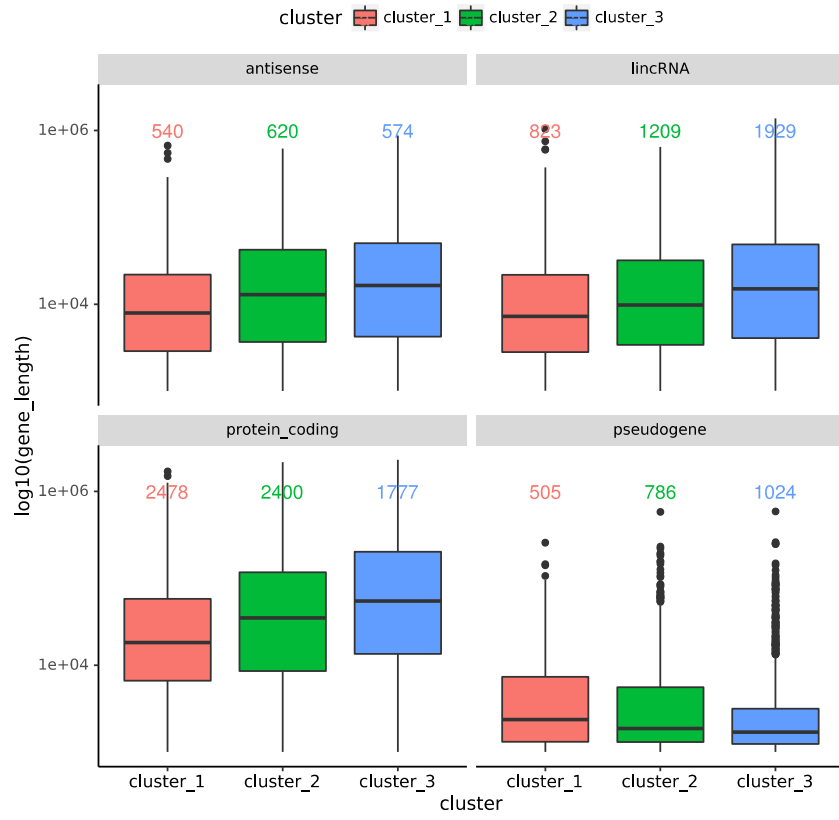
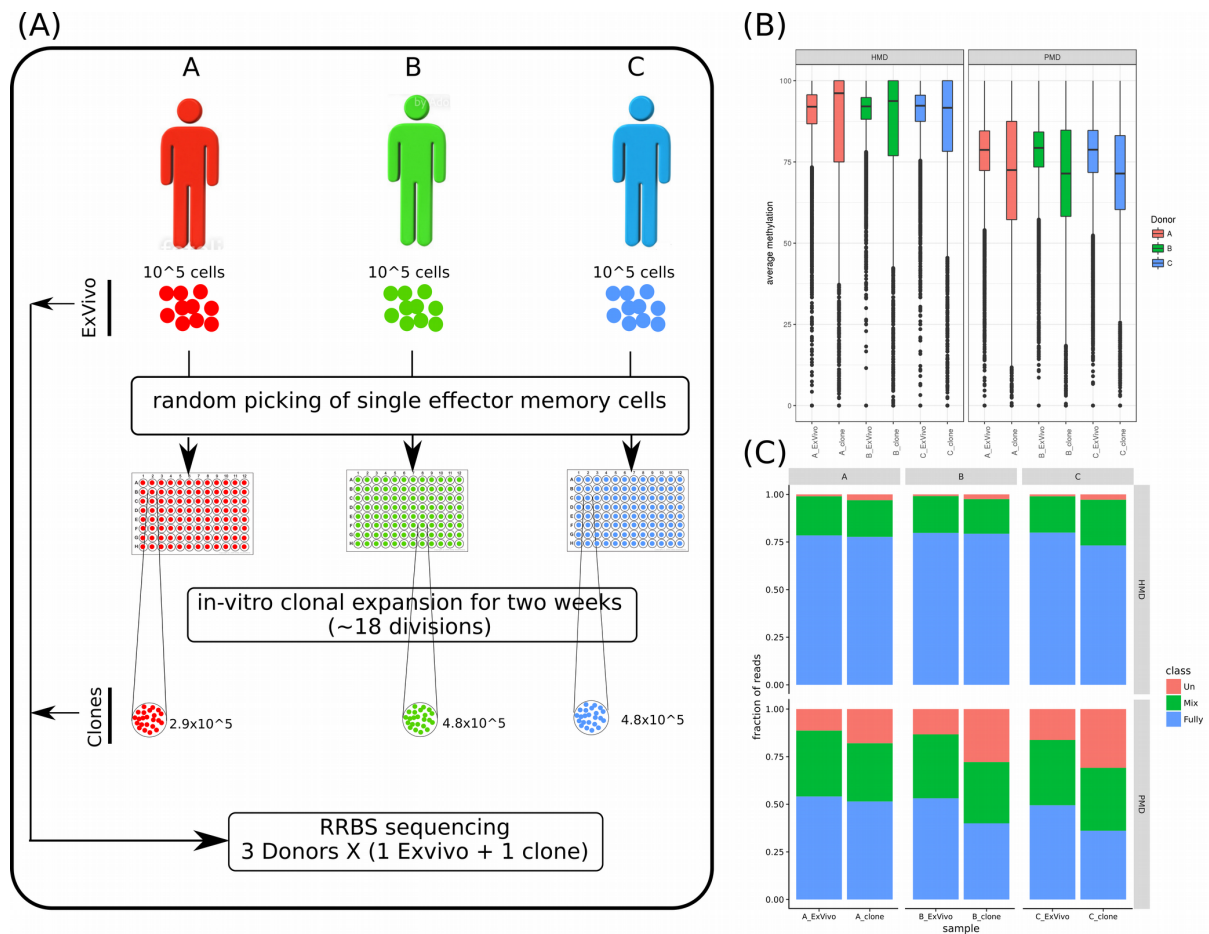


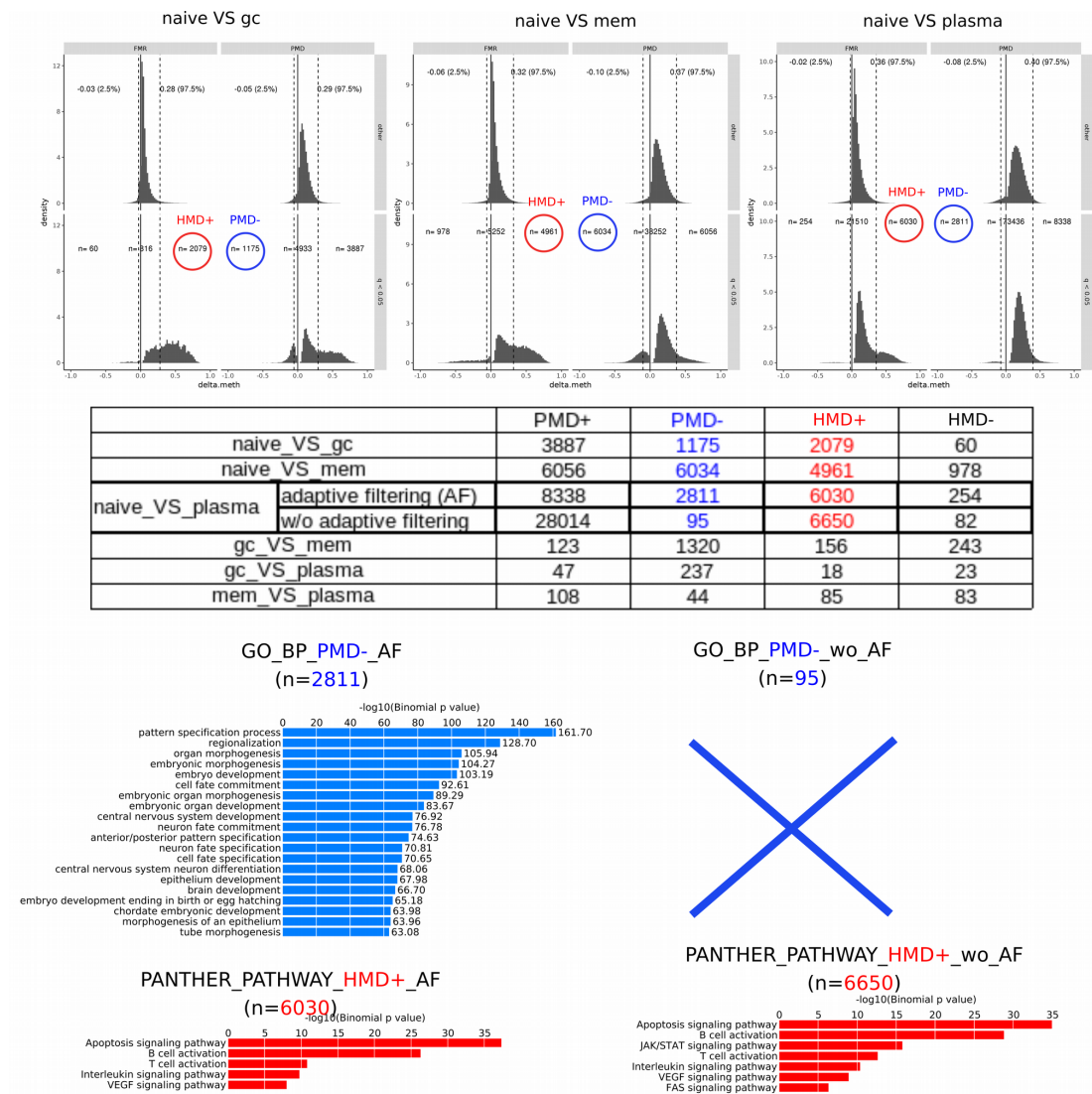
Figure S18. Gene lengths in some gene classes in each cluster

Figure S19



**Figure S19: Clonal cultures of T memory cells.** (A) Workflow of the designed experiment for T memory single cell expansion (details in the supplementary materials). (B) Average methylation across the three donors (ex vivo + in vitro cloned single cells) in PMDs/HMDs (as defined by TEM sample in Durek et al). A clear loss of methylation in PMDs of in vitro samples was observed. (C) CpG pattern class distributions in the PMDs/HMDs/others (as described in the supplementary materials). The fraction of mixed patterns (green), in PMDs, does not change due to the expansion process. The fraction of fully methylated patterns (blue) decreases and is compensated by an increase in the unmethylated patterns (red).

Figure S20



**Figure S20. Adaptive filtering.** DMRs analysis of B-cells during differentiation using the adaptive filtering method from Durek et al 2016. naive=Naive B-cells, gc=germinal center B-cells, mem=memory B-cells. Significant DMRs are annotated as HMDs+/- or PMDs+/- . Gene enrichment analysis for HMD+ and PMD- was performed using GREAT tool.



## Supplementary methods

### Clonal cultures of T memory cells

CD4<sup>+</sup> T memory cells (CD3<sup>+</sup> CD4<sup>+</sup> CD45RA<sup>-</sup> CD45RO<sup>+</sup> CD25<sup>-</sup>) from three different donors were sorted by flow-cytometry either as a bulk culture ('ex vivo' sample) or in a single-cell format into 96 well-plates. Single cells were cultured in the presence of a TCR stimulus (human T Cell Activation/Expansion Kit, Miltenyi Biotech) and human interleukin-2. After expansion, single clonal cultures were picked and treated with bisulfite for RRBS analysis. All three ex vivo and the matching three clones were sequenced.

### CpG patterns analysis:

We considered four consecutive CpGs to be in the same read and classified the patterns into three classes; fully methylated patterns (Fully), fully unmethylated patterns (Un) and the remaining pattern combinations are "mix". We calculated the fraction of each pattern genome wide in each ex vivo and the matched cloned sample. The patterns were stratified across HMDs and PMDs as defined from Figure 3A in the MS. We considered cluster 1 as "HMDs", cluster 5 as "PMDs" and the remaining clusters as "others".

## References

Durek, Pawel, et al. "Epigenomic profiling of human CD4<sup>+</sup> T cells supports a linear differentiation model and highlights molecular regulators of memory development." *Immunity* 45.5 (2016): 1148-1161.

## Chapter 3

# Epigenomic Profiling of Human CD4<sup>+</sup> T Cells Supports a Linear Differentiation Model and Highlights Molecular Regulators of Memory Development

The full text of this chapter was originally published as:

Durek, P.<sup>§</sup>, Nordström, K.<sup>§</sup>, Gasparoni, G.<sup>§</sup>, Salhab, A.<sup>§</sup>, Kressler, C.<sup>§</sup>, De Almeida, M.<sup>§</sup>, Bassler, K., Ulas, T., Schmidt, F., Xiong, J., et al. (2016). Epigenomic profiling of human CD4<sup>+</sup> T cells supports a linear differentiation model and highlights molecular regulators of memory development. *Immunity*, 45(5):1148-1161.

The author of this thesis contributed to the DNA-methylation analysis, PMDs calling. He generated the main Figure 1C and 1D, and the supplementary figures S2A, S2C and S5B. He contributed in writing the manuscript with other authors.

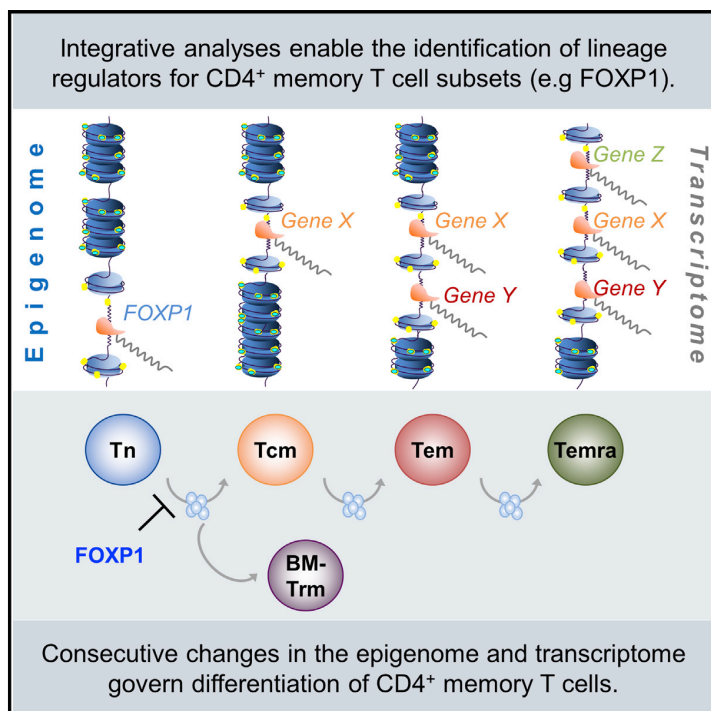
---

<sup>§</sup>Co-first authors

# Immunity

## Epigenomic Profiling of Human CD4<sup>+</sup> T Cells Supports a Linear Differentiation Model and Highlights Molecular Regulators of Memory Development

### Graphical Abstract



### Authors

Pawel Durek, Karl Nordström, Gilles Gasparoni, ..., Jörn Walter, Alf Hamann, Julia K. Polansky

### Correspondence

julia.polansky@drfz.de

### In Brief

As part of the IHEC consortium, Durek et al. (2016) generated deep epigenomes and transcriptomes of CD4<sup>+</sup> memory T cell subsets to infer their lineage relationships and to demonstrate the impact of epigenetic regulation on known and novel molecular regulators involved in memory generation. Explore the Cell Press IHEC webportal at [www.cell.com/consortium/IHEC](http://www.cell.com/consortium/IHEC).

### Highlights

- Comprehensive epigenomes for human CD4<sup>+</sup> T memory subsets generated and analyzed
- Integrative analyses support a linear model of memory T cell differentiation
- Epigenetic control of transcriptional regulators of memory differentiation revealed
- Chromatin changes highlight novel regulators for T memory cell differentiation



Durek et al., 2016, *Immunity* 45, 1148–1161  
 November 15, 2016 © 2016 Elsevier Inc.  
<http://dx.doi.org/10.1016/j.immuni.2016.10.022>

CellPress

# Epigenomic Profiling of Human CD4<sup>+</sup> T Cells Supports a Linear Differentiation Model and Highlights Molecular Regulators of Memory Development

Pawel Durek,<sup>1,23</sup> Karl Nordström,<sup>2,23</sup> Gilles Gasparoni,<sup>2,23</sup> Abdulrahman Salhab,<sup>2,23</sup> Christopher Kressler,<sup>1,23</sup> Melanie de Almeida,<sup>1,23</sup> Kevin Bassler,<sup>3</sup> Thomas Ulas,<sup>3</sup> Florian Schmidt,<sup>4,5</sup> Jieyi Xiong,<sup>6</sup> Petar Glazar,<sup>7</sup> Filippus Klironomos,<sup>7</sup> Anupam Sinha,<sup>8</sup> Sarah Kinkley,<sup>9</sup> Xinyi Yang,<sup>9</sup> Laura Arrigoni,<sup>10</sup> Azim Dehghani Amirabad,<sup>4,5</sup> Fatemeh Behjati Ardakani,<sup>4,5</sup> Lars Feuerbach,<sup>11</sup> Oliver Gorka,<sup>12</sup> Peter Ebert,<sup>4</sup> Fabian Müller,<sup>4</sup> Na Li,<sup>9</sup> Stefan Frischbutter,<sup>1</sup> Stephan Schlickeiser,<sup>13</sup> Carla Cendon,<sup>14</sup> Sebastian Fröhler,<sup>6</sup> Bärbel Felder,<sup>15</sup> Nina Gasparoni,<sup>2</sup> Charles D. Imbusch,<sup>11</sup> Barbara Hutter,<sup>11</sup> Gideon Zipprich,<sup>15</sup> Yvonne Tauchmann,<sup>16</sup> Simon Reinke,<sup>17</sup> Georgi Wassilew,<sup>18</sup> Ute Hoffmann,<sup>1</sup> Andreas S. Richter,<sup>10</sup> Lina Sieverling,<sup>11</sup> DEEP Consortium,<sup>19</sup> Hyun-Dong Chang,<sup>14</sup> Uta Syrbe,<sup>20</sup> Ulrich Kalus,<sup>16</sup> Jürgen Eils,<sup>15</sup> Benedikt Brors,<sup>11</sup> Thomas Manke,<sup>10</sup> Jürgen Ruland,<sup>12,21,22</sup> Thomas Lengauer,<sup>4</sup> Nikolaus Rajewsky,<sup>7</sup> Wei Chen,<sup>6</sup> Jun Dong,<sup>14</sup> Birgit Sawitzki,<sup>13</sup> Ho-Ryun Chung,<sup>9</sup> Philip Rosenstiel,<sup>8</sup> Marcel H. Schulz,<sup>4,5</sup> Joachim L. Schultze,<sup>3</sup> Andreas Radbruch,<sup>14</sup> Jörn Walter,<sup>2</sup> Alf Hamann,<sup>1</sup> and Julia K. Polansky<sup>1,24,\*</sup>

<sup>1</sup>Experimental Rheumatology, German Rheumatism Research Centre, 10117 Berlin, Germany

<sup>2</sup>Department of Genetics, University of Saarland, 66123 Saarbrücken, Germany

<sup>3</sup>Life and Medical Sciences Institute, Genomics and Immunoregulation, University of Bonn, 53115 Bonn, Germany

<sup>4</sup>Department of Computational Biology and Applied Algorithmics, Max Planck Institute for Informatics, 66123 Saarbrücken, Germany

<sup>5</sup>Excellence Cluster on Multimodal Computing and Interaction, Saarland University, 66123 Saarbrücken, Germany

<sup>6</sup>Berlin Institute for Medical Systems Biology, Max-Delbrück Center for Molecular Medicine, 13125 Berlin, Germany

<sup>7</sup>Systems Biology of Gene Regulatory Elements, Max-Delbrück Center for Molecular Medicine, 13125 Berlin, Germany

<sup>8</sup>Institute of Clinical Molecular Biology, Christian-Albrechts-University, 24105 Kiel, Germany

<sup>9</sup>Otto Warburg Laboratories: Epigenomics at Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany

<sup>10</sup>Max Planck Institute of Immunobiology and Epigenetics, 78108 Freiburg, Germany

<sup>11</sup>Applied Bioinformatics, Deutsches Krebsforschungszentrum, 59120 Heidelberg, Germany

<sup>12</sup>Institute for Clinical Chemistry and Pathobiochemistry, Klinikum rechts der Isar, Technical University 81675 Munich, Germany

<sup>13</sup>Institute of Medical Immunology, Charité University Medicine, 13353 Berlin, Germany

<sup>14</sup>Cell Biology, German Rheumatism Research Centre, 10117 Berlin, Germany

<sup>15</sup>Data Management and Genomics IT, Deutsches Krebsforschungszentrum, 69120 Heidelberg, Germany

<sup>16</sup>Institut für Transfusionsmedizin, Charité University Medicine, 12203 Berlin, Germany

<sup>17</sup>Berlin-Brandenburg Center for Regenerative Therapies, 13353 Berlin, Germany

<sup>18</sup>Center for Musculoskeletal Surgery, Charité University Medicine, 10117 Berlin, Germany

<sup>19</sup><http://www.deutsches-epigenom-programm.de/>

<sup>20</sup>Medizinische Klinik für Gastroenterologie, Infektiologie und Rheumatologie, Charité University Medicine, 12000 Berlin, Germany

<sup>21</sup>German Cancer Consortium (DKTK), 59120 Heidelberg, Germany

<sup>22</sup>German Center for Infection Research (DZIF), partner site 81675 Munich, Germany

<sup>23</sup>Co-first authors

<sup>24</sup>Lead Contact

\*Correspondence: [julia.polansky@drfz.de](mailto:julia.polansky@drfz.de)

<http://dx.doi.org/10.1016/j.immuni.2016.10.022>

## SUMMARY

The impact of epigenetics on the differentiation of memory T (Tmem) cells is poorly defined. We generated deep epigenomes comprising genome-wide profiles of DNA methylation, histone modifications, DNA accessibility, and coding and non-coding RNA expression in naive, central-, effector-, and terminally differentiated CD45RA<sup>+</sup> CD4<sup>+</sup> Tmem cells from blood and CD69<sup>+</sup> Tmem cells from bone marrow (BM-Tmem). We observed a progressive and proliferation-associated global loss of DNA methylation in heterochromatic parts of the genome during Tmem cell differentiation. Furthermore, distinct gradually changing signatures in the epigenome and the tran-

scriptome supported a linear model of memory development in circulating T cells, while tissue-resident BM-Tmem branched off with a unique epigenetic profile. Integrative analyses identified candidate master regulators of Tmem cell differentiation, including the transcription factor FOXP1. This study highlights the importance of epigenomic changes for Tmem cell biology and demonstrates the value of epigenetic data for the identification of lineage regulators.

## INTRODUCTION

CD4<sup>+</sup> T helper (Th) cells orchestrate the quality and quantity of an adaptive immune reaction and contribute to immunity by

generating a pool of long-lived memory (Tmem) cells, which arise from naive T (Tn) cells after activation by primary antigen encounter. Tmem cells are per se resting, almost non-dividing cells, which can be subdivided into subpopulations based on marker expression, tissue localization and functional properties. Central memory (Tcm) cells appear most similar to Tn cells with respect to their ability to recirculate through blood and lymphoid tissues, the limited effector cytokine commitment, and their high proliferative capacity (Sallusto et al., 1999). In contrast, T effector memory (Tem) cells preferentially home to peripheral tissues and show commitment for the selective production of effector cytokine panels (e.g., IFN- $\gamma$ , IL-4, and IL-17) characteristic of their functional subtype (Th1, Th2, and Th17, respectively). Their capacity to expand and differentiate is more limited than that of Tcm cells—a feature also found for the so far poorly characterized CD4<sup>+</sup> terminally differentiated CD45RA<sup>+</sup> memory (Temra) cells (Henson et al., 2012), which feature expression of selected markers of Tn cells (e.g., CD45RA). In addition to these populations circulating through the blood, recent studies have highlighted the importance of tissue-resident memory cells (Carbone et al., 2013; Schenkel and Masopust, 2014). CD4<sup>+</sup> Tmem cells from the bone marrow (BM-Tmem) have been shown to constitute a major part of long-term memory in mouse and man (Okhrimenko et al., 2014; Tokoyoda et al., 2009).

The developmental relationship of Tmem cell subsets is not well defined. The question whether different Tmem subtypes represent stages in a sequential linear differentiation process, or whether they branch into different sublineages from early activation stages is still a subject of controversy (Ahmed et al., 2009; Flossdorf et al., 2015; Harrington et al., 2008; Kaech and Cui, 2012). Similarly, master regulators controlling the transit from naive to memory stages, particularly in the human system, are largely unknown, partially due to the lack of suitable experimental systems.

Epigenetic mechanisms play a key role in cell differentiation by controlling expression programs that are stable over time and through cellular generations and hence are prime candidates for the imprinting of stable, heritable expression profiles. Because Tmem cells do not revert to the naive stage, their cellular program seems to be permanently switched, pointing toward epigenetic regulation. Main players in epigenetic regulation are DNA methylation (DNA-meth), histone modifications, and non-coding RNAs, which together direct the rearrangement of the chromatin to promote or to prevent expression of the affected genes. Genome-wide analysis of such epigenetic marks therefore allows for conclusions not only on the current gene expression status but also facilitates insights into the history and the future potential of cells. To date only a few studies on mouse Tmem cells have been published, reporting limited datasets (Crompton et al., 2016; Hashimoto et al., 2013; Komori et al., 2015; Russ et al., 2014). A deep and systematic genome-wide analysis of the epigenetic landscape during human CD4<sup>+</sup> Tmem cell differentiation is currently lacking.

As part of the International Human Epigenome Consortium (IHEC) and the German Epigenome Programme (DEEP), we generated comprehensive epigenomic maps of ex vivo isolated isogenic human CD4<sup>+</sup> Tn cells and several Tmem cell subsets from the blood and the bone marrow to address the question

of whether and how the epigenome contributes to the formation, maintenance, and function of Tmem cell populations in humans. Our data support a model of linear differentiation for circulating human Tmem cells—a topic so far studied only in the murine system. In addition, we find that many known molecular regulators of Tmem cells are under epigenetic control and that epigenetic changes point to novel regulator candidates, which are likely to be involved in Tmem cell differentiation.

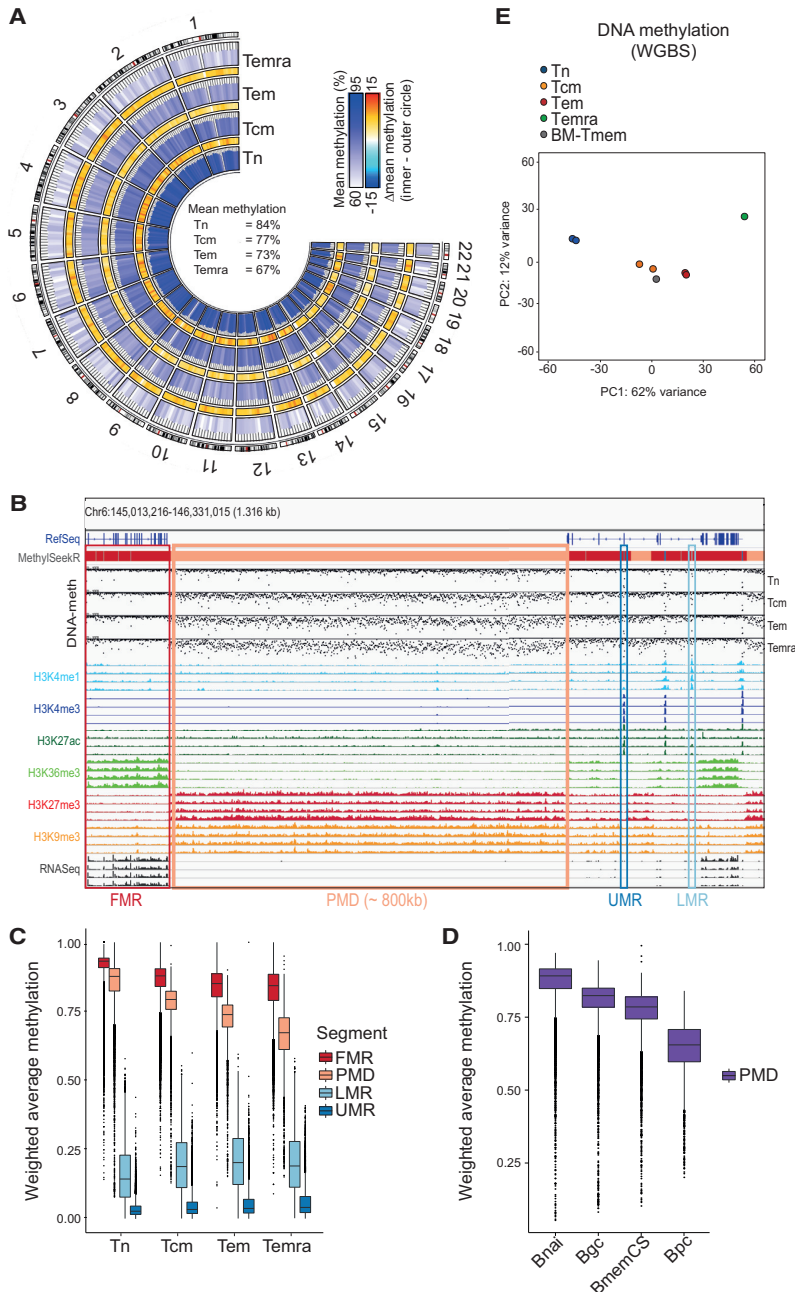
## RESULTS

### Generation of Genome-Wide Epigenetic Datasets of Human CD4<sup>+</sup> Tn Cells and Tmem Cell Subsets

To generate comprehensive epigenomic datasets (i.e., class I epigenomes according to the IHEC standards) for the key differentiation stages of human CD4<sup>+</sup> Th cells, we sorted CD4<sup>+</sup> Tn, Tcm, Tem, and Temra cells from the peripheral blood of healthy human donors by flow cytometry (Figure S1A). To obtain sufficient cell numbers for the subsequent analyses and to mitigate potential inter-donor variations, we used pooled samples of 3–10 female donors (Table S1). For Tn, Tcm, and Tem, analyses of all epigenetic parameters within one replicate were carried out in parallel, i.e., were derived from the same genetic donor pool and therefore represent isogenic samples. For each sample we determined (1) genome-wide DNA-meth profiles, by whole-genome bisulfite sequencing (WGBS) or by reduced representation bisulfite sequencing (RRBS), (2) DNA accessibility maps by nucleosome occupancy and methylome sequencing (NOME-seq), (3) high-resolution histone modification maps (by Chromatin Immune-Precipitation sequencing, ChIP-seq) for H3K4me1 (= mono-methylation of lysine 4 on histone 3), H3K4me3, H3K9me3, H3K27ac, H3K27me3, and H3K36me3 and, (4) transcriptomes for total RNA (depleted from ribosomal RNAs), messenger RNAs (mRNAs), long non-coding RNAs (lncRNAs), micro RNAs (miRNA), and circular RNAs (circRNA) by deep sequencing of three different RNA libraries (polyadenylated RNAs, small RNAs, and total RNAs depleted from ribosomal RNAs). A selection of these datasets (Figure S1B) was generated for CD4<sup>+</sup> BM-Tmem cells, which were separated into the CD69<sup>+</sup> tissue-resident and the circulating CD69<sup>-</sup> subsets (Figure S1A).

### Progressive Segmented Loss of DNA-Meth Correlates with Tmem Cell Differentiation

We profiled the DNA-meth landscape in Tn, Tcm, Tem, and Temra cells using WGBS and observed a strong progressive loss of DNA-meth in the order Tn-Tcm-Tem-Temra with mean methylation levels for the entire genome dropping from 84% in Tn to 67% in Temra (Figure 1A). Loss of methylation predominantly occurred in large domains of up to several hundreds of kilobases (kb), which were decorated with the repressive histone marks H3K27me3 and H3K9me3 (Figure 1B and Figure S2A). Such regions are referred to as “partially methylated domains, PMDs” and can be identified using established software packages (“MethylSeekR,” Burger et al., 2013). PMDs contrasted with broad regions that were uniformly fully methylated (“fully methylated regions,” FMRs, by MethylSeekR) and to peaks of strong consistent de-methylation typically found in CpG islands



**Figure 1. Global Loss of DNA-Meth in Tmem Cells Occurs in Large Heterochromatic Regions**

(A) Circle plots of WGBS data (Tn, Tcm, Tem of donor pool Hf03, and Temra, Hf05) are shown. Mean methylation levels in 10 Mb blocks are depicted as color-coded (white-blue) bars. Heat-maps (blue-red) indicate the methylation difference between the adjacent subsets. Total mean methylation for each cell type is given in the center. (B) Exemplary genomic view of Tn, Tcm, Tem (Hf03 samples), and Temra (Hf05), displaying examples of the genomic segments called by the MethylSeekR software from WGBS data (indicated with boxes): PMD, partially methylated domain; FMR, fully methylated region; LMR, low methylated region; UMR, unmethylated region. The following tracks are shown (top to bottom, each for Tn, Tcm, Tem, Temra): Genes annotated in RefSeq; MethylSeekR-segments; DNA-meth (WGBS); 6 indicated histone modifications; total RNA.

(C) Weighted average DNA-meth across the MethylSeekR segments.

(D) Weighted average methylation across PMDs in B cells (data from the BLUEPRINT project, see Accession codes in [Experimental Procedures](#)). BnaI, naive B cells (ERX625136); Bgc, germinal center B cells (ERX715129); BmemCS, class-switched memory B cells (ERX625127); Bpc, plasma cells (ERX301127).

(E) PCA of DNA-meth data (based on WGBS). CpGs with min. coverage = 5 were considered; only CpGs with calls in all indicated samples were used.

We observed a similar segmented loss of global DNA-meth when re-examining DNA-meth profiles from B cells published by the BLUEPRINT consortium ([Kulis et al., 2015](#)). Here too, PMDs were the genomic segments that displayed progressive loss of DNA-meth with differentiation into memory B cells and antibody-secreting plasma cells ([Figure 1D](#)), indicating that this phenomenon is shared during lymphocyte development.

In a principal-component analysis (PCA), the blood-derived T cell subsets were placed along the main principal component 1 (PC1), in the order Tn-Tcm-Tem-Temra ([Figure 1E](#)), which

(“unmethylated regions,” UMRs) and transcriptional control regions (e.g., CpG-low promoters and enhancers, “low methylated regions,” LMRs). PMDs showed the strongest loss of methylation of all MethylSeekR segments ([Figure 1C](#)) and covered up to 67% of the genome (in Tem cells; [Figure S2B](#)). Hence, PMDs were responsible for the bulk of the observed global DNA de-methylation in Tmem cell populations. PMD-associated genes generally showed low expression levels compared to FMR-associated genes and were fewer in number ([Figure S2C](#)).

mirrors the DNA de-methylation in PMDs. Temra cells fell at the extreme position along PC1 in relation to Tn cells, suggesting that they are the most differentiated population. However, their inter-donor pool variation was larger compared to other cell types ([Figure S2D](#)). In contrast to Temra cells, BM-Tmem cells took an “intermediate” position on PC1 close to circulating Tcm and Tem cells, indicating that their epigenetic imprint toward terminal differentiation is less pronounced ([Figure 1E](#)).

These data show that DNA de-methylation in heterochromatic parts of the genome accompanies Tmem cell differentiation in the order of Tn-Tcm-Tem-Temra with BM-Tmem cells clustering with the Tcm and Tem cell populations.

### **Comprehensive Transcriptome Analyses Reveal a Progressive Change with Tmem Cell Differentiation in the Order of Tn-Tcm-Tem-Temra**

We generated full transcriptomes by RNA-seq and determined expression profiles for total RNA, mRNAs, miRNAs, lncRNA, and circRNA for Tn cells and Tmem subsets. Our analysis identified previously described RNAs, as well as previously unknown RNAs (including 981 novel miRNAs, 173 lncRNAs, and 4,826 candidate circRNAs) and many differentially expressed RNAs between the T cell subtypes (Tables S2–S4).

We performed PCA on each of these functionally independent RNA species. Our analysis revealed a consistent pattern with respect to the main component PC1: for all RNA species, the cell types fell along this axis in the strict order of Tn-Tcm-Tem-Temra (Figure 2A). As observed for DNA-meth, BM-Tmem cells took an intermediate position close to Tcm and Tem cells from the blood rather than resembling the most terminally differentiated Temra cell population. For total RNA and lncRNAs, PC2 indicated properties of Tn that were recapitulated in Temra cells and distinguished them from the other memory subsets. Inter-donor pool differences were generally small, except in the miRNA datasets, in which one donor pool became separated by PC2 from all others. Thus, the consistent arrangement of the T cell subtypes on PC1 for all RNA species indicated a progressive change of the transcriptome during Tmem cell differentiation (Tn-Tcm-Tem-Temra).

To validate this further, we performed a co-expression network analysis using the Tn, Tcm, and Tem samples and focused on the 700 most variable genes or on transcriptional regulators (TRs). The topology of both networks showed similar features, with two major gene clusters and a smaller number of genes connecting these two clusters (Figure 2B). Overlaying the expression differences of the included genes revealed that one major cluster was defined by Tn-, the other by Tem cell-associated genes. Tcm cell-associated genes mainly connected the two main clusters, indicating that this population indeed represents an intermediate stage of T cell differentiation.

An additional bioinformatic approach was used to evaluate the mode of differentiation. We used the degree of similarity of the entire transcriptomes between Tn, Tcm, and Tem cells and calculated the likelihood of possible differentiation models: two linear models in the order of Tn-Tcm-Tem or Tn-Tem-Tcm and one bifurcated model in which Tcm and Tem cells arise independently from Tn. As shown in Figure 2C, the linear Tn-Tcm-Tem model had the highest cosine similarity score of 0.98 (max = 1) and was significantly different from the other two models ( $p < 10^{-16}$ ).

These data show that the transcriptome changes progressively during Tmem cell differentiation in the order of Tn-Tcm-Tem-Temra.

### **Chromatin Accessibility and DNA-Meth Analyses Support a Linear Model of Differentiation for Circulating Tmem Cells**

We wanted to clarify whether the linear relationship between the Tmem cell subsets (Tn-Tcm-Tem-Temra) apparent from the

DNA-meth and transcriptome data (Figures 1E and 2A), could also be deduced from epigenetic imprints in the chromatin structure. For this, we first analyzed genome-wide DNA accessibility maps, which were generated by NOME-seq. In a PCA, again a linear arrangement of the blood-derived T cell subsets in the order of Tn-Tcm-Tem-Temra was visible on the 2<sup>nd</sup> most important component PC2 (Figure 2D). However, the different populations were generally less stringently separated. The main component PC1 separated the replicates Hf03 and Hf04 from Hf06, which reflected a slight change in the NOME protocol between these samples (Figure S2E and Supplemental Experimental Procedures). In addition, when we called accessible (= nucleosome-depleted) regions (NOME-peaks) from Tn, Tcm, and Tem cells and compared their degree of accessibility between the cellular subtypes, the vast majority of sites gained or lost accessibility in the order Tn-Tcm-Tem (Figure 2E).

Next, we analyzed global DNA-meth profiles (by RRBS) of blood- and bone-marrow-derived Tmem cell subsets, with the latter population subdivided into a tissue-resident CD69<sup>+</sup> and a circulating CD69<sup>-</sup> fraction (CD69 being a regulator of tissue egress and marker for tissue-resident cells; Sathaliyawala et al., 2013). While the CD69<sup>-</sup> fraction clustered closely to Tcm and Tem cells from the blood, the CD69<sup>+</sup> tissue-resident BM-Tmem subfraction deviated from its CD69<sup>-</sup> counterpart, as well as from blood-derived circulating populations in PC2 (Figure 2F), indicating a major epigenetic imprint for their tissue residency and specialized function.

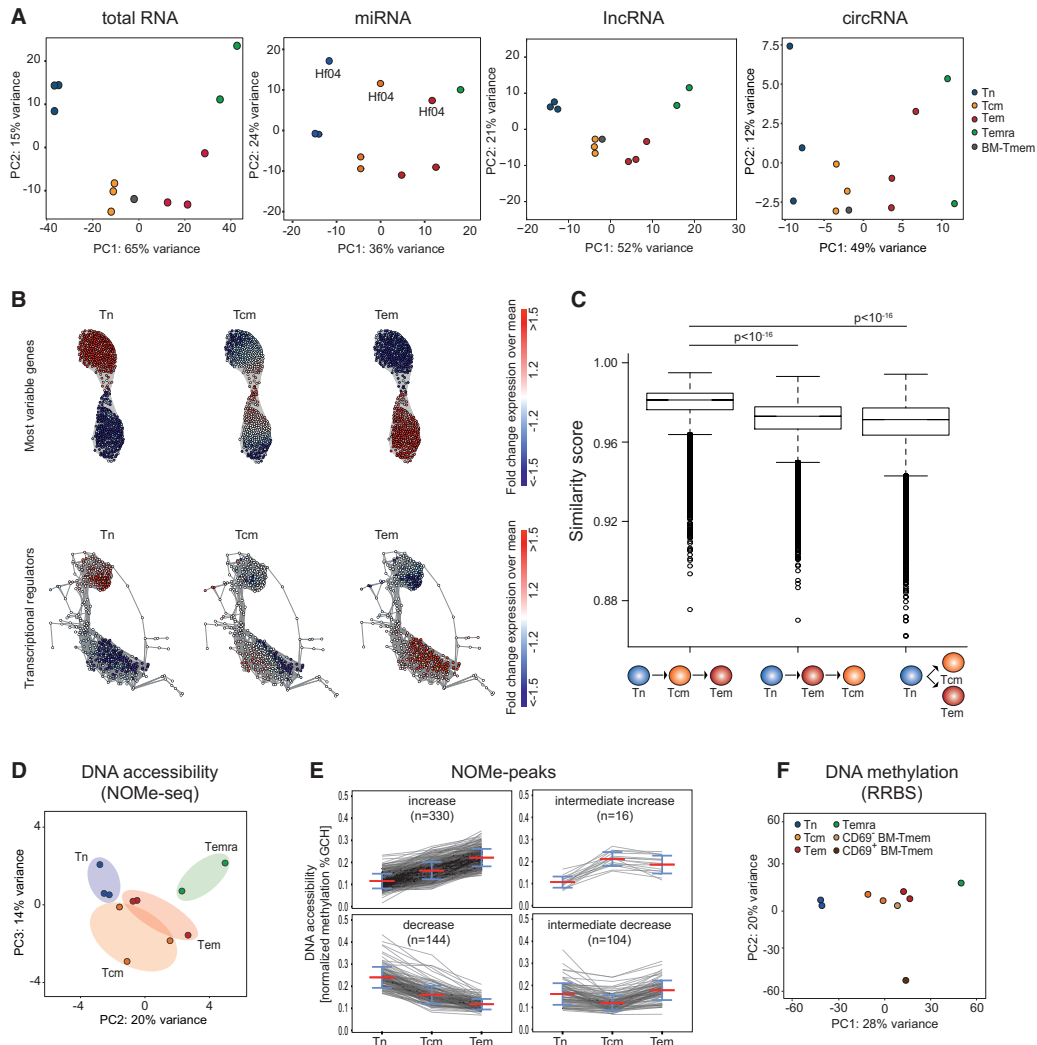
Taken together, our results from epigenomic and transcriptomic analyses support a linear model of differentiation for circulating Tmem cells from the blood with the bone-marrow-resident (CD69<sup>+</sup>) T cell population deviating early and displaying a specific epigenetic imprint (Figure S2F).

### **Changes in DNA-Meth of Transcriptional Control Elements Are Associated with Tmem Cell Differentiation**

DNA-meth can control the expression of genes, which are required for the maintenance of lineage identity, as found for *Foxp3* in regulatory T cells (Huehn et al., 2009). This holds true also for CD4<sup>+</sup> Tmem cells, as we found a correlation between DNA-meth and gene-expression changes, when we used an integrative sparse linear regression model measuring DNA-meth in promoters and gene bodies (Figure S2G). While the highest impact on gene expression was computed for predicted TF binding in accessible chromatin sites (NOME peaks around genes), DNA-meth had a higher regulation potential than miRNAs.

With our genome-wide epigenetic datasets we therefore strived to (1) elucidate to what extent DNA-meth is involved in the regulation of known key Tmem cell checkpoint regulators, and (2) investigate whether epigenomic data could identify novel transcriptional regulators of Tmem differentiation.

For this, we called differentially methylated regions (DMRs) from the WGBS datasets, using the Metilene software (Jühling et al., 2016) applying strict selection criteria (min. #CpGs = 5; min. coverage = 5 reads) and a context-sensitive filtering step to reduce the contribution of the global de-methylation effect observed in PMDs (“adaptive filtering,” Supplemental Experimental Procedures). This approach resulted in 1670 DMRs between Tn, Tcm and Tem cells (Table S5) associated with



**Figure 2. Progressive Changes in the Transcriptomes and in the DNA Accessibility Profiles Support a Linear Differentiation Model for Circulating Human CD4<sup>+</sup> Tmem Subsets**

(A) PCAs of different RNA species for Tn, Tcm, Tem, Temra cells from blood, and Tmem cells from the bone-marrow (BM-Tmem).

(B) Co-regulation network based on the top 700 most variable genes in the dataset (top) or based on transcriptional regulators (TRs, bottom). Nodes represent genes colored according to the corresponding fold-change to mean expression. Links are unweighted and represent significant correlations.

(C) Three possible differentiation models (x axis) were compared using a designed similarity score (y axis), based on the hypothesis, that T cells that are closer to each other in the differentiation order should show more similar gene-expression profiles. The plot shows the distribution of similarity scores obtained (error bars denote SD estimated from 100.000 bootstrap samples). A significantly higher score was obtained for Tn-Tcm-Tem compared to the other models (bootstrapped t test p value).

(D) PCA of DNA accessibility data (based on NOME-seq data).

(E) Visualization of the degree of DNA accessibility (quantile-normalized GCH methylation levels) in consistent nucleosome depleted regions (NOME-peaks) with a statistical difference between at least two cell types. Bars denote mean and SD.

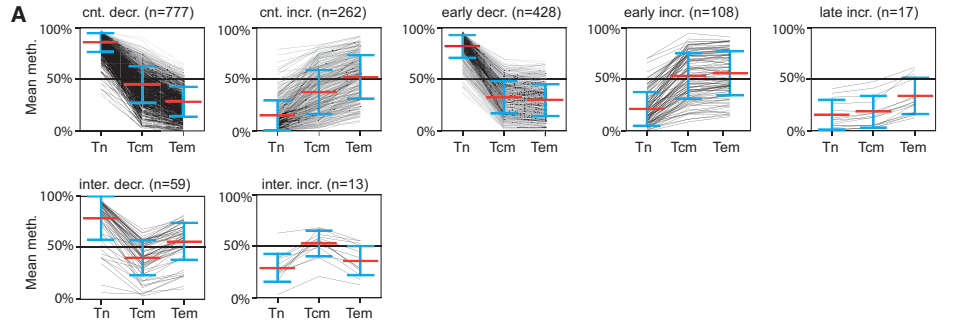
(F) PCA of DNA-meth data (based on RRBS data). CpGs with min. coverage of 5 were considered; only CpGs with calls in all indicated samples were used.

970 protein-coding genes. These DMRs seemed functionally relevant for the regulation of gene expression as most of them were located within or proximal to genes and were classified as promoters or enhancers according to their histone modification profile (Figure S3A). The majority of these DMRs showed a continuous (Tn > Tcm > Tem, 47%) or early (Tn > Tcm and

Tem, 26%) decrease in DNA-meth with Tmem cell generation (Figure 3A).

Next, we analyzed the correlation between DNA-meth changes and gene expression and found that 516 of the DMRs (36%) displayed an inverse correlation to gene expression (Figure S3B). Such DMRs showed the paradigm mode of gene

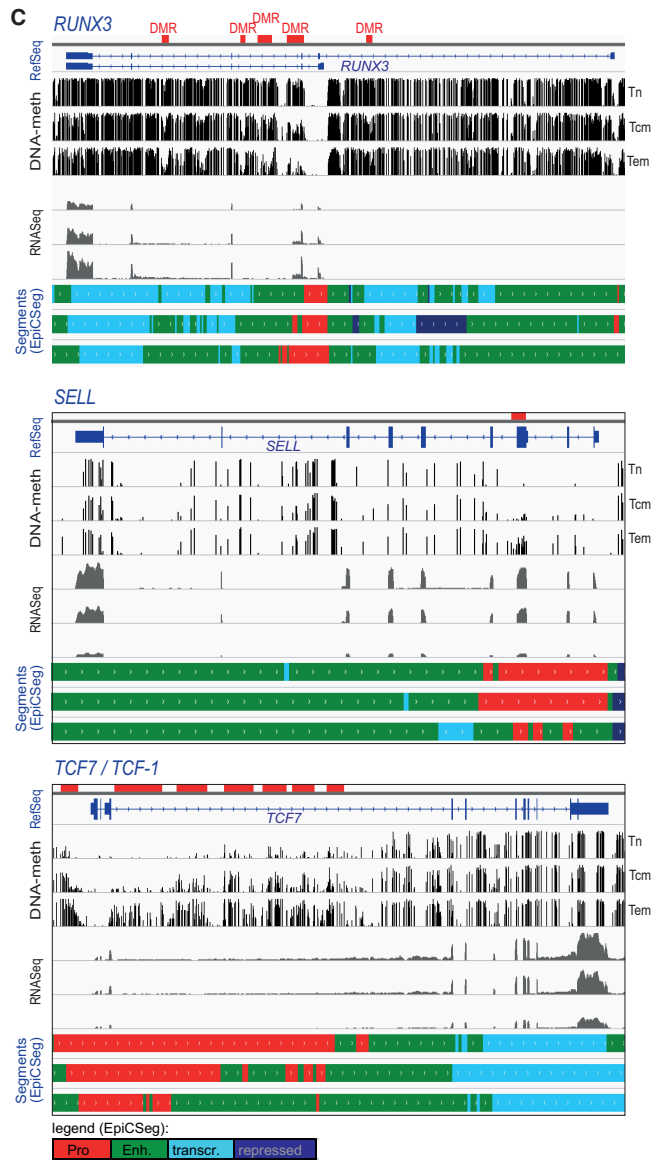




**B**

Gene Symbol	# of DMRs	methylation change (Tn-Tcm-Tem)	expression change (Tn-Tcm-Tem)
ARNT2	1	blue	grey
BACH2	1	red	blue
BATF	3	blue	yellow
BCL2	1	red	yellow
BCOR	7	red	blue
CCR7	1	red	blue
CDKN2A	1	red	blue
DUSP4	1	blue	red
DUSP5	1	blue	red
ELF4	1	red	blue
FASN	1	blue	red
FOXD1	2	blue	red
HNRNP1LL	3	red	blue
IL10RA	1	red	blue
IL2RA	1	red	blue
IL2RB	3	red	blue
LAG3	2	red	blue
LEF1	4	red	blue
MAF	2	red	blue
NCOR2	1	red	yellow
NFATC2	7	blue	red
NOD2	1	red	blue
NOTCH1	1	red	yellow
PDCD1	3	blue	red
PPP3CB	1	blue	yellow
PRDM1	2	red	blue
RBPJ	1	blue	yellow
RPTOR	4	blue	yellow
RUNX3	5	blue	yellow
SELL	1	red	blue
SLAMF1	1	blue	red
SLFN12L	1	blue	yellow
STAM	1	blue	red
TBX21	2	red	blue
TCF4	1	blue	red
TCF7/TCF-1	7	red	blue
TNFRSF1B	3	red	blue
TOX	2	red	blue
ZBTB32	1	blue	yellow
ZEB2	2	red	blue

legend:  
incr. decr. no expr. no change



(legend on next page)

regulation by DNA-meth, in which transcriptional control elements, such as promoter and enhancers, are repressed by increased DNA-meth, whereas a loss of DNA-meth at these elements leads to gene activation. Other DMRs were associated with genes by location, which (1) were not expressed in any cell type, (2) the expression of which did not change, or (3) the expression change correlated to the methylation change (Figure S3B). These classes of DMRs might serve different functions such as (1) preparing a locus for gene expression upon additional environmental signals or locking a locus to prevent alternative cellular fates, (2) stabilizing otherwise transient gene expression, or (3) affecting sites acting as silencers. Furthermore, it cannot be excluded that some DMRs might also act as long-range regulators for distant genes.

These data show that in addition to the large-scale DNA demethylation in PMDs, transcriptional control elements such as promoters and enhancers are targets of epigenetic regulation during Tmem cell differentiation.

### Known Regulators of Tmem Differentiation and Function Display DNA-Meth Changes in Transcriptional Control Regions

To test the assumption that key factors regulating Tmem differentiation and function are under epigenetic control, we extracted a list of 144 known memory-related genes according to recent reviews (Figure S3C) and checked for the occurrence of DMRs in their loci. One quarter of these Tmem cell-related genes displayed one or several DMRs (Figure 3B), 95% of which were associated with a promoter or enhancer histone signature (Table S5). The largest group lost DNA-meth with progressive differentiation, which correlated with an increase in expression (Figure 3B). Among them were genes upregulated upon differentiation from naive to memory states, including surface or intracellular receptors such as *PDCD1* (PD-1), *IL2RA*, and *IL2RB*, *NOD2*, *SLAMF1*, and *TNFRSF1B*, but also many transcriptional regulators such as *RUNX3* (Figure 3C, top), *NFATC2*, *BATF*, *MAF*, *TBX21* (T-BET), the CD45-splicing regulator *HNRNPLL*, *PRDM1* (BLIMP-1), *DUSP4*, *DUSP5*, *STAM*, *TOX*, and *ZEB2*. In a smaller group, increased methylation was linked with decreased expression. This group included the signature markers of Tn and Tcm *SELL* (L-SELECTIN; Figure 3C, middle) and *CCR7*, but also several key transcriptional regulators, namely *TCF7* (encodes TCF-1; Figure 3C, bottom), *LEFT1*, and *BACH2*, which are known to control the development or maintenance of Tmem cells. In a few genes (*FOXO1* and *BACH2*), loss of DNA-meth was associated with a decrease in expression; accordingly, these DMRs might control transcriptional silencers. In other cases (*NOTCH1*, *SLFN12L*, *RPTOR*, *ZBTB32*, *RBPJ*), a change in methylation was not correlated to changes in expression. It is of high interest to investigate whether loss of DNA-meth in genes of this group is not a requirement for expres-

sion but might act by stabilizing transcription, as was found previously for the TSDR (CNS2) enhancer region in the *FOXP3* locus (Huehn et al., 2009; Polansky et al., 2008).

While a causal role of DNA-meth in the regulation of these genes remains to be experimentally demonstrated, these findings provide evidence that epigenetic mechanisms contribute to the developmental regulation of Tmem cells by controlling the expression of key genes.

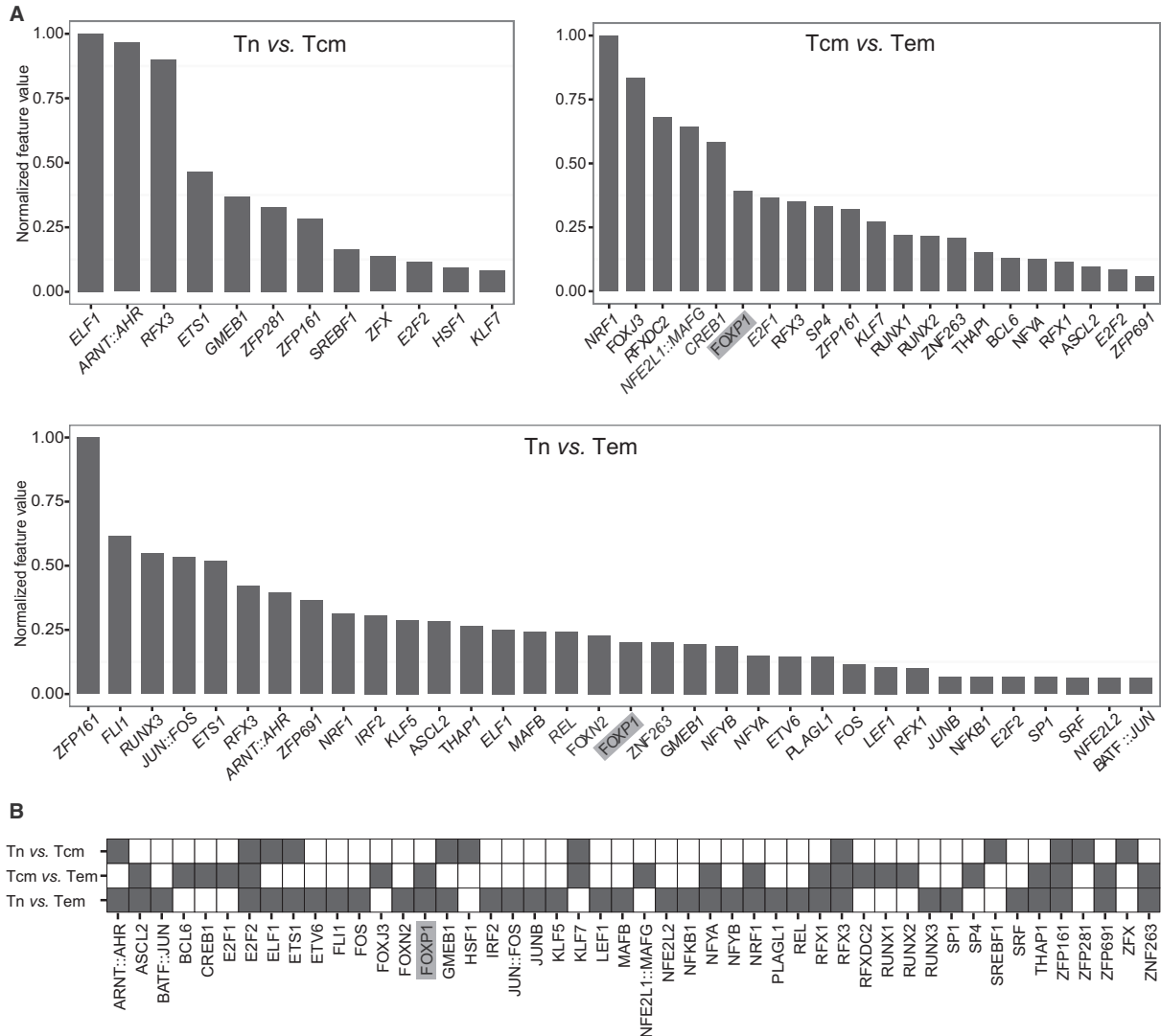
### Integrative Analyses of Epigenomic and Transcriptomic Datasets Facilitate the Identification of Functional Regulator Candidates for Tmem Differentiation

In addition to screening for known developmental regulators, a reciprocal approach can be applied, in which the occurrence of DMRs is used to identify novel candidate genes that might be involved in the control of Tmem differentiation and which undergo direct epigenetic expression control. To this end, 171 DMRs associated with 104 transcriptional regulator genes were identified (Table S5), indicating that these genes might contribute to memory development and maintenance. This list included the gene *FOXP1*.

While these candidates seem to undergo direct expression control by DNA-meth, a different class of regulators might gain or lose functional importance for Tmem cell development and/or function because their binding sites in target genes are being exposed (or blocked) by chromatin remodeling. To identify such “functional epigenetic” regulators, we used an alternative approach combining DNA accessibility data and transcriptomic data. In this, the impact of a given TF on the transcriptional profile is determined by the accessibility of its binding sites within promoters and enhancers in its target genes. We used our DNA accessibility dataset (NOME-seq data) and computed TF binding affinities to open chromatin regions (NOME-peaks). Using a machine learning approach (Schmidt et al., 2016), we modeled differential gene expression between T cell subsets based on these TF binding predictions. In this, the impact of each TF to the differential transcriptional profile is calculated and TFs with a strong influence can be extracted. Indeed, comparison of modeled to observed gene-expression changes, as measured by RNA-seq, displayed a high accuracy (Figure S4A) supporting the validity of the approach. A number of regulatory TF candidates for Tn cells and Tmem subsets could be extracted (Figure 4A and 4B). The lists comprised TFs known to control Tmem cells, such as *BCL6*, *E2F2*, and *RUNX3*, as well as new candidates, including *AHR*, *CREB1*, *ETS1*, *FLI1*, *FOXP1*, *FOXJ3*, *NFEL2*, *NRF1*, *RFX3*, and *ZFP161*. For a number of these (e.g., *AHR*, *FLI1*, *FOXP1*, and *RUNX3*), we also found an associated DMR in their genes and differential expression during Tmem cell differentiation (Table S5), indicating that these factors not only drive transcriptional profiles during Tmem differentiation but are under epigenetic regulation themselves.

### Figure 3. Epigenetic Changes in Known Tmem Cell-Related Genes

(A) Patterns of DNA-meth changes in differentially methylated regions (DMRs). DMRs changing in the order Tn-Tcm-Tem are shown in the upper row. cnt., continuous; inter., intermediate; decr., decrease; incr., increase. Bars denote mean and SD.  
 (B) List of known Tmem cell-related genes (based on Figure S3C) which display at least one DMR in their loci (color legend shown on the bottom).  
 (C) Examples of DMR-containing Tmem regulator genes (*RUNX3*, top; *SELL*, middle; *TCF7/TCF-1*, bottom) showing the location of the identified DMRs (red, top track). The following tracks are shown in each panel (top to bottom, each for Tn, Tcm, Tem): Gene annotation from RefSeq; DNA-meth (WGBS); polyA-RNA; genome segmentation by EpicCSeq (color legend shown on the bottom).



**Figure 4. Selection of Tmem Cell Regulator Candidates Based on Their Predicted Binding Affinities in Open Chromatin Regions of Differentially Expressed Genes**

(A) Bar plots showing normalized feature values (y axis) for each TF (x axis) computed using a machine learning approach (based on logistic regression classifiers) to predict differentially expressed genes in pairwise comparisons of two cellular subtypes. Differences in predicted TF affinities, calculated from open chromatin regions (NOMe-peaks) in the vicinity of a gene, were used as features in the classification. Large feature values denote a higher impact of the TF on differential gene expression.

(B) Summarized representation of all selected TFs shown in (A). Filled boxes reflect that a TF (column) has been selected as a feature in the respective comparison (row). TFs joint by double colons indicate that both TFs are predicted to bind as a complex. The TF FOXP1 is highlighted in gray.

With these analyses we identified several promising new TFs from epigenomic data which are likely to be involved in Tmem cell generation and function.

### FOXP1 Is an Epigenetically Controlled “Naive-Keeping” Checkpoint Regulator

We found the TF FOXP1 to be a particularly interesting candidate for Tmem cell regulation due to several reasons: First, a DMR in the FOXP1 locus displayed increasing DNA-meth, concomitant

with decreased mRNA levels from Tn to Tem (Table S5); second, FOXP1 was predicted to bind to accessible chromatin regions and thus contribute to differential gene expression in Tn versus Tem (Figure 4); and third, FOXP1 was among the top predicted Tn cell-specific regulators according to an iRegulon (Janky et al., 2014) analysis (Figure S4B), which is based on the enrichment of TF binding sites in genes contributing to the cell-type-specific clusters shown in Figure 2B. Therefore, we selected the TF FOXP1 for a more detailed investigation.

Our data suggested that FOXP1 might act as an important regulator for the Tn-to-Tmem transition. Confirming this, we found that T cell-specific depletion of Foxp1 protein expression in Foxp1 conditional-deficient mice resulted in loss of the naive CD44<sup>low</sup> phenotype in T cells (Figure 5A). These findings together with published data (Feng et al., 2011; Wei et al., 2016), support the view that Foxp1 acts as a “naive-keeping” factor for T cells. Analyses of DNA-meth in our datasets revealed a DMR in the *FOXP1* locus, which displayed a strong progressive gain of methylation with differentiation (Tn < Tcm < Tem), which was classified as a selective active promoter in Tn cells (Figure 5B) based on the displayed histone modification patterns (by EpiCSeq, Mammana and Chung, 2015). Indeed, a methylation-sensitive promoter activity of the *FOXP1-DMR* was confirmed in luciferase reporter gene assays in primary human CD4<sup>+</sup> T cells, as the *FOXP1-DMR* was able to drive luciferase expression when cloned upstream of the reporter gene in the sense orientation, but not when the orientation of the *FOXP1-DMR* was inverted or when the *FOXP1-DMR* was methylated (Figure 5C).

Consistent with the occurrence of a Tn cell-specific promoter, we found the *FOXP1* protein expression in human CD4<sup>+</sup> T cells to be highest in Tn cells and to be decreased in Tcm, Tem, and Temra cells (Figure 5D). In addition, we found indications in our RNA-seq datasets for three alternative shorter RNA isoforms, which started within or directly downstream of the *FOXP1-DMR* promoter (Figure S4C). All three isoforms showed preferential expression in Tn compared to Tcm and Tem cells as measured by qPCR (Figure 5E). In addition, two of them contain the complete protein coding sequence (Figure 5E), which we verified by single molecule real-time sequencing (data not shown).

Taken together, these results validate FOXP1 as an important gate-keeper for the naive-to-memory transition, which was identified by integrative analyses of epigenomic data. In addition, these analyses also enabled the identification of the epigenetic control mechanisms regulating differential *FOXP1* expression during Tmem cell generation.

## DISCUSSION

This study reveals that the differentiation of Tn cells into distinct types of memory cells and their long-term maintenance is connected to major epigenetic and transcriptional reprogramming. This is manifested on a global scale with a genome-wide segmented loss of DNA-meth during differentiation and in gene-specific epigenetic changes, which control the stage-specific expression and/or function of transcriptional regulators.

As our first major finding, we documented a progressive genome-wide loss of DNA-meth upon transition from the naive to the memory stages. This de-methylation was most prominent in “partially methylated domains” (PMDs, Hon et al., 2012; Lister et al., 2009), a feature shared in memory differentiation of B cells, but absent during the differentiation of monocytes into macrophages (Wallner et al., 2016). PMDs have been associated with heterochromatic histone signatures and correlate to regions, which are replicated late during S phase and progressively lose methylation during strong proliferation (Aran et al., 2011). Consistent with this, T and B cells, but not monocytes, undergo extensive proliferation during differentiation as a result of TCR-

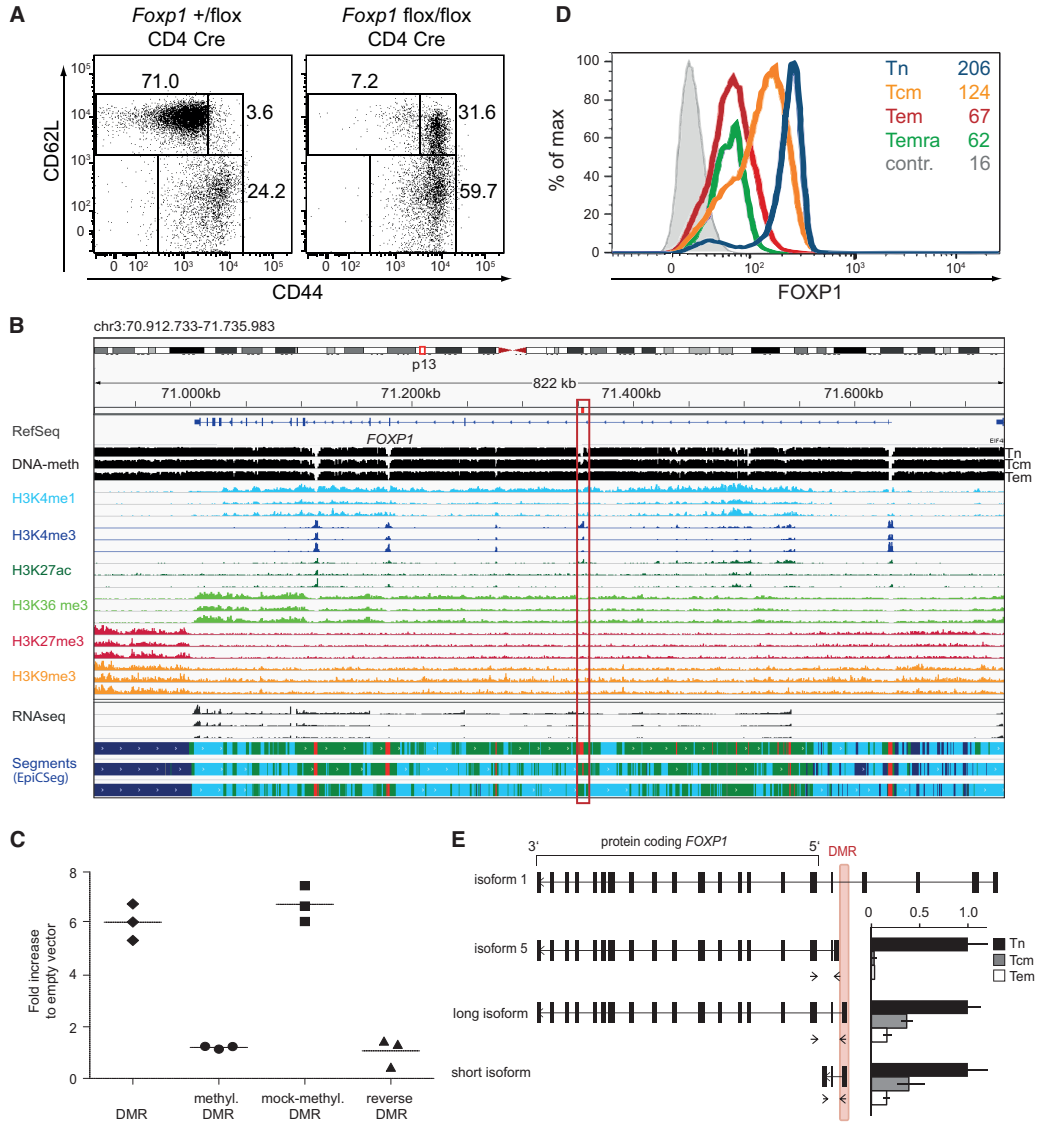
(BCR-) mediated activation. It is therefore feasible that the observed PMD-associated loss of methylation is a consequence and a signature of highly proliferative episodes in the history of these cells.

This interpretation is supported for CD4<sup>+</sup> Tmem cells by two additional observations in our study: (1) the progressive shortening of telomere length in the order of Tn-Tcm-Tem (Figure S5A), and (2) the progressive loss of methylation in PMDs observed in short-term culture of Tn cells proliferating in vitro after TCR-mediated activation (Figures S5B and S5C). It remains to be investigated whether the global de-methylation is just a tolerated bystander effect of proliferation or whether it constitutes a telomere-independent senescence signal for the cell as suggested by studies on hematopoietic stem cells under proliferative stress (Beerman et al., 2013).

These findings are relevant for the interpretation and functional assignment of DNA-meth changes found by gene-specific DNA-meth assays. Using the epigenomic maps of this study as a reference, gene-specific differentially methylated regions (DMRs), which might have (direct) functional relevance for gene expression, can be discriminated from DMRs in regions that are likely to represent a mere imprint from the proliferation history. This discrimination could be of particular relevance when studying epigenetic changes in T cells isolated from chronic stimulatory conditions, e.g., inflammatory diseases such as rheumatoid arthritis or lupus erythematosus where disease-associated methylation differences have been reported (de Andres et al., 2015; Javierre et al., 2010).

The second major conclusion from our global epigenomic analyses sheds light onto the still-controversial subject of the mode of memory differentiation of human CD4<sup>+</sup> T cells: Do memory cells originate (1) early after antigen encounter independently of (but in parallel to) the massive expansion of short-lived effector cells (parallel or bifurcative differentiation model, Arsenio et al., 2015) or (2) do they develop from effector cells, which adopt Tmem stages toward the end of the primary effector phase (linear model)? While we could not directly address the positioning of effector cells in relation to memory development with the present dataset, all our findings are more consistent with the linear progression model for circulating Tmem cells: We observed a strong loss of DNA-meth in PMDs upon transition from Tn to Tcm cells, which was further reduced in the more differentiated Tem and Temra phenotypes. These data indicate that Tcm cells would have already passed through a phase of intense proliferation during initial activation and prior to converting into resting memory cells. While the parallel model cannot be formally excluded by this, additional proliferation-independent datasets were similarly more consistent with the linear model, including: (1) patterns of DNA-meth at single-DMR resolution, as well as patterns of DNA accessibility (NOME-peaks), showed almost exclusively changes in the order of Tn-Tcm-Tem, (2) changes in the transcriptomes grouped the samples along a progressive Tn-Tcm-Tem-Temra cell differentiation axis, (3) network analysis of co-expressed genes placed the Tcm phenotype as intermediate to the Tn- and Tem-associated clusters, and (4) calculation of the similarity of the transcriptomic profiles revealed the linear Tn-Tcm-Tem model as the most likely one.

These conclusions are in part in contrast to conclusions from restricted expression analyses of murine CD8<sup>+</sup> single cells



**Figure 5. A Newly Identified Methylation-Sensitive Promoter Drives Alternative Coding mRNA Isoforms of the “Naive-Keeping” Transcription Factor FOXP1 in Tn**

(A) CD4<sup>+</sup> T cells isolated from spleens of *Foxp1*<sup>+/-lox</sup> CD4 Cre and *Foxp1*<sup>flox/flox</sup> CD4 Cre mice at the age of 6–10 weeks were analyzed by flow cytometry for a CD44<sup>high</sup> memory phenotype. Dot plots show CD62L and CD44 surface expression after gating on CD4<sup>+</sup> living cells.

(B) Genomic view of the human *FOXP1* locus indicating a distinct *FOXP1*-DMR (red box) gaining DNA-meth from Tn to Tcm to Tem cells. The following tracks are shown (each for Tn, Tcm, and Tem cells): RefSeq annotation, DNA-meth (WGBS), 6 histone modifications, total RNA coverage and segmentation by EpiCseg (red, promoter; green, enhancer; light blue, transcribed; dark blue, repressed).

(C) Luciferase reporter assay testing a methylation-sensitive and orientation-dependent promoter activity of the *FOXP1*-DMR in primary human CD4<sup>+</sup> T cells. The Firefly luciferase signal was normalized to the signal of the Renilla transfection control and is shown relative to the empty vector control. One representative experiment performed in triplicates out of two independent experiments is shown.

(D) FOXP1 protein expression in gated human Tn, Tcm, Tem, and Temra cells as assessed by intracellular staining and flow cytometry. Numbers: geometric mean of the FOXP1 signal. Control staining (contr.) was done using the fluorescently labeled secondary antibody only.

(E) Schematic depiction of different human *FOXP1* RNA isoforms. The position of the *FOXP1*-DMR is shown and the protein coding exons are indicated as boxes. Three isoforms start within or shortly downstream of the *FOXP1*-DMR. Their relative expression values are shown for the different cell types (normalized to Tn cells) measured by qPCR using the indicated primers (arrowheads). One representative experiment performed in technical triplicates (mean and SD) out of two independent experiments is shown.

(Arsenio et al., 2015). However, elegant *in vivo* approaches using adoptive transfer systems of single murine CD8<sup>+</sup> memory cells (Gerlach et al., 2013a; Graef et al., 2014) also argue in favor of a linear differentiation model. As similar experiments have not been conducted for CD4<sup>+</sup> cells yet and are impossible in the human system, our data represent new insights into this topic.

The distinct positioning of Temra cells in the analysis of genome-wide DNA-meth and transcriptomic data suggests that they represent a late stage of Tmem differentiation and have undergone extensive proliferation. Alternatively, circulating Temra cells might represent survivors of prolonged effector proliferation due to chronic re-activation. This could also explain the enhanced heterogeneity of the Temra samples, since their epigenetic imprint might have been specialized over time. Indeed, increased frequencies of Temra cells have e.g., been reported in response to persistent CMV infection (Derhovanessian et al., 2011) and in liver disease, where high Temra cell numbers represent a significant risk of organ rejection after organ transplantation (Gerlach et al., 2013b). In contrast, Tmem cells isolated from the bone marrow did not display a Temra-like epigenetic phenotype but were positioned between Tcm and Tem cells, indicating that they have preserved significant expansion and differentiation capacity. In addition, the CD69<sup>+</sup> tissue-resident subset of BM-Tmem cells displayed a distinct DNA-meth profile, indicating that acquisition of a resident phenotype, too, is linked to significant epigenetic reprogramming.

The third pillar of our study is dedicated to the identification of factors, which drive and/or maintain Tmem cells. In this endeavor, we identified many non-coding RNAs (ncRNAs), which are highly and/or differentially expressed in Tn and Tmem cells. Among them, we identified numerous circRNAs for which the normal linear host transcript is barely detectable (Rybak-Wolf et al., 2015), thus, activity of these genes would have been missed in standard polyA- selected RNA-seq and/or normal linear splice analysis. These and other ncRNA molecules (lncRNAs, miRNAs) may not only reflect but also induce functional consequences during Tmem differentiation. The expression of the ncRNAs appears to be coordinated and finely tuned during Tmem differentiation with a remarkable link to other epigenomic changes. Therefore, our dataset provides a deep basis to further investigate the direct contribution of ncRNAs to Tmem differentiation and to clarify the mutual regulatory impact between ncRNAs and chromatin structure.

In addition to RNAs, we report two classes of protein regulators, which include known and potentially new factors controlling Tmem cell generation and function: (1) TFs, which undergo epigenetic expression control during Tmem cell formation, and (2) TFs, which gain or lose functional importance as their binding sites in target genes are being exposed or closed, respectively, independently of their own expression change. For the first class, we found several widely discussed regulators of Tmem differentiation, which displayed differential DNA-meth in promoter or enhancer regions that anti-correlated with differences in gene expression levels, following the classical paradigm of methylation-controlled gene repression. For several of them (e.g., *IL2RA*, *RUNX3*, *NFATC2*, *MAF*, *BACH2*, *FOXO1*), epigenetic control in Tmem differentiation has not been reported so far and awaits experimental confirmation. Interestingly, among differentially methylated genes was also *HNRNPLL*, involved in

alternative splicing of CD45 to the Tmem signature isoform CD45RO, and the two homing-related receptors, *SELL* (L-selectin) and *CCR7*, suggesting that the permanent change in the recirculation pattern with transition from Tcm to Tem is epigenetically fixed. For others, concordant epigenetic changes have already been described in murine CD8<sup>+</sup> T cells (e.g., *BATF*, *LEF1*, *PDCD1*, *TBX21*, *TCF7*, *ZBTB32*; Scharer et al., 2013; Youngblood et al., 2011; Hashimoto et al., 2013).

As for the second class of TFs, we identified FOXP1 as one of the top candidates, a less well known but functionally confirmed Tmem regulator in mice (Feng et al., 2011; Wei et al., 2016). Our present analyses support a similar function in human CD4<sup>+</sup> Tmem and additionally unravel the epigenetic control of the *FOXP1* gene. Other prime TF candidates include TFs previously implicated in Tmem regulation such as RUNX3, E2F2, LEF1, BCL6, or members of the ELF-, KLF-, or FOXJ- families, as well as CREB1, ETS-1, and JUN-FOS, which are known to be involved in multiple cellular processes of differentiation and activation. Additional interesting candidates include (1) the aryl hydrocarbon receptor, AHR, which has been implicated in differentiation of CD4<sup>+</sup> T cells into pro- or anti-inflammatory subsets and, hence, to modulate autoimmune diseases in various animal models (reviewed in Esser et al., 2009; Hanieh, 2014) and (2) the ets-family member FLI-1, which has been reported to affect thymic T cell development, TCR signaling, glycosphingolipid metabolism, and cytokine expression and has been implicated in autoimmune diseases, too (Richard et al., 2013; Sato et al., 2014). Others of the top-predicted TFs have not been directly associated to regulation of Tmem differentiation yet, but are players in potentially relevant cellular processes, such as intracellular signaling (RFX3, via the RAS-MAPK pathway), metabolic processes (NRF1 and SREBF1, associated with mTORC1 signaling), and chromatin remodeling (ZFP161, targeting of the repressive Polycomp complex). Thus, important Tmem cell properties might be under the control of yet neglected transcriptional regulators that could be revealed by the integrated analysis of transcriptomic and epigenomic features.

In conclusion, the comprehensive epigenomic analysis of several human CD4<sup>+</sup> Tmem subsets in this study revealed insights into the Tmem differentiation pathway and allowed the identification of relevant epigenetically controlled transcriptional regulators. In addition, these data constitute a resource of normal T cell differentiation, which can serve as a reference for the identification of altered epigenetic signatures in T cells from pathological situations such as chronic inflammatory disease. The challenging task for the future will be the application of “epigenetic engineering” to achieve therapeutic re-programming of pathogenic T cells or to optimize T cells for their use in cellular therapy.

## EXPERIMENTAL PROCEDURES

### T Cell Isolation

PBMCs from blood of healthy female donors or from bone marrow samples of female donors undergoing hip replacements were isolated and enriched for CD4<sup>+</sup> T cells using the MACS-technology (Miltenyi Biotech). Tn cells and Tmem subsets were purified by flow cytometry using markers shown in Figure S1A. Donors gave their written and informed consent prior to participating in the study (Ethics committee of the Charite Universitaetsmedizin Berlin, application numbers EA1/116/13 and EA1/105/09).

### Epigenomic Data Generation

WGBS was carried out by the combined analysis of two bisulfite-converted libraries using the pre-bisulfite library protocol (Ulrich et al., 2015) and the TruSeq DNA Methylation kit (Illumina, San Diego, USA). RRBS libraries were prepared as previously published (Boyle et al., 2012). For NOME-seq, nuclei of fixed cells were extracted and DNA-meth on GpC motifs in accessible chromatin regions was introduced using the M. CviPI methyltransferase, followed by WGBS analysis. ChIP-seq for histone modifications was carried out as previously described (Arrighoni et al., 2016; Kinkley et al., 2016). RNA was extracted using the miRNeasy Micro Kit (QIAGEN) and three Illumina sequencing libraries were prepared (small RNA sequencing library, one stranded total RNA, and one stranded mRNA library). Sequencing was carried out on HiSeq 2000 and HiSeq2500 machines (Illumina). Bioinformatical processing of the sequencing reads including mapping to the hg19 reference genome is outlined in the Supplemental Experimental Procedures section.

### DNA Methylation Analyses

Genome segmentation based on WGBS data was performed using MethylSeekR (Burger et al., 2013). The methylation levels from both strands were aggregated and weighted average methylation levels were plotted. WGBS data from B cells (Blueprint consortium) was converted to hg19 coordinates using the liftOver tool (Rosenbloom et al., 2015) and segmentation was carried out. Differentially methylated regions (DMRs) were predicted with Metilene (Jühling et al., 2016) in de-novo mode among sites with at least 5x coverage.

### Calling of Accessible Chromatin Region

Nucleosome-depleted regions (NOME-peaks) were identified by segmenting the GCH-methylation signal with a binomial hidden Markov model with two states (1 open/NDR, 0 background) in each sample separately and consistent NOME-peaks confirmed in all three replicates were selected.

### Identification of mRNAs, miRNAs, lncRNAs, and circRNAs

Expression values for total RNA were quantified using TopHat, Htseq-count, and DESeq2 (Anders and Huber, 2010). Cufflinks (Trapnell et al., 2010) was used for the identification of novel lncRNAs. To remove possible coding genes, we estimated the coding potential of novel transcripts using PhyloCSF (Lin et al., 2011) and CPAT (Wang et al., 2013). Mature miRNA read counts were estimated for each sample using miRDeep2 (Friedländer et al., 2012) and miRBase (version 21) annotations. CircRNAs were detected, filtered, and annotated as described before (Memczak et al., 2013).

### Co-Expression Network Construction

Expression data of Tn, Tcm, and Tem cells (3 replicates each) was filtered using either a list of human transcriptional regulators (TRs) or the 700 most variable genes (i.e., most significant p values in an ANOVA-based analysis) to get a reduced expression table of present genes. The group of TRs contained transcription factors (TFs), co-factors, RNA-binding proteins and chromatin remodelers originating from the TFcat data base (Fulton et al., 2009). The expression matrices were loaded into BioLayout Express3D (Theocharidis et al., 2009) and co-regulation networks were generated with a Pearson correlation cutoff of 0.9. The predicted gene-gene pairs were visualized by Cytoscape (Shannon et al., 2003) and fold change expression values calculated against the group mean were mapped to the network.

### Prediction of Transcriptional Regulators Using a Machine-Learning Approach to Model Differential Gene Expression

We used a machine-learning approach based on a logistic regression classifier with the elastic net penalty (Zou and Hastie, 2005) to model differential gene expression between the Tn, Tcm, and Tem subsets. Because the TF features for the logistic regression classifier, we used the ratio of TF gene scores, which were computed using TEPIC (Schmidt et al., 2016). TFs predicted to contribute to differential gene expression were selected.

### Functional Analyses on the TF FOXP1 and the FOXP1-DMR

For the generation of a T cell-specific Foxp1 deletion, a conditional Foxp1 knock-out allele was generated using standard gene targeting techniques in murine ESCs by introducing loxP sites into intronic regions flanking exons 10–12

(T. Patzelt, O. Gorka, and J. Ruland, manuscript in preparation). The generated Foxp1-floxed mice were crossed to CD4-Cre animals (Lee et al., 2001).

Intracellular staining of FOXP1 protein in human CD4<sup>+</sup> T cells was performed using the Fixation/Permeabilization Buffer set for intracellular Foxp3 staining (eBioscience) in a two-step staining procedure (primary FOXP1 antibody polyclonal #2005, Cell Signaling Technology, DyeLight-649-labeled donkey anti-rabbit secondary antibody #406406, BioLegend). Samples were acquired on a BD LSRFortessa instrument (BDBioscience).

The FOXP1-DMR was cloned into the CpG-free Firefly luciferase vector pCpGL (Klug and Rehli, 2006). Treatment with the M.SssI CpG methyltransferase (NEB) allowed selective methylation of the FOXP1-DMR. Ex vivo isolated CD4<sup>+</sup> T cells were TCR-stimulated for 48 hr and transfected with the FOXP1-DMR Firefly vector and a pRL-TK Renilla control vector (Promega) using the Neon<sup>TM</sup> Transfection System (Life Technologies). Firefly and Renilla luciferase activity were assessed using the Dual Luciferase Assay Kit (Promega) after 24 hr. The Firefly luciferase signal was normalized to the Renilla reporter signal.

Expression levels of the FOXP1 RNA isoforms were quantified using the platinum SYBR green qPCR superMix-UDG (Thermo Fisher Scientific) on a Step One instrument (Thermo Fisher Scientific). Relative transcript levels were normalized to hRPS18. Primer sequences are given in the Supplemental Experimental Procedures section.

### ACCESSION NUMBERS

All sequencing data have been deposited at the European Genome-Phenome Archive under the accession number EGAS00001001624. WGBS Blueprint data of B cells are available from the EGA under the accessions EGAD00001001590, EGAD00001001587, EGAD00001001548, and EGAD00001001160.

### SUPPLEMENTAL INFORMATION

Supplemental Information includes five figures, five tables, and Supplemental Experimental Procedures and can be found with this article online at <http://dx.doi.org/10.1016/j.immuni.2016.10.022>.

### AUTHOR CONTRIBUTIONS

Sample preparation: S. Frischbutter, S.S., U.H., C.C., Y.T., S.R., G.W., U.K., U.S., J.D., H.-D.C., B.S., A.H., J.K.P.; WGBS and NOME-seq: G.G., N.G., J.W.; ChIP-seq: S.K., N.L., H.-R.C., L.A., A.S.R., T.M.; RNA-seq: S. Fröhler, W.C.; mapping & management of sequencing data: B.F., G.Z., K.N., P.E., C.D.I., B.H., B.B., J.E.; Data analysis: P.D., K.N., A. Salhab, F.M., J.W., T.L. (DNA-meth); P.D., J.P.K. (integrative DMR analysis), K.N., G.G. (NOME-seq); X.Y., H.-R.C., P.E., T.L. (ChIP-seq); J.X., W.C. (lncRNAs); P.G., N.R. (circRNAs); F.K., N.R. (miRNAs); A. Sinha, P.R. (total RNAs); L.F., L.S., B.B. (telomere length); T.U., K.B., J.L.S. (co-expression network); F.S., A.D.A., F.B.A., M.H.S. (gene expression prediction and modeling of TF binding, similarity score); O.G., J.R. (Foxp1-KO mice); C.K., MdA, A.H., J.K.P. (functional FOXP1 analyses); C.K., MdA, A.R., J.D., B.S., A.H., J.K.P. (functional data interpretation). J.K.P., J.W., A.H. designed and coordinated the study supported by N.G.; J.K.P., J.W., A.H. wrote the manuscript with contributions from other authors.

### ACKNOWLEDGMENTS

We thank René Maier for expert technical assistance, the FCCF of the DRFZ for expert cell sorting, staff from EURICE for project management, and Ulrike Biskup for help with figure layout. Selected data generated by the Blueprint Consortium ([www.blueprint-epigenome.eu](http://www.blueprint-epigenome.eu); EU FP7/2007-2013, grant agreement no. 282510 BLUEPRINT) were used. Funding is as follows: German Epigenome Programme (DEEP) of the Federal Ministry of Education and Research in Germany (BMBF), DFG HA 1505/10-1 to A.H., DFG-SFB650-TP1 to A.H., ERC grant 322865 to J.R., EU FP7 “ONE Study” to B.S., DFG SFB650 to B.S., DFG SFB704 to J.L.S., J.L.S. is a member of the Excellence Cluster Immunosenescence, ERC-2010-AdG\_20100317 Grant 268987 to A.R. and Priority Programme 1468 Immunobone to A.R.

Received: December 30, 2015  
 Revised: June 22, 2016  
 Accepted: July 22, 2016  
 Published: November 15, 2016

## REFERENCES

- Ahmed, R., Bevan, M.J., Reiner, S.L., and Fearon, D.T. (2009). The precursors of memory: models and controversies. *Nat. Rev. Immunol.* **9**, 662–668.
- Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106.
- Aran, D., Toperoff, G., Rosenberg, M., and Hellman, A. (2011). Replication timing-related and gene body-specific methylation of active human genes. *Hum. Mol. Genet.* **20**, 670–680.
- Arrigoni, L., Richter, A.S., Betancourt, E., Bruder, K., Diehl, S., Manke, T., and Bönisch, U. (2016). Standardizing chromatin research: a simple and universal method for ChIP-seq. *Nucleic Acids Res.* **44**, e67.
- Arsenio, J., Metz, P.J., and Chang, J.T. (2015). Asymmetric Cell Division in T Lymphocyte Fate Diversification. *Trends Immunol.* **36**, 670–683.
- Beerman, I., Bock, C., Garrison, B.S., Smith, Z.D., Gu, H., Meissner, A., and Rossi, D.J. (2013). Proliferation-dependent alterations of the DNA methylation landscape underlie hematopoietic stem cell aging. *Cell Stem Cell* **12**, 413–425.
- Boyle, P., Clement, K., Gu, H., Smith, Z.D., Ziller, M., Fostel, J.L., Holmes, L., Meldrim, J., Kelley, F., Gnirke, A., and Meissner, A. (2012). Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation profiling. *Genome Biol.* **13**, R92.
- Burger, L., Gaidatzis, D., Schübeler, D., and Stadler, M.B. (2013). Identification of active regulatory regions from DNA methylation data. *Nucleic Acids Res.* **41**, e155.
- Carbone, F.R., Mackay, L.K., Heath, W.R., and Gebhardt, T. (2013). Distinct resident and recirculating memory T cell subsets in non-lymphoid tissues. *Curr. Opin. Immunol.* **25**, 329–333.
- Crompton, J.G., Narayanan, M., Cuddapah, S., Roychoudhuri, R., Ji, Y., Yang, W., Patel, S.J., Sukumar, M., Palmer, D.C., Peng, W., et al. (2016). Lineage relationship of CD8(+) T cell subsets is revealed by progressive changes in the epigenetic landscape. *Cell. Mol. Immunol.* **13**, 502–513.
- de Andres, M.C., Perez-Pampin, E., Calaza, M., Santaclara, F.J., Ortea, I., Gomez-Reino, J.J., and Gonzalez, A. (2015). Assessment of global DNA methylation in peripheral blood cell subpopulations of early rheumatoid arthritis before and after methotrexate. *Arthritis Res. Ther.* **17**, 233.
- Derhovanessian, E., Maier, A.B., Hähnel, K., Beck, R., de Craen, A.J., Slagboom, E.P., Westendorp, R.G., and Pawelec, G. (2011). Infection with cytomegalovirus but not herpes simplex virus induces the accumulation of late-differentiated CD4+ and CD8+ T-cells in humans. *J. Gen. Virol.* **92**, 2746–2756.
- Esser, C., Rannug, A., and Stockinger, B. (2009). The aryl hydrocarbon receptor in immunity. *Trends Immunol.* **30**, 447–454.
- Feng, X., Wang, H., Takata, H., Day, T.J., Willen, J., and Hu, H. (2011). Transcription factor Foxp1 exerts essential cell-intrinsic regulation of the quiescence of naive T cells. *Nat. Immunol.* **12**, 544–550.
- Flossdorf, M., Rössler, J., Buchholz, V.R., Busch, D.H., and Höfer, T. (2015). CD8(+) T cell diversification by asymmetric cell division. *Nat. Immunol.* **16**, 891–893.
- Friedländer, M.R., Mackowiak, S.D., Li, N., Chen, W., and Rajewsky, N. (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.* **40**, 37–52.
- Fulton, D.L., Sundararajan, S., Badis, G., Hughes, T.R., Wasserman, W.W., Roach, J.C., and Sladec, R. (2009). TFCat: the curated catalog of mouse and human transcription factors. *Genome Biol.* **10**, R29.
- Gerlach, C., Rohr, J.C., Perié, L., van Rooij, N., van Heijst, J.W., Velds, A., Urbanus, J., Naik, S.H., Jacobs, H., Beltman, J.B., et al. (2013a). Heterogeneous differentiation patterns of individual CD8+ T cells. *Science* **340**, 635–639.
- Gerlach, U.A., Vogt, K., Schlickeiser, S., Meisel, C., Streitz, M., Kunkel, D., Appelt, C., Ahrlich, S., Lachmann, N., Neuhaus, P., et al. (2013b). Elevation of CD4+ differentiated memory T cells is associated with acute cellular and antibody-mediated rejection after liver transplantation. *Transplantation* **95**, 1512–1520.
- Graef, P., Buchholz, V.R., Stemberger, C., Flossdorf, M., Henkel, L., Schiemann, M., Drexler, I., Höfer, T., Riddell, S.R., and Busch, D.H. (2014). Serial transfer of single-cell-derived immunocompetence reveals stemness of CD8(+) central memory T cells. *Immunity* **41**, 116–126.
- Hanieh, H. (2014). Toward understanding the role of aryl hydrocarbon receptor in the immune system: current progress and future trends. *BioMed Res. Int.* **2014**, 520763.
- Harrington, L.E., Janowski, K.M., Oliver, J.R., Zajac, A.J., and Weaver, C.T. (2008). Memory CD4 T cells emerge from effector T-cell progenitors. *Nature* **452**, 356–360.
- Hashimoto, S., Ogoshi, K., Sasaki, A., Abe, J., Qu, W., Nakatani, Y., Ahsan, B., Oshima, K., Shand, F.H., Ametani, A., et al. (2013). Coordinated changes in DNA methylation in antigen-specific memory CD4 T cells. *J. Immunol.* **190**, 4076–4091.
- Henson, S.M., Riddell, N.E., and Akbar, A.N. (2012). Properties of end-stage human T cells defined by CD45RA re-expression. *Curr. Opin. Immunol.* **24**, 476–481.
- Hon, G.C., Hawkins, R.D., Caballero, O.L., Lo, C., Lister, R., Pelizzola, M., Valsesia, A., Ye, Z., Kuan, S., Edsall, L.E., et al. (2012). Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Res.* **22**, 246–258.
- Huehn, J., Polansky, J.K., and Hamann, A. (2009). Epigenetic control of FOXP3 expression: the key to a stable regulatory T-cell lineage? *Nat. Rev. Immunol.* **9**, 83–89.
- Janky, R., Verfaillie, A., Imrichová, H., Van de Sande, B., Standaert, L., Christiaens, V., Hulselmans, G., Hertel, K., Naval Sanchez, M., Potier, D., et al. (2014). iRegulon: from a gene list to a gene regulatory network using large motif and track collections. *PLoS Comput. Biol.* **10**, e1003731.
- Javierre, B.M., Fernandez, A.F., Richter, J., Al-Shahrour, F., Martin-Subero, J.I., Rodríguez-Ubrea, J., Berdasco, M., Fraga, M.F., O'Hanlon, T.P., Rider, L.G., et al. (2010). Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. *Genome Res.* **20**, 170–179.
- Jühling, F., Kretzmer, H., Bernhart, S.H., Otto, C., Stadler, P.F., and Hoffmann, S. (2016). metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome Res.* **26**, 256–262.
- Kaech, S.M., and Cui, W. (2012). Transcriptional control of effector and memory CD8+ T cell differentiation. *Nat. Rev. Immunol.* **12**, 749–761.
- Kinkley, S., Helmuth, J., Polansky, J.K., Dunkel, I., Gasparoni, G., Fröhler, S., Chen, W., Walter, J., Hamann, A., and Chung, H.R. (2016). reChIP-seq reveals widespread bivalency of H3K4me3 and H3K27me3 in CD4(+) memory T cells. *Nat. Commun.* **7**, 12514.
- Klug, M., and Rehli, M. (2006). Functional analysis of promoter CpG methylation using a CpG-free luciferase reporter vector. *Epigenetics* **1**, 127–130.
- Komori, H.K., Hart, T., LaMere, S.A., Chew, P.V., and Salomon, D.R. (2015). Defining CD4 T cell memory by the epigenetic landscape of CpG DNA methylation. *J. Immunol.* **194**, 1565–1579.
- Kulis, M., Merkl, A., Heath, S., Queirós, A.C., Schuyler, R.P., Castellano, G., Beekman, R., Raineri, E., Esteve, A., Clot, G., et al. (2015). Whole-genome fingerprint of the DNA methylome during human B cell differentiation. *Nat. Genet.* **47**, 746–756.
- Lee, P.P., Fitzpatrick, D.R., Beard, C., Jessup, H.K., Lehar, S., Makar, K.W., Pérez-Melgosa, M., Sweetser, M.T., Schlissel, M.S., Nguyen, S., et al. (2001). A critical role for Dnmt1 and DNA methylation in T cell development, function, and survival. *Immunity* **15**, 763–774.
- Lin, M.F., Jungreis, I., and Kellis, M. (2011). PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275–i282.



- Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322.
- Mammana, A., and Chung, H.R. (2015). Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. *Genome Biol.* **16**, 151.
- Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., Maier, L., Mackowiak, S.D., Gregersen, L.H., Munschauer, M., et al. (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495**, 333–338.
- Okhrimenko, A., Grün, J.R., Westendorf, K., Fang, Z., Reinke, S., von Roth, P., Wassilew, G., Kühn, A.A., Kudernatsch, R., Demski, S., et al. (2014). Human memory T cells from the bone marrow are resting and maintain long-lasting systemic memory. *Proc. Natl. Acad. Sci. USA* **111**, 9229–9234.
- Polansky, J.K., Kretschmer, K., Freyer, J., Floess, S., Garbe, A., Baron, U., Olek, S., Hamann, A., von Boehmer, H., and Huehn, J. (2008). DNA methylation controls Foxp3 gene expression. *Eur. J. Immunol.* **38**, 1654–1663.
- Richard, E.M., Thiyagarajan, T., Bunni, M.A., Basher, F., Roddy, P.O., Siskind, L.J., Nietert, P.J., and Nowling, T.K. (2013). Reducing FLI1 levels in the MRL/lpr lupus mouse model impacts T cell function by modulating glycosphingolipid metabolism. *PLoS ONE* **8**, e75175.
- Rosenbloom, K.R., Armstrong, J., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M., et al. (2015). The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.* **43**, D670–D681.
- Russ, B.E., Olshanksy, M., Smallwood, H.S., Li, J., Denton, A.E., Prier, J.E., Stock, A.T., Croom, H.A., Cullen, J.G., Nguyen, M.L., et al. (2014). Distinct epigenetic signatures delineate transcriptional programs during virus-specific CD8(+) T cell differentiation. *Immunity* **41**, 853–865.
- Rybak-Wolf, A., Stottmeister, C., Glazar, P., Jens, M., Pino, N., Giusti, S., Hanan, M., Behm, M., Bartok, O., Ashwal-Fluss, R., et al. (2015). Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed. *Mol. Cell* **58**, 870–885.
- Sallusto, F., Lenig, D., Förster, R., Lipp, M., and Lanzavecchia, A. (1999). Two subsets of memory T lymphocytes with distinct homing potentials and effector functions. *Nature* **401**, 708–712.
- Sathaliyawala, T., Kubota, M., Yudanin, N., Turner, D., Camp, P., Thome, J.J., Bickham, K.L., Lerner, H., Goldstein, M., Sykes, M., et al. (2013). Distribution and compartmentalization of human circulating and tissue-resident memory T cell subsets. *Immunity* **38**, 187–197.
- Sato, S., Lennard Richard, M., Brandon, D., Jones Buie, J.N., Oates, J.C., Gilkeson, G.S., and Zhang, X.K. (2014). A critical role of the transcription factor fli-1 in murine lupus development by regulation of interleukin-6 expression. *Arthritis Rheumatol.* **66**, 3436–3444.
- Scharer, C.D., Barwick, B.G., Youngblood, B.A., Ahmed, R., and Boss, J.M. (2013). Global DNA methylation remodeling accompanies CD8 T cell effector function. *J. Immunol.* **191**, 3419–3429.
- Schenkel, J.M., and Masopust, D. (2014). Tissue-resident memory T cells. *Immunity* **41**, 886–897.
- Schmidt, F., Gasparoni, N., Ebert, P., Gianmoena, K., Cadenas, C., Polansky, J.K., Barann, M., Sinha, A., Froehler, S., Gasparoni, G., et al. (2016). Combining transcription factor affinities with open-chromatin data for accurate gene expression prediction. *bioRxiv*. Published online October 19, 2016. <http://dx.doi.org/10.1101/081935>.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504.
- Theocharidis, A., van Dongen, S., Enright, A.J., and Freeman, T.C. (2009). Network visualization and analysis of gene expression data using BioLayout Express(3D). *Nat. Protoc.* **4**, 1535–1550.
- Tokoyoda, K., Zehentmeier, S., Hegazy, A.N., Albrecht, I., Grün, J.R., Löhning, M., and Radbruch, A. (2009). Professional memory CD4+ T lymphocytes preferentially reside and rest in the bone marrow. *Immunity* **30**, 721–730.
- Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515.
- Urich, M.A., Nery, J.R., Lister, R., Schmitz, R.J., and Ecker, J.R. (2015). MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. *Nat. Protoc.* **10**, 475–483.
- Wallner, S., Schröder, C., Leitão, E., Berulava, T., Haak, C., Beißer, D., Rahmann, S., Richter, A.S., Manke, T., Bönisch, U., et al. (2016). Epigenetic dynamics of monocyte-to-macrophage differentiation. *Epigenetics Chromatin* **9**, 33.
- Wang, L., Park, H.J., Dasari, S., Wang, S., Kocher, J.P., and Li, W. (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **41**, e74.
- Wei, H., Geng, J., Shi, B., Liu, Z., Wang, Y.H., Stevens, A.C., Sprout, S.L., Yao, M., Wang, H., and Hu, H. (2016). Cutting Edge: Foxp1 Controls Naive CD8+ T Cell Quiescence by Simultaneously Repressing Key Pathways in Cellular Metabolism and Cell Cycle Progression. *J. Immunol.* **196**, 3537–3541.
- Youngblood, B., Oestreich, K.J., Ha, S.J., Duraiswamy, J., Akondy, R.S., West, E.E., Wei, Z., Lu, P., Austin, J.W., Riley, J.L., et al. (2011). Chronic virus infection enforces demethylation of the locus that encodes PD-1 in antigen-specific CD8(+) T cells. *Immunity* **35**, 400–412.
- Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.* **67**, 301–320.

## Supplemental Information

### Epigenomic Profiling of Human CD4<sup>+</sup> T Cells Supports

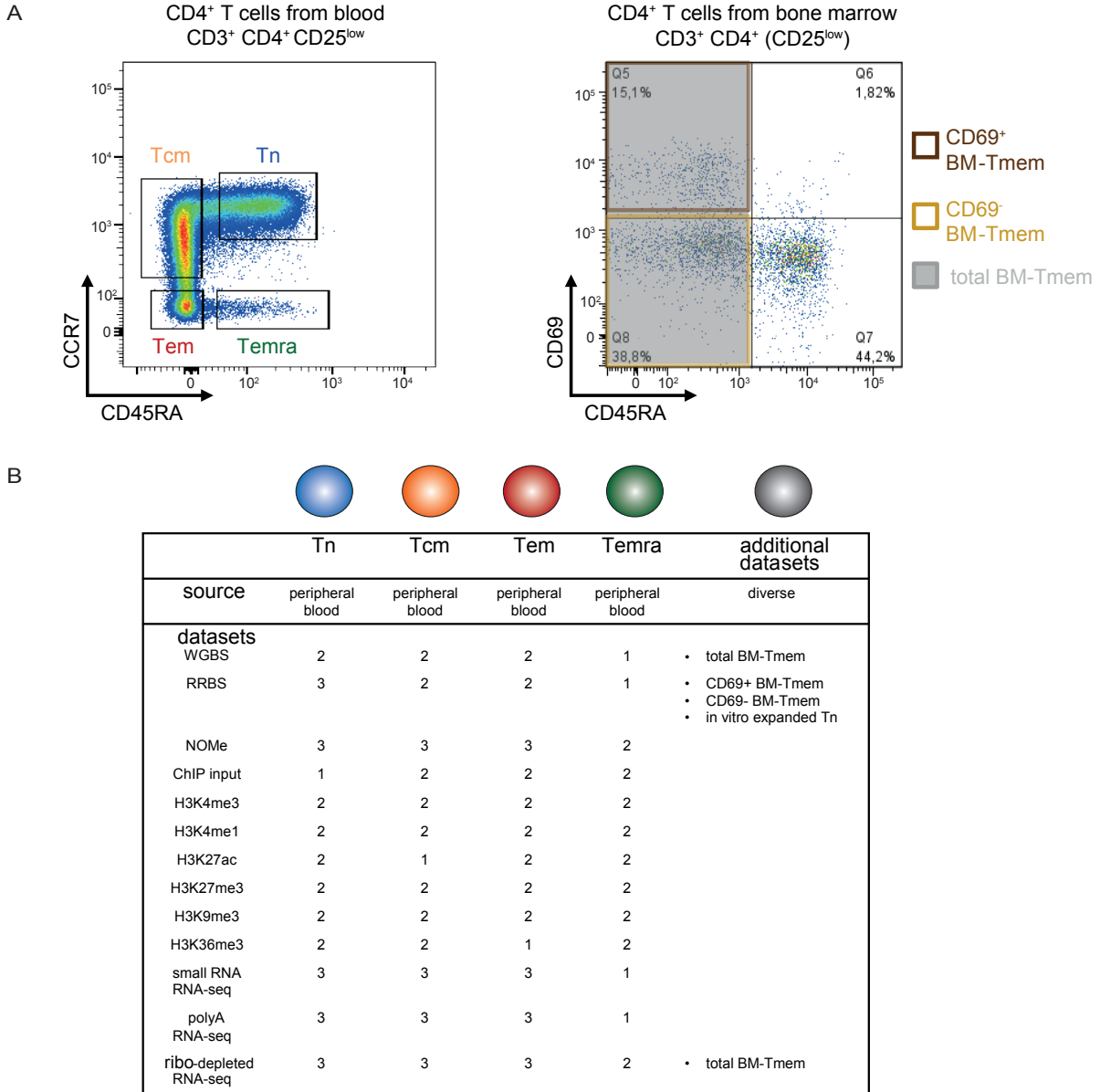
#### a Linear Differentiation Model and Highlights

#### Molecular Regulators of Memory Development

Pawel Durek, Karl Nordström, Gilles Gasparoni, Abdulrahman Salhab, Christopher Kressler, Melanie de Almeida, Kevin Bassler, Thomas Ulas, Florian Schmidt, Jieyi Xiong, Petar Glazar, Filippos Klironomos, Anupam Sinha, Sarah Kinkley, Xinyi Yang, Laura Arrigoni, Azim Dehghani Amirabad, Fatemeh Behjati Ardakani, Lars Feuerbach, Oliver Gorka, Peter Ebert, Fabian Müller, Na Li, Stefan Frischbutter, Stephan Schlickeiser, Carla Cendon, Sebastian Fröhler, Bärbel Felder, Nina Gasparoni, Charles D. Imbusch, Barbara Hutter, Gideon Zipprich, Yvonne Tauchmann, Simon Reinke, Georgi Wassilew, Ute Hoffmann, Andreas S. Richter, Lina Sieverling, DEEP Consortium, Hyun-Dong Chang, Uta Syrbe, Ulrich Kalus, Jürgen Eils, Benedikt Brors, Thomas Manke, Jürgen Ruland, Thomas Lengauer, Nikolaus Rajewsky, Wei Chen, Jun Dong, Birgit Sawitzki, Ho-Ryun Chung, Philip Rosenstiel, Marcel H. Schulz, Joachim L. Schultze, Andreas Radbruch, Jörn Walter, Alf Hamann, and Julia K. Polansky

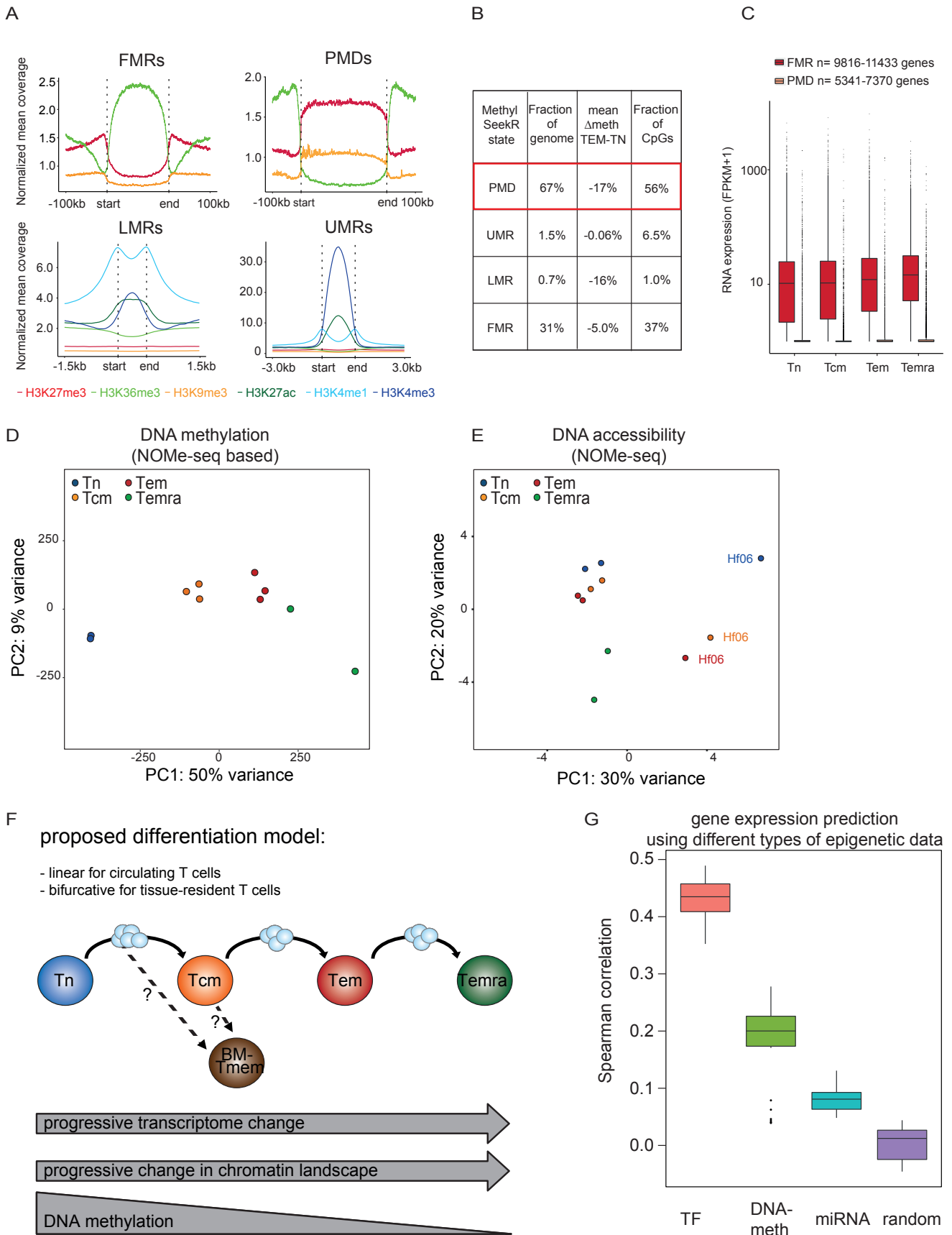
## Supplemental Figures

**Figure S1: T cell populations and data types generated and analyzed in the present study.**



**Figure S1, related to Figure 1-5: T cell populations and data types generated and analyzed in the present study.** A) Sorting strategy of blood- or bone marrow-derived T cell populations. Samples from blood (left) were pre-gated on CD3<sup>+</sup> CD4<sup>+</sup> CD25<sup>low</sup> cells in the lymphocyte gate. Samples from the bone-marrow (right) were pre-gated on CD3<sup>+</sup> CD4<sup>+</sup> CD25<sup>low</sup> cells in the lymphocyte gate for total BM-Tmem and on CD3<sup>+</sup> CD4<sup>+</sup> cells for the CD69<sup>+</sup> and CD69<sup>-</sup> subsets. B) Overview of the epigenomic data sets generated in this study.

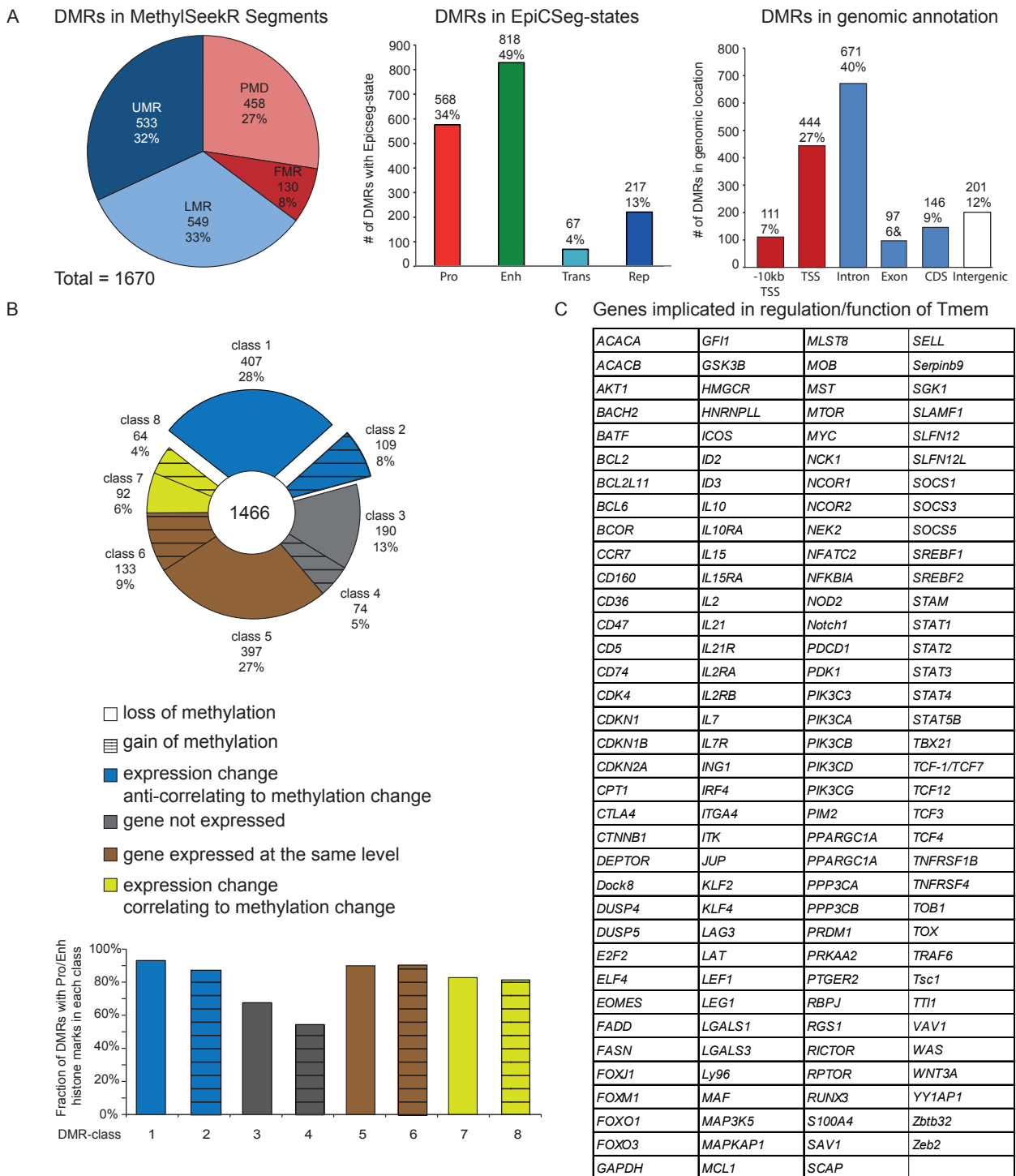
**Figure S2: Progressive changes in the DNA-meth levels in PMDs, in the transcriptomes and in the DNA accessibility profiles support a linear differentiation model for circulating human CD4+ Tmem subsets.**



**Figure S2, related to Figure 2 and 3: Progressive changes in the DNA-meth levels in PMDs, in the transcriptomes and in the DNA accessibility profiles support a linear differentiation model for circulating human CD4<sup>+</sup> Tmem subsets.**

A) Normalized mean coverage of different histone modification marks across the MethylSeekR segments in Tem cells (donor pool Hf03). PMD=partially methylated domain, FMR=fully methylated region, LMR=low methylated region, UMR=unmethylated region. B) Characteristics of MethylSeekR states (derived from Tem Hf03): Fraction of genome covered, mean methylation loss in Tem vs. Tn cells and fraction of CpG sites covered. C) Expression values for protein-coding genes across the MethylSeekR segments in Tn, Tcm, Tem and Temra cells. The range of the numbers of genes included in each segment is shown. D) PCA on DNA-meth in CpG context (excluding GpCpG context) from NOME-seq data. Only sites with a minimum read coverage of five in all indicated samples were considered. E) PCA on NOME-seq data showing PC1 and PC2. Note that PC1 is separating one replicate (donor pool Hf06) due to a slight change in the NOME-seq protocol. F) Scheme for the proposed model of CD4<sup>+</sup> Tmem cell differentiation suggested by epigenomic and transcriptomic analyses: Circulating Tmem cell subsets from the blood (Tn, Tcm, Tem, Temra cells) are generated successively (= linear model) with multiple rounds of proliferation accompanying each differentiation step, leading to the observed progressive loss of DNA-meth and the progressive change in transcriptomes and chromatin landscape. In contrast, tissue-resident BM-Tmem branch off early and display a particular epigenomic imprint. G) Using either TF binding in open-chromatin peaks (TF), DNA-meth at the promoter or gene body (DNA-meth) or miRNA target sites with miRNA expression (miRNA), the ability to predict gene expression using a regression approach was calculated for each feature. Each boxplot shows the Spearman correlation values of all 6 samples (2x Tn, 2x Tcm, 2x Tem) and the result of 6-fold nested cross validation.

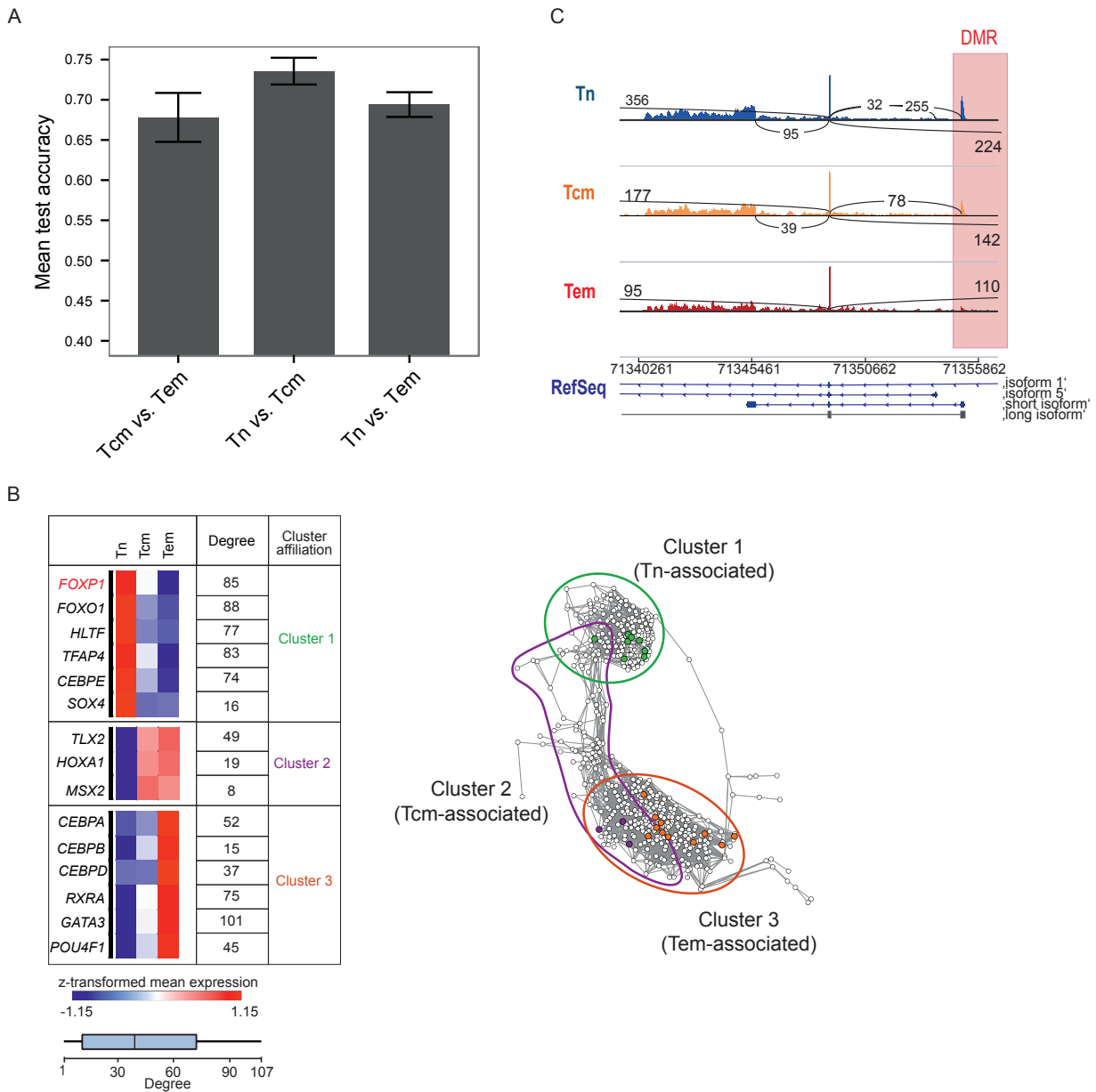
**Figure S3: Identification and characterization of differentially methylated regions (DMRs) in Tn, Tcm and Tem cells which were used to assess the epigenetic contribution to known Tmem cell-related genes.**



**Figure S3, related to Figure 3: Identification and characterization of differentially methylated regions (DMRs) in Tn, Tcm and Tem cells which were used to assess the epigenetic contribution to known Tmem cell-related genes.**

A) DMRs were called using the Metilene software and filtered using the 'adaptive filtering' approach (see *Supplemental Experimental Procedures*) which resulted in a total of 1670 DMRs. The number and proportion of DMRs falling into each MethylSeekR segment (left, PMD=partially methylated domain, FMR=fully methylated region, LMR=low methylated region, UMR=unmethylated region), into each functional chromatin state (called by EpiCseg using the Hf03 Tem sample, Pro=promoter, Enh=enhancer, Trans=transcribed, Rep=repressed, middle) or into each genic location (right, TSS = transcription start site, CDS = coding sequence) are shown. B) Classification of DMRs according to their correlation between change of DNA-meth and gene expression of the associated gene. Number and proportion of DMRs are shown for each of the 8 DMR classes (top). Proportion of DMRs in each class which show a promoter (Pro) or enhancer (Enh) histone signature (in the Hf03 Tem sample by EpiCseg, bottom). C) List of 144 Tmem cell-related genes which were extracted from recent publications on T cell memory (see *Supplemental Experimental Procedures*) and analyzed for their association to DMRs.

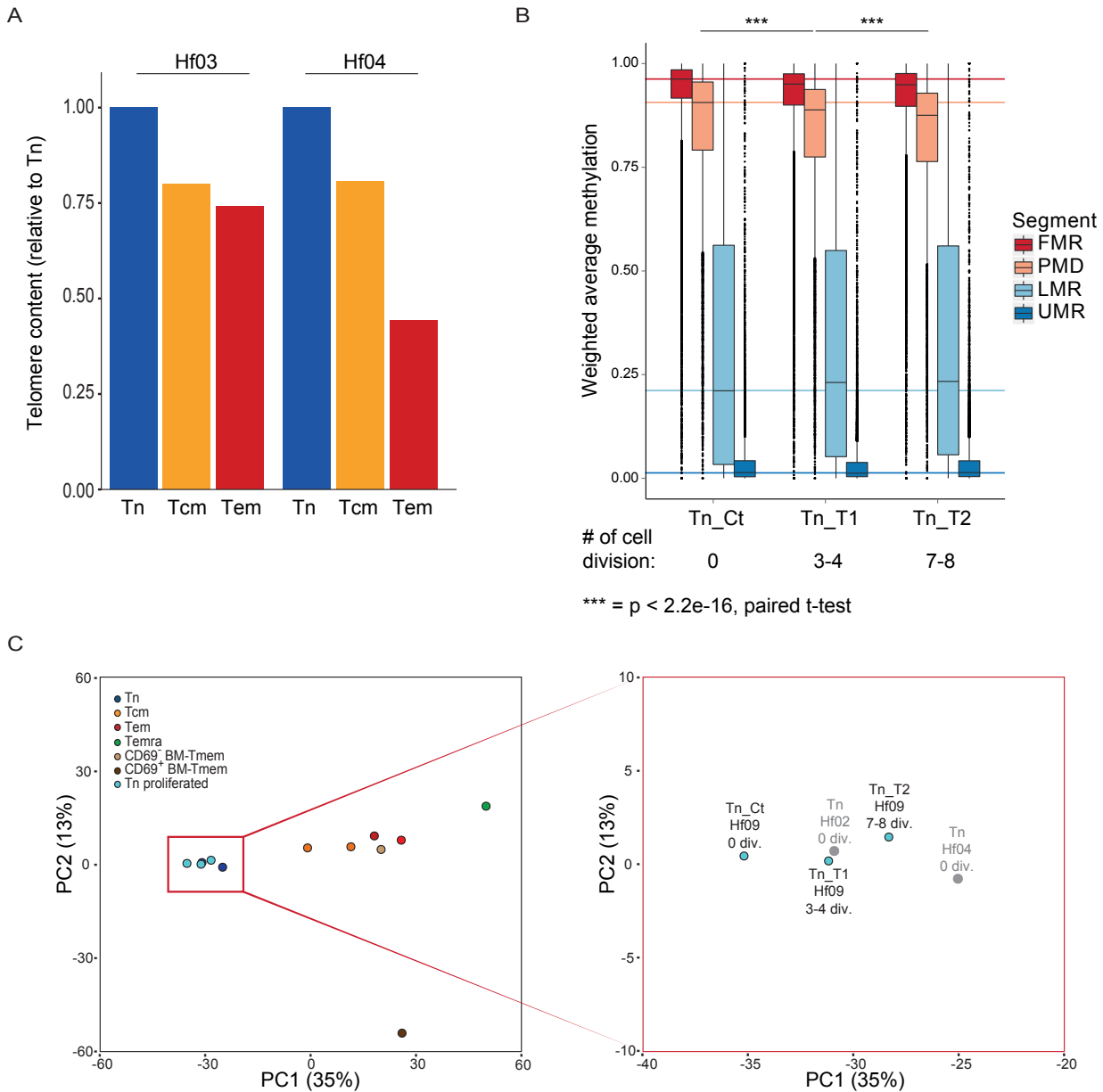
**Figure S4: Identification of Tmem regulator candidates from transcriptomic and epigenomic datasets using two different approaches.**



**Figure S4, related to Figure 4 and 5: Identification of Tmem regulator candidates from transcriptomic and epigenomic datasets using two different approaches.**

A) Test accuracy of the applied machine learning approach to model differential gene expression between Tn, Tcm and Tem cells. This approach was based on the TF binding affinities in open chromatin regions (NOME-peaks), which allows the identification of TFs contributing to the differential gene expression profile (TFs shown in Figure 4). B) Putative master regulators for Tmem cell differentiation were predicted using iRegulon (Janky et al., 2014). The analysis is based on predicted TF-binding site enrichments in promoters of co-regulated genes within the Tn-, Tcm- and Tem cell-associated clusters of transcriptional regulators (shown in Figure 2B, bottom). For the TFs with most significantly enriched predicted binding sites, we visualized their expression, inferred their connectivity (left) and marked their location within the TF network (right). Heatmaps represent z-scored mean expression values and the corresponding degree (indicating the number of connections within the network). The boxplot indicates the overall distribution of degrees within the network. C) Sashimi plot of mRNA-seq data zoomed into the *FOXP1* locus, displaying the region downstream of the *FOXP1*-DMR. Arcs represent reads covering splicing events with the corresponding counts. Only arcs representing at least 25 reads are shown. RNA isoforms inferred from the mRNA-seq data are shown at the bottom, three of which (in blue) are annotated in RefSeq.

**Figure S5: Loss of DNA-meth in PMDs correlates to past episodes of proliferation in the CD4+ T cells.**



**Figure S5, related to Figure 1: Loss of DNA-meth in PMDs correlates to past episodes of proliferation in the CD4+ T cells.**

A) Reduced telomere length in Tmem compared to Tn cells. Quantification of telomere content was performed using the TelomereHunter python package on WGBS-alignments. To account for the variable read length in the WGBS data, the repeatThreshold was set to  $0.06 \times \text{read length}$ . Results were normalized to the value of Tn cells for each donor pool (Hf03 and Hf04). Similar results were obtained when ChIP-input data from the same cellular samples were used (data not shown). B) In vitro activated Tn cells lose DNA-meth in PMDs corresponding to their number of cell divisions. Weighted average DNA-meth, aggregated from both strands, across the MethyseekR segments in ex vivo Tn cells (Tn\_Ct, donor pool Hf09) as well as Tn cells, which have undergone 3-4 (Tn\_T1) or 7-8 (Tn\_T2) cell divisions in vitro (sorted according to CFSE dilution). The median of weighted average methylation levels for ex vivo TNs are marked with colored lines for each segment. Statistically significant differences (paired t-test) were found in PMDs between Tn\_Ct and Tn\_T1 as well as Tn\_T1 and Tn\_T2. C) PCA of DNA-meth profiles (based on RRBS data, left, zoom-in on the right) containing ex vivo isolated T cell samples (Tn, Tcm, Tem, Temra, Tn\_Ct, CD69<sup>+</sup> and CD69<sup>-</sup> BM-Tmem) as well as Tn cells after proliferation in vitro (Tn\_T1 and Tn\_T2).



## Supplemental Tables

**Table S1: Overview of the genome-wide datasets generated in this study** (related to Figure 1-5). The list includes information on the donor pools, the data type and the number of mapped reads for each dataset.

**Table S2: List of identified miRNAs** (related to Figure 2). The list includes known and new miRNAs as well as differentially expressed miRNAs between the different T cell subtypes.

**Table S3: List of identified lncRNAs** (related to Figure 2). The list includes all identified lncRNAs and their expression values (FPKM) for all samples.

**Table S4: List of identified circRNA candidates** (related to Figure 2). The list includes all identified circRNAs including a classification into highly expressed circRNAs and genes for which the circular isoforms seems to be dominant over the linear spliced isoform.

**Table S5: List of DMRs** (related to Figure 3). The list includes all DMRs (Tn vs. Tcm vs. Tem cells) after 'adaptive filtering' including information on their DNA-meth level, location within MethylSeekR segments and CpG-Islands and the correlation to the expression changes of the associated genes.

## Supplemental Experimental Procedures

### *T cell isolation from human peripheral blood and bone-marrow*

PBMC from peripheral blood of healthy female donors were isolated directly from whole blood (donor pools Hf02 and Hf05) or from leukocyte filters (Hf03-04, Hf06) obtained from the blood bank of the Charité University Hospital, Institute of Transfusion Medicine. Bone marrow samples were obtained from systemically healthy female donors undergoing hip replacements.

Bone marrow-derived T cell subsets were sorted as described earlier (Okhrimenko et al., 2014). For blood-derived T cell subsets, PBMCs were isolated by density gradient centrifugation using Lymphocyte Separation Medium LSM 1077 (PAA). Remaining erythrocytes were lysed in Erythrocyte-Lysis-Buffer (Buffer EL, Qiagen). CD4<sup>+</sup> T lymphocytes were enriched using the MACS-technology and human CD4 MicroBeads (Miltenyi Biotec) on an AutoMACS-instrument (Miltenyi Biotec). The enriched population was stained using antibodies targeting CD3 (UCHT1), CD4 (OKT4), CD127 (A019D5), CD25 (M-A251), CCR7 (GO43H7, all from Biolegend) and CD45RA (2H4LDH11LDB9, Beckmann-Coulter) and sorted on a FACS-Aria instrument (BD Bioscience). Regulatory T cells were excluded by gating out the CD3<sup>+</sup> CD4<sup>+</sup> CD25<sup>high</sup> CD127<sup>low</sup> population. The remaining cells were sorted into naive (Tn), central memory (Tcm), effector memory (Tem) and Temra populations using the following marker combinations: Tn = CD3<sup>+</sup> CD4<sup>+</sup> CD45RA<sup>-</sup> CCR7<sup>+</sup>, Tcm = CD3<sup>+</sup> CD4<sup>+</sup> CD45RA<sup>-</sup> CCR7<sup>+</sup>, Tem = CD3<sup>+</sup> CD4<sup>+</sup> CD45RA<sup>+</sup> CCR7<sup>-</sup> and Temra = CD3<sup>+</sup> CD4<sup>+</sup> CD45RA<sup>+</sup> CCR7<sup>+</sup>. Purity of the sorted populations was confirmed by flow-cytometry. Each sample was split into 4 fractions, snap-frozen and stored as a cell pellet at -80°C. For RNA-seq, one aliquot was pelleted, resuspended and homogenized in QIAzol Lysis Reagent (Qiagen) using a 21G needle before snap-freezing. Aliquots of several donors were pooled for the analysis of genome-wide epigenetic signatures (Table S1). Pools for ChIP-seq and NOME-seq were fixed with 1% methanol-free formaldehyde for 5 min at RT, quenched with 0.125 M glycine for 5 min RT and washed with PBS.

### *T cell culture*

Naïve or total CD4<sup>+</sup> T lymphocytes were cultured in RPMI 1640 medium with Glutamax (Thermo Fisher Scientific) supplemented with 10% FCS (Biochrom, Merck Millipore), 25 mM HEPES, 1 mM sodium pyruvate, 50 µM β-mercaptoethanol, 100 U/ml penicillin/streptomycin (all Merck Millipore). In vitro culture was initiated by TCR-mediated stimulation with plate-bound anti-CD3 (UCHT1; 4 µg/ml) and anti-CD28 (CD28.2; 2 µg/ml, both BD Bioscience) in the presence of 20 ng/ml recombinant human IL-2 (R&D Systems) for 2-3 days. Cells were rested in uncoated plates until day 7 and then re-stimulated on coated plates over night. Culture was terminated on day 14.

### *WGBS*

For each sample two types of WGBS libraries were prepared to obtain a well balanced coverage across the genome. For the first type, a pre-bisulfite library protocol as described by Urich et al. (Urich et al., 2015) was followed, using

2 µg of DNA. For the second type, a post-bisulfite library was prepared using 100 ng of DNA with the TruSeq DNA Methylation kit (Illumina, San Diego, USA) according to the manufacturer's instructions. All libraries were quality-checked and quantified on an Agilent Bioanalyzer (Santa Clara) and by qPCR using the Perfecta qPCR FastMix (Quanta Biosciences). Samples were sequenced for three lanes (two lanes pre-bisulfite and one lane post-bisulfite library) on a HiSeq2500 machine (Illumina) resulting in ~30x fold raw genome coverage per sample.

#### *RRBS*

RRBS libraries were prepared according to the procedure described by Boyle et al., 2012, with small modifications: Briefly, 10 to 100 ng of genomic DNA were digested overnight using 50 U of MspI or HaeIII enzyme (NEB) followed by end-repair, A-tailing, NGS-adaptor ligation and purification with Ampure XP beads (Beckman Coulter). The libraries were then bisulfite converted with the EZ-DNA Methylation Gold kit (Zymo Research) and PCR-amplified for 12-14 cycles using Hot Star Taq polymerase (Qiagen) followed by a final Ampure beads purification step.

#### *NOMe-seq*

Fixed frozen (Hf06 only) or fresh frozen cells (200-500k) were thawed in nuclei extraction buffer (60 mM KCl; 15 mM Tris-HCl, pH 8.0; 15 mM NaCl; 1 mM EDTA, pH 8.0; 0.5 mM EGTA, pH 8.0; 0.5 mM spermidine free base) supplemented with complete protease inhibitor cocktail (Roche) and 0.1% NP40 (Sigma-Aldrich) and incubated on ice for 30 min. During incubation, fixed samples were dounced 10-20 times with a douncing pistil (Qiagen). Nuclei were centrifuged (500 g, 4 °C, 8 min) and the pellet was washed with NP40-free nuclei buffer. After centrifugation, nuclei were resuspended in 90 µl of 1x GpC-buffer (NEB) followed by addition of 70 µl of NOME reaction mix (7 µl 10x GpC buffer, NEB), 1.5 µl of 32 mM SAM (NEB), 45 µl of 1 M Sucrose, 60 U of M. CviPI (NEB), 0.5 µl of aqua bidest. The reaction was incubated 3 h at 37 °C and another 0.5 µl of SAM were added hourly. The reaction was stopped with 160 µl NOME stop buffer (20 mM Tris-HCl, pH 8.0; 600 mM NaCl; 1 % SDS, 10 mM EDTA) and 10 µl proteinase K (20 mg/ml, Sigma-Aldrich) and genomic DNA was extracted. NOME libraries were prepared from 100 ng of DNA using the post-bisulfite library protocol (see WGBS section) and sequenced for two lanes on a HiSeq2500 instrument.

#### *ChIP-seq*

Chromatin Preparation: Three million fixed cells were lysed in 300 µl chromatin lysis buffer (50 mM Tris-HCl pH8.1, 100 mM NaCl, 1% SDS, 3% Triton X-100, 5 mM EDTA, 0.2% NaN<sub>3</sub>) supplemented with 3x protease inhibitors (Roche complete protease inhibitor cocktail) on ice for 10 minutes. Lysates were then diluted 3x with a dilution buffer (50 mM Tris-HCl pH 8.6, 100 mM NaCl, 5 mM EDTA, 0.2% NaN<sub>3</sub>) and homogenized ten times with a syringe (271/2 gauge). The lysate was then aliquoted (200 µl) into 1.5 ml TPX polymethylpentene tubes (Diagenode) and sheared at 4 °C in a Bioruptor Pico for 3x10 cycles. The chromatin was then pooled and centrifuged for 10 minutes at 14,000 rpm at 4°C to pellet debris. For Temra cells, one million cells were formaldehyde-fixed and nuclei were extracted as previously described (Arrigoni et al., 2016). Nuclei were resuspended in 1 ml of shearing buffer (10 mM Tris-HCl pH 8, 1 mM EDTA, 0.1% SDS, supplemented with 1x protease inhibitor cocktail) and chromatin was sonicated using Covaris E220 ultrasonicator for 12 minutes, using the instrument settings 140W peak power, 5% duty factor and 200 cycles/burst.

5% of the chromatin was collected to analyze the shearing efficiency and the remaining 840 µl of chromatin was diluted with 800 µl of dilution buffer supplemented with 1x protease inhibitors. Chromatin from Temra cells has been diluted at 1:1 ratio with Diagenode ChIP buffer H (Auto histone ChIP-seq kit), supplemented with 1x protease inhibitor cocktail, prior ChIP.DNA preparation and Chromatin Shearing Efficiency: 40 µl of sheared chromatin was collected and diluted up to 100 µl with TE buffer pH 9.5 and supplemented with 4 µl 5M NaCl. The chromatin was de-crosslinked at 65°C for 2h, followed by an RNaseA treatment (0.2 mM) at 60°C for 30 minutes and proteinase K treatment (3 µl; Sigma) for 2h at 55°C. The DNA was then isolated using ChIP DNA concentrator columns (Zymo Research D5205) according to the manufacturer's instructions. Shearing efficiency was monitored by loading 1 ng DNA on a high sensitivity DNA chip using high sensitivity DNA reagents according to the manufacturer's instructions (Agilent) and analyzed on an Agilent 2100 Bioanalyzer.

Chromatin Immunoprecipitation: The ChIP was performed using the Diagenode Auto Histone ChIPseq kit on an IPstar SX-8G compact automated system (Diagenode) using their indirect method (Ag + Ab → Beads). The IP reaction was performed for 11 h, the incubation with beads for 7h and the washes were performed for 5 minutes each. 100 µl of chromatin (165,000 cells, 60,000 for Temra cells), 10 µl of protein A magnetic beads and 1 µg of Diagenode ChIPseq grade rabbit polyclonal antibody (H3K4me1; pAb-194-050, H3K4me3; pAb-003-050, H3K9me3; pAb-193-050, H3K27Ac; pAb-196-050, H3K27me3; pAb-195-050 and H3K36me3; pAb-192-050) were used per ChIP. 20 µl of chromatin (33,000 cells) and 1 µl for Temra cells (600 cells) was used for the input. ChIPs

were eluted in 100 µl TE buffer pH 9.5 and the input was diluted up to 100 µl with TE buffer pH 9.5. 4 µl of 5 M NaCl was added to each sample and the DNA was isolated as described above after de-crosslinking, RNaseA and proteinase K digestion.

**Library Preparation:** The libraries were generated using a Diagenode Microplex Library preparation Kit (C05010010) according to manufacturer's instructions. In brief, 2 ng or less of ChIP DNA was used to generate sequencing libraries. After adapter ligation, 10 rounds of PCR amplification were used to amplify library DNA. The libraries were size selected using Agencourt AMPure XP beads (Beckman Coulter) to remove ligated adaptors and DNA fragments greater than 1000bp, using the calculated ratios of 0.56 first followed by 0.95. Libraries for Temra cells ChIP DNA were prepared using the NEBNext Ultra library preparation kit (NEB, E7370S) following manufacturer's instructions and skipping the size selection. Adapter-ligated DNA fragments were amplified using 13 PCR cycles.

**ChIP Library Sequencing:** All libraries were sequenced on an Illumina HiSeq 2500 using version 3 chemistry and 50bp paired end sequencing according to Illumina suggested protocols.

#### *RNA-seq incl. RNA isolation*

Pooled samples for RNA-seq were homogenized in QIAzol Lysis Reagent (Qiagen) using a 21G needle. After addition of chloroform and phase separation by centrifugation, total RNA, including small RNAs, was extracted from the aqueous phase by using the miRNeasy Micro Kit (Qiagen) following the manufacturer's recommendations.

**Sequencing of long RNA libraries:** Starting from 2x500ng total RNA of RIN 9.6, one stranded total RNA and one stranded mRNA library were prepared according to the manufacturer's instructions (Illumina). Both libraries were sequenced for 2x101nt on an Illumina HiSeq 2000, yielding about 100 million paired-end reads for each library.

**Sequencing of short RNA libraries:** Starting from 1-5 µg total RNA of RIN 9.6, the small RNA sequencing library was prepared according to the manufacturer's instructions (Illumina) and sequenced for 1x51nt on an Illumina HiSeq 2000, yielding about 10 million single end reads for each library.

#### *Mapping of WGBS and NOME-seq*

The WGBS data were processed as described in Wang et al., 2013b. The hg19 reference genome (37d5) was transformed in silico for both the top strand (C to T) and bottom strand (G to A) using MethylTools (Hovestadt et al., 2014). Before alignment, adaptor sequences were trimmed using SeqPrep (<https://github.com/jstjohn/SeqPrep>). The first read in each read pair was then C-to-T converted and the 2nd read in the pair was G-to-A converted. The converted reads were aligned to a combined reference of the transformed top (C to T) and bottom (G to A) strands using BWA (bwa-0.6.2-tpx) with default parameters, yet, disabling the quality threshold for read trimming (-q) of 20 and the Smith-Waterman for the unmapped mate (-s). After alignment, reads were converted back to the original states, and reads mapped to the antisense strand of the respective reference were removed. Duplicate reads were removed, and the complexity determined using Picard MarkDuplicates (<http://picard.sourceforge.net/>). Reads with alignment scores less than 1 were filtered before subsequent analysis. Total genome coverage was calculated using the total number of bases aligned from uniquely mapped reads over the total number of mappable bases in the genome.

#### *Mapping of RRBS*

RRBS data were trimmed with the Cutadapt (<http://dx.doi.org/10.14806/ej.17.1.200>) wrapper

Trim Galore! ([http://www.bioinformatics.babraham.ac.uk/projects/trim\\_galore](http://www.bioinformatics.babraham.ac.uk/projects/trim_galore)) in RRBS mode, and subsequently mapped with MethylTools as described for WGBS, excluding the removal of duplicates.

#### *Mapping of ChIPseq data*

Reads were mapped to the 1000 genomes phase 2 assembly of the human reference genome (NCBI build 37.1, downloaded from [ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2\\_reference\\_assembly\\_sequence/](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/)) with a hardware-accelerated implementation of Burrows-Wheeler Aligner BWA aln version 0.6.2 (Li and Durbin, 2009) with -q 20, and BWA 0.6.2 sampe with -a 1000. Merging and duplicate marking was performed with Picard version 1.125 (<http://broadinstitute.github.io/picard>).

#### *Mapping of RNA data*

BAM files of RNA-seq reads were produced with TopHat 2.0.11 (Kim et al., 2013), with Bowtie 2.2.1 (Langmead and Salzberg, 2012) and NCBI build 37.1 in --library-type fr-firststrand and --b2-very-sensitive setting.

#### *Calling of DNA-methylation values*

Alignments of bisulfite treated reads were analyzed with a Bis-SNP (Liu et al., 2012) based pipeline comprising Picard tools (<http://broadinstitute.github.io/picard>), samtools (Li et al., 2009), bamUtil (<http://genome.sph.umich.edu/wiki/BamUtil>) and UCSC tools (Kent et al., 2010). The pipeline includes SNP-aware realignment, trimming of overlapping read pairs and re-calibration of quality values before methylation ratios were calculated for all cytosines. In the case of RRBS data, the maximum coverage allowed was adjusted from the default 250 reads to 2000 reads per loci. For WGBS and RRBS data, cytosines in CG-context were extracted for further processing, while GCH and HCG-data were gathered from the NOME data sets.

#### *Genome segmentation based on DNA-methylation (MethylSeekR)*

The WGBS data was used to segment the genome into four different states using MethylSeekR (Burger et al., 2013); fully methylated regions (FMRs), partially methylated domains (PMDs), lowly methylated regions (LMRs) and unmethylated regions (UMRs). The tool ran with a methylation level threshold at 0.5, a coverage cutoff at five reads per CpG and maximum FDR of 0.05, resulting in a threshold of at least four CpGs per LMR. The methylation levels from both strands were aggregated and weighted average methylation levels were plotted across the four segments resulting in very few outliers in LMRs/UMRs with methylation values > 0.5 due to the smoothing in MethylSeekR.

For the downstream analyses, we merged the large scale regions (PMDs and FMRs) from each replicate, allowing them to span gaps introduced by shorter LMRs or UMRs. The merged PMDs were calculated as the difference between the original PMDs after joining neighboring regions allowing for 5kb gaps and joined original FMRs, excluding regions shorter than 5kb. The merged FMRs were then calculated as the complement to the merged PMDs, excluding genomic gaps as annotated by UCSC (Rosenbloom et al., 2015), e.g., telomeres and centromeres.

The normalized mean coverage of three broad histone marks (H3K27me3, H3K36me3 and H3K9me3) were plotted genome wide across the merged FMRs/PMDs and the original LMRs/UMRs with the appropriate flanking regions using deepTools (Ramirez et al., 2014), with the addition of the other three histone marks in LMRs/UMRs. Only merged PMDs and FMRs longer than 20kb and 10kb were considered, respectively.

The number of protein coding genes falling within merged PMDs and FMRs was calculated, demanding a minimum of 80% of the gene length to be overlapped by the segment. A pseudo count (1) was added to FPKM to avoid zeros in the box-plots. The aforementioned analyses were restricted to chromosomes 1-22.

WGBS Blueprint data from B cells was converted to hg19 coordinates using the liftOver tool (Rosenbloom et al., 2015) and segmentation was carried out afterward.

#### *DMR calling and 'adaptive filtering'*

Quality control and primary analysis of methylation data was done using custom R scripting and the RnBeads software package (Assenov et al., 2014). Differentially methylated regions (DMRs) were predicted with Metilene (Juhling et al., 2015) in de-novo mode among sites with at least 5x coverage. The tool ran without cutoffs for q-value and absolute methylation difference.

For 'adaptive filtering', the resulting loci were split with regard to overlap of merged PMDs or not, due to the different global effects within and outside merged PMDs. In each group, cutoffs for the methylation difference were determined as the 2.5% and 97.5% quantiles. Extracting the tails of the  $\Delta$ methylation distribution and requiring a q-value below 0.05 resulted in the set of DMRs to be considered.

#### *NDR calling (Nome-peaks)*

Nucleosome depleted regions (NDRs) were identified by segmenting the GCH-methylation signal with a binomial hidden Markov model with two states (1 open/NDR, 0 background). Each putative NDR was contrasted to the closest 4kb of flanking background regions up- and down-stream by computing a significance with Fisher's exact test. Empirical false discovery rates were calculated using permutation analysis in which shuffled GCH methylation values genome-wide were run through the same process (HMM and Fisher's exact test) and subsequently used to control for multiple testing with a cutoff at 0.01.

#### *Differential NDRs*

Accessible regions were called in each sample separately, but only consistent NDRs (cNDRs) confirmed in all three replicates (Hf03/Hf04/Hf06) were kept. The union of overlapping, cNDRs was used for the differential analysis. One major reason for the deviation between replicates was deviating global average GCH-methylation levels. Given this, the GCH-methylation signal was normalized to the third quartile and after removing sites with a coverage below five, Metilene (Juhling et al., 2015) was applied in regional mode to conserved NDRs with a q-value cutoff at 0.05 which yielded 596 differential cNDRs.

#### *Generation of DMRs in Tn, Tcm, Tem cells*

DMRs from pairwise comparisons were overlapped and joined into master regions. Significance levels as well as directions of the methylation changes were taken over from the underlying DMRs, by prioritizing DMRs with the highest absolute methylation difference, if more than one DMR from the respective pairwise comparison was present. Mean methylation levels were computed by averaging methylation levels of CpG consistently covered by at least 5 reads in Hf03 and Hf04 samples. MethylSeekR regions were assigned by prioritizing overlaps of the DMRs with PMD, UMR, LMR and FMR regions. CpG Island annotations were based on UCSC hg19 annotations, by prioritizing Islands, Shores (+/- 2kb from Island) and Shelves (+/- 2kb - 4kb from Islands). Not overlapping DMRs were defined as Opensea. EpiCSeq states were assigned by prioritizing overlapping DMRs with promoter, enhancer, transcribed, repressed and unmarked states. GENCODE annotation was used to assign DMR to annotated locations and functions by prioritizing overlapping promoters (+/- 500bp from TSSs), protein coding sequence (CDS), (non-coding) exons, introns and regions upstream the TSS (-10kb TSSs). DMRs with no overlaps were defined as intragenic. Pairwise differential transcription analyses between cell-types were performed for mRNA samples Hf02, Hf03 and Hf04. Significant changes were defined by p-value < 0.05 and fold change > 1.3 of the averaged normalized read counts. A minimal read count of 10 was set prior to the fold change computation to prevent extreme values of lowly expressed genes. Genes with averaged normalized read counts < 10 in all cell-types were defined as not expressed.

#### *Transcriptome analyses*

Htseq-count from HTSeq-0.6.1p1 (Anders et al., 2015) was used to count reads mapping to genes from GENCODE release 19 (GRCh37.p13) in '-f bam -s reverse -m union -a 20' setting.

Pairwise differential expression analyses were performed with DESeq2\_1.8.1 (Anders and Huber, 2010). An FDR cutoff of 0.01 was used to select differentially expressed genes. Principal component analyses were performed on rlog-normalized reads, as implemented in DESeq2.

#### *Identification and analysis of lncRNAs*

10 Tophat aligned bam files (ribo-depleted RNA, Hf02-04 for Tn, Tcm, Tem and Hf05 for Temra) were used for a reference-assisted transcriptome assembly by Cufflinks (v2.2.1) (Trapnell et al., 2010), using the Ensembl gene annotation. The output GTF files were merged by cuffmerge. Among the novel transcripts, only the ones with 'u' (intergenic transcription) or 'x' (Exonic overlap with reference on the opposite strand) as 'class\_code' were kept. Only spliced novel transcripts were considered. Fusion genes (3'-extensions of upstream genes) were filtered, if more than one paired-end read partly aligned to its upstream genes. To remove lowly expressed genes from novel and known lncRNAs as well as from coding RNAs, only genes with an average of min. 20 reads and a mean FPKM of at least 1 in at least one of four cell types were selected for further analyses.

To remove possible coding genes from novel transcripts, the coding potential was predicted using PhyloCSF (Lin et al., 2011) and CPAT (Wang et al., 2013a). PhyloCSF predicts protein coding potential of transcripts using codon substitution frequencies based on multi-species nucleotide sequence alignments. The pre-calculated human hg19 PhyloCSF values based on 29 mammalian alignments were downloaded from <https://github.com/mlin/PhyloCSF/wiki>. CPAT integrates the coding information of four sequence features: "open reading frame size, open reading frame coverage, Fickett TESTCODE statistic and hexamer usage bias" for each predicted open reading frame (ORF). CPAT was fed with self-predicted ORFs based on canonical transcript start/stop codons, and the recommended cutoff of 0.364. Novel transcripts predicted as protein coding by either PhyloCSF or CPAT were removed from further analysis. The method was validated by applying it to 10 616 annotated and expressed coding genes as well as 669 lncRNAs. 10 455 (98.5%) coding genes and 77 (11.5%) lncRNAs were reported as protein coding.

#### *Identification and analysis of miRNAs*

Mature miRNA read counts were estimated for each sample using miRDeep2 (Friedlander et al., 2012) and miRBase (version 21) annotations. Identical mature miRNA counts originating from different precursor miRNAs were summed in order to produce total mature miRNA counts per sample and miRNAs with either less than five counts or identical counts in all samples were removed from the analysis. Principal component analysis was based on centered variance-stabilized counts (Huber et al., 2002). Clustering of samples was based on the Euclidean distance of Spearman correlations computed over the variance-stabilized counts. Differential expression analysis was done on the summarized counts using the DESeq2 package (Love et al., 2014).

#### *circRNA detection, annotation and quantification*

CircRNAs were detected, filtered and annotated as described before (Memczak et al., 2013), using hg19 genome assembly as a reference. Two additional quality filtering steps were performed. We remapped all reads supporting head-to-tail junctions to the human genome using STAR (Dobin et al., 2013), and discarded the ones that mapped successfully. We also corrected head-to-tail read counts in cases where both mates of a paired-end read supported the same splice junction.

We labeled circRNAs with 50 or more head-to-tail junction reads in at least one sample as “highly expressed”, resulting in a subset of 344 circRNAs. Head-to-tail read counts were increased by a pseudocount of 2 (equal to the detection threshold), normalized to library depth (total number of head-to-tail reads), and log-transformed. Principal component analysis was performed on these values using standard R functions.

#### *Co-expression network construction and meta-information visualization*

Expression data of Tn, Tcm and Tem cells (3 replicates each) was filtered by either a list of human transcriptional regulators (TRs) or 700 most variable genes (i.e. most significant p-values in an ANOVA-based analysis) to get a reduced expression table of present genes. The group of TRs contained transcription factors (TFs), co-factors, RNA-binding proteins and chromatin remodelers originating from the TFCat data base (Fulton et al., 2009). The expression matrices were loaded into BioLayout Express3D (Theocharidis et al., 2009) and co-regulation networks were generated with a Pearson correlation cutoff of 0.9 resulting in a network of 553 nodes for TRs and 700 nodes for the most variable genes. For further analysis, the predicted gene-gene pairs were visualized by Cytoscape (Shannon et al., 2003) using organic layout. Further information at transcriptional level such as fold change values calculated against the group mean (defined as the mean of all samples within the dataset) were mapped to the network one by one.

#### *Transcription factor binding site (TFBS) prediction by iRegulon*

To predict TFBS enrichment for a cluster of co-expressed genes, the Cytoscape plugin iRegulon (<http://iregulon.aertslab.org/index.html>; Janky et al., 2014) was applied. The co-expression network was classified into three clusters according to the regulation patterns of Tn, Tcm and Tem cells. For each subtype separately, the cluster genes were used as input for iRegulon. The corresponding species *Homo sapiens* and putative regulatory region of 500 bp upstream were selected and default parameters (e.g. enrichment threshold of 3, ROC threshold for AUC calculation of 0.03 and rank threshold of 5000, etc.) were used. The resulting output file was used to filter for TFs which were also present in the respective network clusters. Predicted TFs with the most significant normalized enrichment scores were visualized in a heatmap.

#### *Gene expression prediction using epigenetic data*

We use Elastic Net regression (Zou and Hastie, 2005) to learn a linear model for each sample to predict log<sub>2</sub> gene expression levels for each sample. As features we use four different setups: 1) each gene can be regulated by a microRNA that has a predicted target site in its 3'UTR, where the feature value is the expression of the microRNA in the sample. That means for each gene a different number of non-zero features is defined based on TargetsCan interactions (version 6.0) (Friedman et al., 2009). 2) We compute the average methylation value in four different intervals, (i) 3kb upstream of the gene's TSS, (ii) 3kb downstream of the gene's TSS, (iii) 1.5 kb upstream and 1.5 kb downstream of the gene's TSS, (iv) average methylation in the whole gene body. We use the TEPIC software (Schmidt et al., 2016) to compute TF binding features, considering NOME peaks that reside within 3 kb of the TSS (1.5 kb upstream and 1.5 kb downstream). TEPIC computes TF affinities within the NOME-peaks and aggregates them to gene TF scores. 4) We shuffle the feature matrices from 1-3 and relearn the model to obtain an estimate for model performance on random data.

In order to make the model learning comparable for the three feature sets, all model estimations are run on the same set of genes, namely the ones that have at least one NOME-peak within their 3kb promoter region, as defined above. For each set of the three features we perform parameter selection using six fold cross validation using log<sub>2</sub> gene expression values as a response in the regression. Then we compute the Spearman correlation coefficient between predicted gene expression levels and 6-fold hold out sets, i.e. we use a nested cross-validation approach, to measure model performance.

#### *Machine-learning approach to model differential gene expression*

We use a logistic regression classifier with the elastic net penalty (Zou and Hastie, 2005) to learn a linear model for three pairwise comparisons (Tn vs Tcm, Tcm vs Tem, Tn vs Tem). In each comparison we aim to predict up and down regulated genes. Differentially expressed genes were identified as mentioned above (Section *Transcriptome*

analyses), using a FDR cut-off of 0.05. As TF features for the logistic regression classifier we use a two-step approach.

First, we compute TF binding features for each gene using the TEPIC software (Schmidt et al., 2016; <https://github.com/SchulzLab/TEPIC>) using NOME-peaks for each replicate. The following command line was used for TEPIC (e.g. for replicate 51\_Hf03\_BITN):

```
bash TEPIC.sh -g hs37d5.fu -b 51_Hf03_BITN_Ct_NOME_S_1.NCSv2.20150513.GRCh37.cpg.filtered.GCH.peaks.fdr001.bed -o 51_Hf03_BITN_Ct_TEPIC -p pwm vertebrates_jaspar_uniprobe_converted.txt -n 6 -a protein_coding_only.gtf -w 50000 -c 16). Shortly, TEPIC computes predicted binding affinities for each TF in each peak region and scales these affinities by NOME signal strength, and then summarizes these peak scores per TF in a 50kb window centered on the TSS of each gene. This gives a TF-gene association map for each biological sample.
```

Second, we compute a mean TF-gene association map  $MA^R$  a matrix with G rows and T columns for a set of biological replicates R, where G is the number of genes and T is the number of TFs. One entry  $ma_{i,j,R}$  of  $MA^R$  is computed as follows:

(1)  $ma_{i,j,R} = \frac{1}{|R|} \sum_{r \in R} a_{i,j,r}$ , where  $ma_{i,j,R}$  is the mean affinity of TF  $j$  for gene  $i$  for the set of all replicates of subtype R, and  $a_{i,j,r}$  is the affinity of TF  $j$ , for gene  $i$  in replicate  $r$ .

Next, the ratio between the means for each TF between the subtypes considered in the comparison is calculated (2).

(2)  $ra_{i,j}^{R1,R2} = \frac{ma_{i,j,R1}}{ma_{i,j,R2}}$ , where  $ra_{i,j}^{R1,R2}$  is the ratio of mean affinity values between replicates of subtype 1 (R1), and subtype 2 (R2), for gene  $i$  and TF  $j$ . We denote as  $RA^{R1,R2}$  an GxT matrix of ratios between mean affinity values for subtypes R1 and R2, such that one entry is computed using formula (2).  $RA^{R1,R2}$  is the feature matrix that is used by the classifier to up and down-regulated genes.

We measure classification performance using accuracy calculated for a 10-fold outer cross validation (CV) loop. A 6-fold inner CV loop is used for parameter learning. 80% of the data are used for training and 20% for testing.

#### Analysis of the likelihood of different differentiation models based on gene expression similarity

Based on the hypothesis, that T cells that are closer to each other in the differentiation order should show more similar gene expression profiles, we designed a cell type similarity score to measure the similarity of two cell types.

$$\cos(\Theta_{r2}^{r1}) = \frac{\sum_{i=1}^G me_i^{R1} me_i^{R2}}{\sqrt{\sum_{i=1}^G me_i^{R1^2}} \sqrt{\sum_{i=1}^G me_i^{R2^2}}},$$

where  $me_i^{r1}$  ( $me_i^{r2}$ ) denotes the mean expression of gene  $i$  in subtype R1 (R2), and  $G$  is the total number of genes.

Then we sum this score to derive a similarity for the investigated differentiation orders:

$$\begin{aligned} sim_{TN-TCM-TEM} &= \cos(\Theta_{TCM}^{TN}) + \cos(\Theta_{TEM}^{TCM}) \\ sim_{TN-TEM-TCM} &= \cos(\Theta_{TEM}^{TN}) + \cos(\Theta_{TCM}^{TEM}) \\ sim_{TN-TCM; TN-TEM} &= \cos(\Theta_{TCM}^{TN}) + \cos(\Theta_{TEM}^{TN}), \end{aligned}$$

where  $sim_x$  denotes the cosine similarity score for differentiation order  $x$ .

The summed scores are normalized between 0 and 1. A large score reflects highly similar gene expression profiles. Using bootstrapping, we did 100,000 resampling steps of 15,000 out of 57,688 genes to assess whether the similarity scores are significantly different between the orderings. Statistical difference between the means of the final similarity-scores was assessed using two-sided t-tests.

#### Estimation of telomere length

Quantification of telomere content was performed using the TelomereHunter software on the aligned BAM files (<https://www.dkfz.de/en/applied-bioinformatics/telomerehunter/telomerehunter.html>). To account for the variable read length in the WGBS data the *repeatThreshold* was set to 0.06x(read length) and applied on the Illumina TruSeq PCR free library data.

#### Genome segmentation based on histone modification (5-states using Epicseg)

Annotation of different chromatin states was performed by Epicseg (Mammana and Chung, 2015) using all six histone marks. To get a robust classification scheme, valid across all cell subsets, the number of states was set to 5, the length of segmentation unit to 200bp and the minimum mapping quality to 20. The five states were defined as: Promoter (mainly enriched by H3K4me3 and H3K27ac), Enhancer (mainly enriched by H3K4me1), Transcribed (mainly enriched by H3K36me3), Repressed (mainly enriched by H3K27me3 and H3K9me3) and Unmarked (regions with very few ChIP-seq reads) as evaluated by the “log of mean read counts” heatmaps.

#### *Analysis of intracellular FOXP1 protein by flow cytometry*

Following staining of surface proteins (see *T cell isolation*), cells were fixed using the 1x Fixation/Permeabilization Buffer for intracellular Foxp3 staining (eBioscience) for 30min at 4°C. Cells were washed in 1x Permeabilization Buffer (eBioscience) and sequentially stained in 1x Permeabilization Buffer using a Foxp1 antibody (polyclonal, Cell Signalling Technology #2005) and a DyeLight-649-labelled donkey anti-rabbit secondary antibody (Biolegend, #406406). Control stainings were performed using the secondary antibody only. Each staining step was performed for 30min at 4°C and was followed by a washing step using 1x Permeabilization Buffer. Cells were resuspended in PBS and stored at 4°C until acquisition on a BD LSRFortessa™ instrument (BD Bioscience). Analysis of flow-cytometric data was performed using the Flowjo software (Flowjo, LLC).

#### *Cloning of luciferase constructs and luciferase reporter assay*

The *FOXP1-DMR* was amplified by PCR from genomic DNA of CD4<sup>+</sup> T cells using *FOXP1-DMR* forward and reverse primers (see below: *Quantitative RT-PCR* section) and cloned into the CpG-free Firefly luciferase vector pCpGL (Klug and Rehli, 2006) by restriction digest with BglII (New England Biolabs). Successful cloning and orientation of the *FOXP1-DMR* was confirmed by sequencing.

CD4<sup>+</sup> T lymphocytes were enriched by magnetic cell sorting using the human CD4 MicroBeads (Miltenyi Biotec) and stimulated for 48h with plate-bound anti-CD3 (UCHT1) anti-CD28 (CD28.2, both BD Bioscience). 1x10<sup>6</sup> cells were transfected with 2 pmol pCpGL plasmid using the Neon™ Transfection System (Life Technologies). Cells were co-transfected with 200ng of pRL-TK Renilla luciferase reporter vector (Promega) as an internal control. The transfected cells were seeded in antibiotic-free medium supplemented with hIL-2 (20ng/ml) and cultured for 24h without stimulation. Firefly and Renilla luciferase activity were assessed using the Dual Luciferase Assay Kit (Promega) and an Orion L Microplate Luminometer (Berthold Technologies) according to the manufacturer's instructions. To calculate reporter activity, the Firefly luciferase signal was divided by the Renilla reporter signal for each sample in order to normalize for differences in transfection efficiency. Normalized luciferase signals were then divided by the value of the empty pCpGL-basic vector in order to calculate the fold change activity of each sample relative to pCpGL-basic.

#### *In vitro methylation of plasmid DNA*

Plasmids were methylated using M.SssI CpG methyltransferase (New England Biolabs) according to the manufacturer's recommendation. Briefly, 30µg plasmid was incubated with M.SssI in the presence of S-adenosylmethionine (SAM; New England Biolabs) for four hours at 37°C. Mock-methylated plasmid was treated in the same way without the addition of M.SssI or SAM. Plasmids were purified using the NucleoSpin Extract II kit (Macherey Nagel).

The efficiency of methylation was verified by digesting both methylated and mock-methylated plasmids using the methylation-sensitive restriction enzyme HpaII and the methylation-insensitive enzyme MspI followed by analysis of the digestion product using gel electrophoresis.

#### *Quantitative RT-PCR*

cDNA was generated by reverse transcription of total RNA using oligo(dT)<sub>20</sub> primers and SuperScript III Reverse Transcriptase (Thermo Fisher Scientific). Quantitative real-time PCR was performed with platinum SYBR green qPCR superMix-UDG (Thermo Fisher Scientific) on a Step One instrument (Thermo Fisher Scientific). Relative transcript levels were normalized to *hRPS18* using the  $\Delta\Delta CT$  method. To quantify the different *FOXP1* isoforms, specific primer pairs for each isoform were used as well as the *hRPS18* primers as a reference.

Primer	Sequence 5' - 3'
<b>Cloning of luciferase vectors</b>	
Foxp1-DMR forward	GACGAGATCTGATTTGTACCCAAG
Foxp1-DMR reverse	GACGAGATCTATTATAGCAACATAACATTTAAC
<b>Primers used for qPCR</b>	
Foxp1 Short Isoform forward	TGCCTCCTCACCATGAACGG
Foxp1 Short Isoform reverse	CTGGATGGCAAGGCTTCTCC
Foxp1 Long Isoform forward	GTGAAAATTGCCTCTCCCGC
Foxp1 Long Isoform reverse	CTGGATGGCTGAACCGTTACT
Foxp1 Isoform 5 forward	CAGTGAGGCTGCTGAAGGTTT
Foxp1 Isoform 5 reverse	CTGGATGGCTGAACCGTTACT
hRPS18 forward	ATTAAGGGTGTGGGCCGAAG
hRPS18 reverse	GGAGCTTGTTGCCAGACCA



## ***Selection of 144 genes associated with Tmem cell differentiation and -function from recent literature***

The following articles were used: Best et al., 2013; Bottcher and Knolle, 2015; Caserta and Zamoyska, 2007; Chang et al., 2014; Gray et al., 2014; Huber and Lohoff, 2014; Kaech and Cui, 2012; Kurtulus et al., 2012; Lochner et al., 2015; Oberdoerffer et al., 2008; Sprent and Surh, 2001; Thaventhiran et al., 2013; Tsukumo et al., 2013; Weng et al., 2012.

## **Supplemental References:**

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* *31*, 166-169.

Assenov, Y., Muller, F., Lutsik, P., Walter, J., Lengauer, T., and Bock, C. (2014). Comprehensive analysis of DNA methylation data with RnBeads. *Nature methods* *11*, 1138-1140.

Best, J.A., Blair, D.A., Knell, J., Yang, E., Mayya, V., Doedens, A., Dustin, M.L., Goldrath, A.W., and Immunological Genome Project, C. (2013). Transcriptional insights into the CD8(+) T cell response to infection and memory T cell formation. *Nat Immunol* *14*, 404-412.

Bottcher, J., and Knolle, P.A. (2015). Global transcriptional characterization of CD8+ T cell memory. *Semin Immunol* *27*, 4-9.

Caserta, S., and Zamoyska, R. (2007). Memories are made of this: synergy of T cell receptor and cytokine signals in CD4(+) central memory cell survival. *Trends Immunol* *28*, 245-248.

Chang, J.T., Wherry, E.J., and Goldrath, A.W. (2014). Molecular regulation of effector and memory T cell differentiation. *Nat Immunol* *15*, 1104-1115.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* *29*, 15-21.

Friedman, R.C., Farh, K.K., Burge, C.B., and Bartel, D.P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* *19*, 92-105.

Gray, S.M., Kaech, S.M., and Staron, M.M. (2014). The interface between transcriptional and epigenetic control of effector and memory CD8(+) T-cell differentiation. *Immunol Rev* *261*, 157-168.

Hovestadt, V., Jones, D.T., Picelli, S., Wang, W., Kool, M., Northcott, P.A., Sultan, M., Stachurski, K., Ryzhova, M., Warnatz, H.J., *et al.* (2014). Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. *Nature* *510*, 537-541.

Huber, M., and Lohoff, M. (2014). IRF4 at the crossroads of effector T-cell fate decision. *Eur J Immunol* *44*, 1886-1895.

Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* *18 Suppl 1*, S96-104.

Kent, W.J., Zweig, A.S., Barber, G., Hinrichs, A.S., and Karolchik, D. (2010). BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* *26*, 2204-2207.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* *14*, R36.

Kurtulus, S., Tripathi, P., and Hildeman, D.A. (2012). Protecting and rescuing the effectors: roles of differentiation and survival in the control of memory T cell development. *Front Immunol* *3*, 404.

- Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods* 9, 357-359.
- Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754-1760.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078-2079.
- Liu, Y., Siegmund, K.D., Laird, P.W., and Berman, B.P. (2012). Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome biology* 13, R61.
- Lochner, M., Berod, L., and Sparwasser, T. (2015). Fatty acid metabolism in the regulation of T cell function. *Trends Immunol* 36, 81-91.
- Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* 15, 550.
- Oberdoerffer, S., Moita, L.F., Neems, D., Freitas, R.P., Hacohen, N., and Rao, A. (2008). Regulation of CD45 alternative splicing by heterogeneous ribonucleoprotein, hnRNPLL. *Science* 321, 686-691.
- Ramirez, F., Dundar, F., Diehl, S., Gruning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* 42, W187-191.
- Sprent, J., and Surh, C.D. (2001). Generation and maintenance of memory T cells. *Curr Opin Immunol* 13, 248-254.
- Thaventhiran, J.E., Fearon, D.T., and Gattinoni, L. (2013). Transcriptional regulation of effector and memory CD8+ T cell fates. *Curr Opin Immunol* 25, 321-328.
- Tsukumo, S., Unno, M., Muto, A., Takeuchi, A., Kometani, K., Kurosaki, T., Igarashi, K., and Saito, T. (2013). Bach2 maintains T cells in a naive state by suppressing effector memory-related genes. *Proc Natl Acad Sci U S A* 110, 10735-10740.
- Wang, Q., Gu, L., Adey, A., Radlwimmer, B., Wang, W., Hovestadt, V., Bahr, M., Wolf, S., Shendure, J., Eils, R., *et al.* (2013b). Tagmentation-based whole-genome bisulfite sequencing. *Nat Protoc* 8, 2022-2032.
- Weng, N.P., Araki, Y., and Subedi, K. (2012). The molecular basis of the memory T cell response: differential gene expression and its epigenetic regulation. *Nat Rev Immunol* 12, 306-315.

## Chapter 4

### **DNA methylation signatures of monozygotic twins clinically discordant for multiple sclerosis**

The full text of this chapter was originally published as:

Souren, N. Y., Gerdes, L. A., Lutsik, P., Gasparoni, G., Beltrán, E., Salhab, A., Kümpfel, T., Weichenhan, D., Plass, C., Hohlfeld, R., and Walter, J. (2019). DNA methylation signatures of monozygotic twins clinically discordant for multiple sclerosis. *Nature Communications*, 10(1):2094.

under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>).

The author of this thesis contributed to data analysis, specifically to PMDs analysis using WGBS data and he generated the supplementary figures S7 and S8. He contributed to manuscript editing and reviewing.

ARTICLE

<https://doi.org/10.1038/s41467-019-09984-3>

OPEN

# DNA methylation signatures of monozygotic twins clinically discordant for multiple sclerosis

Nicole Y. Souren<sup>1</sup>, Lisa A. Gerdes<sup>2</sup>, Pavlo Lutsik<sup>3</sup>, Gilles Gasparoni<sup>1</sup>, Eduardo Beltrán<sup>2</sup>, Abdulrahman Salhab<sup>1</sup>, Tania Kümpfel<sup>2</sup>, Dieter Weichenhan<sup>3</sup>, Christoph Plass<sup>3</sup>, Reinhard Hohlfeld<sup>2,4</sup> & Jörn Walter<sup>1</sup>

Multiple sclerosis (MS) is an inflammatory, demyelinating disease of the central nervous system with a modest concordance rate in monozygotic twins, which strongly argues for involvement of epigenetic factors. We observe highly similar peripheral blood mononuclear cell-based methylomes in 45 MS-discordant monozygotic twins. Nevertheless, we identify seven MS-associated differentially methylated positions (DMPs) of which we validate two, including a region in the *TMEM232* promoter and *ZBTB16* enhancer. In CD4+ T cells we find an MS-associated differentially methylated region in *FIRRE*. Additionally, 45 regions show large methylation differences in individual pairs, but they do not clearly associate with MS. Furthermore, we present epigenetic biomarkers for current interferon-beta treatment, and extensive validation shows that the *ZBTB16* DMP is a signature for prior glucocorticoid treatment. Taken together, this study represents an important reference for epigenomic MS studies, identifies new candidate epigenetic markers, and highlights treatment effects and genetic background as major confounders.

<sup>1</sup>Department of Genetics/Epigenetics, Saarland University, 66123 Saarbrücken, Germany. <sup>2</sup>Institute of Clinical Neuroimmunology, University Hospital and Biomedical Center, Ludwig-Maximilians University Munich, 81377 Munich, Germany. <sup>3</sup>Division of Cancer Epigenomics, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany. <sup>4</sup>Munich Cluster for Systems Neurology (SyNergy), 80336 Munich, Germany. Correspondence and requests for materials should be addressed to N.Y.S. (email: [nicole.souren@hotmail.com](mailto:nicole.souren@hotmail.com)) or to J.W. (email: [j.walter@mx.uni-saarland.de](mailto:j.walter@mx.uni-saarland.de))

**M**ultiple sclerosis (MS), a leading cause of neurological disability in young adults, is considered to be an auto-immune disease, characterized by chronic inflammatory demyelination of the central nervous system<sup>1,2</sup>. Although nuclear genetic factors contribute to the development of MS<sup>3</sup>, a maximum concordance rate for MS in monozygotic (MZ) twins of 25%<sup>4,5</sup>, indicates that interaction with other risk factors is compulsory for clinical symptoms to develop. While various studies suggested mitochondrial DNA variants as plausible MS susceptibility factors, we recently showed that mitochondrial DNA variation (e.g., skewed heteroplasmy) does not play a major role in the discordant clinical manifestation of MS in MZ twins<sup>6</sup>.

DNA methylation differences represent another source of molecular variation that can cause discordant phenotypes within MZ twins<sup>7–13</sup>. As DNA methylation changes can cause transcriptional alterations, aberrant DNA methylation has been observed in various human diseases<sup>14,15</sup>. Discordant DNA methylation profiles within MZ twins have been reported quite frequently at imprinted regions<sup>7–9</sup>, which are characterized by parent-of-origin-specific methylation patterns resulting in mono-allelic expression. As a maternal parent-of-origin effect in MS susceptibility has been reported<sup>16,17</sup>, and several imprinted genes have been linked to immune system development and functioning (reviewed by Ruhrmann et al.<sup>18</sup>), genomic imprinting errors might be involved in the pathogenesis of MS<sup>18</sup>. Additionally, environmental risk factors such as smoking, history of symptomatic Epstein-Barr virus infection, and vitamin D deficiency have been associated with an increased MS risk<sup>19–21</sup>. Although the molecular mechanisms underlying these associations remain unknown, evidence that these environmental factors can induce DNA methylation changes is accumulating<sup>22–25</sup>.

Thus far, several epigenome-wide association studies (EWAS) for MS have been carried out<sup>26–31</sup>, and a number of differentially methylated CpG positions (DMPs) have been reported, including DMPs in the *HLA-DRB1* locus. Although these studies used the same array platform (i.e., Infinium HumanMethylation450 (450 K)), the results are inconsistent. Since these studies used genetically unmatched cases and controls, they are potentially hampered by DNA sequence variation. As genetic factors predispose to MS, these studies cannot determine whether MS is due to genetic or epigenetic susceptibility. In addition, SNP-containing probes give rise to biased DNA methylation measurements<sup>32</sup>, and DNA methylation changes are also often the

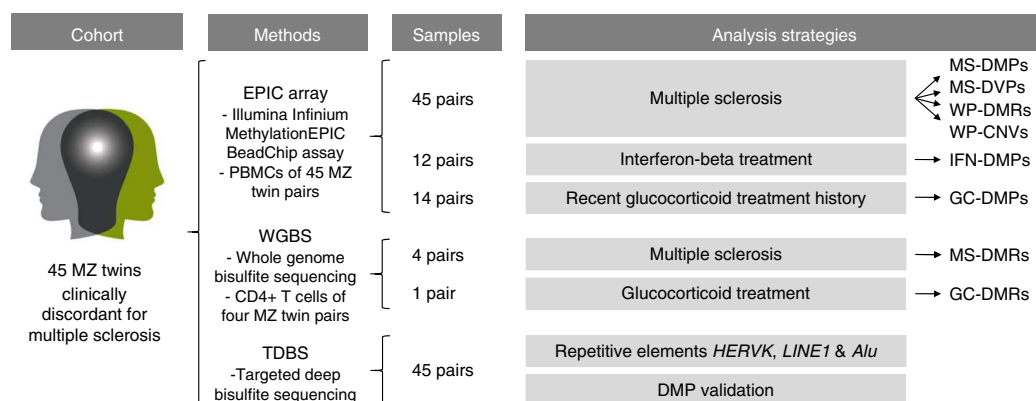
result of *cis*- or *trans*-acting genetic variants (methylation quantitative trait loci or mQTLs)<sup>33</sup>. A MZ twin-based design controls for these genetic differences and for other factors (potentially affecting the methylome, including gender, age, and a broad range of environmental factors). Thus far, one EWAS in MS-discordant MZ twins has been reported, but no DNA methylation differences were identified<sup>34</sup>. Since this study included only three pairs and exclusively aimed at identifying very large methylation differences (i.e.,  $\geq 80\%$  methylation in one co-twin and  $\leq 20\%$  in the other), further studies in larger cohorts are required.

Here, we describe an EWAS comprising a unique cohort of 45 MZ twins clinically discordant for MS in which we aim to identify MS-associated DNA methylation changes in peripheral blood mononuclear cells (PBMCs) and to study the effect of MS treatments on the methylome (Fig. 1). Although we confirm that MS-discordant MZ twins have very similar methylomes, we identify a few new MS-associated candidate loci and observe DNA methylation changes associated with current interferon-beta (IFN) and prior glucocorticoid (GC) treatment.

## Results

**PBMC-based methylomes.** PBMCs of 46 MZ twins clinically discordant for MS were accessible and genome-wide DNA methylation profiles were established using Illumina Infinium MethylationEPIC BeadChips (EPIC arrays). After quality control and filtering, methylation data of 849,832 sites were available for 45 twin pairs. As expected, within-pair array-wide correlation coefficients were very high (mean = 0.995), indicating high-quality data. Clinical characteristics of the 45 MS-discordant MZ twins are shown in Table 1 and Supplementary Fig. 1.

**Detection and validation of MS-DMPs in *TMEM232* and *ZBTB16*.** To identify DMPs associated with the clinical manifestation of MS (MS-DMPs), first a pair-wise analysis was carried out on the EPIC array data of the 45 pairs without adjusting for cell-type composition (Supplementary Fig. 2a, b). The Q-Q plot in Supplementary Fig. 2b shows that the obtained *p*-values (Wilcoxon signed-rank test) clearly deviate from the null expectation. This inflation was eliminated after adjusting for cell-type composition (Fig. 2a, b), indicating that many differences are due to variation in cellular composition.



**Fig. 1** Schematic overview of the study design and analysis strategies. DMPs, differentially methylated CpG positions; DMRs, differentially methylated regions; DVP, differentially variable CpG positions; GC-DMPs, glucocorticoid treatment-associated DMPs; HERVK, human endogenous retrovirus type K; IFN-DMPs, interferon-beta treatment-associated DMPs; LINE1, long interspersed nuclear element-1; MS-DMPs, multiple sclerosis-associated DMPs; MS-DVPs, multiple sclerosis-associated DVPs; MZ, monozygotic; PBMCs, peripheral blood mononuclear cells; WP-CNVs, within-pair copy-number variations; WP-DMRs, within-pair differentially methylated regions. The logo of the MS/TWIN/STUDY is not covered by the article CC BY license. Image credit goes to Lisa Ann Gerdes. All rights reserved, used with permission

**Table 1 Characteristics of the MZ twins clinically discordant for MS**

Characteristic	MS-affected MZ co-twins	Non-affected MZ co-twins	Range	$p^b$
Number of pairs	45	45		
Gender (female/male)	32/13	32/13		
Age at study entry (years)	42.3 ± 12.1	42.3 ± 12.1	(21–67)	
Age of disease onset (years) <sup>a</sup>	27.9 ± 8.4		(14–46)	
Years clinically discordant for MS at sample collection <sup>a</sup>	15.3 ± 11.1		(1–45)	
EDSS at study entry	3.3 ± 2.3		(0–9.5)	
Pairs longer than 10 years clinically discordant for MS	25 (56%)			
Pairs with a positive family history of MS	13 (29%)			
MS type				
– RRMS	31 (69%)			
– SPMS	12 (27%)			
– PPMS	2 (4%)			
Smoking status				
Smoking at disease onset	23 (51%)	19 (42%)		0.53
Pack-years at disease onset	0.03 (0–3.5)	0 (0–3.8)		0.81
Smoking at sample collection	14 (31%)	12 (27%)		0.82
Pack-years at sample collection	0.6 (0–10.8)	0 (0–6.3)		0.24

Continuous data expressed as: mean ± standard deviation or median (interquartile range). Categorical data expressed as: number of observations (%).  
EDSS Expanded Disability Status Scale, PPMS primary-progressive MS, RRMS relapsing-remitting MS, SPMS secondary-progressive MS  
<sup>a</sup>See Supplementary Fig. 1 for boxplots (with all data points) showing the distribution of the age of disease onset and the years that the MZ twins were clinically discordant for MS at sample collection  
<sup>b</sup>MS-affected versus non-affected MZ co-twins calculated using a two-tailed Wilcoxon signed-rank test for continuous data and two-tailed Fisher's exact test for categorical data

Mean within-pair  $\beta$ -value differences ( $\Delta\beta$ -values) were small (Fig. 2a and Supplementary Fig. 2a). The largest differences were observed for *ECT2* (cg12393503), *SELP3* (cg02520593), and *IL34* (cg01447350), with mean  $\Delta\beta$ -values of 0.15, 0.06, and  $-0.09$ , respectively, but they did not reach statistical significance. In several twins, these CpGs showed very large  $\Delta\beta$ -values ( $\sim 0.8$ ) (Supplementary Fig. 3), but these differences were not confirmed by validation using targeted, deep bisulfite sequencing (TDBS) (Supplementary Fig. 4). This indicates that some EPIC probes are prone to technical artefacts, as reported by others<sup>35</sup>, and that validation using independent assays is required.

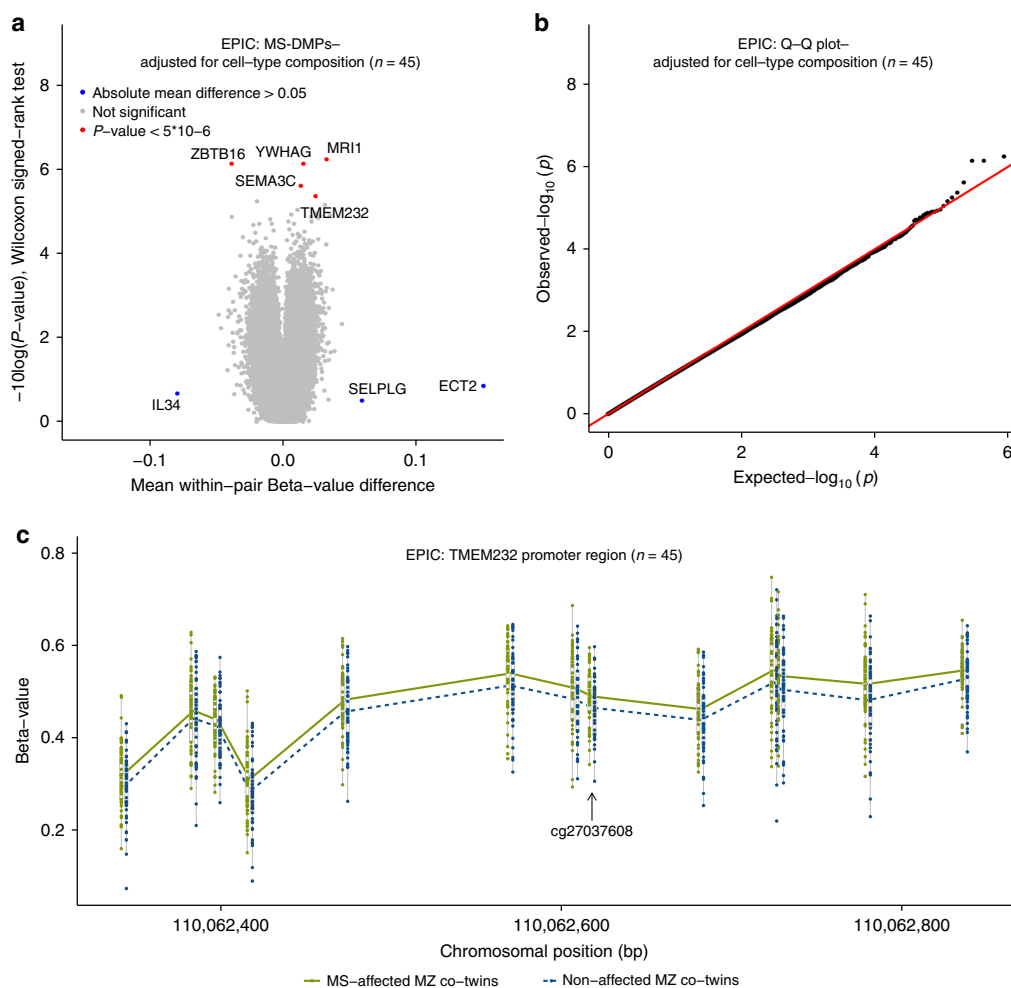
The unadjusted analysis revealed 39 MS-DMPs with a suggestive  $p < 5 \times 10^{-6}$  (Wilcoxon signed-rank test). After correcting for multiple testing six MS-DMPs remained genome-wide significant (false discovery rate (FDR)  $< 0.05$ ) (Supplementary Fig. 2a). After adjusting for cell-type composition, no MS-DMP had FDR  $< 0.05$ , but five MS-DMPs had a suggestive  $p < 5 \times 10^{-6}$  (Fig. 2a and Table 2). One of these MS-DMPs is located in the promoter of the *TMEM232* gene (cg27037608, mean  $\Delta\beta$ -value = 0.024), encoding for a transmembrane protein of unknown function. Genetic variants in *TMEM232* have been associated with atopic dermatitis and allergic rhinitis in GWAS<sup>36,37</sup>. For this MS-DMP, EPIC array data for 12 neighboring CpGs were also available, which all showed a similar effect, and  $p < 0.01$  was calculated for eight CpGs (Fig. 2c). A second solitary MS-DMP was observed in the gene body of *SEMA3C* (cg00232450, mean  $\Delta\beta$ -value = 0.013), which has been suggested to promote dendritic cell migration during innate and adaptive immune responses, and to be involved in axonal guidance and growth<sup>38</sup>. A third MS-DMP is located in the *YWHAG* gene (cg01708711, mean  $\Delta\beta$ -value = 0.015), which has been associated with MS severity in a GWAS<sup>39</sup>. This MS-DMP has five neighboring CpGs on the array, but all are non-significant, questioning the significance of this MS-DMP. A fourth MS-DMP (cg25345365) showed the largest methylation difference (mean  $\Delta\beta$ -value =  $-0.039$ ) and is located in an enhancer within *ZBTB16*, which has been reported to be essential for natural killer T (NKT) cell development<sup>40</sup>. The fifth MS-DMP (cg25755428, mean  $\Delta\beta$ -value = 0.033) is located in the *MRII* gene, in which mutations have been associated with vanishing white matter disease<sup>41</sup>. The  $\beta$ -value distribution of this and

neighboring CpGs suggests that this concerns a mQTL (Table 2). Additional adjustment for smoking status did not alter the  $p$ -values of these 5 MS-DMPs, indicating that they are not confounded by smoking status.

Next, all 698 MS-DMPs with  $p < 0.001$  (Wilcoxon signed-rank test) after adjusting for cell-type composition were functionally annotated using the GREAT tool, which assigns biological meaning to a set of non-coding genomic regions by analyzing the annotations of nearby genes<sup>42</sup>. This analysis revealed that *TMEM232* is enriched for MS-DMPs. Other annotation categories were not significant.

Based on their significance, effect size and/or whether neighboring probes were also differentially methylated, the *TMEM232* (cg27037608) and *ZBTB16* (cg25345365) MS-DMPs were selected for validation using TDBS. The TDBS data correlated highly with the array data ( $r_{\text{Pearson-TMEM232}} = 0.84$  and  $r_{\text{Pearson-ZBTB16}} = 0.89$ , Supplementary Figs. 5a and 6a), and both MS-DMPs as well as the surrounding CpGs were significantly differentially methylated between the MS-discordant twins (Supplementary Figs. 5b and 6b). This confirms that these MS-DMPs represent true effects in our cohort.

***TMEM232* and *ZBTB16* MS-DMPs associated with long-standing MS.** To verify whether the identified methylation differences are dependent on disease duration, we performed a pairwise analysis including only the EPIC array data of the 25 pairs that have been clinically discordant for MS longer than 10 years (Supplementary Table 1). In the analysis adjusted for cell-type composition, two DMPs had a suggestive  $p < 5 \times 10^{-6}$  (Wilcoxon signed-rank test), one of which is located in the promoter of the *TACSTD2* gene encoding tumor-associated calcium signal transducer 2 (mean  $\Delta\beta$ -value =  $-0.022$ )<sup>43</sup>. The second DMP is located in the promoter of the *RCL1* gene (mean  $\Delta\beta$ -value =  $-0.011$ ), which has been linked to depression<sup>44</sup>. For these two long-standing MS-DMPs, EPIC array data of neighboring CpGs were available ( $< 500$  bp), and two neighboring CpGs of the *TACSTD2* MS-DMP showed a similar trend ( $p < 0.10$ ). Additionally, the *TMEM232* MS-DMPs cg27037608 (mean  $\Delta\beta$ -value = 0.026,  $p = 3.2 \times 10^{-5}$ ) and cg26583412 (mean  $\Delta\beta$ -value = 0.038,  $p = 1.8 \times 10^{-5}$ ) were among the top 15 most significant DMPs associated with long-standing MS.



**Fig. 2** DNA methylation changes associated with the clinical manifestation of MS. Results of the differential DNA methylation analysis including the EPIC array data of the 45 MZ twin pairs clinically discordant for MS. **a** Volcano plot of the  $p$ -values resulting from the nonparametric two-tailed Wilcoxon signed-rank test against the mean within-pair  $\beta$ -value difference for each CpG. Data were adjusted for cell-type composition. **b** Q-Q plot of the  $p$ -values resulting from the nonparametric two-tailed Wilcoxon signed-rank test shown in Fig. 2a. Data were adjusted for cell-type composition. Within-pair  $\beta$ -value difference ( $\Delta\beta$ -value) = clinically MS-affected MZ co-twin—non-affected MZ co-twin. **c** Overview of the *TMEM232* promoter region. Data are presented as Tukey boxplots including the individual data points that represent the (adjusted)  $\beta$ -values of the significant MS-associated differentially methylated CpG position (MS-DMP) cg27037608 and 12 neighboring CpGs present on the EPIC array. The lines connect the mean methylation values of each CpG site for the MS-affected and clinically non-affected MZ co-twins separately. Boxplots represent the interquartile range or IQR (bottom and top of the box) and 1.5 times the IQR (whiskers). Source data are provided as a Source Data file.  $n$  = number of twin pairs

Furthermore, the *ZBTB16* MS-DMP had a mean  $\Delta\beta$ -value difference of  $-0.036$  ( $p = 0.002$ , Supplementary Table 1). Additional adjustment for smoking status did not change the results. Hence, the *TMEM232* and *ZBTB16* MS-DMPs are also associated with long-standing MS.

**Evaluation of the *TMEM232* MS-DMPs in a case-control cohort.** Next, we evaluated the selected MS-DMPs in whole blood-based 450 K EWAS data of 140 unrelated MS patients and 139 controls from Kular et al.<sup>31</sup>. Unfortunately, the cg25345365 *ZBTB16*, cg27037608 and cg26583412 *TMEM232* MS-DMPs were not present on the 450 K array, but data from seven neighboring CpGs in *TMEM232* were available. Although none of these CpGs were significantly differentially methylated between the MS cases and controls ( $p > 0.05$ , linear regression), methylation levels were always higher in the MS patients, confirming the directionality of the effect observed in the twins (Supplementary Table 2).

**Whole-genome bisulfite sequencing reveals MS-DMR in *FIRRE*.** To identify additional MS-associated differentially methylated regions (MS-DMRs), we performed whole-genome bisulfite sequencing (WGBS) for a subset of four MS-discordant female twin pairs on CD4<sup>+</sup> memory T cells, which have been implicated in the pathogenesis of MS<sup>45</sup>. First, genome-wide DNA methylation changes were evaluated by identifying and comparing partially methylated domains (PMD), fully methylated regions (FMRs), low methylated regions (LMRs), and unmethylated regions (UMRs). However, no significant differences were observed between the MS-discordant co-twins ( $p > 0.05$ , paired  $t$ -test) (Supplementary Figs. 7, 8).

Next, a DMR analysis was carried out and MS-DMRs were defined as  $\geq 3$  CpGs (max. distance 500 bp), each having  $p < 0.05$  (paired  $t$ -test) and an absolute mean methylation difference  $> 0.2$ . The DMR analysis revealed a prominent MS-DMR located in an intronic CTCF/YY1 bound regulatory region in *FIRRE*, which

**Table 2 DMPs associated with the clinical manifestation of MS (n = 45 twin pairs)<sup>a</sup>**

Probe ID	Gene/ Location <sup>b</sup>	Functional region <sup>c</sup>	$\beta$ -value MS	(U/A) non-MS	$\Delta\beta$ value (95% CI) (U/A)	$\beta$ -value range	$p_{W-U}/p_{W-A}$	$FDR_{W-U}/$ $FDR_{W-A}$	Close probes <sup>d</sup>	450 K	Full name & reported function
cg27037608	<i>TMEM232</i> / chr5: 110062618	TSS200/ TFBS	0.488/ 0.489	0.466/ 0.465	0.022 (0.012,0.032)/ 0.024 (0.014,0.034)	0.31-0.59	$4.8 \times 10^{-5}/$ $4.3 \times 10^{-6}$	0.13/ 0.74	13 within 500 bp, 8 with $p$ < 0.01	N	Transmembrane protein 232: associated with atopic dermatitis and allergic rhinitis in GWAS <sup>36,37</sup>
cg00232450	<i>SEMA3C</i> / chr7: 80421169	Body/DHS	0.813/ 0.810	0.793/ 0.797	0.020 (0.012,0.027)/ 0.013 (0.008,0.019)	0.73-0.86	$1.6 \times 10^{-7}/$ $2.5 \times 10^{-6}$	0.03/ 0.52	0 within 2 kb	N	Semaphorin 3 C: involved in axonal guidance and growth. Promotes dendritic cell migration during innate and adaptive immune responses <sup>38</sup>
cg01708711	<i>YWHAG</i> / chr7: 75959031	Body/CpG island/ TFBS	0.855/ 0.854	0.839/ 0.839	0.016 (0.011,0.021)/ 0.015 (0.010,0.020)	0.79-0.89	$1.8 \times 10^{-7}/$ $7.3 \times 10^{-7}$	0.03/ 0.21	5 within 350 bp, $p > 0.01$	Y	Tyrosine 3- monooxygenase/ tryptophan 5-mono- oxygenase activation protein, gamma: associated with MS severity in GWAS <sup>39</sup>
cg25345365	<i>ZBTB16</i> / chr11: 114050114	Body/ DHS/ FANTOM5 enhancer	0.540/ 0.544	0.587/ 0.583	-0.047 (-0.063, -0.031)/ -0.039 (-0.053, -0.024)	0.36-0.72	$1.5 \times 10^{-7}/$ $7.3 \times 10^{-7}$	0.03/ 0.21	0 within 2 kb	N	Zinc Finger And BTB Domain Containing 16: transcription factor essential for NKT cell development <sup>40</sup>
cg25755428	<i>MRI1</i> / chr19: 13875111	TSS1500/ CpG island/ DHS	0.328/ 0.336	0.311/ 0.303	0.017 (0.006,0.027)/ 0.033 (0.021,0.044)	0.05-0.89 <sup>e</sup>	$8.6 \times 10^{-4}/$ $5.8 \times 10^{-7}$	0.17/ 0.21	mQTL <sup>e</sup>	Y	Methylthioribose-1- phosphate isomerase 1: includes mutation associated with vanishing white matter disease <sup>41</sup>

Source data are provided as a Source Data file

<sup>a</sup>A adjusted for cell-type composition, CI confidence interval, DHS DNase I hypersensitive site,  $FDR_{W-A}$  FDR two-tailed Wilcoxon signed-rank test adjusted for cell-type composition,  $FDR_{W-U}$  FDR two-tailed Wilcoxon signed-rank test unadjusted for cell-type composition, GWAS genome-wide association study,  $p_{W-A}$  p-value two-tailed Wilcoxon signed-rank test adjusted for cell-type composition,  $p_{W-U}$  p-value two-tailed Wilcoxon signed-rank test unadjusted for cell-type composition, TFBS transcription factor-binding site, TSS200 the region from transcription start site (TSS) to -200 nt upstream of TSS, TSS1500 -200 to -1500 nt upstream of TSS, U unadjusted for cell-type composition, 450 K CpG present on 450 K array (N no, Y yes),  $\Delta\beta$ -value within-pair  $\beta$ -value difference (clinically MS-affected MZ co-twin-non-affected MZ co-twin)

<sup>b</sup>Listed are the five MS-DMPs with a suggestive  $p < 5 \times 10^{-6}$  (two-sided Wilcoxon signed-rank test) in the pair-wise analysis carried out using the EPIC array data of the 45 MZ twin pairs adjusted for cell-type composition

<sup>c</sup>Genome coordinates are human genome build GRCh37/hg19

<sup>d</sup>Based on information provided by the Illumina manifest

<sup>e</sup>Number of EPIC probes mapping close to the DMPs are listed and whether these probes have a  $p < 0.01$  (two-sided Wilcoxon signed-rank test)

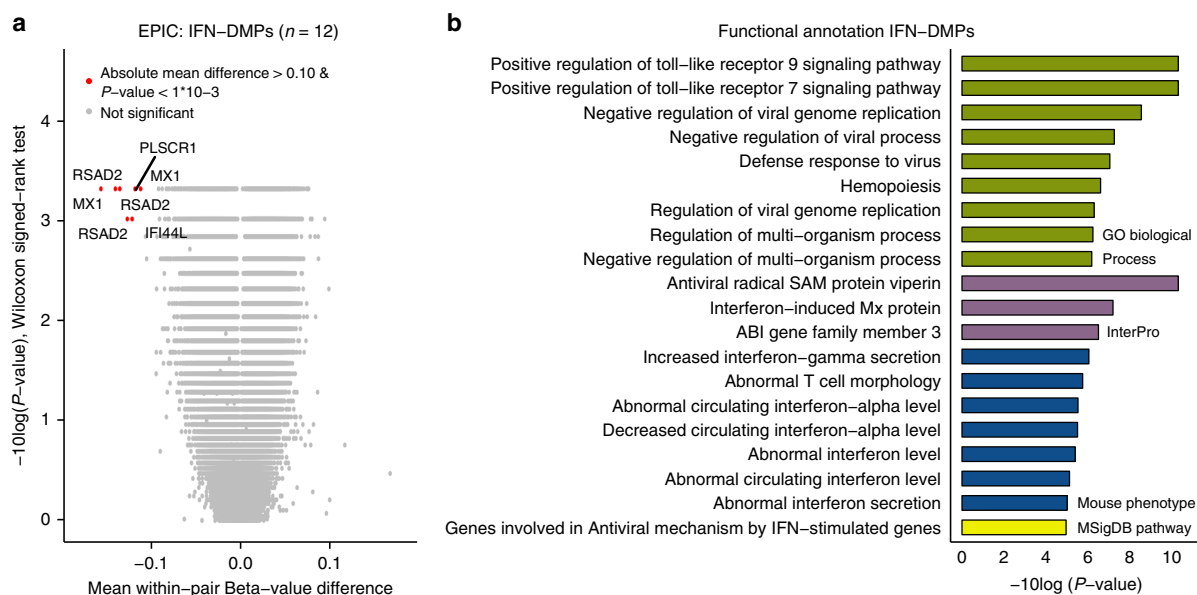
<sup>f</sup>Behaves like a methylation quantitative trait loci

is located on the X-chromosome (chrX:130863481-130863509) and encodes a circular, long, non-coding RNA (Supplementary Figs. 9 and 10)<sup>46</sup>. This MS-DMR is not covered by the EPIC array, but a probe (cg08117231) located 6 bp upstream of this DMR was not significant in the PBMC-based EWAS, nor in the females-only analysis ( $p > 0.05$ , Wilcoxon signed-rank test). When performing the analysis using a less robust methylation difference of  $>0.15$ , then 19 additional MS-DMRs were identified (Supplementary Table 3). Five of these were also covered by the EPIC array, but were not significant in the PBMC-based EWAS. Only 11 of these additional MS-DMRs showed overall consistent methylation differences across the entire DMR, including an MS-DMR in the *DDAH1* gene that also contains an established MS-associated SNP<sup>3</sup>. Unfortunately, the *TMEM232* and *ZBTB16* loci did not fulfill the filtering criterion of  $\geq 10$  reads coverage across all samples, but note that they were validated by TDBS.

**Within-pair DMPs are common among MZ twins.** Our PBMC-based EWAS concentrated on the identification of MS-DMPs showing differences across many twin pairs. However, since MS is a heterogeneous disease, DNA methylation changes present in only a few cases should also be considered. Therefore the EPIC

array data was used to identify within-pair DMRs (WP-DMRs). To detect robust methylation changes in individual twin pairs, WP-DMRs were defined as  $\geq 3$  CpGs within 1 kb having a  $\Delta\beta$ -value (adjusted for cell-type composition)  $> 0.20$  and the aberrant methylated co-twin a  $\beta$ -value greater than  $\pm 3$  standard deviations from the mean. Overall, 45 WP-DMRs were identified in 17 of the 45 twin pairs, ranging from one to 11 WP-DMRs per pair (Supplementary Table 4 and Supplementary Figs. 11-14). Of the 45 WP-DMRs, 43 were solitary and pair-specific and only two WP-DMRs (*ISOC2* and *HIST1H3E*) were found in two independent pairs (Supplementary Fig. 11), but the aberrant methylation pattern did not correlate with the MS phenotype (Supplementary Table 4). Of the 43 pair-specific WP-DMRs, 16 showed an aberrant methylation pattern in the non-affected co-twin and 27 in the MS-affected co-twin. These WP-DMRs have not been associated with MS in other EWAS, nor do they overlap with MS-associated genes reported in the GWAS Catalog (accessed May 2018)<sup>47</sup>. Two pair-specific WP-DMRs were located in reported imprinted DMRs<sup>48</sup> in *SVOPL* and *HML13/MCTS2P* (Supplementary Fig. 12), but in both cases the non-affected co-twin showed an abnormal methylation pattern. Lowering the WP-DMR  $\Delta\beta$ -value threshold to 0.15 revealed that of the 27 WP-DMRs, which were aberrantly methylated in the MS-affected





**Fig. 3** Interferon-beta (IFN) treatment-associated DNA methylation changes. **a** Results of the differential DNA methylation analysis including only the EPIC array data of the 12 pairs, of which the MS-affected MZ co-twin was treated with IFN at the moment of blood collection. The volcano plot presents the  $p$ -values resulting from the nonparametric two-tailed Wilcoxon signed-rank test vs. the mean within-pair  $\beta$ -value difference for each CpG. Data were not adjusted for cell-type composition. Within-pair  $\beta$ -value difference ( $\Delta\beta$ -value) = MS-affected IFN-treated MZ co-twin - clinically non-affected MZ co-twin.  $n$  = number of twin pairs. **b** Summary of the functional annotation analysis using GREAT<sup>42</sup>, on the 257 IFN-associated differentially methylated CpG positions (IFN-DMPs) (absolute mean within-pair  $\beta$ -value difference  $>0.05$  and two-sided Wilcoxon signed-rank test  $p < 0.001$ ). Annotation terms are ranked according to their enrichment  $p$ -values calculated by GREAT<sup>42</sup>. GO Biological Process terms (Hyper raw  $p < 1 \times 10^{-6}$ ) and the other presented terms (Hyper raw  $p < 1 \times 10^{-5}$ )

co-twins, four WP-DMRs were also identified in other twins. Of these one intergenic WP-DMR was present in four pairs and always hypermethylated in the MS-affected co-twins (Supplementary Table 5). Furthermore, among the 23 pair-specific WP-DMRs, which were aberrantly methylated in the MS-affected co-twins, four WP-DMRs were identified in one pair in the protocadherin gamma (*PCDHG*) gene cluster (Supplementary Fig. 13), and another was observed in the promoter of the non-clustered protocadherin 10 (*PCDH10*) gene (Supplementary Fig. 14). Protocadherins are highly expressed in the brain and involved in neuronal development<sup>49</sup>.

**Methylation variability not enhanced in MS-discordant twins.** Increased DNA methylation variability has been observed in MZ twins discordant for the autoimmune diseases type 1 diabetes (T1D) and rheumatoid arthritis (RA)<sup>12,13</sup>. Hence, we tested whether DNA methylation variability is also implicated in MS using the iEVORA algorithm<sup>50</sup>. Applying the default FDR  $< 0.001$  resulted in only 25 differentially variable CpG positions (DVPs) of which the majority (88%) was hypervariable in the non-affected co-twins (Supplementary Table 6 and Supplementary Fig. 15). Hence, our PBMC-based EPIC array data does not support the presence of an MS-associated DNA methylation variability signature in these MS-discordant MZ twins.

**IFN treatment induces robust DNA methylation changes.** Our study design also allows to identify MS treatment-related DNA methylation changes. In our cohort, IFN is the most common disease-modifying treatment, and although IFN-induced transcriptomic alterations in blood cells of MS patients have been studied previously<sup>51–53</sup>, DNA methylation changes have not been reported so far. We performed a pair-wise analysis including the EPIC array data of the 12 pairs of which the MS-affected co-twins were treated with IFN at blood collection. The mean  $\Delta\beta$ -values

were larger in this subcohort (Fig. 3a), as we identified 257 DMPs with an absolute mean  $\Delta\beta$ -value  $> 0.05$  and  $p < 0.001$  (Wilcoxon signed-rank test). None of the MS-DMPs listed in Table 2 were among these 257 IFN-associated DMPs (IFN-DMPs). The 257 IFN-DMPs were annotated to 212 genes, of which 124 genes (58%) overlap with IFN-regulated genes recorded in the INTERFEROME gene expression database (accessed May 2018)<sup>51</sup>. Functional annotation analysis revealed clear enrichment for genes involved in antiviral defense and interferon homeostasis (Fig. 3b). Moreover, seven IFN-DMPs had an absolute mean  $\Delta\beta$ -value  $> 0.10$  and  $p < 0.001$ , due to strong hypomethylation in the IFN-treated MS-affected co-twins (Supplementary Table 7 and Supplementary Fig. 16). These seven DMPs were located in *RSAD2* ( $n = 3$ ), *MX1* ( $n = 2$ ), *IFI44L* ( $n = 1$ ) and *PLSCR1* ( $n = 1$ ), i.e., genes reported to be up-regulated in blood cells of IFN-treated MS patients<sup>51–53</sup>. Although the estimated NK and B cells proportions differed significantly between the IFN-treated MS-affected and non-affected co-twins (Supplementary Table 8), adjusting the data for cell-type composition resulted in only a slight attenuation of the IFN-effect (Supplementary Table 7). Hence, our results indicate that these seven DMPs are robust markers for monitoring IFN treatment effects in PBMCs.

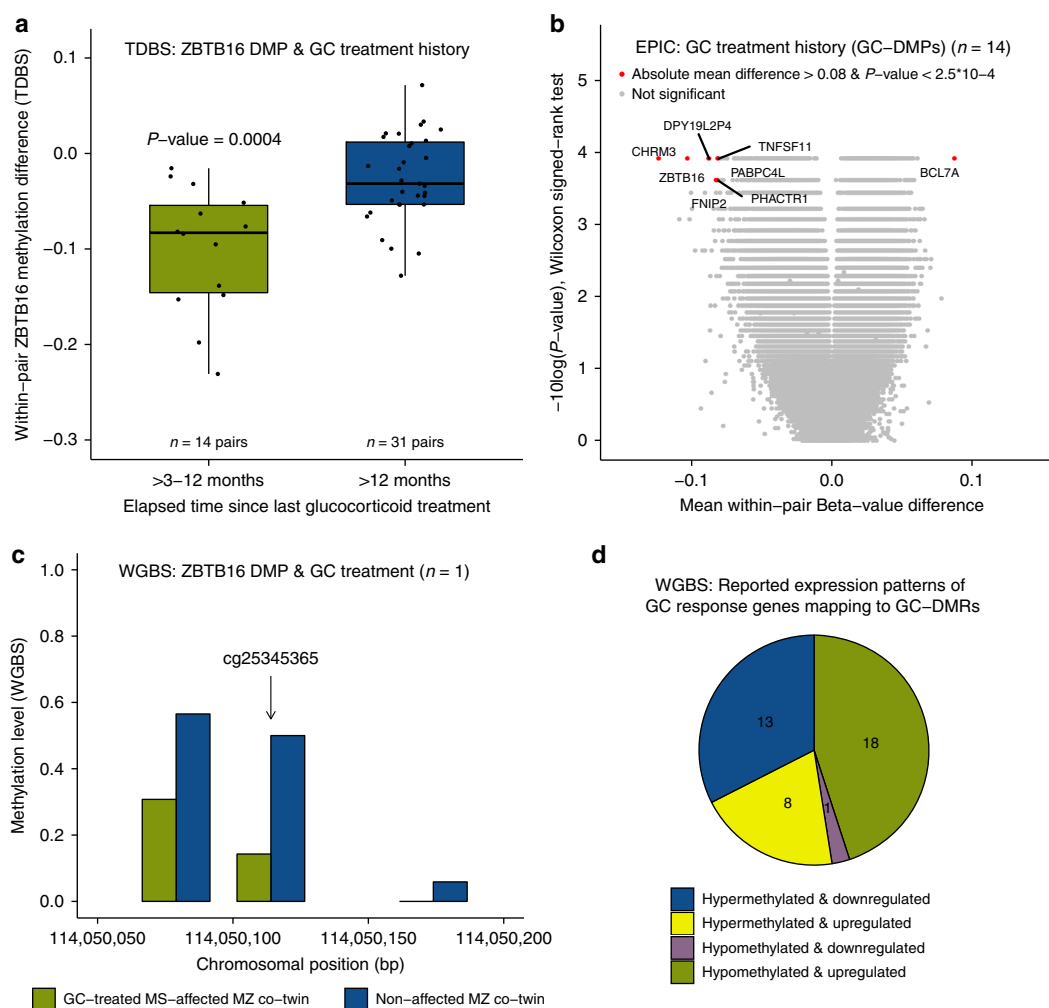
**GC treatment induces hypomethylation at *ZBTB16* enhancer DMP.** Among the MS-DMPs, the *ZBTB16* DMP (cg25345365) had the largest effect size ( $\sim 4\%$ ) and is located in an enhancer in intron 3 of *ZBTB16*, which encodes for a transcription factor also known as promyelocytic leukemia zinc finger (PLZF). *ZBTB16*/PLZF has been reported to be essential for NKT cell development<sup>40</sup>, and to contribute to T-helper 17 (Th17) cell differentiation and phenotype maintenance<sup>54</sup>. However, *ZBTB16* is also known as a major GC response gene, being highly up-regulated after GC exposure<sup>55</sup>, and several days of high-dose intravenous GC therapy is generally used to treat relapses in MS.

None of the MS-affected co-twins included in the array-based EWAS received GCs within three months prior blood collection. Nevertheless, GC treatment constitutes a serious confounder, because 43 of the 45 MS-affected co-twins have a GC treatment history (and the healthy co-twins not). Of these 14 received GCs within >3-12 months prior to blood collection (i.e. high-dose intravenous methylprednisolone (IVMP) 1 g/day for at least 3 days and on average 6 days). In those 14 pairs, the within-pair methylation differences at the *ZBTB16* DMP are significantly larger (more negative) than the pairs in which the MS-affected co-twin had received GCs longer than 12 months ago ( $p = 0.0004$ , Wilcoxon rank-sum test) (Fig. 4a and Supplementary Fig. 17).

This indicates that the strong association between the *ZBTB16* DMP and the MS phenotype is due to the GC treatment history of the MS-affected co-twins.

**GC-induced DMPs are not widespread in GC-response genes.**

Then, the EPIC array data of the 14 pairs of which the MS-affected co-twin had received GCs within >3-12 months prior to blood collection was analyzed to study the effect of recent GC treatment history on the PBMC methylomes (Fig. 4b). 320 potential GC-DMPs had an absolute mean  $\Delta\beta$ -value > 0.05 and  $P < 0.001$  (Wilcoxon signed-rank test) and were annotated to 279



**Fig. 4** Prior glucocorticoid (GC) treatment-associated DNA methylation changes. **a** Within-pair methylation differences of the *ZBTB16* DMP (cg25345365), determined using TDBS in the 14 pairs of which the MS-affected co-twin received GCs within >3-12 months prior blood collection, compared to the 31 pairs of which the MS-affected co-twin was treated with GCs more than 1 year ago. Boxplots represent the median (central line), the interquartile range or IQR (bottom and top of the box), and 1.5 times the IQR (whiskers).  $p$ -value = nonparametric two-tailed Wilcoxon rank-sum test result. Source data are provided as a Source Data file. **b** Results of the differential DNA methylation analysis including only the EPIC array data of the 14 pairs of which the MS-affected co-twins received GCs >3-12 months prior to blood collection. The volcano plot presents the  $p$ -values resulting from the nonparametric two-tailed Wilcoxon signed-rank test vs. the mean within-pair  $\beta$ -value difference for each CpG. Data were unadjusted for cell-type composition. **a-b** Within-pair methylation/ $\beta$ -value difference = MS-affected MZ co-twin receiving GCs >3-12 months prior to blood collection - clinically non-affected MZ co-twin. **c** Methylation level of the *ZBTB16* DMP (cg25345365) region determined using WGBS in CD4+ memory T cells of one MS-discordant MZ twin pair of which the MS-affected MZ twin was treated very recently with GCs at the time of blood collection. Coverage at cg25345365 is >20 reads in each co-twin. Source data are provided as a Source Data file. **d** Methylation and reported expression patterns of the 41 GC-DMRs that overlap with GC-response (dexamethasone) genes recorded in the EMBL-EBI Expression Atlas (accessed May 2018). One GC-DMR was excluded because it was reported to be down- and upregulated after dexamethasone treatment (Supplementary Table 9).  $n$  = number of twin pairs

genes. Of these, only five genes (1.8%), including *CCNA1*, *GMPR*, *ITGA6*, *LSPI*, and *ZBTB16*, overlap with the 721 GC-response (dexamethasone) genes recorded in the EMBL-EBI Expression Atlas (accessed May 2018). The other four MS-DMPs listed in Table 2 were not among these 320 GC-DMPs.

To study acute GC treatment effects, a WGBS analysis was performed on CD4+ memory T cells of a twin pair of which the MS-affected co-twin was very recently treated with two courses of GCs (i.e., 2 months and 10 days before blood collection with IVMP 1 g/day for 3 and 5 days, respectively). The WGBS data confirmed the strong hypomethylation of the *ZBTB16* DMP (cg25345365) in the GC-treated MS-affected co-twin (36% methylation difference) (Fig. 4c). In addition, 1424 other potential GC-DMRs were identified in the WGBS data, consisting of at least 2 CpGs, absolute mean methylation difference >0.25 and  $p < 0.01$  (Wald test). These GC-DMRs were annotated to 682 genes. Only 41 GC-DMRs overlap with 39 (5.7%) GC-response genes reported in the EMBL-EBI Expression Atlas (Supplementary Table 9), which represent potential GC-treatment epigenetic biomarkers. The majority of these 41 GC-DMRs were hypomethylated and the corresponding GC-response gene was recorded as upregulated due to GC treatment (Fig. 4d).

#### **ZBTB16 methylation and EPIC array-wide hypermethylation.**

DNA methylation of the repetitive elements *Alu*, human endogenous retrovirus type K (*HERVK*), and the long interspersed nuclear element-1 (*LINE1*) in the PBMC-derived samples were assessed by TDBS. Although *Alu* methylation correlated significantly with *LINE1* methylation ( $r_{\text{Pearson}} = 0.43$ ,  $p = 0.003$ ), methylation levels were overall very similar, showing maximum absolute within-pair differences for *Alu*, *HERVK*, and *LINE1* of only 0.015, 0.024, and 0.025, respectively. Hence, *Alu*, *HERVK* and *LINE1* methylation levels did not differ between the MS-discordant co-twins ( $p > 0.05$ , Wilcoxon signed-rank test). Although *Alu* and *HERVK* methylation were affected by cell-type composition differences, adjusting for cell-type composition did not change the results. For *Alu* generic primers were used, and since the element has ~1 million copies/genome, with a minimum sequencing coverage of 2000 reads/sample about 0.2% of the elements were analyzed. In contrast, *HERVK* and *LINE1* primers were designed to amplify the youngest subfamilies, which gives with a minimum sequencing coverage of 2000 reads >6 fold coverage per individual *HERVK* and *LINE1* element (see legend of Supplementary Table 12).

The volcano plots of the EPIC array data in Supplementary Fig. 2a and Fig. 2 are slightly unbalanced because 59.1% and 55.9% of the CpGs have a positive mean within-pair  $\beta$ -value difference before and after cell-type adjustment, respectively. Hence, the EPIC array data suggest an overall hypermethylation in the MS-affected co-twins (Supplementary Table 10). Although this EPIC array-wide hypermethylation in the MS-affected co-twins was not significantly associated with GC treatment history (Supplementary Fig. 18a), the number of hypermethylated CpGs in the MS-affected co-twins correlated significantly with the within-pair *ZBTB16* DMP methylation differences ( $r_{\text{Pearson}} = -0.36$ ,  $p = 0.02$ ) (Supplementary Fig. 18b).

**No evidence for within-pair copy-number variations.** Finally, discordant phenotypes within MZ twins can also be due to within-pair copy-number variations (WP-CNVs)<sup>56</sup>. Hence, we checked the EPIC array data for CNVs, but within the MZ twin pairs no chromosomal gains and losses were observed (Supplementary Fig. 19).

#### **Discussion**

Here, we present the largest EWAS in MZ twins clinically discordant for MS to date. Although the PBMC-based methylomes

of the 45 MS-discordant MZ twins were highly similar (mean  $\Delta\beta$ -values < 0.05), a few new MS-associated candidate loci were identified.

The most prominent MS-DMP was the technically replicated cg25345365 DMP in *ZBTB16*, which has thus far not been reported, probably because the 450 K array used in other MS EWAS studies does not cover this CpG<sup>26–31</sup>. The transcription factor *ZBTB16* is a GC-response gene that becomes highly upregulated after GC exposure<sup>55</sup>, and we show that the strong association between the *ZBTB16* DMP and MS in our EWAS is due to GC treatment history. Since none of the MS-affected co-twins had received GCs within three months prior to blood collection, our results indicate that for epigenomic and transcriptomic studies in MS a more stringent inclusion criterion is required (Supplementary Fig. 17). Our results might also have broader implications, because GCs are used in a variety of inflammatory and autoimmune diseases, but dosage and administration route vary per disorder. The GC-glucocorticoid receptor (GCR) complex regulates transcription by binding to glucocorticoid-response elements (GREs) in the genome<sup>57</sup>. GC-GCR binding has been associated with DNA demethylation at enhancer elements, supposedly due to active demethylation<sup>58</sup>. Indeed the *ZBTB16* DMP is hypomethylated in the (GC-treated) MS-affected co-twins, and is located in an enhancer and flanked (<100 bp) by two consensus GRE downstream half-sites (TGTTCT) (Supplementary Fig. 20), which are believed to be sufficient for binding of the GC-GCR complex<sup>57</sup>. In contrast to IFN, a strong GC signature was not observed, because the EPIC array and the WGBS data did not show methylation differences in other common, GC-regulated genes, like *FKBP5* and *TXNIP*<sup>55,59</sup>. However, IFN treatment was ongoing, while GC treatment had been given >3–12 months prior to blood collection. In the WGBS analysis a very recently GC-treated MS-affected co-twin was included, but as this concerned a single-replicate experiment, very stringent analyses criteria had to be applied (e.g. coverage threshold  $\geq 15$  reads). Hence, our data reveals the *ZBTB16* DMP as a prominent epigenetic biomarker for GC treatment, and future studies should assess its utility in predicting clinical GC response in patients with inflammatory or autoimmune diseases receiving GC therapy.

Our PBMC-based EWAS also revealed a DMR enriched for MS-DMPs in the *TMEM232* promoter region, which shows enrichment for the chromatin activation mark H3K4me3 in different immune cell types (Supplementary Fig. 21). Despite the small effect size (mean  $\Delta\beta$ -value = 0.024), this MS-DMR was technically replicated using TDBS, indicating a true effect. *TMEM232* MS-DMPs were also strongly associated with long-standing MS, and no evidence indicated that the association is confounded by treatment history. In whole blood-based case-control 450 K data, no significant difference was observed, but the two most prominent *TMEM232* MS-DMPs were not present on the 450 K array. Nevertheless, neighboring CpGs present on the 450 K array confirmed the directionality of the effect, which might indicate that the MS-DMR is restricted to a PBMC subtype and is diluted in whole blood, in which neutrophils are the predominant cell type. *TMEM232* is a member of the transmembrane (TMEM) protein family, which is predicted to be part of mitochondrial, endoplasmic reticulum, lysosome, and Golgi apparatus membranes<sup>60</sup>. While the function of *TMEM232* is still unknown, variants in this gene have been associated with atopic dermatitis and allergic rhinitis in GWAS<sup>36,37</sup>. Although this might point towards a common immunologic pathway involving *TMEM232*, robust evidence that supports an association between atopic diseases and MS is lacking<sup>61</sup>. Further studies in PBMCs and sorted immune cells are needed to verify the association between the *TMEM232* DMR and MS.

We also carried out a WGBS analysis on CD4+ memory T cells of four MS-discordant female MZ twins. Although this pilot did not reveal widespread global or site-specific MS-associated methylation differences, one potential MS-DMR was identified in an intronic regulatory region in the X-linked *FIRRE* gene. This encodes for a circular, long, non-coding RNA reported to be involved in positioning the inactive X-chromosome to the nucleolus and to maintain histone H3K27me3 methylation<sup>46,62</sup>. While the CpGs within this MS-DMR are not covered by the EPIC array, a probe located 6 bp upstream of the DMR was not significant in the PBMC-based EWAS. This might indicate that this MS-DMR is CD4+ T cell-specific, but can also be the result of stochastic variation caused, e.g., by molecular processes such as X-inactivation. Although our results are preliminary, MS is more common in women<sup>3</sup> and a role of X-inactivation in the pathogenesis of MS has been proposed (reviewed by Brooks et al.<sup>63</sup>); therefore, this DMR represents a possible candidate.

Since MS is a heterogeneous disease, DNA methylation changes present in only a few patients might also contribute to disease manifestation. To identify such rare methylation differences, a WP-DMR analysis was performed, revealing 45 WP-DMRs in 17 twin pairs. This suggests that WP-DMRs are quite common among MZ twins, but, as our analysis is restricted to disease-discordant MZ twins, this cannot be extrapolated to healthy MZ twins. Additional filtering revealed that 24 of these WP-DMRs were associated with the MS phenotype, of which 23 were pair-specific and one intergenic WP-DMR was present in 4 twin pairs. Although these WP-DMRs have not previously been associated with MS, two WP-DMRs were related to genes encoding protocadherins that are involved in neuronal development<sup>49</sup>. Hence, a contribution of these WP-DMRs to the discordant phenotype cannot be excluded, but since they are mainly pair-specific, these results should be interpreted very cautiously.

Several observations suggest a role of genomic imprinting in the etiology of MS<sup>18</sup>. In this context, MZ twins are of particular interest because MZ twins discordant for imprinting defects have been described relatively frequently<sup>7–9</sup>. Although we detected two WP-DMRs in reported imprinted DMRs (*SVOPL* and *HM13/MCTS2P*), the aberrant methylation profile was observed in the non-affected co-twin in both cases. Consequently, our PBMC-based analysis does not support the hypothesis that genomic imprinting errors contribute to the discordant clinical manifestation of MS in these MZ twins.

Neven et al.<sup>64</sup> reported hypermethylation of the repetitive elements *Alu*, *LINE1*, and *Sat-α* in blood of MS patients. We also assessed methylation of the repetitive elements *Alu*, *HERVK*, and *LINE1* but observed no differences. However, our EPIC array data does suggest a slight hypermethylation in the MS-affected co-twins. Also Bos et al.<sup>26</sup> observed in their 450 K data evidence for hypermethylation in CD8+ T cells of MS patients, but not for CD4+ T cells or whole blood. While GC treatment history was not directly associated with EPIC array-wide hypermethylation, we observed a rather weak, but significant association between increased within-pair *ZBTB16* methylation differences and the number of hypermethylated CpGs in the MS-affected co-twins. Although further confirmation is needed, this association might indirectly indicate that GCs also affect global DNA methylation levels. This might also explain the strong repetitive element hypermethylation in MS patients reported by Neven et al.<sup>64</sup>, who applied an inclusion criterion of only >1 month after GC treatment. However, in our study, hypermethylation in the MS-affected co-twins was only observed in the EPIC array data, and repetitive elements are strongly underrepresented on this array. Accordingly, additional studies are warranted to assess the association between DNA hypermethylation and MS and whether it is confounded by GC treatment history.

All MS-DMPs observed in our study have remained undetected in previous MS EWAS studies<sup>26–31</sup>. However, those studies observed much larger methylation differences and applied absolute mean  $\beta$ -value difference thresholds of  $>0.05$ <sup>26</sup> or  $>0.10$ <sup>27–30</sup>. As those studies used genetically unmatched cases and controls<sup>26–31</sup>, the reported large methylation differences might mainly be driven by genetic variation. Hannon et al.<sup>65</sup> recently showed that, in particular, sites with variable DNA methylation levels and sites robustly associated with environmental exposures are influenced by genetic effects, highlighting the need to control for genetic background in EWAS. Although our discordant MZ twin design perfectly controls for genetic variation, there are also limitations. MS-discordant MZ twins are scarce and therefore it is not possible to control for treatment effects without losing statistical power. Furthermore, the healthy co-twins are at risk to develop MS in the future and subsequently some of the pairs included in this EWAS will get clinically concordant for MS. Since the evolution of MS is supposed to be a continuum it is likely that prior to the clinical onset there is a prodromal phase of undefined duration, with subclinical subtle changes in CSF or MRI pointing to latent neuroinflammation. However, the onset of this postulated prodromal phase is impossible to define and, therefore, we aimed to identify DNA methylation differences that contribute to the discordant clinical manifestation of MS in MZ twins. For the EPIC array analysis, only DNA extracted from PBMCs was available, although it might be more informative to profile distinctive subtypes, such as CD4+ T cells, CD8+ T cells and B cells that are believed to be involved in the pathophysiology of MS<sup>45,66</sup>. Nevertheless, Paul et al.<sup>12</sup> profiled CD4+ T cells, B cells, and monocytes of 50 T1D-discordant MZ twins using the 450 K array, and observed only one genome-wide significant DMP in T cells (mean  $\Delta\beta$ -value = 0.023)<sup>12</sup>. This might indicate that for detecting robust MS-DMPs even rarer subpopulations such as Th1, Th17, and regulatory T cells need to be profiled, or immune cells in the cerebrospinal fluid. In contrast, Paul et al.<sup>12</sup> identified 10,548 differentially variable CpG positions (DVPs) in B cells, 4314 in CD4+ T cells, and 6508 in monocytes, and the T1D-affected MZ co-twins were enriched for DVPs. In addition, Webster et al.<sup>13</sup> identified 1107 DVPs in whole-blood 450 K data of 79 RA-discordant MZ twins, of which 763 DVPs were hypervariable in the RA-affected MZ co-twins. Although, we used the same method and significance threshold as applied in these studies, we only identified 25 DVPs of which the majority was hypervariable in the non-affected co-twins. Hence, our PBMC-based EPIC data does not reveal an MS-associated DNA methylation variability signature.

In conclusion, our EWAS shows that PBMC-based methylomes of MS-discordant MZ twins are highly similar, and no evidence was found that genomic imprinting errors or CNVs explain the discordant phenotype. However, a few candidate loci were identified, including a MS-DMR in the *TMEM232* promoter. Furthermore, epigenetic biomarkers for MS treatments were identified, revealing that besides short-term also medium-term treatment effects are detectable in blood cells, which should be considered in epigenomic and transcriptomic studies. Overall, we believe that this study represents an important first step in elucidating epigenetic mechanisms underlying the pathogenesis of MS.

## Methods

**Participants.** Twins were recruited by launching a nationally televised appeal and internet notification in Germany with support from the German Multiple Sclerosis Society (DMSG, regional and national division). Inclusion criteria for study participation were met for MZ twins with an MS diagnosis according to the revised McDonald criteria or clinically isolated syndrome in one co-twin only<sup>67</sup>. In total, 55 MZ twin pairs visited the outpatient department at the Institute of Clinical Neuroimmunology in Munich for a detailed interview and neurological examination. To confirm MS diagnosis, medical records including MRI scans from the

patients' treating neurologists were obtained and reviewed. For inclusion in the present analysis, PBMCs had to be available from both co-twins, resulting in 46 MZ twin pairs. The pair that carries the Leber's hereditary optic neuropathy-specific mutation m.11778G>A was not included in this analysis<sup>6</sup>. At blood collection, 23 MS-affected co-twins were receiving disease-modifying treatments, including interferon-beta (IFN,  $n = 12$ ), natalizumab ( $n = 5$ ), glatiramer acetate ( $n = 3$ ), teriflunomide ( $n = 1$ ), and dimethyl fumarate ( $n = 2$ ). None of the MS-affected co-twins included in the array-based EWAS received GCs within three months prior to blood collection. The non-affected co-twins underwent a detailed interview, including a comprehensive history of past and present complaints. In addition, non-affected co-twins were asked in detail for any occurrence of neurological symptoms in the past and an experienced MS neurologist (LAG) performed a neurological examination, including the EDSS. All previous patient registry information, including MRI scans if existing, were obtained and critically reviewed. The study was approved by the local ethics committee of the Ludwig Maximilians University of Munich and all participants gave written informed consent.

**DNA extraction and zygosity determination.** PBMCs were isolated from whole blood using Ficoll density gradient centrifugation and DNA was extracted using the QIAamp DNA Blood Midi Kit (Qiagen, Hilden, Germany). Extracted DNA was treated with RNase A/T1 Mix (Thermo Scientific, Oberhausen, Germany) and subsequently purified using the Genomic DNA Clean & Concentrator™-10 Kit (Zymo Research, CA, USA). As previously described<sup>6</sup>, zygosity was confirmed by genotyping 17 highly polymorphic microsatellite markers and by next-generation sequencing of 33 SNPs.

**Infinium MethylationEPIC BeadChip assay.** Genomic DNA was treated with bisulfite using the EZ DNA Methylation kit (D5002, Zymo Research), of which a detailed description is provided in the Supplementary Methods. Both members of a twin pair were always processed in the same batch. Genome-wide DNA methylation profiles of 46 MZ twin pairs clinically discordant for MS were generated using Illumina's Infinium MethylationEPIC BeadChip assay (EPIC array) (Illumina, San Diego, CA, USA) at the Department of Psychiatry and Psychotherapy of the Saarland University Hospital. The assay determines DNA methylation levels at >850,000 CpG sites and provides coverage of CpG islands, RefSeq genes, ENCODE open chromatin, ENCODE transcription factor-binding sites, and FANTOM5 enhancers. The assay was performed according to the manufacturer's instructions and scanned on an Illumina HiScan. To avoid batch effects, both members of a twin pair were always assayed on the same array.

**EPIC array data processing and DMP identification.** Raw EPIC array data were preprocessed using the RnBeads R/Bioconductor package<sup>68</sup>. Low-quality samples and probes were removed using the GreedyCut algorithm, based on a detection  $p$ -value threshold of 0.05, as implemented in the RnBeads package. In addition, probes with less than three beads and probes with a missing value in at least 5% of the samples were removed. For each CpG site, a  $\beta$ -value was calculated, which represents the fraction of methylated cytosines at that particular CpG site (0 = unmethylated, 1 = fully methylated). Subsequently,  $\beta$ -values were normalized using Illumina's default normalization method. In total, methylation data of 849,832 sites (866,895 in total) were available for 45 MS-discordant MZ twins. The relatively large number of excluded probes is due to inclusion of early access EPIC arrays, which have 11,652 fewer probes than the final release EPIC arrays. The EPIC array includes 59 SNP sites, which were used for quality control. All MZ twin pairs, except one, shared the same genotypes. The exceptional pair showed only a discordant genotype for SNP rs6471533. However, validation using targeted deep sequencing (TDS) revealed that both co-twins have the same genotype for the rs6471533 SNP (Supplementary Fig. 22), which indicates a technical artifact in the corresponding EPIC probe rather than a true genetic difference.

To identify differentially methylated CpG positions (DMPs) a two-sided non-parametric Wilcoxon signed-rank test was carried out. For the MS EWAS, an arbitrary significance level  $\alpha < 5 \times 10^{-6}$  was considered suggestive and genome-wide significance was defined as false discovery rate (FDR)  $< 0.05$ . All statistical analyses were performed in R. A functional annotation analysis was performed using the Genomic Regions Enrichment of Annotations Tool (GREAT v3.0.0) with default settings and the EPIC array CpGs, which passed quality control, as background<sup>42</sup>.

**Power calculation.** Since the power function of the Wilcoxon signed-rank test is difficult to express<sup>69</sup>, we used its closest parametric equivalent (paired T-test) to estimate the power of our MS EWAS. With a sample size of 45 twin pairs, >98% power is achieved to detect a mean  $\beta$ -value difference of at least 0.05 with a (genome-wide) significance threshold of  $1 \times 10^{-7}$ , using a two-sided paired T-test and assuming a standard deviation of 0.0266 (which is the true median standard deviation observed in our data). Details of this power calculation and calculations using smaller mean  $\beta$ -value differences are presented in Supplementary Table 11. The power analysis was performed using SAS University Edition.

**Estimation of cell-type composition.** A detailed description of the cell-type composition estimation is provided in the Supplementary Methods. In brief,

cell-type composition of each PBMC sample was estimated using the Houseman algorithm implemented in the *minfi* R/Bioconductor package<sup>70</sup>. The obtained *minfi* estimates were used to adjust the  $\beta$ -values for cellular composition using linear regression and the residuals were used for downstream analysis. To obtain interpretable, adjusted  $\beta$ -values, the unadjusted mean  $\beta$ -value of each CpG site was added to the residuals. To check the quality of the adjustment, the adjusted  $\beta$ -values were used to recalculate the within-pair correlations. As a result, Supplementary Fig. 23 shows that the overall within-pair correlations are, as expected, higher after adjusting for cell-type composition.

**Within-pair DMR analysis.** To identify WP-DMRs in the EPIC array data, the  $\beta$ -value differences ( $\Delta\beta$ -values) (adjusted for cell-type composition) per CpG were calculated for each twin pair (the 257 IFN-associated CpGs were excluded). To avoid false positives caused by single probes, WP-DMRs were defined as  $\geq 3$  CpGs, each having an absolute  $\Delta\beta$ -value  $> 0.2$  with a maximum 1 kb distance between neighboring CpGs. To exclude regions that are characterized by overall variable methylation levels, WP-DMRs were only considered when the  $\beta$ -value of the aberrant methylated co-twin was more than three standard deviations away from the mean.

**DVP identification.** For the DVP analysis, probes containing a SNP within five bases of the measured CpG site, probes mapping to the sex chromosomes, and probes with at least one missing value were excluded, resulting in methylation data of 759,291 sites. DVPs were identified using the iEVORA algorithm<sup>50</sup>, which measures differential variability between two groups by utilizing the Bartlett's test to detect differences in variance and an unpaired T-test to identify difference in means. CpGs with a FDR-corrected Bartlett's  $p < 0.001$  and raw  $t$ -test  $p < 0.05$  were defined as DVPs.

**Copy-number variation analysis.** CNV analysis with the EPIC array data was performed using the *conumee* R/Bioconductor package with default settings (<http://bioconductor.org/packages/conumee/>, R package version 1.6.0, Accessed 1 Nov 2016). Individual profiles and output were manually assessed. To define chromosomal gains and losses within the MZ twin pairs, an absolute segment mean threshold  $\geq 0.3$  was applied.

**Targeted deep bisulfite sequencing.** TDBS was used to validate DMPs resulting from the EPIC array analysis and to determine methylation levels of the repetitive elements *HERVK*, *L1NE1*, and Alu. Amplicons were generated on bisulfite-treated DNA using region-specific primers with TruSeq adaptor sequences on their 5'-ends (Illumina). Reaction conditions and primer sequences are described in Supplementary Table 12. Purified PCR products were quantified, pooled, amplified using index primers (five cycles), and sequenced in a 300-bp paired-end MiSeq run (Illumina). After demultiplexing, adaptor trimming, and clipping overlapping mates, the resulting FASTQ files were imported into BiQ Analyzer HiMod<sup>71</sup> to filter out low-quality reads and call the methylation levels. Final coverage was >1500 reads/base.

**Targeted deep sequencing.** The rs6471533 SNP was genotyped using TDS (see Supplementary Table 12 for reaction conditions and primer sequences). The workflow is similar to that described for TDBS, except that genomic DNA was used and that the resulting FASTQ files were aligned to the reference sequence using Bowtie 2. Subsequently, variants were called with SAMtools mpileup and variant information was extracted using filter pileup. Final coverage was >1500 reads/base.

**Third-party MS case-control cohort data analysis.** The whole blood-based 450 K data of 140 unrelated MS patients and 139 controls from Kular et al.<sup>31</sup> (GSE106648) were available as intensity matrices of methylated and unmethylated probe intensities, which were imported into R using RnBeads<sup>68</sup>. As no quality information was provided, GreedyCut and bead-based filtering was not possible. Proportions of major cell types were estimated as described above. In line with the original study<sup>31</sup>, MS-DMPs were determined using linear regression with *limma*<sup>72</sup> as implemented in RnBeads adjusting for gender, age, smoking status, and the first two principal components of the estimated cell proportion matrix. Adjustment for the batch effect was not possible as the corresponding variable was not provided.

**Whole-genome bisulfite sequencing in CD4+ T cells.** WGBS was used to profile CD4+ central and effector memory T cells of four MS-discordant female MZ twin pairs (mean age 43.3 years, discordant for MS > 12 years, Supplementary Table 13). Of one pair, the MS-affected co-twin had been treated very recently with GCs at the time of blood collection (but never received any immune-modulating therapy), while the MS-affected co-twins of the other three pairs had not received GCs or other immune-modulating therapies within at least 12 months prior to blood collection. The cell sorting procedure, the preparation of the WGBS libraries, and the preprocessing of the WGBS sequencing data are described in detail in the Supplementary Methods. The coverage statistics of the samples are summarized in Supplementary Table 13.

**DMR identification in the WGBS data.** To identify MS-associated DMRs (MS-DMRs), the WGBS data of all four pairs were analyzed using the RnBeads package, in which a paired *t*-test was performed for every CpG. Only CpGs with a coverage  $\geq 10$  reads in all samples were included, resulting in methylation information of about 2.7 million CpGs (Supplementary Table 13). MS-DMRs were defined as  $\geq 3$  CpGs, each having  $p < 0.05$  (two-sided paired *t*-test) and an absolute mean methylation difference  $> 0.2$ , and a maximum of 500 bp distance between neighboring significant CpGs.

To identify GC treatment-associated DMRs (GC-DMRs), the WGBS data of the pair with the GC-treated MS-affected co-twin were analyzed using DSS-single<sup>73</sup>, which is designed to detect DMRs from WGBS data without replicates. To increase the quality of this single-replicate DMR analysis, only CpGs with a coverage  $\geq 15$  reads in both samples were included and the sex chromosomes were excluded, resulting in methylation data of up to 2.8 million CpGs (Supplementary Table 13). It has been reported that binding of the GCR complex is rare within CpG islands and predominantly occurs at distal regulatory elements<sup>58</sup>. To detect DMRs in such CpG poor regions, the DSS settings included a smoothing span of 100 bp and minimum DMR length of 25 bp with  $\geq 2$  CpGs and  $p < 0.01$  (Wald test). The absolute mean methylation difference had to be larger than 0.25, and to limit the number of false positives only GC-DMRs located in reported GC-response genes were considered. The GC and MS-DMRs were annotated using the ChIPseeker R/Bioconductor package (v1.14.2)<sup>74</sup>.

**Partially methylated domain analysis in WGBS data.** The WGBS data was segmented into PMDs, low methylated regions (LMRs), and unmethylated regions (UMRs) using the MethylSeekR R/Bioconductor package<sup>75</sup>. After filtering gaps annotated by UCSC, the rest of the genome was designated as fully methylated regions (FMRs). As input for MethylSeekR the aggregated strand information per CpG was used, and the MethylSeekR settings included coverage  $\geq 5$  reads per CpG, 50% methylation and an FDR  $\leq 0.05$  for calling hypomethylated regions, resulting in a cut-off of  $\geq 4$  CpGs per LMR. For each segment, the methylation levels between the non-affected and MS-affected co-twins were compared using a paired *t*-test on the median weighted average methylation values. In addition, to assess the genome-wide PMD similarity across the eight samples, the genome was binned into 1 kb windows and each was annotated with 1 if the bin overlapped with a PMD and with 0 otherwise. Based on this binarized matrix a hierarchical clustering was performed in R using ward.D2 as agglomeration method and euclidean as a distance measurement. The very same procedure was performed for FMRs.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

The epigenomic data has been deposited at the European Genome-phenome Archive (EGA, <http://www.ebi.ac.uk/ega/>), which is hosted at the EBI, under accession number EGAS00001003147. The source data underlying Table 2, Supplementary Tables 1–9, Figs. 2c, 4a, c, and Supplementary Figs. 5, 6, 9, and 11–18 are provided as a Source Data file. The authors declare that all other data are contained within the article and its supplementary files or available from the author upon request.

Received: 17 July 2018 Accepted: 3 April 2019

Published online: 07 May 2019

## References

- Browne, P. et al. Atlas of multiple sclerosis 2013: a growing global problem with widespread inequity. *Neurology* **83**, 1022–1024 (2014).
- Dendrou, C. A., Fugger, L. & Friese, M. A. Immunopathology of multiple sclerosis. *Nat. Rev. Immunol.* **15**, 545–558 (2015).
- Sawcer, S., Franklin, R. J. & Ban, M. Multiple sclerosis genetics. *Lancet Neurol.* **13**, 700–709 (2014).
- Hawkes, C. H. & Macgregor, A. J. Twin studies and the heritability of MS: a conclusion. *Mult. Scler.* **15**, 661–667 (2009).
- Westerlind, H. et al. Modest familial risks for multiple sclerosis: a registry-based study of the population of Sweden. *Brain* **137**, 770–778 (2014).
- Souren, N. Y. et al. Mitochondrial DNA variation and heteroplasmy in monozygotic twins clinically discordant for multiple sclerosis. *Hum. Mutat.* **37**, 765–775 (2016).
- Blik, J. et al. Lessons from BWS twins: complex maternal and paternal hypomethylation and a common source of haematopoietic stem cells. *Eur. J. Hum. Genet.* **17**, 1625–1634 (2009).
- Inoue, T. et al. Continuous hypomethylation of the KCNQ1OT1:TSS-DMR in monozygotic twins discordant for Beckwith-Wiedemann syndrome. *Am. J. Med Genet A* **173**, 2847–2850 (2017).
- Riess, A. et al. First report on concordant monozygotic twins with Silver-Russell syndrome and ICR1 hypomethylation. *Eur. J. Med Genet* **59**, 1–4 (2016).
- Javierre, B. M. et al. Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. *Genome Res.* **20**, 170–179 (2010).
- Laborie, L. B. et al. DNA hypomethylation, transient neonatal diabetes, and prune belly sequence in one of two identical twins. *Eur. J. Pediatr.* **169**, 207–213 (2010).
- Paul, D. S. et al. Increased DNA methylation variability in type 1 diabetes across three immune effector cell types. *Nat. Commun.* **7**, 13555 (2016).
- Webster, A. P. et al. Increased DNA methylation variability in rheumatoid arthritis-discordant monozygotic twins. *Genome Med.* **10**, 64 (2018).
- Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14**, 204–220 (2013).
- Feinberg, A. P. The key role of epigenetics in human disease prevention and mitigation. *N. Engl. J. Med.* **378**, 1323–1334 (2018).
- Ebers, G. C. et al. Parent-of-origin effect in multiple sclerosis: observations in half-siblings. *Lancet* **363**, 1773–1774 (2004).
- Hoppenbrouwers, I. A. et al. Maternal transmission of multiple sclerosis in a dutch population. *Arch. Neurol.* **65**, 345–348 (2008).
- Ruhrmann, S., Stridh, P., Kular, L. & Jagodic, M. Genomic imprinting: a missing piece of the multiple sclerosis puzzle? *Int J. Biochem Cell Biol.* **67**, 49–57 (2015).
- Zhang, P. et al. The risk of smoking on multiple sclerosis: a meta-analysis based on 20,626 cases from case-control and cohort studies. *PeerJ* **4**, e1797 (2016).
- Handel, A. E. et al. An updated meta-analysis of risk of multiple sclerosis following infectious mononucleosis. *PLoS ONE* **5**, e12496 (2010).
- Duan, S. et al. Vitamin D status and the risk of multiple sclerosis: a systematic review and meta-analysis. *Neurosci. Lett.* **570**, 108–113 (2014).
- Scott, R. S. Epstein-Barr virus: a master epigenetic manipulator. *Curr. Opin. Virol.* **26**, 74–80 (2017).
- Joeanes, R. et al. Epigenetic signatures of cigarette smoking. *Circ. Cardiovasc Genet* **9**, 436–447 (2016).
- Marabita, F. et al. Smoking induces DNA methylation changes in Multiple Sclerosis patients with exposure-response relationship. *Sci. Rep.* **7**, 14589 (2017).
- Zeitelhofer, M. et al. Functional genomics analysis of vitamin D effects on CD4+ T cells in vivo in experimental autoimmune encephalomyelitis. *Proc. Natl Acad. Sci. USA* **114**, E1678–E1687 (2017).
- Bos, S. D. et al. Genome-wide DNA methylation profiles indicate CD8+ T cell hypermethylation in multiple sclerosis. *PLoS ONE* **10**, e0117403 (2015).
- Graves, M. et al. Methylation differences at the HLA-DRB1 locus in CD4+ T-Cells are associated with multiple sclerosis. *Mult. Scler.* **20**, 1033–1041 (2014).
- Maltby, V. E. et al. Genome-wide DNA methylation profiling of CD8+ T cells shows a distinct epigenetic signature to CD4+ T cells in multiple sclerosis patients. *Clin. Epigenetics* **7**, 118 (2015).
- Kulakova, O. G. et al. Whole-genome DNA methylation analysis of peripheral blood mononuclear cells in multiple sclerosis patients with different disease courses. *Acta Nat.* **8**, 103–110 (2016).
- Maltby, V. E. et al. Differential methylation at MHC in CD4+ T cells is associated with multiple sclerosis independently of HLA-DRB1. *Clin. Epigenetics* **9**, 71 (2017).
- Kular, L. et al. DNA methylation as a mediator of HLA-DRB1\*15:01 and a protective variant in multiple sclerosis. *Nat. Commun.* **9**, 2397 (2018).
- Price, M. E. et al. Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenet. Chromatin* **6**, 4 (2013).
- Gaunt, T. R. et al. Systematic identification of genetic influences on methylation across the human life course. *Genome Biol.* **17**, 61 (2016).
- Baranzini, S. E. et al. Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. *Nature* **464**, 1351–1356 (2010).
- Pidsley, R. et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol.* **17**, 208 (2016).
- Hirota, T. et al. Genome-wide association study identifies eight new susceptibility loci for atopic dermatitis in the Japanese population. *Nat. Genet.* **44**, 1222–1226 (2012).
- Ramasamy, A. et al. A genome-wide meta-analysis of genetic variants associated with allergic rhinitis and grass sensitization and their interaction with birth order. *J. Allergy Clin. Immunol.* **128**, 996–1005 (2011).
- Curreli, S., Wong, B. S., Latinovic, O., Konstantopoulos, K. & Stamatou, N. M. Class 3 semaphorins induce F-actin reorganization in human dendritic cells: Role in cell migration. *J. Leukoc. Biol.* **100**, 1323–1334 (2016).
- International Multiple Sclerosis Genetics C. Genome-wide association study of severity in multiple sclerosis. *Genes Immun.* **12**, 615–625 (2011).

40. Mao, A. P. et al. Multiple layers of transcriptional regulation by PLZF in NKT-cell development. *Proc. Natl Acad. Sci. USA* **113**, 7602–7607 (2016).
41. Sunker, A. & Alkuraya, F. S. Identification of MR11, encoding translation initiation factor eIF-2B subunit alpha/beta/delta-like protein, as a candidate locus for infantile epilepsy with severe cystic degeneration of the brain. *Gene* **512**, 450–452 (2013).
42. McLean, C. Y. et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
43. Goldenberg, D. M., Stein, R. & Sharkey, R. M. The emergence of trophoblast cell-surface antigen 2 (TROP-2) as a novel cancer target. *Oncotarget* **9**, 28989–29006 (2018).
44. Amin, N. et al. A rare missense variant in RCL1 segregates with depression in extended families. *Mol. Psychiatry* **23**, 1120–1126 (2018).
45. Hohlfeld, R., Dormmair, K., Meinel, E. & Wekerle, H. The search for the target antigens of multiple sclerosis, part 1: autoreactive CD4+ T lymphocytes as pathogenic effectors and therapeutic targets. *Lancet Neurol.* **15**, 198–209 (2016).
46. Izuogu, O. G. et al. Analysis of human ES cell differentiation establishes that the dominant isoforms of the lncRNAs RMST and FIRRE are circular. *BMC Genom.* **19**, 276 (2018).
47. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
48. Joshi, R. S. et al. DNA methylation profiling of uniparental disomy subjects provides a map of parental epigenetic bias in the human genome. *Am. J. Hum. Genet.* **99**, 555–566 (2016).
49. Peek, S. L., Mah, K. M. & Weiner, J. A. Regulation of neural circuit formation by protocadherins. *Cell Mol. Life Sci.* **74**, 4133 (2017).
50. Teschendorff, A. E. et al. DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer. *Nat. Commun.* **7**, 10478 (2016).
51. Rusinova, I. et al. Interferome v2.0: an updated database of annotated interferon-regulated genes. *Nucleic Acids Res.* **41**, D1040–D1046 (2013).
52. Singh, M. K. et al. Gene expression changes in peripheral blood mononuclear cells from multiple sclerosis patients undergoing beta-interferon therapy. *J. Neurol. Sci.* **258**, 52–59 (2007).
53. Nickles, D. et al. Blood RNA profiling in a large cohort of multiple sclerosis patients and healthy controls. *Hum. Mol. Genet.* **22**, 4194–4205 (2013).
54. Singh, S. P. et al. PLZF regulates CCR6 and is critical for the acquisition and maintenance of the Th17 phenotype in human cells. *J. Immunol.* **194**, 4350–4361 (2015).
55. Tissing, W. J. et al. Genomewide identification of prednisolone-responsive genes in acute lymphoblastic leukemia cells. *Blood* **109**, 3929–3935 (2007).
56. Zwijnenburg, P. J., Meijers-Heijboer, H. & Boomsma, D. I. Identical but not the same: the value of discordant monozygotic twins in genetic research. *Am. J. Med Genet B Neuropsychiatr. Genet.* **153B**, 1134–1149 (2010).
57. Del Monaco, M. et al. Identification of novel glucocorticoid-response elements in human elastin promoter and demonstration of nucleotide sequence specificity of the receptor binding. *J. Invest Dermatol.* **108**, 938–942 (1997).
58. Wiench, M. et al. DNA methylation status predicts cell type-specific enhancer activity. *EMBO J.* **30**, 3028–3039 (2011).
59. Winkler, B. K., Lehnert, H., Oster, H., Kirchner, H. & Harbeck, B. FKBP5 methylation as a possible marker for cortisol state and transient cortisol exposure in healthy human subjects. *Epigenomics* **9**, 1279–1286 (2017).
60. Wrzesinski, T. et al. Expression of pre-selected TMEMs with predicted ER localization as potential classifiers of ccRCC tumors. *BMC Cancer* **15**, 518 (2015).
61. Monteiro, L., Souza-Machado, A., Menezes, C. & Melo, A. Association between allergies and multiple sclerosis: a systematic review and meta-analysis. *Acta Neurol. Scand.* **123**, 1–7 (2011).
62. Yang, F. et al. The lncRNA Firre anchors the inactive X chromosome to the nucleolus by binding CTCF and maintains H3K27me3 methylation. *Genome Biol.* **16**, 52 (2015).
63. Brooks, W. H. & Renaudineau, Y. Epigenetics and autoimmune diseases: the X chromosome-nucleolus nexus. *Front Genet* **6**, 22 (2015).
64. Neven, K. Y. et al. Repetitive element hypermethylation in multiple sclerosis patients. *BMC Genet* **17**, 84 (2016).
65. Hannon, E. et al. Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins. *PLoS Genet* **14**, e1007544 (2018).
66. Hohlfeld, R., Dormmair, K., Meinel, E. & Wekerle, H. The search for the target antigens of multiple sclerosis, part 2: CD8+ T cells, B cells, and antibodies in the focus of reverse-translational research. *Lancet Neurol.* **15**, 317–331 (2016).
67. Polman, C. H. et al. Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. *Ann. Neurol.* **69**, 292–302 (2011).
68. Assenov, Y. et al. Comprehensive analysis of DNA methylation data with RnBeads. *Nat. Methods* **11**, 1138–1140 (2014).
69. Shieh, G., Jan, S. L. & Randles, R. H. Power and sample size determinations for the Wilcoxon signed-rank test. *J. Stat. Comput. Simul.* **77**, 717–724 (2007).
70. Aryee, M. J. et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
71. Becker, D. et al. BiQ Analyzer HiMod: an interactive software tool for high-throughput locus-specific analysis of 5-methylcytosine and its oxidized derivatives. *Nucleic Acids Res.* **42**, W501–W507 (2014).
72. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-seq and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
73. Wu, H. et al. Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic Acids Res.* **43**, e141 (2015).
74. Yu, G., Wang, L. G. & He, Q. Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).
75. Burger, L., Gaidatzis, D., Schubeler, D. & Stadler, M. B. Identification of active regulatory regions from DNA methylation data. *Nucleic Acids Res.* **41**, e155 (2013).

### Acknowledgements

We are grateful to all the twins, who participated in this study. We thank Jasmin Gries for performing MiSeq sequencing, Karl Nordström for preprocessing the reads, Kathrin Kattler for assisting tWGBS protocol optimization, and Katja Anslinger for zygosity determination. We acknowledge the use of the CF FlowCyt at the Biomedical Center of the Ludwig-Maximilians-Universität München. This work was supported by the Gemeinnützige Hertie Stiftung; German MS Foundation (regional and national division); German Research Council [SFB-TR 128 SyNergy]; Krankheitsbezogenes Kompetenznetz Multiple Sklerose, Cyliax Stiftung, Verein zur Therapieforschung für MS Kranke, and the German Federal Ministry of Education and Research (BMBF) funded program DEEP (01KU1216F).

### Author contributions

Study design: N.Y.S., L.A.G., T.K., R.H., J.W. Patient recruitment and care: L.A.G., T.K. Clinical data collection: L.A.G. Experimental work and data collection: N.Y.S., G.G., E.B. Data processing: N.Y.S., P.L. Data analysis: N.Y.S., P.L., A.S. Supervising data processing and analysis: P.L. Facilitating technical and material support: C.P., D.W., J.W. Supervision: R.H., J.W. Manuscript writing: N.Y.S. Manuscript editing: all authors.

### Additional information

**Supplementary Information** accompanies this paper at <https://doi.org/10.1038/s41467-019-09984-3>.

**Competing interests:** The authors declare no competing interests.

**Reprints and permission** information is available online at <http://npg.nature.com/reprintsandpermissions/>

**Journal peer review information:** *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019

**Supplementary Information:**

**DNA methylation signatures of  
monozygotic twins clinically discordant  
for multiple sclerosis**

Souren et al.



## Supplementary Tables

**Supplementary Table 1. Characteristics of the top 15 most significantly differentially methylated positions (DMPs) associated with long-standing MS, identified by a pair-wise analysis including only the EPIC array data of the 25 MZ twins pairs that have been clinically discordant for MS for more than 10 years (n = 25 twin pairs).** The first two DMPs had a suggestive P-value <  $5 \times 10^{-6}$  (adjusted for cell type composition).

Probe ID	Location <sup>a</sup>	Gene	Functional region <sup>b</sup>	450k	Mean $\beta$ -value (U)		Mean $\Delta\beta$ -value (95% CI) (U)	Mean $\Delta\beta$ -value (95% CI) (A)	$\beta$ -value range	$P_{W-U}/P_{W-A}$	
					MS-affected co-twins (n=25)	non-affected co-twins (n=25)					
1	cg11243634	chr1:59044320	<b>TACSTD2</b>	TSS1500/DHS	Y	0.667	0.689	-0.022 (-0.029,-0.015)	-0.022 (-0.029,-0.015)	0.62-0.73	<b>1.13*10<sup>-6</sup>/2.56*10<sup>-6</sup></b>
2	cg23896094	chr9:4839083	<b>RCL1</b>	TSS1500/Body/TFBS	N	0.890	0.901	-0.011 (-0.016,-0.007)	-0.011 (-0.016,-0.007)	0.87-0.93	<b>4.17*10<sup>-6</sup>/4.17*10<sup>-6</sup></b>
3	cg06958567	chr18:52495541	<b>RAB27B</b>	TSS1500/DHS	Y	0.120	0.129	-0.009 (-0.015,-0.004)	-0.010 (-0.015,-0.006)	0.08-0.18	3.29*10 <sup>-4</sup> /8.17*10 <sup>-6</sup>
4	cg07850221	chr19:36235109	<b>UZAF1L4/PSENFEN</b>	Body/TSS1500/TFBS	Y	0.825	0.834	-0.009 (-0.012,-0.005)	-0.009 (-0.013,-0.005)	0.78-0.87	1.83*10 <sup>-5</sup> /8.17*10 <sup>-6</sup>
5	cg22891413	chr5:1407569	<b>SLC6A3</b>	Body	Y	0.901	0.910	-0.009 (-0.013,-0.005)	-0.009 (-0.013,-0.006)	0.85-0.94	2.50*10 <sup>-4</sup> /1.01*10 <sup>-5</sup>
6	cg04669407	chr5:148521450	<b>ABLIM3</b>	5'UTR/CpG island/DHS	Y	0.377	0.404	-0.027 (-0.037,-0.017)	-0.025 (-0.036,-0.015)	0.28-0.48	1.97*10 <sup>-6</sup> /1.23*10 <sup>-5</sup>
7	cg17514766	chr19:39826927	<b>GMFG</b>	TSS1500/TFBS	Y	0.105	0.115	-0.010 (-0.014,-0.005)	-0.010 (-0.014,-0.006)	0.08-0.16	3.81*10 <sup>-5</sup> /1.83*10 <sup>-5</sup>
8	cg22008026	chr4:155413392	<b>DCHS2</b>	TSS1500/CpG island/DHS	Y	0.213	0.237	-0.024 (-0.031,-0.016)	-0.021 (-0.029,-0.014)	0.16-0.31	1.01*10 <sup>-5</sup> /1.83*10 <sup>-5</sup>
9	cg26583412	chr5:110062780	<b>TMEM232</b>	TSS1500/DHS	N	0.528	0.490	0.038 (0.022,0.054)	0.038 (0.023, 0.052)	0.26-0.68	5.39*10 <sup>-5</sup> /1.83*10 <sup>-5</sup>
10	cg26630171	chr19:3459270	<b>NFIC</b>	Body/DHS	Y	0.906	0.897	0.008 (0.004,0.012)	0.008 (0.005,0.012)	0.86-0.93	1.20*10 <sup>-4</sup> /1.83*10 <sup>-5</sup>
11	cg23516613	chr19:12595698	<b>ZNF709</b>	TSS200/CpG island/DHS	Y	0.121	0.131	-0.010 (-0.014,-0.006)	-0.009 (-0.013,-0.006)	0.10-0.16	1.01*10 <sup>-5</sup> /2.21*10 <sup>-5</sup>
12	cg16468417	chr3:35835606	<b>ARPP-21</b>	3'UTR	Y	0.788	0.769	0.019 (0.011-0.026)	0.016 (0.009,0.022)	0.73-0.84	3.19*10 <sup>-5</sup> /2.66*10 <sup>-5</sup>
13	cg27037608	chr5:110062618	<b>TMEM232</b>	TSS200/TFBS	N	0.498	0.472	0.026 (0.014,0.038)	0.027 (0.015,0.038)	0.34-0.59	8.80*10 <sup>-5</sup> /3.19*10 <sup>-5</sup>
14	cg11079989	chr1:17222715	<b>CROCC</b>	Body/DHS	Y	0.135	0.148	-0.013 (-0.019,-0.008)	-0.013 (-0.018,-0.008)	0.11-0.19	2.66*10 <sup>-5</sup> /4.54*10 <sup>-5</sup>
15	cg10344516	chr10:6689948		DHS	N	0.152	0.14	0.012 (0.007,0.018)	0.012 (0.006,0.018)	0.10-0.20	1.23*10 <sup>-5</sup> /5.39*10 <sup>-5</sup>
756	cg25345365	chr11:114050114	<b>ZBTB16</b>	Body/DHS/enhancer	N	0.526	0.566	-0.040 (-0.065,-0.015)	-0.036 (-0.057,-0.014)	0.36-0.72	0.002/0.0018

Source data are provided as a Source Data file. <sup>a</sup>All the genome coordinates are based on human genome build GRCh37/hg19. <sup>b</sup>Based on information provided by the Illumina manifest. Since all genes have multiple transcripts, the "UCSC\_RefGene\_Group" gene-related location is listed. A = adjusted for cell-type composition, CI = confidence interval, DHS = DNase I hypersensitive site, n = number of MS-discordant MZ twin pairs,  $P_{W-A}$  = P-value two-tailed Wilcoxon signed-rank test adjusted for cell-type composition,  $P_{W-U}$  = P-value two-tailed Wilcoxon signed-rank test unadjusted for cell-type composition, TFBS = transcription factor binding site, TSS200 = the region from transcription start site (TSS) to -200 nt upstream of TSS, TSS1500 = -200 to -1500 nt upstream of TSS, U = unadjusted for cell-type composition, 450k = probe present on the 450k array (Y = yes, N = no), 5'UTR = 5' untranslated region,  $\Delta\beta$ -value = within-pair  $\beta$ -value difference (clinically MS-affected MZ co-twin - non-affected MZ co-twin).

**Supplementary Table 2. Evaluation of the *TMEM232* MS-DMPs in blood-based 450K EWAS data of 140 MS patients and 139 unrelated controls from Kular et al<sup>1</sup>. The results of the PBMC-based EPIC array EWAS in the 45 MZ twins clinically discordant for MS are shown as well.**

Probe ID <sup>a</sup>	Location at chr 5:	Blood-based 450K EWAS of unrelated cases and controls						PBMC-based EPIC EWAS of MS discordant MZ twins							
		Mean $\beta$ -value ( $\pm$ SD)		Mean $\Delta\beta$ -value	$\beta$ -value range	$P_U^b$	$P_A^b$	$P_{A2}^b$	Mean $\beta$ -value		Mean $\Delta\beta$ -value (95% CI)	$\beta$ -value range	$P_{W-U}$	$P_{W-A}$	
		MS patients (n=140)	Controls (n=139)						MS-affected co-twins (n=45)	Non-affected co-twins (n=45)					
cg23279021	110062343	1 <sup>st</sup> exon	0.309 $\pm$ 0.076	0.297 $\pm$ 0.084	0.012	0.11-0.67	0.20	0.13	0.24	0.320	0.295	0.025 (0.011,0.040)	0.08-0.49	0.001	2.6*10 <sup>-4</sup>
cg17248924	110062384	TSS200	0.327 $\pm$ 0.098	0.314 $\pm$ 0.107	0.012	0.12-0.76	0.25	0.15	0.23	0.457	0.444	0.013 (-0.005,0.031)	0.22-0.62	0.16	0.08
cg11641395	110062398	TSS200	0.276 $\pm$ 0.080	0.268 $\pm$ 0.086	0.008	0.10-0.55	0.32	0.22	0.35	0.437	0.425	0.012 (-0.001,0.025)	0.26-0.58	0.16	0.12
cg06429214	110062417	TSS200	0.276 $\pm$ 0.099	0.259 $\pm$ 0.108	0.017	0.05-0.76	0.12	0.07	0.19	0.309	0.284	0.025 (0.008,0.041)	0.09-0.50	0.008	0.005
cg25259944	110062473	TSS200	0.325 $\pm$ 0.089	0.308 $\pm$ 0.099	0.017	0.11-0.70	0.10	0.05	0.16	0.483	0.455	0.028 (0.012,0.044)	0.27-0.61	0.001	0.002
cg22429640	110062570	TSS200								0.539	0.514	0.025 (0.009,0.041)	0.33-0.65	0.005	0.003
cg19398821	110062608	TSS200								0.507	0.484	0.023 (0.006,0.041)	0.29-0.68	0.017	0.016
cg27037608	110062618	TSS200								0.488	0.466	0.022 (0.012,0.032)	0.31-0.59	4.8*10 <sup>-5</sup>	4.3*10 <sup>-6</sup>
cg17946588	110062682	TSS1500								0.461	0.439	0.022 (0.007,0.037)	0.26-0.60	0.006	0.005
cg10597099	110062725	TSS1500								0.544	0.522	0.022 (0.002,0.041)	0.22-0.73	0.032	0.013
cg19526166	110062729	TSS1500	0.388 $\pm$ 0.098	0.374 $\pm$ 0.112	0.014	0.15-0.80	0.23	0.14	0.18	0.532	0.505	0.027 (0.007,0.048)	0.31-0.71	0.013	0.007
cg26583412	110062780	TSS1500								0.516	0.482	0.034 (0.020,0.048)	0.23-0.68	2.0*10 <sup>-5</sup>	1.3*10 <sup>-5</sup>
cg06414816	110062837	TSS1500	0.403 $\pm$ 0.078	0.391 $\pm$ 0.087	0.012	0.20-0.69	0.21	0.15	0.31	0.547	0.526	0.021 (0.007,0.035)	0.35-0.64	0.010	0.010

Source data are provided as a Source Data file. <sup>a</sup>The 450K array contains only 7 of the 13 *TMEM232* promoter probes that are present on the EPIC array. <sup>b</sup>Significance estimated using linear regression (see Methods for details). CI = confidence interval, n = number of individuals,  $P_U$  = P-value unadjusted,  $P_A$  = P-value adjusted for sex, age and smoking status,  $P_{A2}$  = P-value adjusted for sex, age, smoking status and cell type composition,  $P_{W-A}$  = P-value Wilcoxon two-tailed signed-rank test adjusted for cell-type composition,  $P_{W-U}$  = P-value Wilcoxon two-tailed signed-rank test unadjusted for cell-type composition, TSS200 = the region from transcription start site (TSS) to -200 nt upstream of TSS, TSS1500 = -200 to -1500 nt upstream of TSS,  $\Delta\beta$ -value = within-pair  $\beta$ -value difference (clinically MS-affected MZ co-twin - non-affected MZ co-twin).

**Supplementary Table 3. MS-associated differentially methylated regions (MS-DMRs) identified in whole genome bisulfite sequencing (WGBS) data of CD4+ memory T cells of four MS discordant female MZ twin pairs (n = 4 twin pairs).** MS-DMRs listed in this table were defined as  $\geq 3$  CpGs, each having  $P < 0.05$  (two-tailed paired T-test) and absolute mean methylation difference  $> 0.15$ , and a maximum of 500 bp distance between neighbouring significant CpGs.

Chr <sup>a</sup>	Start	End	# Width CG	Location	Mean methylation		Mean methylation difference	Gene	Full gene name	Robust DMR	Distance to closest EPIC probe and $P_{W-A}^b$	EWAS
					MS-affected MZ co-twins	Non-affected MZ co-twins						
1	17711436	17712116	681 3	Intron	0.74	0.91	-0.17	<i>PADI6</i>	Peptidyl arginine deiminase 6	No	900 bp, $P > 0.05$	
1	85791031	85791047	17 5	Intron	0.85	0.67	0.18	<i>DDAH1</i>	Dimethylarginine dimethylaminohydrolase 1 <sup>2,3</sup>	Yes	4 kb	
1	87644875	87644911	37 5	Distal Intergenic	0.56	0.76	-0.20			Yes	1 kb	
1	228755378	228755464	87 3	Distal Intergenic	0.71	0.51	0.20			Yes	1.5 kb	
2	87568953	87569230	278 3	Intron	0.40	0.59	-0.18	<i>RMND5A</i>	Required for meiotic nuclear division 5 homolog A	No	<b>0 bp</b> (cg21997198), $P > 0.05$	
5	176007126	176007142	17 3	Intron	0.75	0.57	0.18	<i>CDHR2</i>	Cadherin related family member 2	Yes	600 bp, $P > 0.05$	
6	57421688	57421697	10 3	Intron	0.71	0.51	0.20	<i>PRIM2</i>	DNA primase subunit 2	Yes	9 kb	
6	170403829	170404107	279 3	Distal Intergenic	0.62	0.84	-0.22			Yes	250 bp, $P > 0.05$	
7	31117742	31118094	353 3	Intron	0.65	0.82	-0.17	<i>ADCYAP1R1</i>	ADCYAP receptor type 1	No	1.5 kb	
7	98424232	98424350	119 3	Distal Intergenic	0.51	0.70	-0.20			No	<b>0 bp</b> (cg16446288/cg11757417), $P = 0.04/0.86$ , $\Delta\beta = -0.01/0.00$	
8	102904079	102904097	19 3	Intron	0.33	0.53	-0.19	<i>NCALD</i>	Neurocalcin delta	Yes	170 bp, $P > 0.05$	
9	43134844	43135115	272 3	Promoter (1-2kb)	0.45	0.62	-0.17	<i>ANKRD20A3</i>	Ankyrin repeat domain 20 family member A3	Yes	90 bp, $P > 0.05$	
9	66493004	66493351	348 3	Promoter (1-2kb)	0.42	0.24	0.18	<i>PTGER4P2-CDK2AP2P2</i>	PTGER4P2-CDK2AP2P2 read through, transcribed pseudogene	No	<b>0 bp</b> (cg17548900), $P > 0.05$	
11	133519484	133519982	499 3	Distal Intergenic	0.35	0.60	-0.25			No	10 kb	
13	27295982	27296125	144 3	Distal Intergenic	0.20	0.43	-0.23			Yes	<b>0 bp</b> (cg16557370 <sup>d</sup> /cg08419873), $P > 0.05$	
14	101291034	101291083	50 3	Promoter (1-2kb)	0.45	0.64	-0.19	<i>MEG3</i>	Maternally expressed 3 (non-protein coding)	Yes	<b>0 bp</b> (cg23870378), $P > 0.05$	
15	21083039	21083557	519 3	Distal Intergenic	0.62	0.42	0.20			No	50 kb	
16	22545670	22545683	14 3	Exon	0.59	0.80	-0.21	<i>NPIP5</i>	Nuclear pore complex interacting protein family member B5	Yes	5 kb	
20	23515851	23516134	284 3	Intron	0.67	0.84	-0.17	<i>CST13P</i>	Cystatin 13, pseudogene	No	4 kb	
<b>X<sup>e</sup></b>	<b>130863481</b>	<b>130863509</b>	<b>29 3</b>	<b>Intron</b>	<b>0.66</b>	<b>0.40</b>	<b>0.26</b>	<b>FIRRE</b>	<b>Firre intergenic repeating RNA element</b>	Yes	<b>6 bp</b> (cg08117231), $P > 0.05$	

Source data are provided as a Source Data file. <sup>a</sup>Genomic coordinates are based on human genome build GRCh37/hg19. The MS-DMRs were annotated using the ChIPseeker R/Bioconductor package (v1.14.2)<sup>4</sup>. <sup>b</sup>In this column the approximate distance to the closest EWAS EPIC probe is listed. When the distance is  $< 1$  kb, then of this EPIC probe the  $P_{W-A}$ -value of the pair-wise analysis using the EPIC array data of the 45 MZ twin pairs adjusted for cell-type composition is listed as well. If the distance is 0 bp, then the EPIC probe is located within the MS-DMR. <sup>c</sup>cg16446288 is exactly located at chr7:98424232-98424233. <sup>d</sup>cg16557370 is

exactly located at chr13:27295982-27295983. <sup>a</sup>This MS-DMR fulfilled the stringent selection criteria of  $\geq 3$  CpGs, each having  $P < 0.05$  (two-tailed paired T-test) and absolute mean methylation difference  $> 0.20$ , and a maximum of 500 bp distance between neighbouring significant CpGs.  $P_{W-A}$  = P-value two-tailed Wilcoxon signed-rank test adjusted for cell-type composition. Robust DMR = DMR that shows overall consistent methylation differences (same direction) across the entire DMR (applying the lower methylation threshold of 0.15 resulted in several MS-DMRs not showing consistent methylation differences (same direction) across the entire DMR). Please see the Source Data File for details.  $\Delta\beta$  = Within-pair  $\beta$ -value difference (clinically MS-affected MZ co-twin – non-affected MZ co-twin).

**Supplementary Table 4. Within-pair differentially methylated regions (WP-DMRs)<sup>a</sup> identified in the EPIC array data of the 45 MZ twins clinically discordant for multiple sclerosis (MS).**

Gene locus	Chr	IR	Location first CpG <sup>b</sup>	Location last CpG <sup>b</sup>	#EPIC probes	#Twin pairs	Abnormal methylation profile	Methylation aberration	Pair	Treatment
<i>RBP7</i>	1		10057303	10057312	3	1	Non-affected co-twin	Hyper	E	GLAT
<i>KIF26B</i>	1		245710332	245710401	3	1	MS co-twin	Hypo	AG	IFN
<i>PAX8-AS1/PAX8/LOC440839/LOC654433</i>	2		113992694	113993313	7	1	MS co-twin	Hypo	V	DMF
<i>DUSP19</i>	2		183943175	183943698	9	1	MS co-twin	Hyper	AD	IFN
<i>PLOD2</i>	3		145878963	145878979	3	1	MS co-twin	Hyper	AN	IFN
<i>LRRC34</i>	3		169531663	169531783	3	1	MS co-twin	Hyper	AD	IFN
<i>RP11-1398P2.1</i>	4		1581921	1582181	4	1	MS co-twin	Hyper	V	DMF
<i>TACR3</i>	4		104640662	104641250	4	1	MS co-twin	Hyper	G	
<i>PCDH10</i>	4		134070433	134070441	3	1	MS co-twin	Hyper	Y	IFN
<i>PCDHG</i> gene cluster	5		140749783	140750160	4	1	MS co-twin	Hyper	P	
<i>PCDHG</i> gene cluster	5		140762261	140762315	3	1	MS co-twin	Hyper	P	
<i>PCDHG</i> gene cluster	5		140792511	140792540	3	1	MS co-twin	Hyper	P	
<i>PCDHG</i> gene cluster	5		140810051	140810137	3	1	MS co-twin	Hyper	P	
<i>DPYSL3</i>	5		146889238	146889275	3	1	Non-affected co-twin	Hyper	AA	IFN
<i>CCNG1</i>	5		162864291	162864633	8	1	MS co-twin	Hyper	Y	IFN
<i>HIST1H3E</i>	6		26224013	26224925	6	2	Non-affected co-twins	Hyper	H/AG	TFM/IFN
<i>HIST1H2AL</i>	6		27833095	27833555	3	1	Non-affected co-twin	Hyper	U	IFN
NA	6		30434109	30434324	5	1	MS co-twin	Hyper	V	DMF
<i>AGPAT1/RNF5/RNF5P1</i>	6		32146466	32146595	4	1	MS co-twin	Hyper	P	
<i>DNAH8</i>	6		38682995	38683221	4	1	MS co-twin	Hyper	AD	IFN
<i>SVOPL</i>	7	Y	138348774	138349443	5	1	Non-affected co-twin	Hypo	AB	IFN
NA	7		158750244	158751184	5	1	MS co-twin	Hyper	V	DMF
<i>DLC1</i>	8		13134144	13134166	3	1	Non-affected co-twin	Hyper	R	
<i>TRMT12</i>	8		125462982	125463066	4	1	MS co-twin	Hyper	W	
<i>NEBL-AS1/NEBL</i>	10		21462747	21462768	3	1	Non-affected co-twin	Hyper	AA	IFN
<i>HSD17B7P2</i>	10		38645376	38645740	3	1	Non-affected co-twin	Hyper	AD	IFN
<i>CAT</i>	11		34460140	34460557	3	1	MS co-twin	Hyper	H	TFM
<i>DIXDC1</i>	11		111847892	111848326	3	1	Non-affected co-twin	Hyper	AA	IFN
<i>WDR66</i>	12		122356316	122356598	5	1	Non-affected co-twin	Hyper	AD	IFN
<i>GPR133</i>	12		131488390	131488726	3	1	MS co-twin	Hypo	AD	IFN
<i>DHRS4L2</i>	14		24438909	24439192	4	1	Non-affected co-twin	Hyper	AD	IFN
<i>CLEC14A</i>	14		38724646	38724675	3	1	MS co-twin	Hyper	W	
<i>PAK6/C15orf56</i>	15		40545050	40545145	3	1	Non-affected co-twin	Hyper	E	GLAT
<i>LOC101928414/CTD-2651B20.3</i>	15		45571526	45571636	4	1	MS co-twin	Hyper	AD	IFN
NA	15		53092788	53093509	3	1	MS co-twin	Hyper	AG	IFN
<i>CLK3</i>	15		74890733	74891207	3	1	Non-affected co-twin	Hyper	AD	IFN
<i>UNC45A</i>	15		91473167	91473569	6	1	Non-affected co-twin	Hyper	P	
<i>ITGAM</i>	16		31342453	31343056	4	1	Non-affected co-twin	Hyper	T	
<i>C17orf97</i>	17		259755	259924	3	1	MS co-twin	Hyper	AG	IFN
<i>L3MBTL4</i>	18		6414958	6414978	4	1	Non-affected co-twin	Hyper	AA	IFN
<i>ZNF254</i>	19		24269919	24270468	4	1	MS co-twin	Hyper	AD	IFN
<i>LYPD5</i>	19		44324903	44325004	3	1	MS co-twin	Hyper	AD	IFN
<i>ISOC2</i>	19		55972646	55973778	11	2	MS & Non-affected co-twin	Hyper	H/BA	TFM/-
<i>AURKC</i>	19		57742345	57742423	4	1	MS co-twin	Hypo	AG	IFN
<i>HM13/MCTS2P</i>	20	Y	30134929	30135362	7	1	Non-affected co-twin	Hyper	B	IFN

Source data are provided as a Source Data file. <sup>a</sup>WP-DMRs were defined as  $\geq 3$  CpGs with a within-pair  $\beta$ -value difference  $> 0.20$  (adjusted for cell-type composition) and a maximum 1 kb distance between neighboring CpGs (the 257 IFN-associated CpGs were excluded from this analysis). In addition, the  $\beta$ -value of the "abnormally methylated" co-twin had to be greater than  $\pm 3$  standard deviations from the mean. <sup>b</sup>All genome coordinates are based on human genome build GRCh37/hg19. Chr = chromosome, DMF = dimethyl fumarate, GLAT = glatiramer acetate, IFN = interferon-beta, IR = imprinted region (Y = yes), TFM = teriflunomide.

**Supplementary Table 5. Results of the evaluation whether the 27 WP-DMRs<sup>a</sup>, that were aberrantly methylated in the MS-affected co-twins (listed in Table 4), were present in other pairs as well by applying a lower  $\Delta\beta$ -value threshold of 0.15.** In total, four WP-DMRs were also identified in other twin pairs, of which one intergenic WP-DMR was present in 4 pairs and always associated with the MS phenotype (in bold). Hence, in total 24 MS-associated WP-DMRs were identified in 11 pairs, of which 23 were pair-specific and one present in 4 twin pairs. Clinical characteristics such as gender, MS course, disease duration at sampling date, age at first disease manifestation, MS treatment, and pack-years at sample collection did not differ between these 11 twin pairs and the 34 other pairs ( $P>0.05$ , two-tailed Wilcoxon rank sum test for continuous data and two-tailed Fisher's exact test for categorical data).

Gene locus	Chr	IR	Location first CpG <sup>b</sup>	Location last CpG <sup>b</sup>	#EPIC probes	#Twin pairs	Abnormal methylation profile	Methylation aberration	Pair	Treatment	#Twin pairs $\Delta\beta$ -value 0.15 <sup>c</sup>	Abnormal methylation profile	Pair <sup>d</sup>	Treatment
<i>KIF26B</i>	1		245710332	245710401	3	1	MS co-twin	Hypo	AG	IFN	1	MS co-twin	AG	IFN
<i>PAX8-AS1/PAX8/LOC440839/LOC654433</i>	2		113992694	113993313	7	1	MS co-twin	Hypo	V	DMF	2	Non-affected co-twin & MS co-twin	U/V	IFN/DMF
<i>DUSP19</i>	2		183943175	183943698	9	1	MS co-twin	Hyper	AD	IFN	1	MS co-twin	AD	IFN
<i>PLOD2</i>	3		145878963	145878979	3	1	MS co-twin	Hyper	AN	IFN	1	MS co-twin	AN	IFN
<i>LRRRC34</i>	3		169531663	169531783	3	1	MS co-twin	Hyper	AD	IFN	1	MS co-twin	AD	IFN
<i>RP11-1398P2.1</i>	4		1581921	1582181	4	1	MS co-twin	Hyper	V	DMF	3	MS co-twin & Non-affected co-twin	L/V/AF	-/DMF/NAT
<i>TACR3</i>	4		104640662	104641250	4	1	MS co-twin	Hyper	G		1	MS co-twin	G	
<i>PCDH10</i>	4		134070433	134070441	3	1	MS co-twin	Hyper	Y	IFN	1	MS co-twin	Y	IFN
<i>PCDHG</i> gene cluster	5		140749783	140750160	4	1	MS co-twin	Hyper	P		1	MS co-twin	P	
<i>PCDHG</i> gene cluster	5		140762261	140762315	3	1	MS co-twin	Hyper	P		1	MS co-twin	P	
<i>PCDHG</i> gene cluster	5		140792511	140792540	3	1	MS co-twin	Hyper	P		1	MS co-twin	P	
<i>PCDHG</i> gene cluster	5		140810051	140810137	3	1	MS co-twin	Hyper	P		1	MS co-twin	P	
<i>-CCNG1</i>	5		162864291	162864633	8	1	MS co-twin	Hyper	Y	IFN	1	MS co-twin	Y	IFN
<i>NA</i>	6		30434109	30434324	5	1	MS co-twin	Hyper	V	DMF	1	MS co-twin	V	DMF
<i>AGPAT1/RNF5/RNF5P1</i>	6		32146466	32146595	4	1	MS co-twin	Hyper	P		1	MS co-twin	P	
<i>DNAH8</i>	6		38682995	38683221	4	1	MS co-twin	Hyper	AD	IFN	1	MS co-twin	AD	IFN
<b>NA</b>	<b>7</b>		<b>158750244</b>	<b>158751184</b>	<b>5</b>	<b>1</b>	<b>MS co-twin</b>	<b>Hyper</b>	<b>V</b>	<b>DMF</b>	<b>4</b>	<b>MS co-twin</b>	<b>U/V/AB/IFN/DMF/IFN/IFN</b>	
<i>TRMT12</i>	8		125462982	125463066	4	1	MS co-twin	Hyper	W		1	MS co-twin	W	
<i>CAT</i>	11		34460140	34460557	3	1	MS co-twin	Hyper	H	TFM	2	MS co-twin & Non-affected co-twin	H/P	TFM/-
<i>GPR133</i>	12		131488390	131488726	3	1	MS co-twin	Hypo	AD	IFN	1	MS co-twin	AD	IFN
<i>CLEC14A</i>	14		38724646	38724675	3	1	MS co-twin	Hyper	W		1	MS co-twin	W	
<i>LOC101928414/CTD-2651B20.3</i>	15		45571526	45571636	4	1	MS co-twin	Hyper	AD	IFN	1	MS co-twin	AD	IFN
<i>NA</i>	15		53092788	53093509	3	1	MS co-twin	Hyper	AG	IFN	1	MS co-twin	AG	IFN
<i>C17orf97</i>	17		259755	259924	3	1	MS co-twin	Hyper	AG	IFN	1	MS co-twin	AG	IFN
<i>ZNF254</i>	19		24269919	24270468	4	1	MS co-twin	Hyper	AD	IFN	1	MS co-twin	AD	IFN
<i>LYPD5</i>	19		44324903	44325004	3	1	MS co-twin	Hyper	AD	IFN	1	MS co-twin	AD	IFN
<i>AURKC</i>	19		57742345	57742423	4	1	MS co-twin	Hypo	AG	IFN	1	MS co-twin	AG	IFN

Source data are provided as a Source Data file. <sup>a</sup>WP-DMRs were defined as  $\geq 3$  CpGs with a within-pair  $\beta$ -value difference  $>0.20$  (adjusted for cell-type composition) and a maximum 1 kb distance between neighboring CpGs (the 257 IFN-associated CpGs were excluded from this analysis). In addition, the  $\beta$ -value of the “abnormally methylated” co-twin had to be greater than  $\pm 3$  standard deviations from the mean. <sup>b</sup>All genome coordinates are based on human genome build GRCh37/hg19. <sup>c</sup>A  $\Delta\beta$ -value threshold of 0.15 was used to evaluate whether the 27 WP-DMRs, that were aberrantly methylated in the MS-affected co-twins, were present in other twin pairs as well. Chr = chromosome, DMF = dimethyl fumarate, GLAT = glatiramer acetate, IFN = interferon-beta, IR = imprinted region (Y = yes), TFM = teriflunomide.

**Supplementary Table 6. The 25 differentially variable positions (DVPs) identified between MS-affected and clinically non-affected MZ co-twins using iEVORA (n = 45 twin pairs).<sup>5</sup>**

Probe ID	Location <sup>a</sup>	Gene	Functional region <sup>b</sup>	P <sub>unpaired T-Test</sub>	P <sub>FDR-corrected Barlett's test</sub>	Hypervariable Group	
1	cg09319843	chr18:25757569	<i>CDH2</i>	TSS200/CpG Island	1.49*10 <sup>-03</sup>	2.63*10 <sup>-04</sup>	Non-affected
2	cg07380496	chr5:71403420	<i>MAP1B</i>	1stExon/CpG Island	2.05*10 <sup>-03</sup>	3.08*10 <sup>-07</sup>	Non-affected
3	cg08927443	chr18:25757565	<i>CDH2</i>	TSS200/CpG Island	2.25*10 <sup>-03</sup>	1.62*10 <sup>-07</sup>	Non-affected
4	cg21303011	chr3:24537177	<i>THRB</i>	TSS1500/CpG Island	4.28*10 <sup>-03</sup>	1.81*10 <sup>-04</sup>	Non-affected
5	cg11732619	chr5:168728076	<i>SLIT3</i>	5'UTR/1stExon/CpG Island	4.92*10 <sup>-03</sup>	2.02*10 <sup>-04</sup>	Non-affected
6	cg11181094	chr9:125093748			6.07*10 <sup>-03</sup>	4.82*10 <sup>-05</sup>	Non-affected
7	cg13913015	chr2:47797963	<i>KCNK12</i>	TSS1500/CpG Island	9.64*10 <sup>-03</sup>	5.09*10 <sup>-06</sup>	Non-affected
8	cg06090660	chr18:25757555	<i>CDH2</i>	TSS200/CpG Island	1.17*10 <sup>-02</sup>	5.39*10 <sup>-04</sup>	Non-affected
9	cg11777419	chr14:104604401	<i>KIF26A</i>	TSS1500/CpG Island	1.20*10 <sup>-02</sup>	5.49*10 <sup>-06</sup>	Non-affected
10	cg23526824	chr17:38245542	<i>THRA</i>	Body	1.29*10 <sup>-02</sup>	5.22*10 <sup>-04</sup>	Non-affected
11	cg26452004	chr14:69726546	<i>GALNT16</i>	TSS200/CpG Island	1.57*10 <sup>-02</sup>	7.33*10 <sup>-04</sup>	Non-affected
12	cg26330510	chr5:1155853			1.62*10 <sup>-02</sup>	2.45*10 <sup>-09</sup>	Non-affected
13	cg07147599	chr16:50502136			1.88*10 <sup>-02</sup>	3.65*10 <sup>-04</sup>	Non-affected
14	cg26245302	chr6:163148501	<i>PARK2/PACRG</i>	TSS1500/Body/5'UTR/CpG Island	2.34*10 <sup>-02</sup>	8.30*10 <sup>-06</sup>	Non-affected
15	cg09936645	chr1:207627581	<i>CR2</i>	TSS200/CpG Island	2.43*10 <sup>-02</sup>	3.02*10 <sup>-08</sup>	Non-affected
16	cg07848601	chr5:170289430	<i>RANBP17</i>	Body/CpG Island	2.52*10 <sup>-02</sup>	4.81*10 <sup>-04</sup>	Non-affected
17	cg08558397	chr7:752149	<i>PRKAR1B</i>	5'UTR/1stExon/CpG Island	2.90*10 <sup>-02</sup>	9.26*10 <sup>-06</sup>	Non-affected
18	cg23307163	chr10:4828732			2.98*10 <sup>-02</sup>	4.68*10 <sup>-04</sup>	Non-affected
19	cg16026114	chr1:232765417			3.69*10 <sup>-02</sup>	2.16*10 <sup>-06</sup>	Non-affected
20	cg25088874	chr4:95678817	<i>BMPR1B</i>	TSS1500/CpG Island	4.19*10 <sup>-02</sup>	4.60*10 <sup>-06</sup>	Non-affected
21	cg23683528	chr2:235860449	<i>SH3BP4</i>	TSS200/CpG Island	4.35*10 <sup>-02</sup>	5.96*10 <sup>-06</sup>	MS-affected
22	cg20928782	chr11:63803364	<i>MACROD1</i>	Body	4.46*10 <sup>-02</sup>	2.22*10 <sup>-04</sup>	MS-affected
23	cg12954230	chr15:100882231	<i>ADAMTS17</i>	TSS200/CpG Island	4.55*10 <sup>-02</sup>	3.57*10 <sup>-07</sup>	Non-affected
24	cg21947590	chr19:620162	<i>POLRMT</i>	Body/CpG Island	4.58*10 <sup>-02</sup>	4.33*10 <sup>-05</sup>	Non-affected
25	cg09272992	chr7:150497601	<i>TMEM176B</i>	5'UTR/TSS1500/1stExon/CpG Island	4.58*10 <sup>-02</sup>	7.97*10 <sup>-06</sup>	MS-affected

Source data are provided as a Source Data file. DVPs were defined as CpGs with a FDR-corrected Barlett's P-value<0.001 and raw T-test P-value<0.05. <sup>a</sup>All genome coordinates are based on human genome build GRCh37/hg19. <sup>b</sup>Based on information provided by the Illumina manifest.



**Supplementary Table 7. Characteristics of the seven most significant interferon-beta-associated differentially methylated positions (IFN-DMPs) (absolute mean  $\Delta\beta$ -value > 0.10 and  $P_{W-U} < 0.001$ ), identified by a pair-wise analysis only including the EPIC array data of the 12 pairs of which the MS-affected co-twins were treated with IFN at the moment of blood collection (n = 12 twin pairs).**

Probe ID	Location <sup>a</sup>	Gene	Functional region <sup>b</sup>	450k	Mean $\beta$ -value (U/A)		Mean $\Delta\beta$ -value (95% CI) (U)	Mean $\Delta\beta$ -value (95% CI) (A)	$\beta$ -value range	$P_{W-U}/P_{W-A}$	Full name
					IFN-treated MS-affected co-twins	non-affected co-twins					
cg03607951	chr1:79085586	<i>IFI44L</i>	TSS1500/DHS	Y	0.57/0.59	0.69/0.68	-0.12 (-0.16,-0.08)	-0.09 (-0.13,-0.05)	0.44-0.77	$9.77 \times 10^{-4}$ $/1.46 \times 10^{-3}$	Interferon-induced protein 44-like
cg06981309	chr3:146260954	<i>PLSCR1</i>	5'UTR/DHS	Y	0.59/0.60	0.71/0.70	-0.12 (-0.16,-0.07)	-0.09 (-0.13,-0.05)	0.50-0.76	$4.88 \times 10^{-4}$ $/4.88 \times 10^{-4}$	Phospholipid scramblase 1
cg10549986	chr2:7018153	<i>RSAD2</i>	1 <sup>st</sup> exon/DHS	Y	0.17/0.18	0.31/0.30	-0.14 (-0.19,-0.09)	-0.12 (-0.16,-0.07)	0.11-0.44	$4.88 \times 10^{-4}$ $/4.88 \times 10^{-4}$	Radical S-adenosyl methionine domain-containing protein 2
cg10771443	chr2:7018855	<i>RSAD2</i>	Body/DHS	N	0.36/0.38	0.49/0.45	-0.13 (-0.18,-0.07)	-0.07 (-0.12,-0.03)	0.26-0.56	$9.77 \times 10^{-4}$ $/9.27 \times 10^{-3}$	
cg15839328	chr2:7018885	<i>RSAD2</i>	Body/DHS	N	0.36/0.38	0.49/0.47	-0.13 (-0.18,-0.08)	-0.09 (-0.12,-0.05)	0.27-0.57	$4.88 \times 10^{-4}$ $/1.46 \times 10^{-3}$	
cg21549285	chr21:42799141	<i>MX1</i>	5'UTR/DHS	Y	0.63/0.64	0.79/0.78	-0.16 (-0.22,-0.09)	-0.14 (-0.20,-0.07)	0.44-0.86	$4.88 \times 10^{-4}$ $/1.46 \times 10^{-3}$	MX dynamin like
cg26312951	chr21:42797847	<i>MX1</i>	TSS200/5'UTR/TFBS/open chromatin	Y	0.32/0.33	0.43/0.42	-0.11 (-0.16,-0.07)	-0.09 (-0.13,-0.05)	0.17-0.48	$4.88 \times 10^{-4}$ $/2.44 \times 10^{-3}$	GTPase 1

Source data are provided as a Source Data file. <sup>a</sup>All the genome coordinates are based on human genome build GRCh37/hg19. <sup>b</sup>Based on information provided by the Illumina manifest. Since all genes have multiple transcripts, the "UCSC\_RefGene\_Group" gene-related location is listed. <sup>c</sup>Other potential IFN-DMPs in *IFI44L* are cg13452062 ( $\Delta\beta$ -value = 0.20,  $P_{W-U}$  = 0.002) and cg05696877 ( $\Delta\beta$ -value = 0.15,  $P_{W-U}$  = 0.001). 450k = probe present on the 450k array (Y = yes, N = no), A = adjusted for cell-type composition, CI = confidence interval, DHS = DNase I hypersensitive site, IFN-DMPs = interferon-beta treatment-associated differentially methylated positions, n = number of pairs,  $P_{W-A}$  = P-value two-tailed Wilcoxon signed-rank test adjusted for cell-type composition,  $P_{W-U}$  = P-value two-tailed Wilcoxon signed-rank test unadjusted for cell-type composition, TFBS = transcription factor binding site, TSS200 = the region from transcription start site (TSS) to -200 nt upstream of TSS, TSS1500 = -200 to -1500 nt upstream of TSS, U = unadjusted for cell-type composition, 5'UTR = 5' untranslated region,  $\Delta\beta$  = within-pair  $\beta$ -value difference (clinically MS-affected, IFN-treated MZ co-twin - non-affected MZ co-twin).

**Supplementary Table 8. Estimated cell type proportions of the 12 pairs of which the clinically MS-affected MZ co-twins were treated with interferon-beta (IFN) at the moment of blood collection (n = 12 twin pairs).<sup>a</sup>**

Cell Types	Cellular proportions IFN-treated MS- affected MZ co-twins <sup>b</sup>	Cellular proportions clinically non-affected MZ co-twins <sup>b</sup>	Mean within-pair difference(95% CI) <sup>b</sup>	P <sub>w</sub>
CD4+ T cells	0.35 ± 0.08	0.34 ± 0.06	0.01 (-0.02,0.04)	0.57
CD8+ T cells	0.18 ± 0.08	0.19 ± 0.06	-0.01 (-0.04,0.03)	0.57
CD19+ B cells	0.14 ± 0.04	0.12 ± 0.03	0.02 (0.01,0.04)	<b>0.01</b>
CD14+ Monocytes	0.18 ± 0.07	0.19 ± 0.07	-0.01 (-0.04,0.02)	0.42
CD56+ NK cells	0.06 ± 0.07	0.12 ± 0.09	-0.06 (-0.10,-0.02)	<b>0.007</b>
Granulocytes	0.09 ± 0.12	0.04 ± 0.06	0.05 (-0.01,0.10)	0.09

Source data are provided as a Source Data file. <sup>a</sup>Cell type proportions were estimated using the DNA methylome reference-based method of Houseman et al.<sup>6</sup> implemented in the *minfi* R/Bioconductor package<sup>7</sup>. <sup>b</sup>Values are expressed as mean ± SD. <sup>c</sup>Within-pair difference = clinically MS-affected, IFN-treated MZ co-twin – non-affected MZ co-twin. CI = confidence interval, n = number of pairs. P<sub>w</sub>=P-value nonparametric two-tailed Wilcoxon signed-rank test. P-values<0.05 are in **bold**.

**Supplementary Table 9. Summary of the 41 GC-DMRs<sup>a</sup> (in 39 genes) that overlap with GC-response (dexamethasone) genes recorded in the EMBL-EBI Expression Atlas (accessed May 2018) (n =1 pair).**

Gene	Chr	Start <sup>b</sup>	End <sup>b</sup>	Width	#CG	Location	Mean	Mean	Mean	Methylation	Expression effect of GCs recorded in EMBL-EBI Expression Atlas
							methylation GC-treated MS co-twin	methylation unaffected co-twin	methylation difference	aberration GC-treatment	
<i>ADAMTS2</i>	5	178652819	178652857	39	2	Intron	0.61	0.87	-0.26	Hypo	Up
<i>ADAMTS2</i>	5	178661985	178662010	26	3	Intron	0.85	0.60	0.25	Hyper	Up
<i>ADORA3</i>	1	112049050	112049096	47	2	Intron	0.64	0.90	-0.26	Hypo	Up
<i>ALK</i>	2	29794637	29794701	65	3	Intron	0.62	0.89	-0.27	Hypo	Up
<i>APOBEC3A_B</i>	22	39363192	39363237	46	2	Intron	0.88	0.55	0.33	Hyper	Down
<i>ARRDC2</i>	19	18124639	18124670	32	2	3' UTR	0.28	0.58	-0.29	Hypo	Up
<i>ATP6V0D2</i>	8	87126109	87126154	46	2	Exon	0.99	0.66	0.33	Hyper	Down
<i>BCL11A</i>	2	60756422	60756472	51	2	Intron	0.73	0.44	0.29	Hyper	Down
<i>CALHM6</i>	6	116783956	116783983	28	2	Intron	0.80	0.51	0.29	Hyper	Up
<i>CCL26</i>	7	75416004	75416035	32	2	Intron	0.61	0.88	-0.26	Hypo	Up
<i>CD83</i>	6	14122060	14122276	217	4	Intron	0.57	0.30	0.27	Hyper	Down
<i>CDH1</i>	16	68816259	68816295	37	2	Intron	0.51	0.05	0.45	Hyper	Down
<i>COL4A2</i>	13	110996989	110997025	37	3	Intron	0.67	0.92	-0.25	Hypo	Up
<i>DDC</i>	7	50597354	50597383	30	2	Intron	0.85	0.58	0.27	Hyper	Up
<i>EGFR</i>	7	55180338	55180378	41	2	Intron	0.52	0.84	-0.32	Hypo	Up
<i>EVL</i>	14	100524807	100524853	47	2	Intron	0.89	0.62	0.27	Hyper	Down
<i>FAM49A</i>	2	16745180	16745220	41	2	Intron	0.46	0.78	-0.32	Hypo	Up
<i>FETUB</i>	3	186367720	186367750	31	2	Intron	0.52	0.21	0.31	Hyper	Up
<i>FGF18</i>	5	170864899	170864934	36	2	Intron	0.87	0.57	0.30	Hyper	Up
<i>FKBP1B</i>	2	24274695	24274736	42	2	Intron	0.69	0.96	-0.28	Hypo	Up
<i>GMPR</i>	6	16260445	16260473	29	2	Intron	0.68	0.93	-0.26	Hypo	Up
<i>HBEGF</i>	5	139717277	139717324	48	2	Intron	0.76	0.47	0.29	Hyper	Down
<i>IP6K3</i>	6	33710284	33710320	37	2	Intron	0.27	0.60	-0.33	Hypo	Up
<i>KALRN</i>	3	124281329	124281373	45	2	Intron	0.77	0.51	0.26	Hyper	Down
<i>KLHL29</i>	2	23750256	23750281	26	2	Intron	0.88	0.52	0.36	Hyper	Up
<i>LIFR</i>	5	38601623	38601671	49	2	Upstream	0.60	0.86	-0.26	Hypo	Up
<i>MGAT4A</i>	2	99264172	99264216	45	2	Intron	0.23	0.53	-0.30	Hypo	Up
<i>MTSS1</i>	8	125727330	125727375	46	2	Intron	0.91	0.60	0.32	Hyper	Up
<i>MYO7A</i>	11	76905938	76908116	2179	3	Intron	0.62	0.97	-0.36	Hypo	Up
<i>MYO7A</i>	11	76898554	76898595	42	2	Intron	0.17	0.48	-0.30	Hypo	Up
<i>NDRG1</i>	8	134261878	134261928	51	3	Intron	0.04	0.46	-0.42	Hypo	Up
<i>P2RY6</i>	11	73000436	73000474	39	3	Intron	0.50	0.23	0.28	Hyper	Down
<i>PGBD5</i>	1	230554455	230554484	30	2	Intron	0.75	0.46	0.29	Hyper	Down
<i>PHACTR3</i>	20	58236433	58236475	43	3	Intron	0.37	0.66	-0.29	Hypo	Up
<i>PHGDH</i>	1	120267645	120267683	39	2	Intron	0.93	0.67	0.25	Hyper	Down
<i>PRSS21</i>	16	2869882	2869916	35	2	Intron	0.92	0.57	0.35	Hyper	Down
<i>RGCC</i>	13	42038266	42038305	40	2	Intron	0.11	0.42	-0.31	Hypo	Up/Down
<i>RUNX2</i>	6	45448204	45448233	30	2	Intron	0.83	0.56	0.28	Hyper	Up
<i>SPATA13</i>	13	24825973	24825999	27	4	Exon	0.06	0.31	-0.25	Hypo	Down
<i>TIMP3</i>	22	33197034	33197061	28	2	Exon	0.75	0.48	0.28	Hyper	Down
<i>ZBTB16</i>	11	114050079	114050114	36	2	Intron	0.24	0.54	-0.29	Hypo	Up

Source data are provided as a Source Data file. <sup>a</sup>The glucocorticoid treatment-associated DMRs (GC-DMRs) result from the DMR analysis of the WGBS data of CD4+ memory T-cells of a MS discordant MZ twin pair of which the MS-affected co-twin was very recently treated with GCs at the moment of blood collection (n=1). The GC-DMRs were identified using the DSS-single package<sup>8</sup>, including a smoothing span of 100 bp, a minimum region length of 25 bp with  $\geq 2$  CpGs and a P-value $<0.01$ . The absolute mean methylation difference had to be larger than 0.25, and to limit the number of false positives only GC-DMRs located in reported GC-response genes were considered. <sup>b</sup>Genomic coordinates are based on human genome build GRCh37/hg19.

**Supplementary Table 10. Number of hyper- and hypomethylated CpGs in the EPIC array EWAS data of the 45 MZ twins clinically discordant for MS, according to different  $\Delta\beta$ -values and P-value thresholds (n = 45 twin pairs).**

Hypermethylated in MS-affected co-twin	#CpGs	%CpGs	Hypomethylated in MS-affected co-twin	#CpGs	%CpGs
<u>Unadjusted for cell type composition</u>			<u>Unadjusted for cell type composition</u>		
Mean $\Delta\beta$ -value>0	502222	59.1	Mean $\Delta\beta$ -values<0	347551	40.9
Mean $\Delta\beta$ -value>0.005	168434	62.2	Mean $\Delta\beta$ -value<-0.005	102184	37.8
Mean $\Delta\beta$ -value>0.01	50255	54.4	Mean $\Delta\beta$ -value<-0.01	42090	45.6
Mean $\Delta\beta$ -value>0 & P<0.001	2913	60.1	Mean $\Delta\beta$ -value<0 & P<0.001	1933	39.9
Mean $\Delta\beta$ -value>0.005 & P<0.001	2880	59.9	Mean $\Delta\beta$ -value<-0.005 & P<0.001	1930	40.1
Mean $\Delta\beta$ -value>0.01 & P<0.001	2496	58.1	Mean $\Delta\beta$ -value<-0.01 & P<0.001	1806	41.9
<u>Adjusted for cell type composition</u>			<u>Adjusted for cell type composition</u>		
Mean $\Delta\beta$ -value>0	475327	55.9	Mean $\Delta\beta$ -values<0	374446	44.1
Mean $\Delta\beta$ -value>0.005	101816	62.8	Mean $\Delta\beta$ -value<-0.005	60242	37.2
Mean $\Delta\beta$ -value>0.01	17129	63.8	Mean $\Delta\beta$ -value<-0.01	9712	36.2
Mean $\Delta\beta$ -value>0 & P<0.001	385	55.2	Mean $\Delta\beta$ -value<0 & P<0.001	313	44.8
Mean $\Delta\beta$ -value>0.005 & P<0.001	354	53.4	Mean $\Delta\beta$ -value<-0.005 & P<0.001	309	46.6
Mean $\Delta\beta$ -value>0.01 & P<0.001	249	53.9	Mean $\Delta\beta$ -value<-0.01 & P<0.001	213	46.1

$\Delta\beta$ -values = clinically MS-affected MZ co-twin – non-affected MZ co-twin. n= number of twin pairs, P = P-value two-tailed Wilcoxon signed-rank test.

**Supplementary Table 11. Statistical power of the multiple sclerosis EWAS that includes 45 MZ twin pairs clinically discordant for multiple sclerosis.**

Magnitude of the correlation	Power to detect a mean $\beta$ -value difference of (at least) <b>0.05</b> at $\alpha = 1 \times 10^{-7}$	Power to detect a mean $\beta$ -value difference of (at least) <b>0.04</b> at $\alpha = 1 \times 10^{-7}$	Power to detect a mean $\beta$ -value difference of (at least) <b>0.03</b> at $\alpha = 1 \times 10^{-7}$
0	0.984	0.749	0.210
0.2	0.999	0.914	0.390
0.4	>0.999	0.991	0.686
0.6	>0.999	>0.999	0.961
0.8	>0.999	>0.999	>0.999
1.0	>0.999	>0.999	>0.999

The table shows the statistical power to detect a mean  $\beta$ -value difference of (at least) 0.05, 0.04 and 0.03 with a (genome-wide) significance threshold of  $1 \times 10^{-7}$ , a sample size of 45 MS discordant MZ twin pairs using a two-sided paired T-test and assuming a standard deviation of 0.0266 (which is the true median standard deviation observed in the EPIC array data).

**Supplementary Table 12. Primer sequences and PCR conditions.<sup>a</sup>**

Method	CpG/SNP number	Gene/element	Forward primer sequence (5'→3')	C	Reverse primer sequence (5'→3')	C	T	Cyc	Product size	# CpGs
TDBS	cg12393503	<i>ECT2</i>	GATTTTGTGTGAGTGAGAGAGGTGT	133	TCITCTATCCAAAAAACAACAATA	133	58	42	252	19
	cg01447350	<i>IL34</i>	TTTTAGTTATTTGGGAGGTTGAAGTAG	133	ATCCATAAATAACTCAAATAAAAAACAAA	133	59	42	340	8
	cg02520593	<i>SELPLG</i>	TTTGTGTTTAAGAGGTAATAATTGAAGTT	133	ATATCCCAACTACAAATCCAATACAAA	133	58	42	230	3
	cg27037608	<i>TMEM232</i>	ATTAGGATTTATAAGTGAATTTTTATTGTTT	133	CAAAACATTCTAAATACITTTTACTCACTA	133	60	40	379	9
	cg25345365	<i>ZBTB16</i>	ATTTTTTGGAGGAAAGAATATATAGTGT	133	AATACAAAATATACCAAAAACAACAACC	133	60	42	191	3
		<i>ALU<sup>b</sup></i>	TTTTAGTATTTTGGGAGGT	100	CCCAAACTAAAAACAATAAC	100	60	30	232	17
		<i>HERVK<sup>c</sup></i>	TATTTTTAATTTTAAGTATTTAGGGAT	100	TTCCTCTTATCTCAACTACAAAAA	100	56	30	233	6
		<i>LINE1<sup>d</sup></i>	GGTTTATTTTATTAGGGAGTGTAGAT	100	AAACCCCTCAAACCAATATAAAATATAA	100	54	30	257	18
BisConAssay		<i>PTPRVP</i>	TGGGGTAATGATGAGAGATGG	100	CTCTCTTTATTTCAAAACCCCTCA	100	58	40	343	NA
TDS	rs6471533		ACCCACTGGTTCTGGGAAG	133	TATGGCATGTTGGCAGAAGA	133	63	35	170	NA

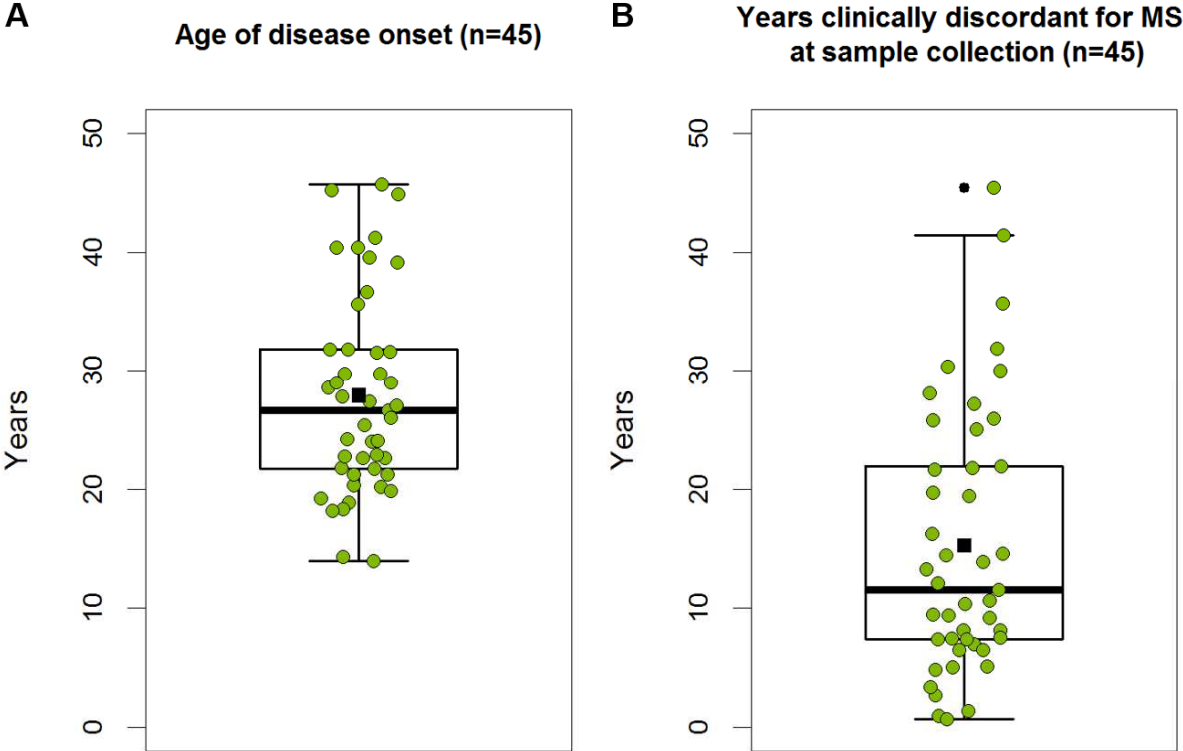
<sup>a</sup>Loci were amplified in 30 µL mixes containing 40 ng bisulfite-treated DNA (TDS: 25 ng genomic DNA), 0.2 mM of each dNTP, *n* nM of each primer, 2.5 mM MgCl<sub>2</sub>, 1.5 U HotStarTaq DNA polymerase (Qiagen) and 1X PCR buffer. DNA was denatured for 15 min at 95°C, followed by *n* cycles of 30 sec at 95°C, 1 min at T°C and 30 sec up to 1 min at 72°C. The reaction was completed by a final extension step of 5 min at 72°C. PCR products were mixed, purified using AMPure XP beads (Agencourt) and quantified using the Qubit Fluorometer and Qubit dsDNA HS assay kit (Invitrogen). *TMEM232* was sequenced with a minimum coverage of 1500 reads. All other amplicons were sequenced with a minimum coverage of at least 2000 reads. <sup>b</sup>Alu primers were designed using the consensus sequence published by Price et al.<sup>9</sup> (nucleotide positions 29-260) and generate *in silico*<sup>10</sup> an "infinite" number of specific PCR products (no mismatches allowed). <sup>c</sup>*HERVK* primers target the youngest subfamily LTR5Hs (nucleotide position 256-487) and generate *in silico*<sup>10</sup> 328 different specific PCR products (no mismatches allowed) of which 98% matches to LTR5Hs according to RepeatMasker (<http://repeatmasker.org/>). <sup>d</sup>*LINE1* Primers were designed using the promoter/5'-UTR consensus sequence (GenBank-Nr. X58075.1, nucleotide positions 105-361) and generate *in silico*<sup>10</sup> 309 different specific PCR products (no mismatches allowed), which mainly comprise the youngest subfamilies L1HS (~64%), L1PA2 (~25%) and L1PA3 (~9%). The repetitive elements were sequenced with a minimum coverage of 2000 reads, giving a >6 fold coverage per individual *HERVK* and *LINE1* element. BisConAssay = bisulfite conversion rate assay (non-bisulfite-dependent primers), C = primer concentration (nM), Cyc = number of cycles, NA = not applicable, T = annealing temperature (°C), TDBS = targeted deep bisulfite sequencing, TDS = targeted deep sequencing, #CpGs = number of CpGs present in the amplicon.

**Supplementary Table 13. Sequencing coverage statistics of the whole genome bisulfite sequencing (WGBS) analysis in CD4+ memory T-cells of four MS discordant female MZ twin pairs.**

Sample	Age	Years discordant for MS	SS	#CpGs	Average coverage	#CpGs with coverage $\geq 5$	#CpGs with coverage $\geq 10$	#CpGs with coverage $\geq 10$ across all samples included in the MS-DMR analysis	Average coverage	#CpGs with coverage $\geq 15$ across all samples included in the GC-DMR analysis	Average coverage
R-MS	39.8	14.4	N	27,951,337	11.0	25,375,958	15,495,007	2,693,926	19.1		
R-H	39.8	14.4	N	27,901,316	9.5	24,183,957	12,021,398	2,693,926	16.6		
AK-MS	46.5	19.1	Y	27,827,126	8.1	22,221,731	8,188,033	2,693,926	15.4		
AK-H	46.5	19.1	Y	27,962,853	11.4	25,245,516	16,062,214	2,693,926	20.4		
AV-MS	45.8	13.9	N	27,984,627	11.9	25,949,059	17,517,610	2,693,926	20.5		
AV-H	45.8	13.9	Y	27,972,888	10.9	25,398,997	15,269,579	2,693,926	19.1		
AY-MS <sup>a</sup>	40.9	12.0	Y	28,002,000	12.1	26,061,610	17,794,246	2,693,926	20.6	2,796,900	21.8
AY-H	40.9	12.0	N	27,970,390	10.8	25,398,195	15,270,455	2,693,926	18.7	2,796,900	20.6

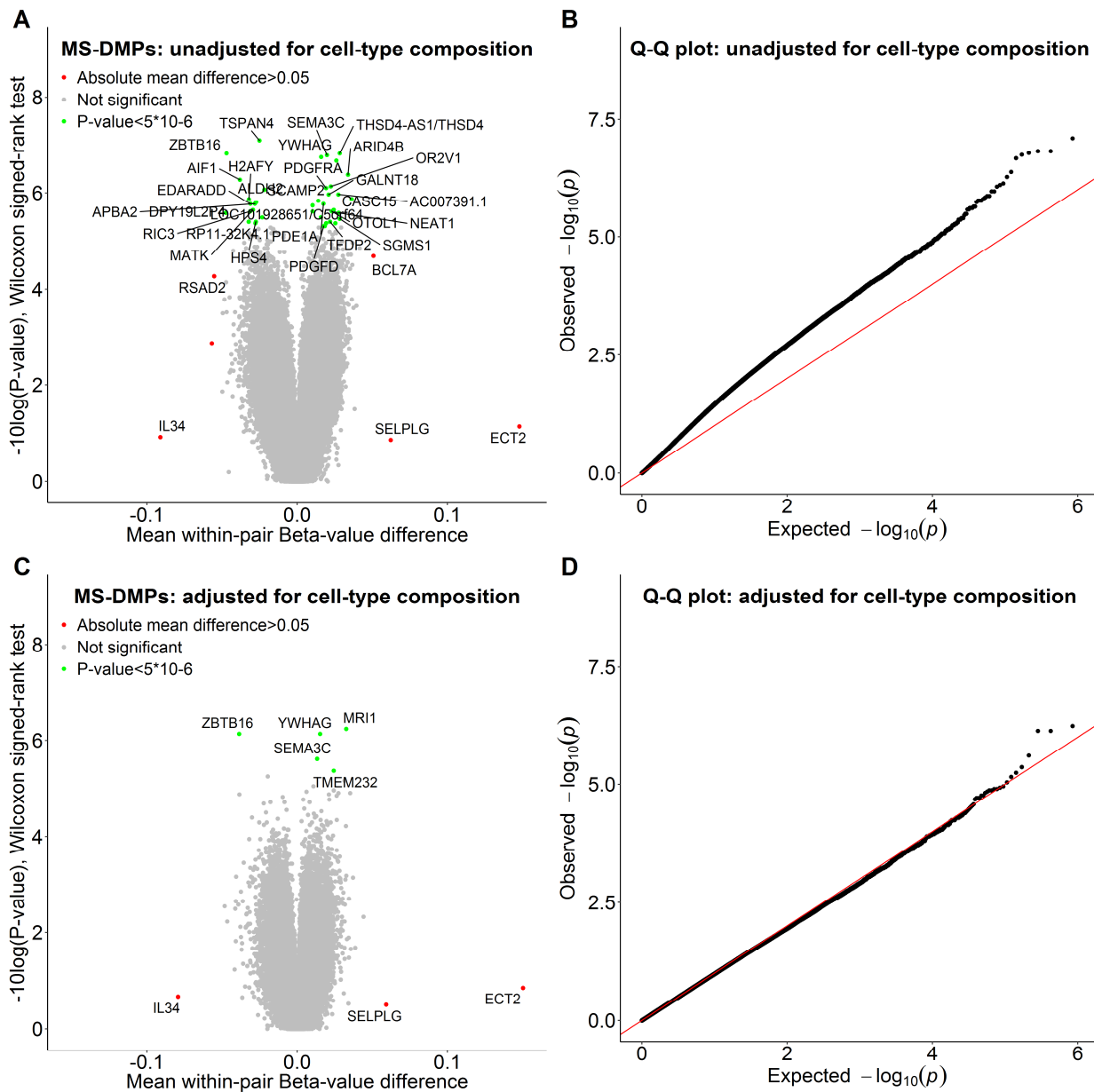
<sup>a</sup>Only the MS-affected co-twin of pair AY had been treated very recently with GCs at the time of blood collection (but never received any immune-modulating therapy), while the MS-affected co-twins of the other three pairs had not received GCs or other immune-modulating therapies within at least 12 months prior to blood collection. GC-DMR = glucocorticoid treatment-associated differentially methylated region, MS-DMR = MS-associated DMR, SS = smoking status at sample collection (Y = yes, N = no).

# Supplementary Figures

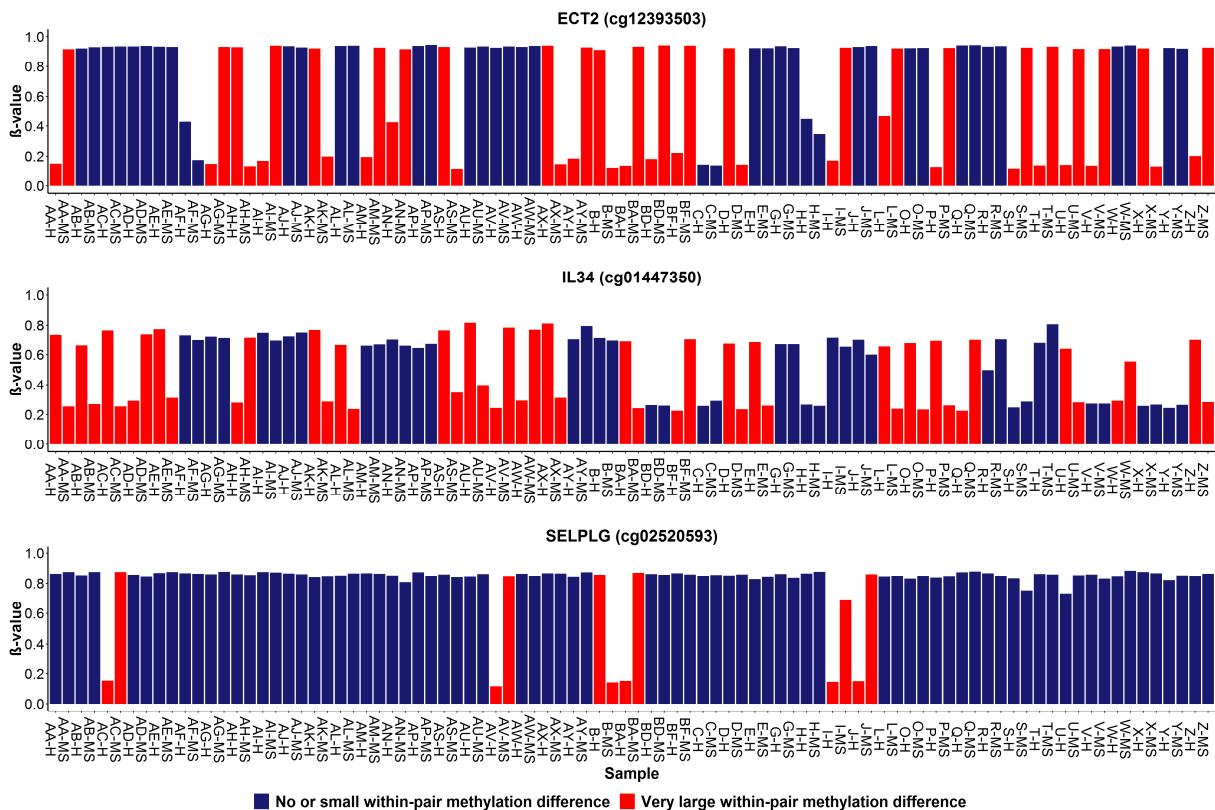


**Supplementary Figure 1. Tukey boxplots (with all data points shown in green) of (A) the age of disease onset and (B) the years that the MZ twins were clinically discordant for MS at sample collection (n = 45 twin pairs).** Our twin cohort has an average age of onset of 28 years, and contains 7 (16%) cases that were younger than 20 years and 6 (13%) cases that were older than 40 years at disease onset. Since MS has an average age of onset of about 30 years and manifests in 70% of the patients between 20 and 40 years of age<sup>11,12</sup>, the age of onset in our cohort is within the normal range. Boxplots represent the median (central line), the interquartile range or IQR (bottom and top of the box), and 1.5 times the IQR (whiskers).

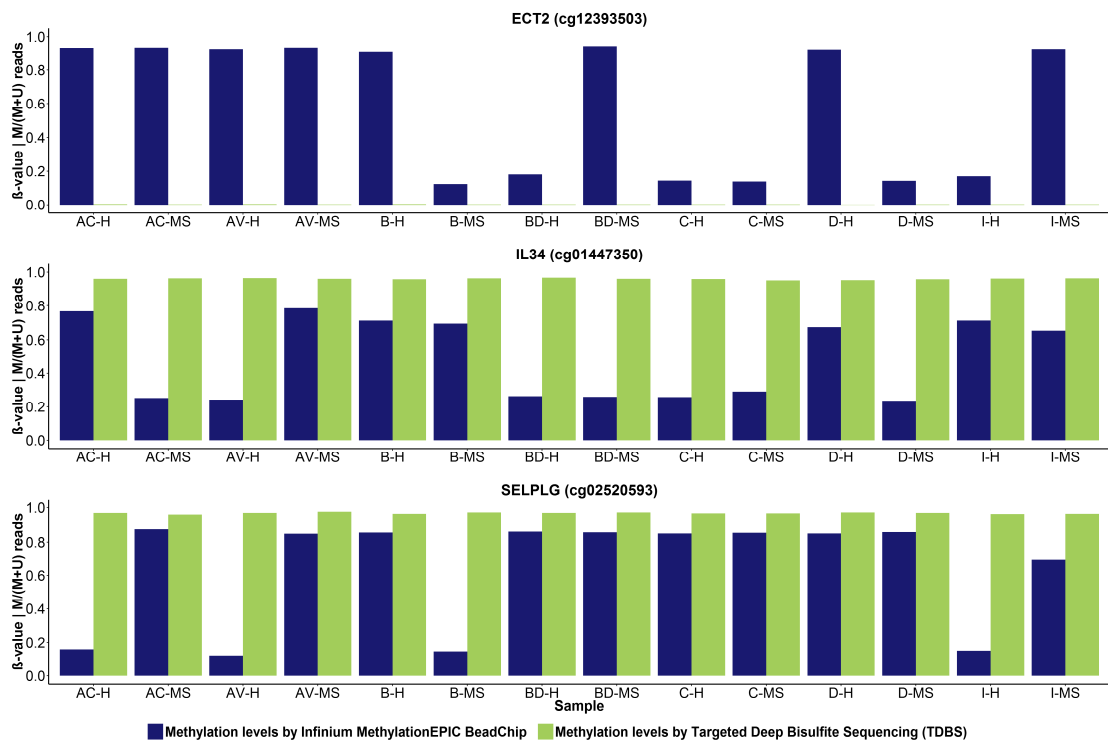




**Supplementary Figure 2. DNA methylation changes associated with the clinical manifestation of MS (n = 45 twin pairs).** Results of the differential DNA methylation analysis including the EPIC array data of all 45 MZ twin pairs clinically discordant for MS. **(A)** Volcano plot of the P-values resulting from the non-parametric two-tailed Wilcoxon signed-rank test versus the mean within-pair  $\beta$ -value difference for each CpG. Data were unadjusted for cell-type composition. **(B)** Q-Q plot of the P-values resulting from the non-parametric two-tailed Wilcoxon signed-rank shown in Figure 3A. Data were unadjusted for cell-type composition. **(C)** Volcano plot of the P-values resulting from the non-parametric two-tailed Wilcoxon signed-rank test against the mean within-pair  $\beta$ -value difference for each CpG. Data were adjusted for cell-type composition. **(D)** Q-Q plot of the P-values resulting from the non-parametric two-tailed Wilcoxon signed-rank shown in Figure 3C. Data were adjusted for cell-type composition. Within-pair  $\beta$ -value difference ( $\Delta\beta$ -value) = clinically MS-affected MZ co-twin – non-affected MZ co-twin.

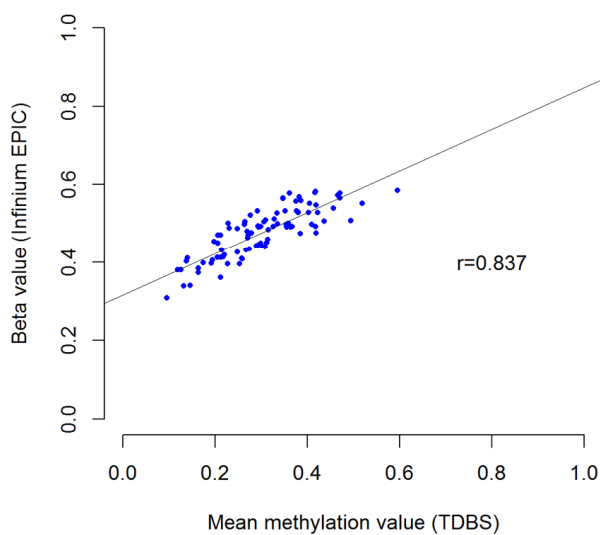


**Supplementary Figure 3. Infinium MethylationEPIC BeadChip  $\beta$ -values of each sample are shown for the *ECT2* (cg12393503), *SELPLG* (cg02520593) and *IL34* (cg01447350) CpGs, which show very large mean within-pair  $\beta$ -value differences ( $n = 45$  twin pairs). Labels indicate: Pair ID - Disease status (i.e. MS = MS-affected MZ co-twin, H = clinically non-affected MZ co-twin).**



**Supplementary Figure 4. Validation of the *ECT2* (cg12393503), *IL34* (cg01447350) and the *SELPLG* (cg02520593) CpGs by targeted deep bisulfite sequencing (TDBS) (n = 7 twin pairs).** On the y-axis the Infinium MethylationEPIC BeadChip  $\beta$ -values as well as the TDBS results are shown, both represented as the fraction of methylated cytosines. Labels indicate: Pair ID - Disease status (i.e. MS = MS-affected MZ co-twin, H = clinically non-affected MZ co-twin), M = methylated, U = unmethylated. In contrast to the Infinium MethylationEPIC BeadChip data, TDBS revealed that *ECT2* (cg12393503) was completely unmethylated, while *IL34* (cg01447350) and *SELPLG* (cg02520593) were highly methylated in all samples.

**A** *TMEM232* (cg27037608): Infinium EPIC vs TDBS



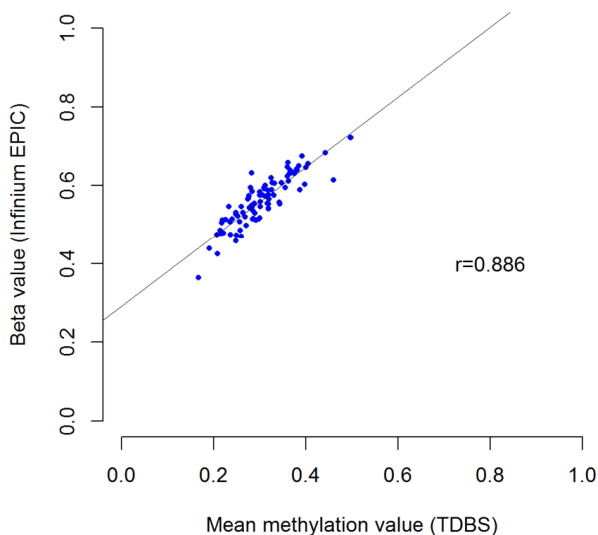
**B**

n = 45		Mean methylation		Mean within-pair methylation difference (U) (95% CI)	P <sub>W-U</sub>	Methylation range
Method	<i>TMEM232</i>	MS-affected co-twin	non-affected co-twin			
EPIC	cg27037608	0.488	0.466	0.022 (0.012,0.032)	4.8*10 <sup>-5</sup>	0.31-0.59
TDBS	CpG-1	0.223	0.206	0.017 (-0.002,0.036)	0.11	0.05-0.42
	CpG-2	0.250	0.226	0.024 (0.006,0.042)	0.01	0.05-0.47
	CpG-3	0.288	0.258	0.030 (0.010,0.050)	7.4*10 <sup>-3</sup>	0.06-0.57
	CpG-4	0.285	0.256	0.029 (0.009,0.049)	4.9*10 <sup>-3</sup>	0.07-0.55
	cg27037608	0.315	0.292	0.023 (0.002,0.043)	0.05	0.09-0.60
	CpG-6	0.275	0.247	0.028 (0.008,0.048)	0.02	0.07-0.49
	CpG-7	0.273	0.240	0.035 (0.015,0.055)	5.4*10 <sup>-4</sup>	0.06-0.53
	CpG-8	0.293	0.260	0.033 (0.014,0.051)	4.1*10 <sup>-4</sup>	0.07-0.57
	CpG-9	0.269	0.248	0.021 (0.002,0.040)	0.06	0.08-0.47
Amplicon	0.274	0.248	0.026 (0.008,0.045)	7.4*10 <sup>-3</sup>	0.07-0.52	

*TMEM232* amplicon contains nine CpGs of which cg27037608 is the fifth and is located at chr5:110,062,618. Genomic coordinates are based on human genome build GRCh37/ hg19. CI = confidence interval (parametric), n = number of MS-discordant twin pairs, TDBS = targeted deep bisulfite sequencing, P<sub>W-U</sub> = P-value nonparametric two-tailed Wilcoxon signed-rank test unadjusted for cell-type composition, U = unadjusted for cell-type composition.

**Supplementary Figure 5. Validation of the *TMEM232* (cg27037608) MS-DMP by targeted deep bisulfite sequencing (TDBS) (n = 45 twin pairs). (A)** Correlation plot of the unadjusted Infinium MethylationEPIC BeadChip data and the TDBS data of the *TMEM232* (cg27037608) MS-DMP of all 45 MZ twin pairs. Infinium MethylationEPIC BeadChip data are expressed as  $\beta$ -value. TDBS data are expressed as mean methylation value, where the methylation level is calculated by dividing the number of reads in which the particular CpG is methylated by the total number of sequenced reads (minimal coverage >1500 reads/base).  $r$  = Pearson's correlation coefficient with P-value. **(B)** Summary of the non-parametric two-tailed Wilcoxon signed-rank test on the unadjusted Infinium MethylationEPIC data and the TDBS data of the *TMEM232* amplicon including the cg27037608 DMP (n = all 45 twin pairs). Within-pair methylation difference = MS-affected MZ co-twin – clinically non-affected MZ co-twin. Source data are provided as a Source Data file.

**A** ZBTB16 (cg25345365): Infinium EPIC vs TDBS

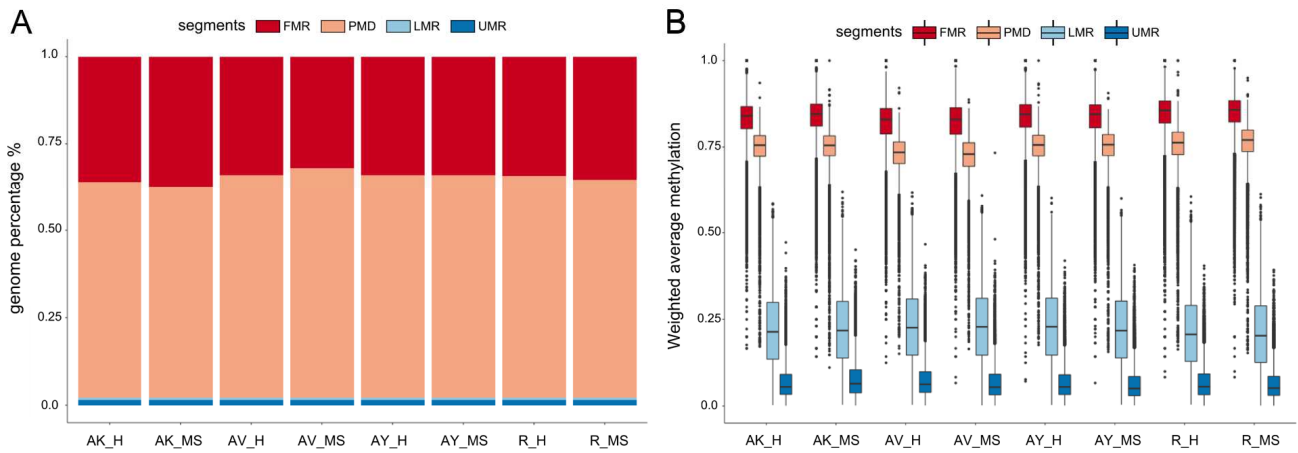


**B**

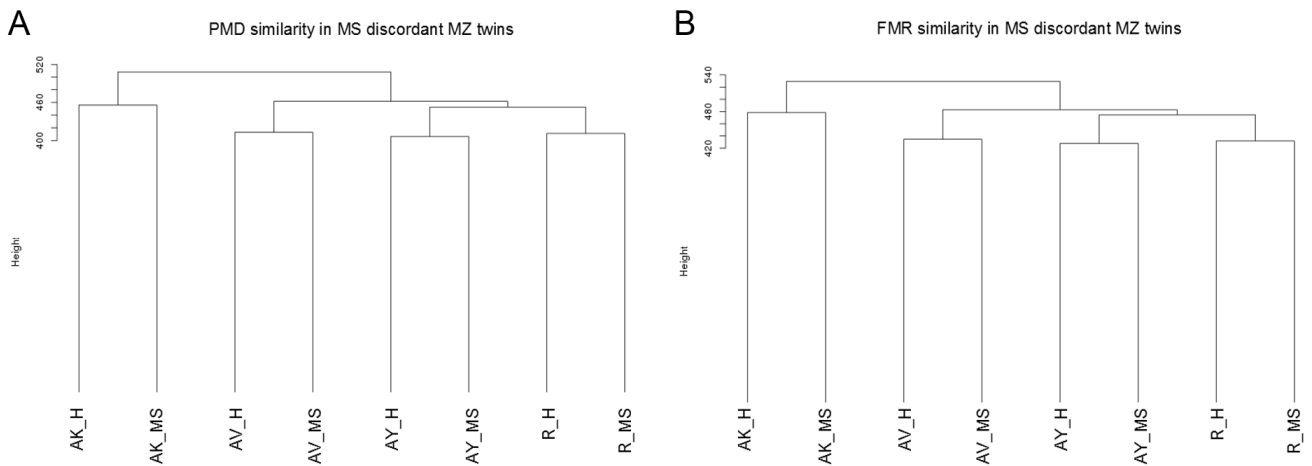
Method	ZBTB16 <sup>a</sup>	Mean methylation		Mean within-pair methylation difference (U) (95% CI)	P <sub>W-U</sub>	Methylation range
		MS-affected co-twin	non-affected co-twin			
EPIC	cg25345365	0.540	0.587	-0.047 (-0.063,-0.031)	1.5*10 <sup>-7</sup>	0.36-0.72
TDBS	CpG-1	0.557	0.604	-0.047 (-0.062,-0.032)	3.8*10 <sup>-7</sup>	0.37-0.75
	cg25345365	0.281	0.330	-0.049 (-0.068,-0.030)	1.4*10 <sup>-6</sup>	0.17-0.50
	CpG-3	0.170	0.195	-0.025 (-0.039,-0.011)	1.2*10 <sup>-3</sup>	0.07-0.32
	Amplicon	0.336	0.377	-0.040 (-0.055,-0.026)	5.3*10 <sup>-7</sup>	0.26-0.51

<sup>a</sup>ZBTB16 amplicon contains three CpGs of which cg25345365 is the second and is located at chr11:114,050,114. CpG-1 is located at chr11:114,050,079 and CpG-3 is located at chr11:114,050,174. Genomic coordinates are based on human genome build GRCh37/ hg19. CI = confidence interval, n = number of MS-discordant MZ twin pairs, TDBS = targeted deep bisulfite sequencing, P<sub>W-U</sub> = P-value two-tailed Wilcoxon signed-rank test unadjusted for cell-type composition, U = unadjusted for cell-type composition.

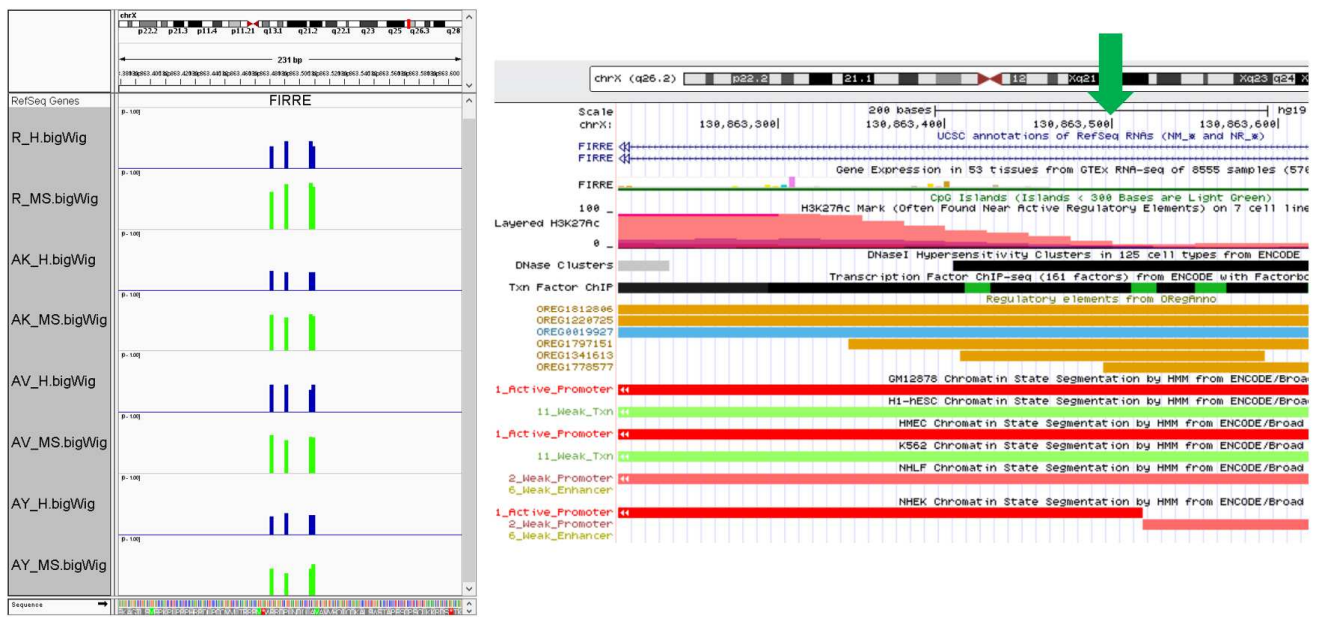
**Supplementary Figure 6. Validation of the ZBTB16 (cg25345365) MS-DMP by targeted deep bisulfite sequencing (TDBS) (n = 45 twin pairs).** (A) Correlation plot of the unadjusted Infinium MethylationEPIC BeadChip data and the TDBS data of the ZBTB16 (cg25345365) MS-DMP of all 45 MZ twin pairs. Infinium MethylationEPIC BeadChip data are expressed as  $\beta$ -value. TDBS data are expressed as mean methylation value, where the methylation level is calculated by dividing the number of reads in which the particular CpG is methylated by the total number of sequenced reads (minimal coverage >2000 reads/base).  $r$  = Pearson's correlation coefficient with P-value. (B) Summary of the nonparametric two-tailed Wilcoxon signed-rank test on the unadjusted Infinium MethylationEPIC BeadChip data and the TDBS data of the ZBTB16 (cg25345365) MS-DMP (n = 45 twin pairs). Within-pair methylation difference = MS-affected MZ co-twin – clinically non-affected MZ co-twin. Source data are provided as a Source Data file.



**Supplementary Figure 7. Results of the PMD analysis on the WGBS data of CD4+ memory T-cells of four MS discordant MZ twin pairs (n = 4 twin pairs).** Median weighted average methylation levels in PMDs, FMRs, LMRs and UMRs were similar between the clinically non-affected and MS-affected MZ co-twins ( $P > 0.05$ , two-tailed paired T-test) **(A)** Segment percentages in the genome plotted as stacked barplots. **(B)** Weighted average methylation levels per segment plotted as boxplots. FMR = fully methylated region, PMD = partially methylated domains, LMR = low methylated region, UMR = unmethylated region. Labels indicate: Pair ID - Disease status (i.e. MS = MS-affected MZ co-twin, H = clinically non-affected MZ co-twin). The MS-affected co-twin of pair AY was treated with glucocorticoid at the moment of blood collection.



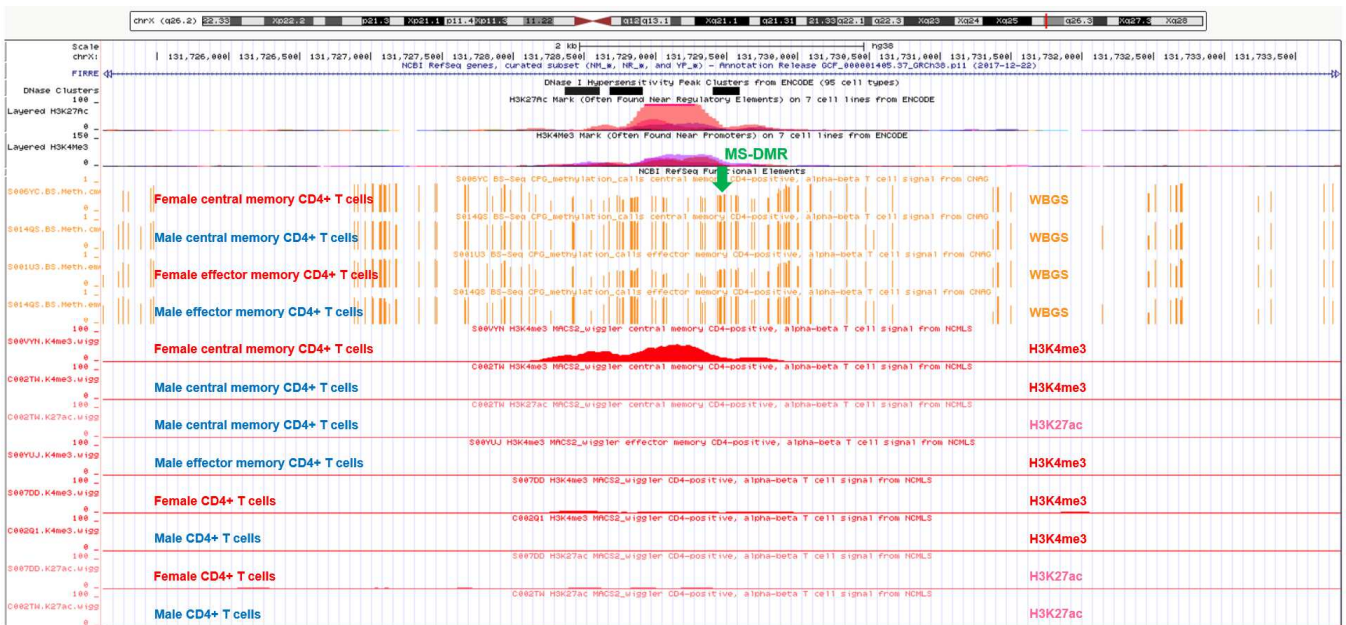
**Supplementary Figure 8. Hierarchical clustering of (A) the partially methylated domains (PMDs) and (B) the fully methylated regions (FMRs), identified in the WGBS data of CD4+ memory T-cells of four MS discordant MZ twin pairs, demonstrating that all co-twins cluster together (n = 4 twin pairs).** Labels indicate: Pair ID - Disease status (i.e. MS = MS-affected MZ co-twin, H = clinically non-affected MZ co-twin). The MS-affected co-twin of pair AY was treated with glucocorticoid at the moment of blood collection.



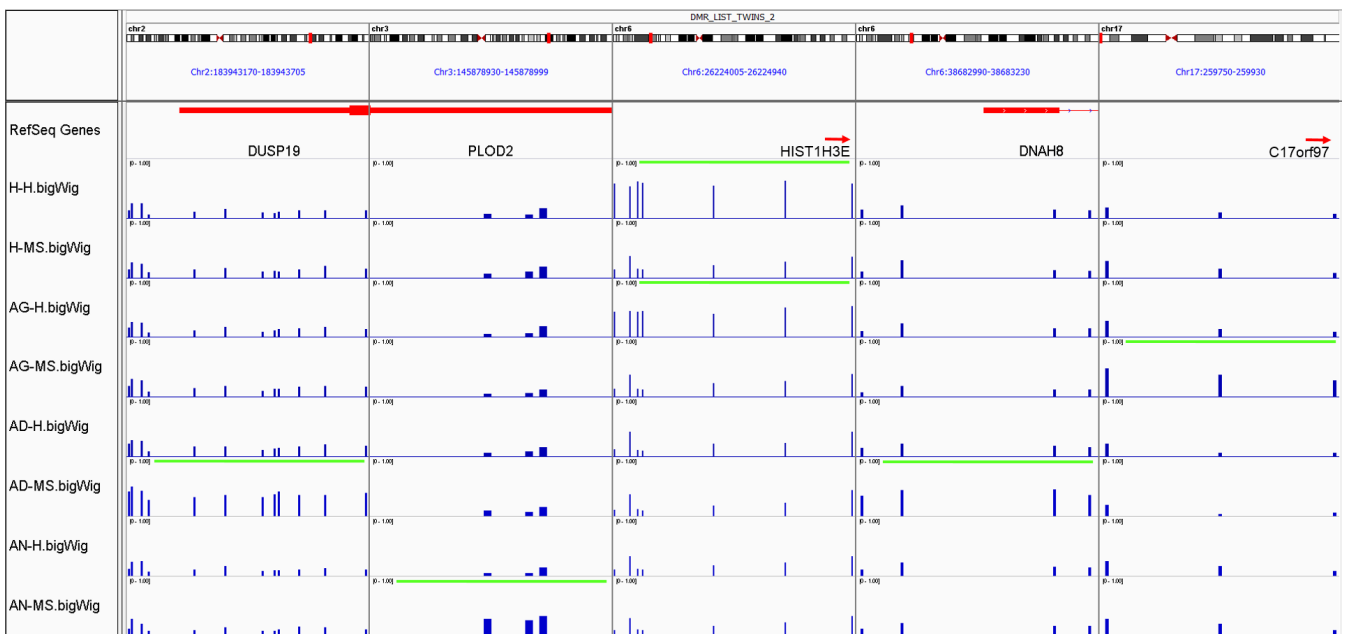
Chr	Start	End	Annotation	Gene	Mean methylation MS co-twins	Mean methylation unaffected co-twins	Mean within-pair methylation difference	Full name
X	130863481	130863509	Intron	<i>FIRRE</i>	0.66	0.40	0.26	Functional intergenic repeating RNA element

**Supplementary Figure 9. MS-DMR in the *FIRRE* gene identified by WGBS of CD4+ memory T-cells of four female MS discordant MZ twin pairs (n = 4 twin pairs).** This MS-DMR is located in an intronic CTCF/YY1 bound regulatory region in the *FIRRE* gene,<sup>13</sup> that is located on the X-chromosome (chrX:130863481-130863509) and encodes a circular long non-coding RNA.<sup>14</sup> MS-DMRs were defined as  $\geq 3$  CpGs, each having P-value < 0.05 (two-tailed paired T test) and absolute mean methylation difference > 0.2, and a maximum 500 bp distance between neighbouring CpGs. The green bars highlights the MS-affected co-twins. All genome coordinates are based on human genome build GRCh37/hg19. Labels indicate: Pair ID - Disease status (i.e. MS = MS-affected MZ co-twin, H = clinically non-affected MZ co-twin). Source data are provided as a Source Data file.

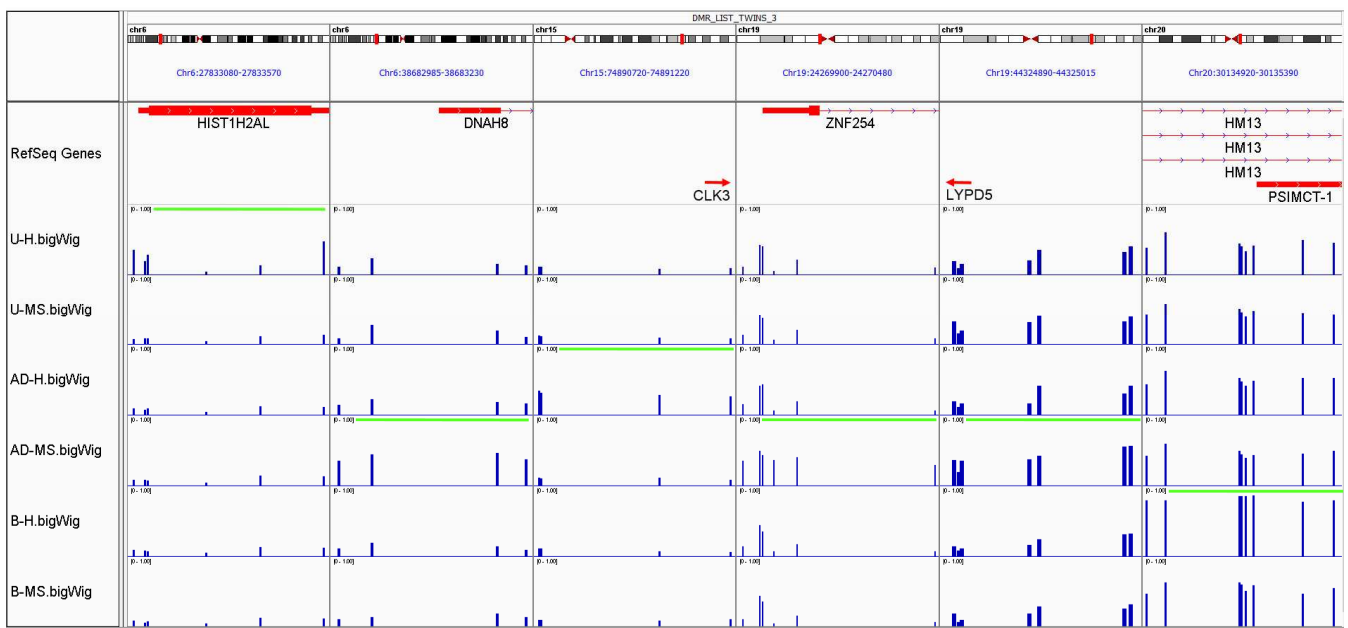




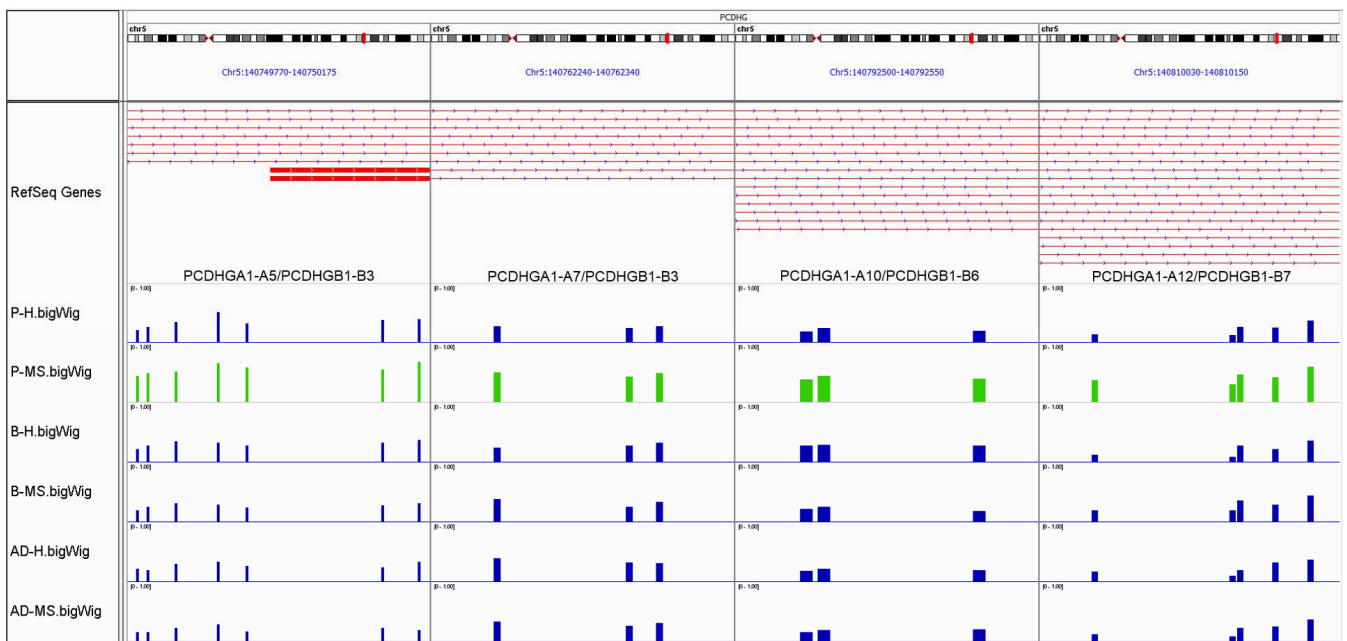
**Supplementary Figure 10. Methylation and chromatin status of the *FIRRE* MS-DMR in various subsets of primary CD4+ T cells in male and female BLUEPRINT samples<sup>15</sup>.** Whole genome bisulfite sequencing (WGBS) data of central memory and effector memory CD4+ T cells shows that in females methylation levels at the *FIRRE* DMR locus are lower compared to males (~50% versus ~100%). In addition, in female central memory CD4+ T cells a H3K4me3 peak is observed at the *FIRRE* DMR locus, but not in males. Unfortunately, H3K27ac central memory and H3K4me3 effector memory CD4+ T-cell data was not available of a female donor. All genome coordinates are based on human genome build GRCh38/hg38.



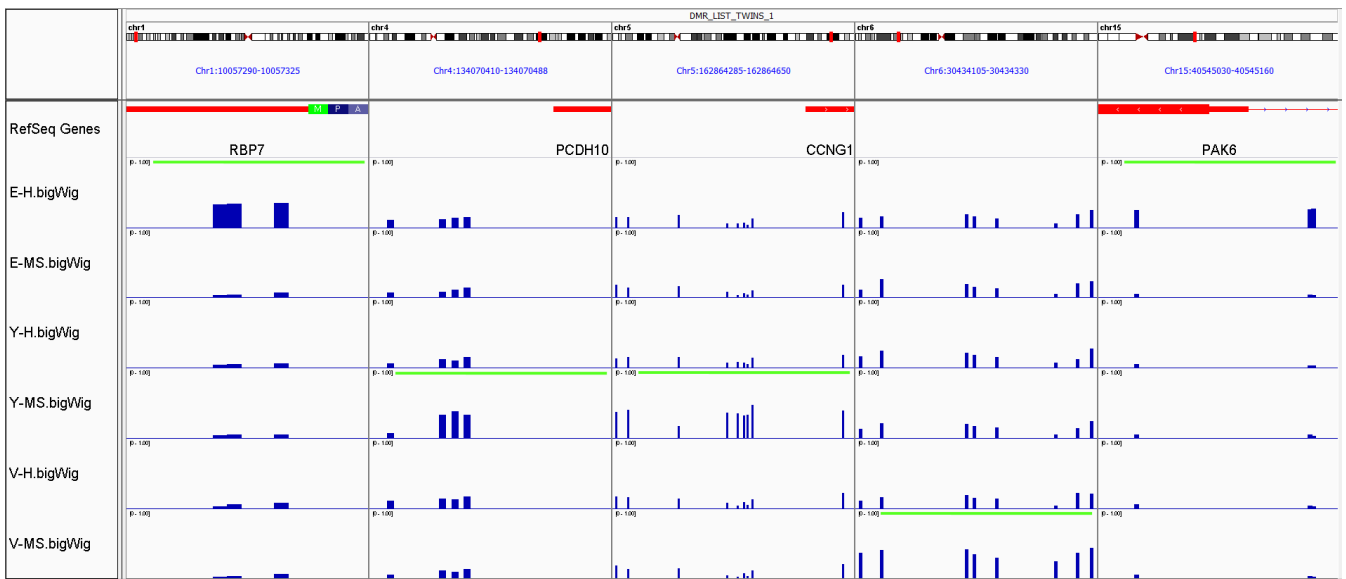
**Supplementary Figure 11. Identified within-pair DMRs (WP-DMRs) in *DUSP19*, *PLOD2*, *HIST1H3E*, *DNAH8* and *C17orf97*.** The green horizontal line highlights the aberrant methylated sample(s). These WP-DMRs were identified in the Infinium MethylationEPIC BeadChip data. Labels indicate: Pair ID - Disease status (i.e. MS = MS-affected MZ co-twin, H = clinically non-affected MZ co-twin). Source data are provided as a Source Data file.



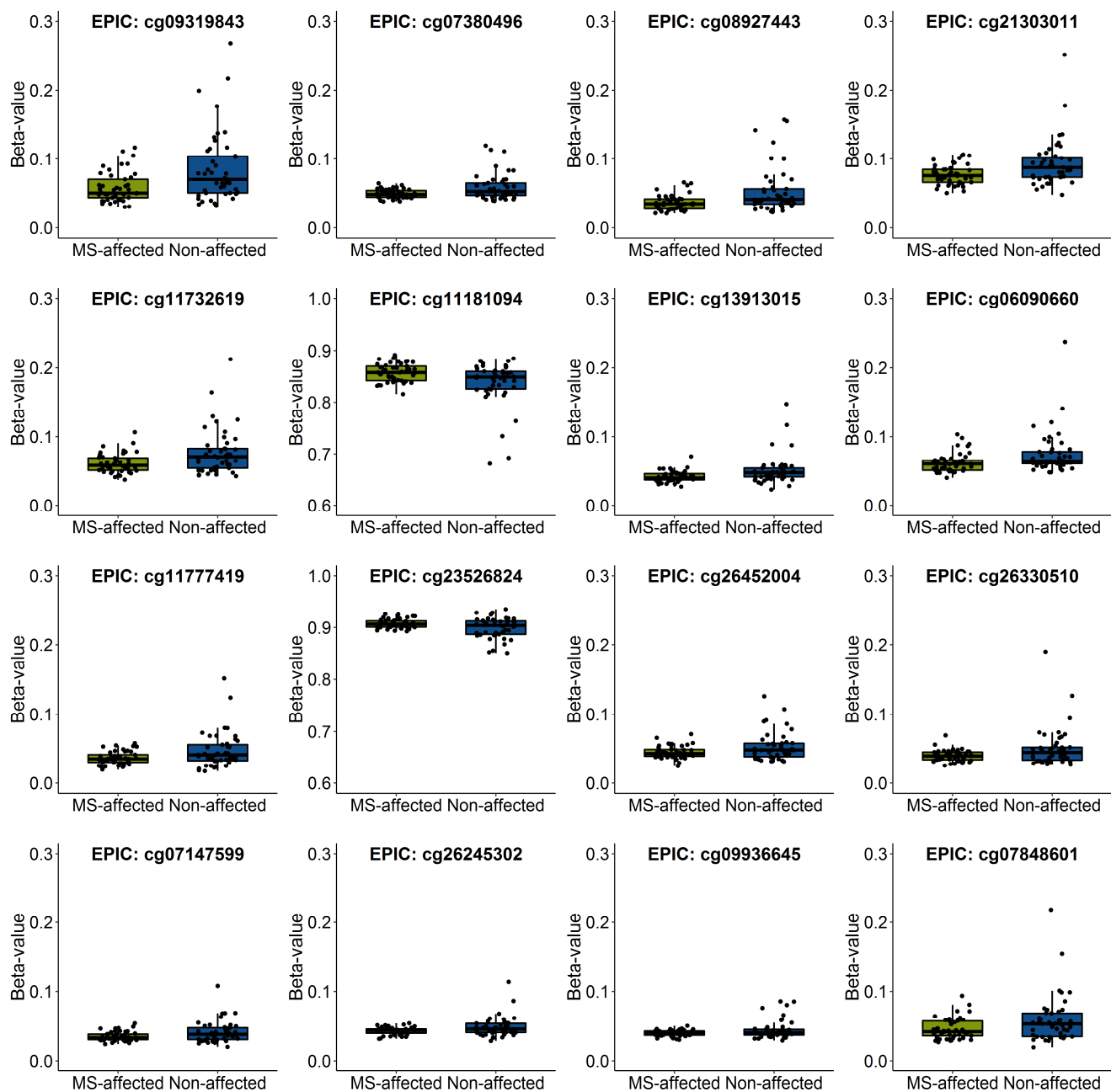
**Supplementary Figure 12. Identified within-pair DMRs (WP-DMRs) in *HIST1H2AL*, *DNAH8*, *CLK3*, *ZNF254*, *LYPD5* and *HM13*/*MCTS2P*.** The green horizontal line highlights the aberrant methylated sample. These WP-DMRs were identified in the Infinium MethylationEPIC BeadChip data. Labels indicate: Pair ID - Disease status (i.e. MS = MS-affected MZ co-twin, H = clinically non-affected MZ co-twin). *MCTS2P* is also called *PSIMCT-1*. Source data are provided as a Source Data file.



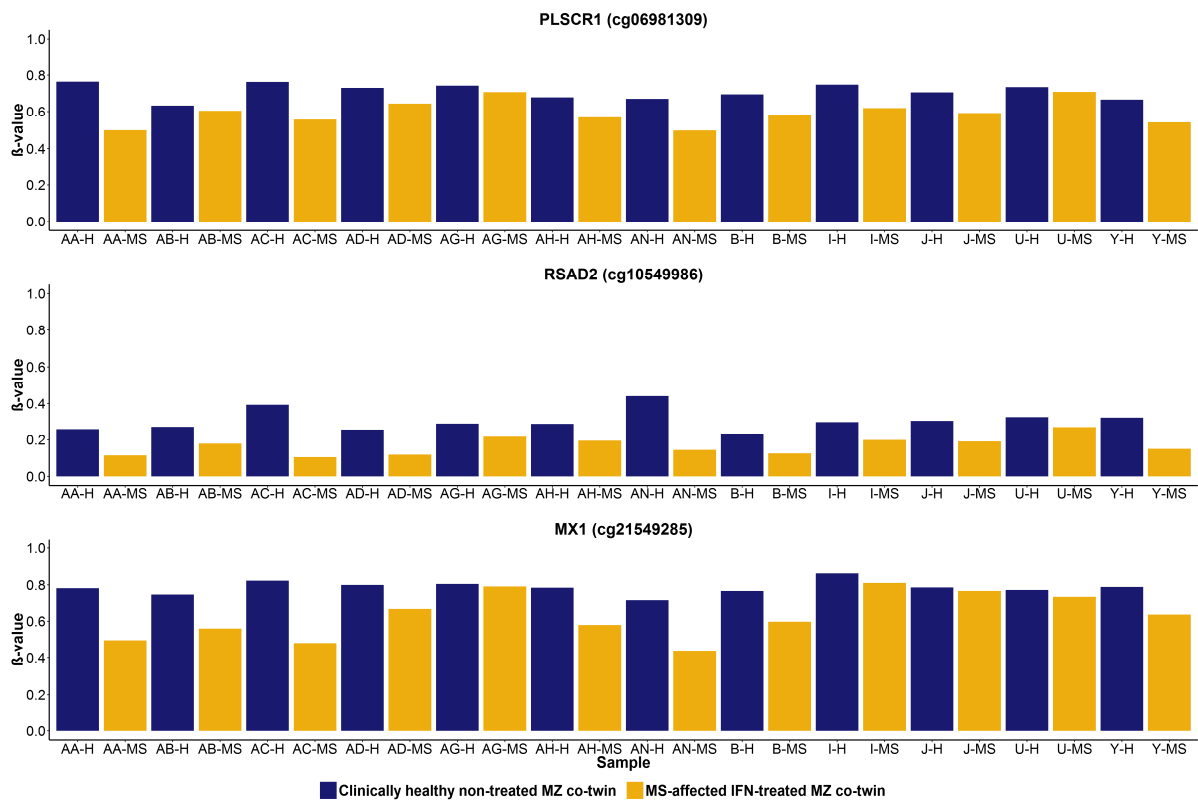
**Supplementary Figure 13. Identified within-pair DMRs (WP-DMRs) in the *PCDHG* gene cluster.** The green bars highlights the aberrant methylated sample. These WP-DMRs were identified in the Infinium MethylationEPIC BeadChip data. Labels indicate: Pair ID - Disease status (i.e. MS = MS-affected MZ co-twin, H = clinically non-affected MZ co-twin). Source data are provided as a Source Data file.



**Supplementary Figure 14. Identified within-pair DMRs (WP-DMRs) in *RBP7*, *PCDH10*, *CCNG1*, Chr6:30434109-30434324 and *PAK6*.** The green horizontal line highlights the aberrant methylated sample. These WP-DMRs were identified in the Infinium MethylationEPIC BeadChip data. Labels indicate: Pair ID - Disease status (i.e. MS = MS-affected MZ co-twin, H = clinically non-affected MZ co-twin). Source data are provided as a Source Data file.

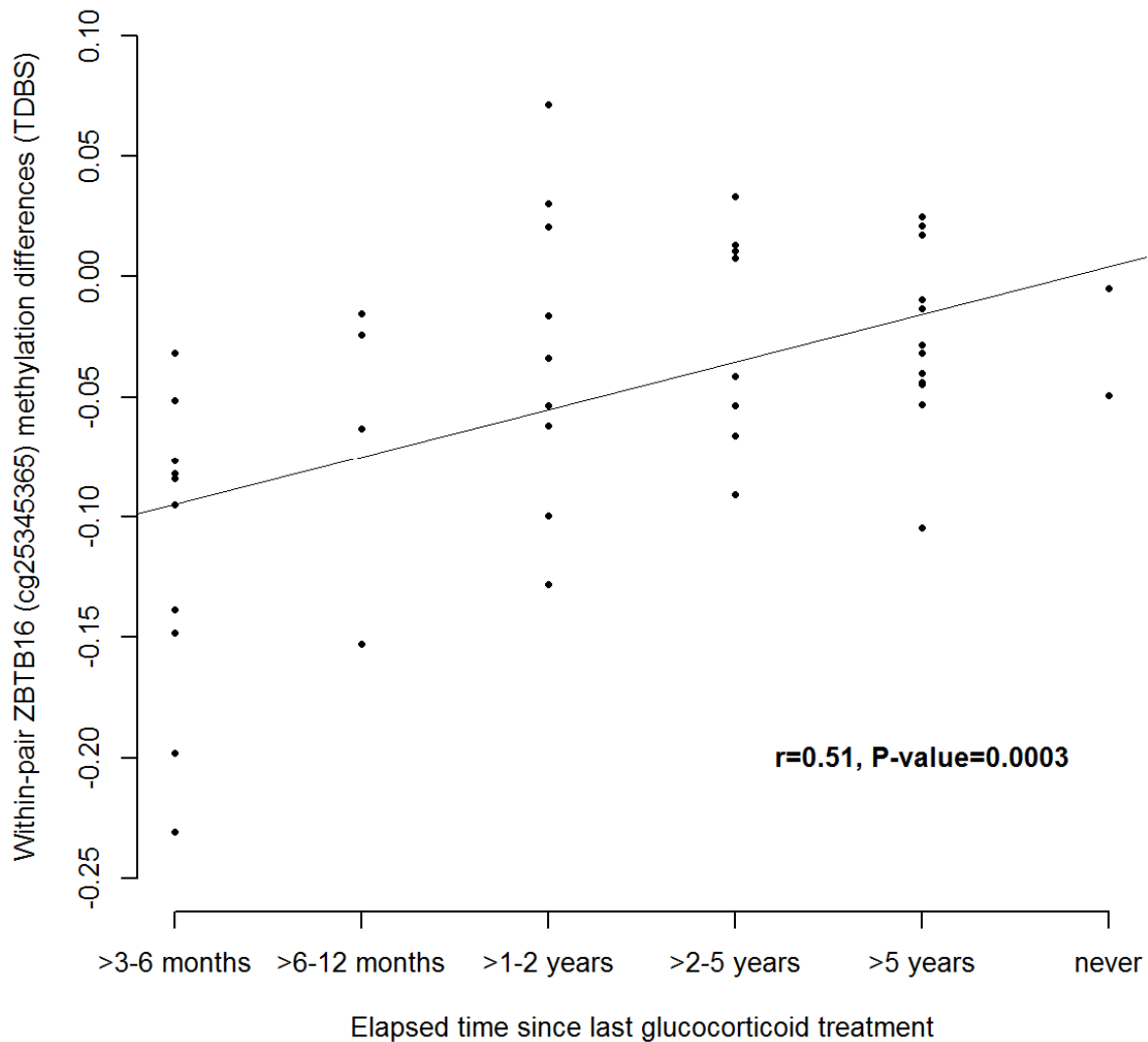


**Supplementary Figure 15. Boxplots of the 16 top-ranked differentially variable positions (DVPs) identified in the 45 MS-discordant MZ twin pairs (n=45 twin pairs).** DVPs were identified using the iEVORA algorithm<sup>5</sup> and were defined as CpGs with a FDR-corrected Barlett's P-value<0.001 and raw T-test P-value<0.05. Boxplots represent the median (central line), the interquartile range or IQR (bottom and top of the box), and 1.5 times the IQR (whiskers). Source data are provided as a Source Data file.



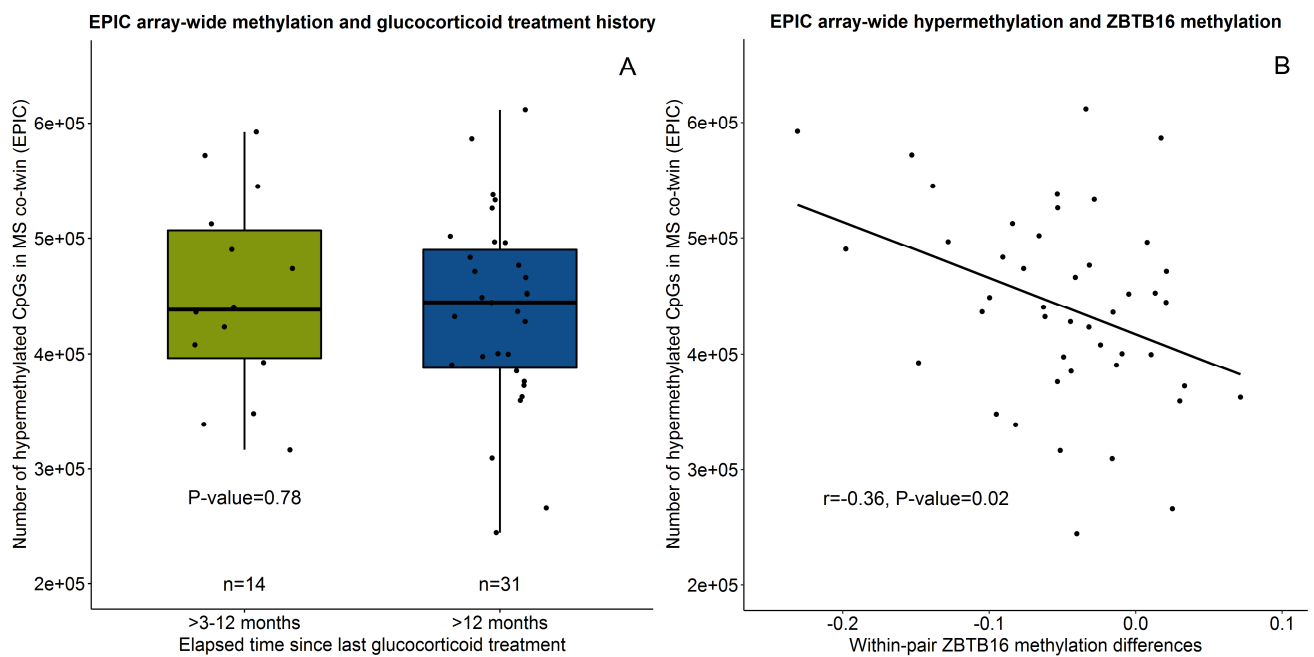
**Supplementary Figure 16. Infinium MethylationEPIC BeadChip  $\beta$ -values of CpGs in *PLSCR1* (cg06981309), *RSAD2* (cg10549986) and *MX1* (cg21549285) that were strongly differentially methylated following interferon-beta (IFN) treatment (IFN-DMPs) (n = 12 twin pairs). Labels indicate: Pair ID - Disease status (i.e. MS = MS-affected IFN-treated MZ co-twin (yellow), H = clinically non-affected non-treated MZ co-twin (blue)). Source data are provided as a Source Data file.**

### ZBTB16 methylation and glucocorticoid treatment history

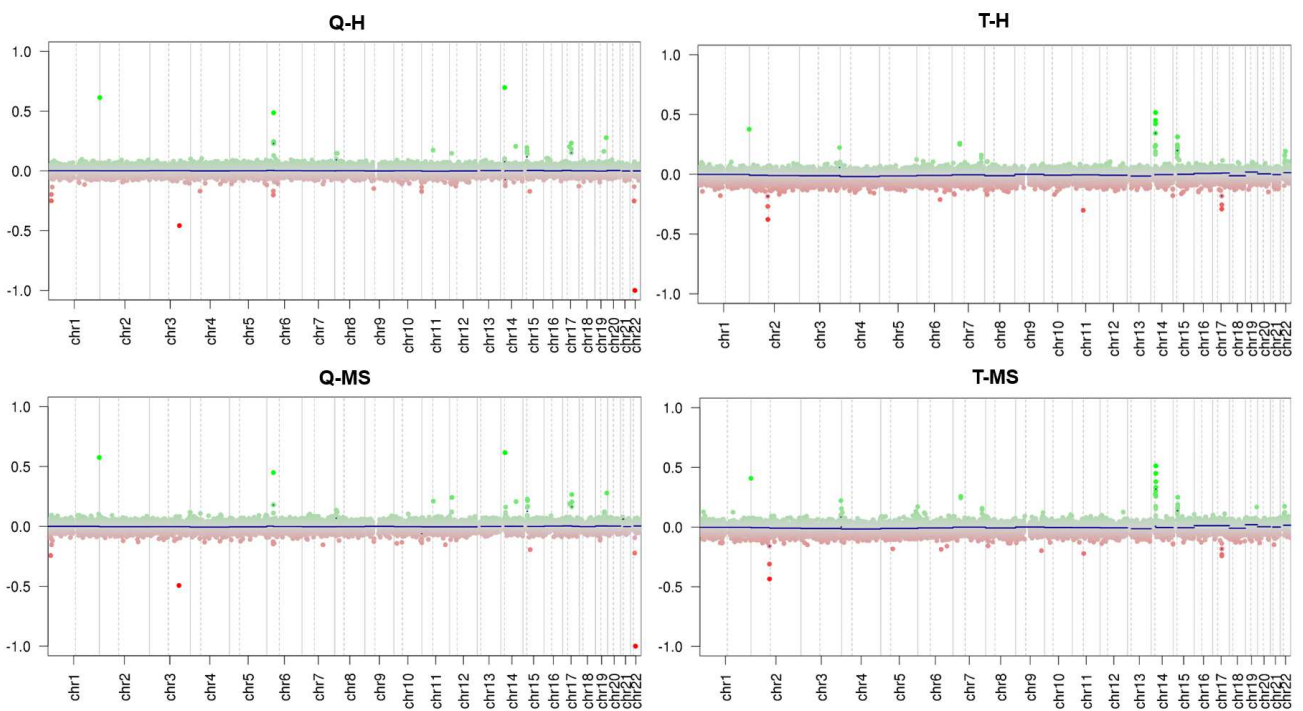


**Supplementary Figure 17. *ZBTB16* MS-DMP (cg25345365, targeted deep bisulfite sequencing (TDBS)) within-pair methylation differences plotted against the elapsed time since the last glucocorticoid treatment in the MS-affected MZ co-twin (n = 45 twin pairs).** r = Pearson's correlation coefficient with P-value, within-pair methylation difference = MS-affected MZ co-twin – clinically non-affected MZ co-twin. Source data are provided as a Source Data file.





**Supplementary Figure 18. EPIC-array wide hypermethylation, glucocorticoid treatment history and *ZBTB16* methylation (n = 45 twin pairs).** (A) Number of hypermethylated CpGs in the MS-affected co-twins and elapsed time since last glucocorticoid treatment in the MS-affected MZ co-twin. Data are presented as Tukey boxplots. P-value = non-parametric two-tailed Wilcoxon rank-sum test result. (B) Number of hypermethylated CpGs in the MS-affected co-twins plotted against the within-pair methylation differences of the *ZBTB16* DMP (cg25345365, TDBS).  $r$  = Pearson's correlation coefficient with P-value, within-pair methylation difference = MS-affected MZ co-twin – clinically non-affected MZ co-twin. Source data are provided as a Source Data file. Boxplots represent the median (central line), the interquartile range or IQR (bottom and top of the box), and 1.5 times the IQR (whiskers).



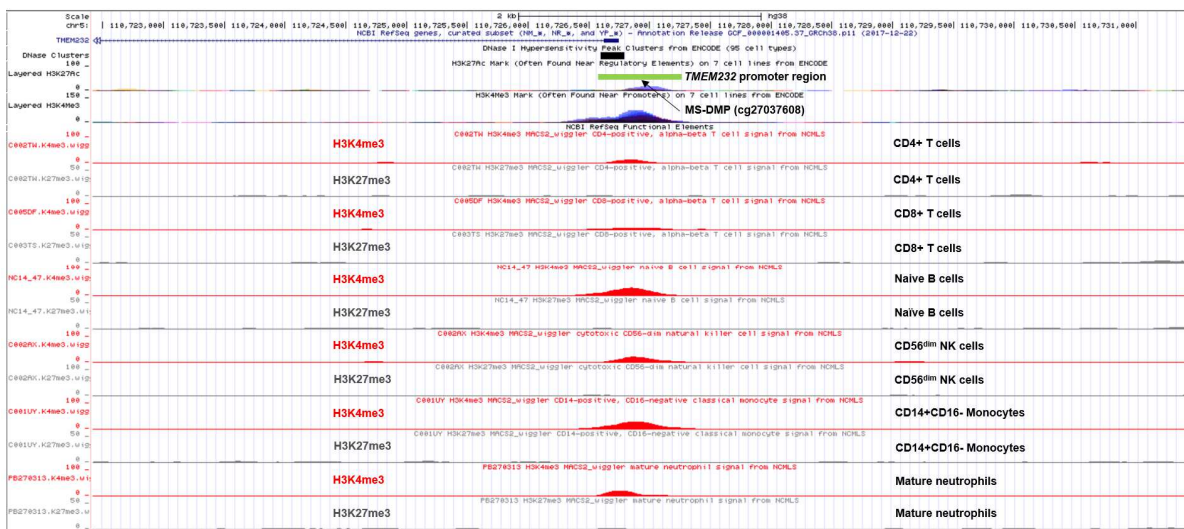
**Supplementary Figure 19. An example of copy number plots generated from the Infinium MethylationEPIC BeadChip data of non-affected co-twins (Q-H & T-H) and their MS-affected MZ co-twins (Q-MS & T-MS).** Gains are indicated in green and losses in red. MZ co-twins show very similar copy number profiles and no within-pair chromosomal gains and losses (defined as absolute segment mean threshold  $\geq 0.3$ ) were observed. Labels indicate: Pair ID - Disease status (i.e. MS = MS-affected MZ co-twin, H = clinically non-affected MZ co-twin).

```

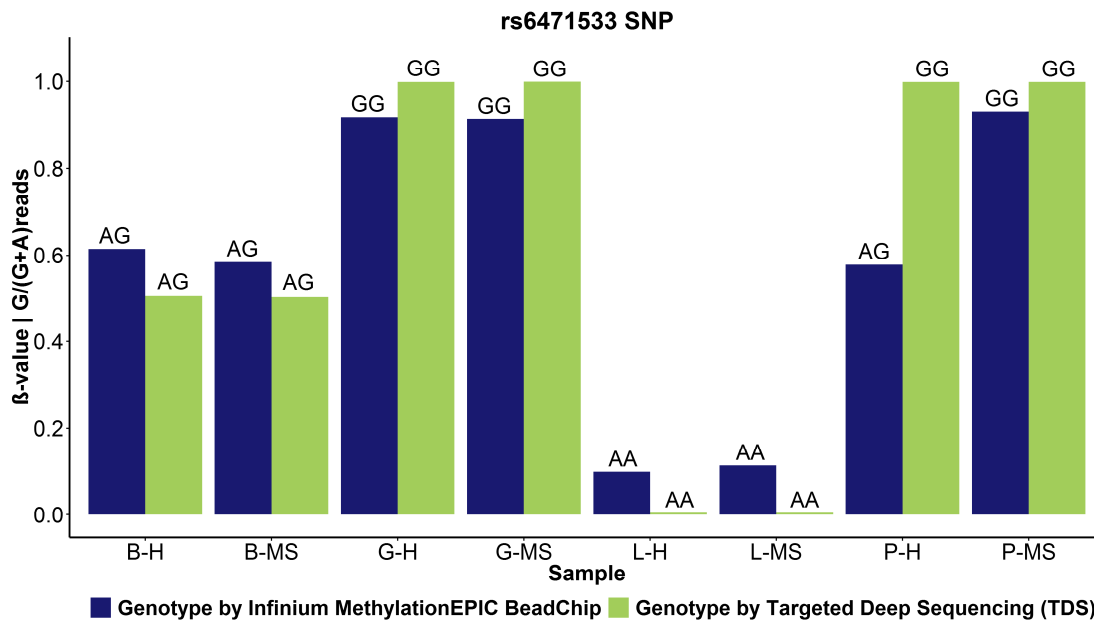
>chr11:114049864-114050363
GTTGTTTTCCCTTTCAGCCAGGATTTGCTCACCTGCTTTCCTTTCCCTCTTCCCTGAAT
CTGTGTATCTTCCAGAACTCTGGTTCTGGATCAGCTTTTTGTTTTGTTGACAAAGGAGA
AAGAGCAAGAGAGAGAGCTAGAGAGAGAGAGAGAGAGAGAGCCCTCAAGCTCCTCTCTGGGA
GTACACATCTCCTTGAGGGAAAGAACACACAGTGCCGGCCTTTGGAATTGGCAGCCAGTG
TGCTGTTCTCGTCTGATAAGAGGTACTGTAAATAAAACTGTACACCATGGCCTGTTGTA
AAATGCCCTGCGTCTGTACTCATTGTTCTGACAGCTTATGCTTTTTTTGGTCTGCTGTT
TTGGTACACTCTGTACTTCCTTATGTAAGCAGGCGTGCAGATCTCATCAGAACATTCAAG
ATGTTTATTTTAAAATCTCAAGGAATTTTGAATAAAAGGGACACACCACTCAACATTAGAT
GCTGGCAAACATTAGGTGTT

```

**Supplementary Figure 20. Genomic sequence of the *ZBTB16* DMP (cg25345365) region.** The cg25345365 CpG is marked in green and the other CpGs in this region are marked yellow. The consensus GRE downstream half-sites (TGTTCT) are in bold and marked blue. The genomic position of the forward and reverse primers used to generate the amplicon for the TDBS analysis are bold and underlined. Genome coordinates are human genome build GRCh37/hg19.

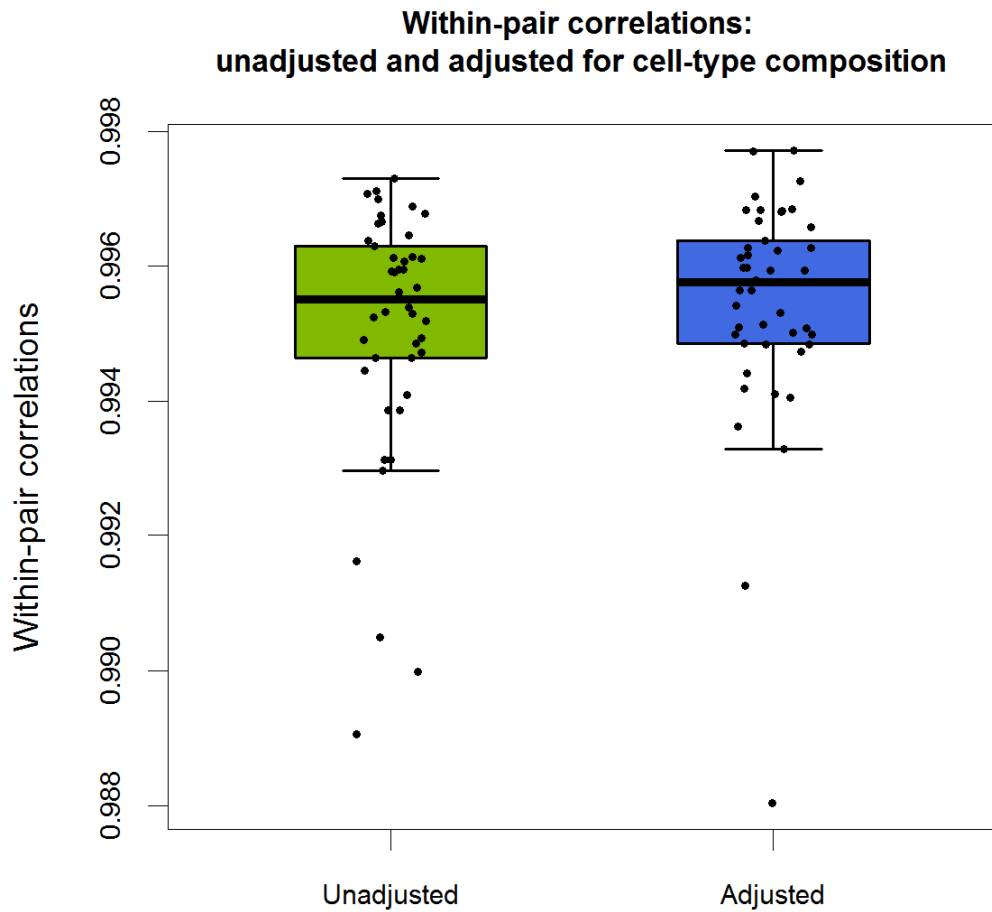


**Supplementary Figure 21. Status of the active H3K4me3 and repressive H3K27me3 chromatin marks at the *TMEM232* promoter region in different immune cell types from BLUEPRINT samples.<sup>15</sup> Note that all genome coordinates are based on human genome build GRCh38/hg38. MS-DMP = MS-associated differentially methylated position.**



**Supplementary Figure 22. Validation of the rs6471533 SNP using targeted deep sequencing (TDS) in pair P, which shows a discordant genotype for the rs6471533 SNP on the Infinium MethylationEPIC BeadChip (n = 4 twin pairs).** The rs6471533 SNP was also genotyped using TDS in pair B heterozygous for the rs6471533 SNP, pair G homozygous for the G allele and pair L homozygous for the A allele, according to the Infinium MethylationEPIC BeadChip. Hence, TDS reveals that both co-twins of pair P are homozygous G allele carriers, and the genotypes of the other pairs agrees with the genotype of the Infinium MethylationEPIC BeadChip.

On the y-axis the Infinium MethylationEPIC BeadChip  $\beta$ -values as well as the TDS results are shown, both represented as the fraction of guanines (0=AA and 1=GG). Labels indicate: Pair ID - Disease status (i.e. MS = multiple sclerosis-affected MZ co-twin, H = clinically non-affected MZ co-twin).



**Supplementary Figure 23. Tukey Box-plots of the within-pair Pearson correlation coefficients before and after adjusting the Infinium MethylationEPIC BeadChip data for cell-type composition (n = 45 twin pairs).** Boxplots represent the median (central line), the interquartile range or IQR (bottom and top of the box), and 1.5 times the IQR (whiskers).

## Supplementary Methods

### Bisulfite treatment

From each PBMC sample, 500 ng DNA was treated with bisulfite using the EZ DNA Methylation kit (D5002, Zymo Research), according to the manufacturer's recommendations for the Illumina Infinium assay. The conversion reaction was incubated at 16 cycles of 95°C for 1 min and 50°C for 60 min. For all six MZ twin pairs that were processed in the first batch, the bisulfite controls present on the Illumina Infinium MethylationEPIC BeadChip (EPIC array) showed suboptimal results, i.e., bisulfite conversion controls (I and II) showed moderate intensities at probes which should be at background level. However, the within-pair array-wide Pearson correlation coefficients for these MZ twin pairs were very high and ranged from 0.996 to 0.997, indicating high quality methylation data. We verified the conversion rate of these samples by targeted, deep bisulfite sequencing (TDBS) of a 343-bp region amplified using non-bisulfite-dependent primers (Chr1:202150908-202151251, see **Supplementary Table 12**). TDBS of these samples showed an average conversion rate of 98.6% (SD=0.37%) (minimal coverage >2000 reads), indicating that the EPIC array bisulfite controls are extremely sensitive. Hence, these EPIC array data were used in the downstream analysis. Nevertheless, the bisulfite treatment procedure was adapted by incubating samples in a programmable ThermoQ Metal Bath with a heated lid (Bioer, Hangzhou, China) instead of a Eppendorf Mastercycler (Eppendorf AG, Hamburg, Germany), and TDBS in 16 samples revealed an average conversion rate of 99.7% (SD=0.10%). Accordingly, the other 40 MZ twin pairs were processed using the adapted bisulfite treatment and the EPIC array bisulfite controls showed normal intensities for those samples. Both members of a twin pair were always processed in the same batch.

### Estimation of cell-type composition

Cell-type composition of each PBMC sample was estimated with the reference-based method first published by Houseman et al.<sup>6</sup>, which employs DNA methylation reference profiles of individual cell types to estimate the cell-type composition of each sample. Several reference-based deconvolution algorithms were compared, including the implementation of the Houseman algorithm in the *minfi* R/Bioconductor package<sup>7</sup>, and the standard constrained projection as well as the two non-constrained, reference-based, cell-type deconvolution approaches recently implemented in the EpiDISH R/Bioconductor package<sup>16</sup>. For a subset of samples (n=61) cell-type proportions determined using immunophenotyping were available, which showed the best correlation with the estimates provided by the *minfi* package. Accordingly, the *minfi* estimates were used to adjust the

$\beta$ -values for cellular composition using linear regression and the residuals were used for downstream analysis. To obtain interpretable, adjusted  $\beta$ -values, the unadjusted mean  $\beta$ -value of each CpG site was added to the residuals. To check the quality of the adjustment, the adjusted  $\beta$ -values were used to recalculate the within-pair correlations. In the final regression model, the proportions of the four major lymphocyte subtypes were included (i.e., CD4+ T, CD8+ T, CD19+ B, and CD56+ NK cells). Myeloid cells (i.e., monocytes and neutrophils) were not included in the model as immunophenotyping data showed that monocyte proportions were not properly estimated and including them resulted in severe adjustment bias in some samples. As a result, **Supplementary Figure 23** shows that the overall within-pair correlations are, as expected, higher after adjusting for cell-type composition.

### **Cell sorting procedure, WGBS library preparation and sequencing data preprocessing**

Whole genome-wide bisulfite sequencing (WGBS) was used to profile CD4+ central and effector memory T cells of four MS-discordant female MZ twin pairs (mean age 43.3 years, discordant for MS >12 years, **Supplementary Table 13**). Of one pair, the MS-affected co-twin had been treated very recently with GCs at the time of blood collection (but never received any immune-modulating therapy), while the MS-affected co-twins of the other three pairs had not received GCs or other immune-modulating therapies within at least 12 months prior to blood collection.

Cryopreserved PBMCs were thawed and gently suspended in 10 ml of pre-cooled FACS buffer (PBS, 2% FCS) and centrifuged at 300 g for 10 min at 4°C. Then, one additional washing step was performed. Cells were stained with the following monoclonal antibodies: CD3-AF700 (OKT3, eBioscience, Frankfurt, Germany); CD4-Pacific-Blue (S3.5, Molecular Probes, Invitrogen, Karlsruhe, Germany); CD8-PerCP (SK1, BioLegend, Fell, Germany); CD45RO-FITC (UCHL1, eBioscience); and CCR7-APC (3D12, eBioscience) on ice for 30 minutes. Cells were then sorted using a FACSria Fusion flow cytometer (BD Biosciences, Heidelberg, Germany) to selectively collect antigen-experienced CD4+ T cells by excluding dead cells, naive CD45RA+CCR7+ T cells, and CD8+ T cells.

WGBS libraries were prepared using a tagmentation-based protocol similar to that described by Weichenhan et al.<sup>17</sup>. Briefly, fresh frozen primary CD4 cell pellets (each 20,000-200,000 cells) were thawed in 50-100  $\mu$ l of 1.1x TD buffer (Illumina) supplemented with 6  $\mu$ l Protease (1 mg/ml; Qiagen) and incubated in a thermomixer at 55°C for 3 h followed by 20 min at 75°C. DNA was quantified using the Qubit HS-DNA kit (Thermo Fisher Scientific, Waltham, USA). From each sample the volume corresponding to 50 ng DNA was transferred in a new 1.5-ml tube and 1x TD buffer was added to a total volume of 47.5  $\mu$ l. Then, the DNA was tagged with 2.5  $\mu$ l of Tn5 from



the Nextera library preparation kit (Illumina) by incubation for 5 min at 55°C. After purification with the MinElute kit (Qiagen) and final elution with 10 µl EB buffer, gaps were repaired by adding 2 µl of 10x CutSmart buffer (NEB, Ipswich, USA), 3 µl of dNTPs (2.5 mM each), 5 U Klenow exo- (NEB) and incubating for 1 h at 30°C. Bisulfite conversion was performed with the EZ Methylation Gold Kit (Zymo Research) with a final 10 µl elution volume. Indexing library enrichment PCR was performed in 40 µl reactions with 1x HotStartBuffer (Qiagen), 0.25 mM of each dNTP, 0.3 µl ssDNA Binding Protein (Affymetrix, Santa Clara, USA), 100 nM of each primer (reverse primer contains sample-specific DNA barcode), 4 U HotStartTaq DNA polymerase (Qiagen), and 10 µl bisulfite-converted DNA. DNA was denatured at 95°C for 15 min, followed by 12 cycles of 30 sec at 95°C, 2 min at 53°C, and 1 min at 72°C, and a final extension step of 7 min at 72°C. Reactions were purified using 0.8x volume AMPure XP Beads (Beckman Coulter, Brea, USA) and eluted in 10 µl Elution Buffer (Qiagen). Library fragment distributions were checked on the Agilent Bioanalyzer (Agilent, Santa Clara, USA).

The WGBS libraries were sequenced in a 100-bp paired-end HiSeq2500 run (Illumina) using custom sequencing primers. After adapter trimming using Trimmomatic v0.36<sup>18</sup>, the read pairs were aligned to the human reference genome (GRCh37) using *bwa-meth* v0.2.0<sup>19</sup>, which is a wrapper of the BWA-MEM1 alignment algorithm suited for bisulfite sequencing data. PCR duplicates were removed using the MarkDuplicates tool of the Picard suite v2.5.0-1 (<http://broadinstitute.github.io/picard>). Methylation levels of the CpG cytosines were determined using MethylDackel v0.2.1 (<https://github.com/dpryan79/MethylDackel.git>). Of both read mates, 10 base pairs were disregarded from both read ends to eliminate the gap repair bias and methylation bias artifacts. The obtained BED files were loaded in the RnBeads package, which aggregated for each CpG the methylation information of both strands. The coverage statistics of the samples are summarized in **Supplementary Table 13**.

## Supplementary References

1. Kular, L. *et al.* DNA methylation as a mediator of HLA-DRB1\*15:01 and a protective variant in multiple sclerosis. *Nat Commun* **9**, 2397 (2018).
2. Beecham, A.H. *et al.* Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat Genet* **45**, 1353-60 (2013).
3. Sawcer, S., Franklin, R.J. & Ban, M. Multiple sclerosis genetics. *Lancet Neurol* **13**, 700-9 (2014).
4. Yu, G., Wang, L.G. & He, Q.Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382-3 (2015).
5. Teschendorff, A.E. *et al.* DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer. *Nat Commun* **7**, 10478 (2016).
6. Houseman, E.A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).
7. Aryee, M.J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363-9 (2014).
8. Wu, H. *et al.* Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic Acids Res* **43**, e141 (2015).
9. Price, A.L., Eskin, E. & Pevzner, P.A. Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. *Genome Res* **14**, 2245-52 (2004).
10. Aranyi, T., Varadi, A., Simon, I. & Tusnady, G.E. The BiSearch web server. *BMC Bioinformatics* **7**, 431 (2006).
11. Liguori, M. *et al.* Age at onset in multiple sclerosis. *Neurol Sci* **21**, S825-9 (2000).
12. O'Connor, P. & Canadian Multiple Sclerosis Working, G. Key issues in the diagnosis and treatment of multiple sclerosis. An overview. *Neurology* **59**, S1-33 (2002).
13. Hacısuleyman, E., Shukla, C.J., Weiner, C.L. & Rinn, J.L. Function and evolution of local repeats in the Firre locus. *Nat Commun* **7**, 11021 (2016).
14. Izuogu, O.G. *et al.* Analysis of human ES cell differentiation establishes that the dominant isoforms of the lncRNAs RMST and FIRRE are circular. *BMC Genomics* **19**, 276 (2018).
15. Stunnenberg, H.G., International Human Epigenome, C. & Hirst, M. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell* **167**, 1145-1149 (2016).
16. Teschendorff, A.E., Breeze, C.E., Zheng, S.C. & Beck, S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC Bioinformatics* **18**, 105 (2017).
17. Weichenhan, D. *et al.* Tagmentation-Based Library Preparation for Low DNA Input Whole Genome Bisulfite Sequencing. *Methods Mol Biol* **1708**, 105-122 (2018).
18. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-20 (2014).
19. Pedersen, B.S., Eyring, K., De, S., Yang, I.V., Schwartz, D.A. Fast and accurate alignment of long bisulfite-seq reads. *arXiv:1401.1129 [qbio.GN]*.

## Chapter 5

### General Discussion, Conclusions and Outlook

In this cumulative work a different aspect of DNA methylation has been investigated together with various epigenetic marks in order to understand the role of DNA methylation topology in cell fate and its association to genome organization. In this chapter the results of the three presented papers are summarized and discussed. Section 5.8 provides a broad perspective on how this work can be further developed, and suggests possible directions for epigenetic data integration.

Since the first published mammalian WGBS data set and the discovery of PMDs, very few studies reported on PMD features. These studies were conducted mainly in cell lines, cancer cells and very limited number of tissues but not yet in isolated specific primary cell types. Additionally, most studies that employed WGBS were focusing, to large degree, on DMRs identification and linking DNA methylation changes of regulatory regions, like promoters, enhancers and insulators, to gene expression. Collectively, this thesis draw the attention to new features of PMDs in different primary cells as well as in cell lines, and provide a new sight to study DNA methylation on a large-scale instead of mining only in the zoomed-in view (i.e., CpG islands, promoters and enhancers).

#### 5.1 DNA methylation and chromatin organization

Part of Chapter 2 highlighted a so far neglected aspect of DNA methylation, i.e., its link to 3D genome organization. DNA methylation segmentation revealed two main epi-organizational units in the genome possessing contrasted features; (i) PMDs characterized by fuzzy methylation patterns and heterochromatic signature, and (ii) HMDs featured by homogeneous, highly methylated CpGs and less packed chromatin. While PMDs overlay with heterochromatic B- compartments, HMDs represent active A- compartments. Such relationship between DNA methylation domains and A/B compartments would be beneficial when detecting TADs, and can be used as a proxy for Hi-C data to define TADs and compartments. As Hi-C data was used before to construct the hierarchical structure of the genome (metaTADs) [Fraser et al., 2015], it would be interesting to check whether this is also reflected in PMDs. It might be that

PMDs which have similar average methylation levels are more close to each other, in the 3D genome configurations, than to those that have more distinct average methylation levels, suggesting an existence of interactions between similarly methylated domains. Additionally, PMDs and HMDs can be used to limit the search space when linking DMRs to their potential target genes, because genes in a particular domain (PMD or HMD) most likely don't interact with genes in different PMDs/HMDs.

Most of DMRs calling methods are dependent on setting a methylation difference cut-off for calling significant DMRs, and this cumulative work showed clearly that the general methylation level is cell-type- and region- dependent, and there is genome wide loss of methylation in highly proliferating cells. Hence, a fixed methylation difference threshold for calling DMRs is not advisable. Instead, a more efficient way for calling and filtering DMRs, termed 'adaptive filtering', was proposed in Chapter 3 and was applied in Chapter 2, where DMRs were first annotated as being in PMDs/HMDs and then filtered according to the distribution of methylation difference of all DMRs in each segment type. Such analysis highlights the importance of utilizing PMDs, not only to capture the large scale changes in DNA methylation but also to enhance the detection of the 'functional DMRs' as local changes.

Chapter 2 represents one of the very early studies that link DNA methylation domains and the higher order chromatin structure. The current studies reported the relationship between histone marks and either the higher order chromatin structure or DNA methylation, without bridging DNA methylation and chromatin structure. Nothjunge et al. showed that PMD formation follows the establishment of B- compartments during the differentiation of cardiac myocytes [Nothjunge et al., 2017]. However, it is still interesting to investigate whether the heterochromatic marks occupy the B- compartments before, after or simultaneously with PMDs establishment. This kind of knowledge will help to understand the intertwined relationship between DNA methylation and heterochromatic marks at B- compartments. A very recent study reported 'Methyl-HiC', a method to profile simultaneously DNA methylation and chromatin architecture [Li et al., 2019]. Such method is useful to dissect the cross talk between DNA methylation and chromatin architecture, and to reveal their heterogeneity when applied on a single cell level.

In summary, it becomes very clear that DNA methylation has many roles other than controlling gene expression programs. It overlays with the 3D chromatin organization of the genome, and more efforts for developing new methods and tools would be needed in the future to dissect their relationship.

## 5.2 PMDs as discriminators for cellular origin

The first pillar of this thesis, Chapter 2, represents a comprehensive study of DNA methylation for a wide spectrum of WGBS human samples covering roughly ten primary cell types and tissues. It provides integrative approach to study PMDs in primary cells, primary tissues and immortalized cell lines, and unravels novel valuable features of PMDs.

The implemented integrative approach, ChromH3M, revealed PMDs as cell type discriminators and showed that gene regulation is existing genome-wide on a large scale level (PMDs/ HMDs), as well as on a smaller scale. PMD-associated genes showed cell-type specificity although they were lowly expressed or unexpressed. This result suggests that low gene activity or deactivation of some genes in PMDs is a regulatory feature, that is important as much as the activation of genes located outside of PMDs, to give a cell its specific functionality. Surprisingly, the average levels of PMD methylation vary between different cell types. For instance, PMDs of the myeloid cells, in general, have higher methylation levels compared to that of lymphoid cells. Moreover, within the lymphoid cells (B- and T-cells) DNA methylation of PMDs decreases with cell differentiation (see Chapter 3). These distinct methylation landscapes between lymphoid and myeloid cells suggest that they are derived from two different progenitors. It remains interesting to investigate PMDs of Multipotent Progenitor (MPP) and compare it to the ones of Common Lymphoid Progenitor (CLP) and Common Myeloid Progenitor (CMP).

Although thymus is the organ where T cells mature, thymocytes located, unexpectedly, closer to monocytes and macrophages than to lymphoid cells (Chapter 2, Figure 2a). This can be partially explained by the results from [Luc et al., 2012], who found that the earliest progenitors in the neonatal thymus possess combined granulocyte - monocyte and lymphoid lineage potential.

Based on the findings that PMD methylation levels are different across different cell types, HMM-based methods for PMD detection are more reliable than threshold-based methods. However, we realized that PMD detection using MethylSeekR (HMM-based) is more challenging for samples with highly methylated landscape like naive B and T-cells, monocytes and macrophages, where the difference of average methylation levels between PMDs and HMDs is small. This might cause loss of some very shallow-short PMDs. To enhance PMD detection, DNA methylation can be integrated with broad histone marks in one model (see Chapter 5.8).

The implemented ChromH3M workflow (see Chapter 2) complements Solo-WCGWs method [Zhou et al., 2018] by defining distinct patterns of PMDs across many samples, not to only identify common-PMDs/HMDs but also cell-type specific PMDs/HMDs. This

workflow can be generalized for any type of segmentation (e.g., LMRs, UMRs or any chromatin states). ChromH3M fits perfectly to the scope of EpiMAP project from IHEC (Section 1.4), and will help integrating large cohort of WGBS samples.

### 5.3 PMDs as indicators of proliferation history of cells

CD4<sup>+</sup> T cells play an important role in the adaptive immune system by generating different sub-types of memory cells (Tmem), which arise from naive T cells (Tn). However, the developmental relationship of Tmem cell sub-types is still under controversial debate. While some suggest a sequential linear model for Tmem differentiation, others reported a parallel model where Tmem branch off into different sub-types at early activation stage.

DNA methylation and other epigenetic marks play an important role in cell differentiation. With the help of such epigenetic data, Chapter 3 reported evidences that support the linear model of Tmem cells differentiation. The major evidence was the observation of genome-wide progressive loss of DNA methylation during differentiation of Tn into differentiated memory sub-types, i.e., Tcm, Tem and Temra. This loss of DNA methylation occurred predominantly at PMDs, which were largely overlapping between Tn and Tmem cells. Moreover, the same phenomenon was confirmed during differentiation of B cells into memory and plasma cells, but not during the differentiation of monocytes to macrophages (they do not proliferate). This feature, demethylation, seems to be shared in the highly proliferating cells, i.e., lymphoid lineage, but absent in non-proliferating cells, i.e., myeloid lineage. Additionally, it suggests PMDs-demethylation as indicator for the proliferation history of cells. Although, the loss of methylation in PMDs was correlated to past episodes of proliferation in CD4<sup>+</sup> T cells, it remains to be investigated, experimentally, whether the loss of methylation is proportional to the number of cell divisions as the model of [Dmitrijeva et al., 2018] proposed. Transcriptome and chromatin accessibility analyses supported the linear differentiation model (Tn > Tcm > Tem > Temra), which is in line with the observed global loss of methylation in PMDs. This is a confirmation that large scale DNA methylation changes are linked to transcriptional control.

To date, there is no concrete evidence that explains the genome-wide hypomethylation in PMDs and the mechanism behind the demethylation process associated with proliferation/replication. [von Meyenn et al., 2016] attributed the global demethylation in naive embryonic stem cells to the loss of DNA methylation maintenance, through impaired recruitment of DNMT1 by UHRF1 to the replication foci, excluding the possibility of active mechanisms by TET or reduction of de novo methylation. Whether this holds for somatic cells still needs more investigations. It is known that

UHRF1 binds to hemimethylated DNA at replication forks during S phase, so it would be interesting to screen the distribution of hemimethylated CpGs in PMDs and HMDs, which can be carried out by utilizing genome-wide hairpin bisulfite assay in combination with long-read sequencing technologies.

Another hypothesis to explain demethylation during differentiation / proliferation is the expansion of pre-existing subpopulations that already contain PMDs. Although our analysis could not exclude this hypothesis, it argued in favour of 'gradual loss of methylation with cell division' model, which is in agreement with [Dmitrijeva et al., 2018]. Alternatively, a TET-associated mechanism may contribute to the genome-wide hypomethylation. A recent study [López-Moyado et al., 2019] postulates that TET deficiency in different cell types might be a fundamental mechanism behind the widespread hypomethylation in heterochromatin compartments, coupled to focal hypermethylation in euchromatic regions, a feature of cancer genomes [Berman et al., 2012]. They explained the cooperation between TET1 and DNMT3A1 in TET-deficient genomes and proposed a model, in which loss of TET1 in mESCs leads to relocalization of DNMT3A1 away from heterochromatic compartments into focal regions in euchromatic compartments that were previously occupied by TET1 (which oxidize 5mC into 5hmC), leading to a global hypomethylation in the heterochromatic compartments (overlapping with PMDs), and regional hypermethylation in the euchromatic compartments.

In summary, two important points need further investigations to elucidate the mechanism behind global demethylation; (i) why and how some domains (PMDs) in the genome are less methylated than others (HMDs), in other words how PMDs are established?, and (ii) by which mechanism(s) the demethylation is propagated during proliferation/replication. A HMM-model developed by [Giehr et al., 2016] to measure the activities of different enzymes (TETs and DNMTs) during proliferation could be applied genome-wide to investigate the potential role of these enzymes in the demethylation phenomenon.

## 5.4 PMDs in cancerous cells and immortalized cell lines

DNA methylation patterns are transmitted during replication into the daughter cells by a copying mechanism utilizing DNMT1 to maintain global DNA methylation. However it is known that cells in culture undergo excessive epigenetic alterations linked to passages and cell replication numbers [Shipony et al., 2014]. Such aberration can lead to erroneous gene expression and alter cell properties. In this respect, the biological interpretation drawn from studies carried out on immortalized cell lines, that are widely used for studying epigenetic control, should be considered

with great care.

Chapter 2 addressed this issue by comparing primary cells to cell lines of the same origin. More specifically, the impact of cultivation and cancer-specific changes of PMD methylation have been investigated. The striking observation was that cancerous liver tissue retained a PMD structure similar to that of primary hepatocytes (PHH) and liver tissue, with mildly reduced methylation level, especially at PMDs, which suggests DNA methylation in PMDs as an epigenetic memory of cancer cell origin. On the other hand, the two cancer cell lines were more similar to each other than to the tissue cancerous cells with (strong) erosion of DNA methylation at PHH-associated PMDs. This suggests that PMDs demethylation in the cancer cell lines might reflect the cell's proliferation history, but less of the 'honest' cancer state.

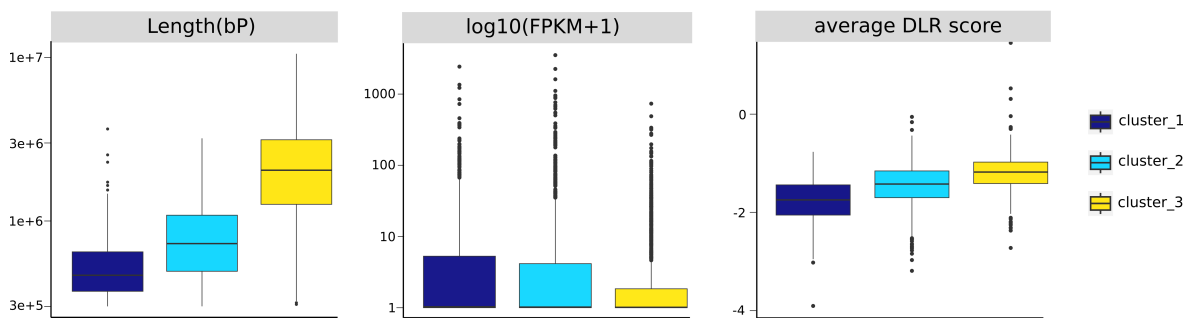
The previously described cancer hallmark, long range hypomethylation [Berman et al., 2012], seems not to be an exclusively cancer characteristic, but it is also an imprint of primary cells as shown in Chapter 2. At which PMD methylation level the cancer starts and deviate from the healthy state needs more investigations. A suggested analysis would be a genome wide screening of DNA methylation levels at PMDs during different stages of the cancer. The outcome from such analysis may serve as a diagnostic tool for cancer or as a biomarker for cancer progression.

## 5.5 Heterochromatic signatures of PMDs and replication domains

PMDs have been associated with repressive histone marks (H3K27me3 and H3K9me3) and linked to late replicating regions. However, detailed analysis of heterochromatic signatures of PMDs in relation to replicating domains is lacking. Utilizing an unsupervised clustering method, *k*-means, PMDs of HepG2 have been classified using replication data into three groups; (i) early/mid S phase (=cluster\_1), (ii) mid/late S phase (=cluster\_2) and (iii) very late S/G2 phase (=cluster\_3) (see Chapter 2, Figure 4a). Surprisingly, each of these domains is characterized by distinct heterochromatic pattern. The most unambiguous cluster of PMDs is cluster\_3, the longest of the three clusters, marked by H3K9me3 and bordered by H3K27me3, while the other two are more similar to each other and can be differentiated by the enrichment level of H3K27me3 (see Chapter 2, Figure 4b). The high base-pair overlapping percentage of cluster\_3 with the shared/common PMDs (as defined by ChromH3M in Chapter 2) suggests these domains as very stable and rigid heterochromatic structure in comparison to the other two. Moreover, they harbor the most lowly expressed genes. The various chromatin states of PMDs also suggest distinct chromatin packaging which



could be confirmed by DLR score <sup>‡</sup> from Hi-C data, where PMDs of the longest cluster, cluster\_3, were the most condensed and compacted in comparison to PMDs of the other two clusters (Figure 5.1 and our recent work [Nordström et al., 2019] (in press in NAR journal). Such packaging may explain why cluster\_1 and cluster\_2 replicate before cluster\_3, or it can simply be due to domain length characteristic, i.e., the longer the region the longer the time it needs for replication. Intriguingly, the packaging was mirrored in the genome wide “background” of NOME signal <sup>‡‡</sup> which has been neglected for long time (see our recent work [Nordström et al., 2019]).



**Figure 5.1: Different chromatin states of PMDs show distinct packaging of chromatin during replication.** PMDs of cluster\_3 (Very late S/G2 phase) are the longest (left panel), and have the lowest transcriptional activity (middle panel), and are more highly condensed in comparison to the other two clusters (early/mid S and mid/late S) according to DLR score (right panel).

Taken together, our findings extend and advance previous studies by defining fine grained heterochromatic signatures in PMDs of HepG2 and IMR90 at different stages of replication (S - G2). The epigenomic plasticity observed in these distinguished domains could be relevant for differentiation and cell fate determination.

## 5.6 PMDs in MS-discordant monozygotic twins

With the notion from the two studies (Chapter 2 and Chapter 3) about global DNA methylation changes in PMDs of cancer and CD4<sup>+</sup> T cells, we investigated whether this hallmark is also existing in MS disease. Indeed, MS co-twins had PMDs as well as the healthy co-twins. However, evaluation of genome-wide DNA methylation changes in PMDs/HMDs of CD4<sup>+</sup> T cells from four MS-discordant monozygotic twins did not

<sup>‡</sup>DLR (Distal-to-Local [log<sub>2</sub>] Ratio) score is defined by Homer software as “Log<sub>2</sub> ratio of distal Hi-C interactions interacting along the chromosome at distances greater than 3 Mb compared to local Hi-C interactions interacting less than 3 Mb” [Heinz et al., 2010]

<sup>‡‡</sup>NOME-Seq stands for “nucleosome occupancy and methylation”. It is a method to profile accessible chromatin regions on a genome wide level by utilizing the enzyme M.CviPI which specifically methylates cytosines in a GpC sequence context. Thus, measuring two signals simultaneously; methylation levels at GpC sites and at endogenous CpG sites (more details in ref. [Nordström et al., 2019]).

reveal significant differences. But, a prominent MS-DMR in a gene located on the X-chromosome was identified. Overall, this study showed how consistent the methylomes of MS-discordant monozygotic twins are with respect to the large domains (PMDs/HMDs), accompanied with few MS-associated DMRs as local changes.

## 5.7 Conclusion

The studies presented in this thesis illuminate diverse roles of DNA methylation in different contexts. Apart from its role in gene expression, it overlays with the 3D chromatin architecture. PMDs are epi-topological units that coincide with the heterochromatic compartments. Distinct heterochromatic states of PMDs in cancer cell lines distinguish different replicating domains and different packaging states. Moreover, PMDs are strong cell type discriminators, and the observed decrease of methylation in PMDs is an indicator of cell proliferation history. The observed decrease in DNA methylation at PMDs is more pronounced in immortalized cell lines, arguing for replication-dependent loss of methylation in PMDs. The segmented loss of methylation in PMDs during CD4<sup>+</sup> T memory cells differentiation supports the linear model of Tmem differentiation.

The computational method and aspects developed within this work are very much suited for the epigenetic community. First, ChromH3M is an easy and straightforward workflow for integration of hundreds of WGBS samples. Second, employing PMDs for stratifying and filtering DMRs is a powerful tool when studying DNA methylation changes in highly proliferating cells. Third, the overall presented epigenetic data integration led to discovering new aspects of DNA methylation that are of great interest to a broad audience of biologists.

## 5.8 Outlook

Epigenetics is a fast growing field in terms of both wet lab and computational methods. Profiling of 1000 epigenomes with high quality standards is an ambitious goal for IHEC, and integrating these data afterward is very challenging. The diversity of the assays make it even formidable task. But our knowledge about the commonalities between the assay's readouts and our expertise about the current tools' limitations can help to direct our thinking about what should be improved in the current tools, what kind of new tools are needed and how can we develop them?

Within the context of this work, we observed that the available DNA methylation segmentation tool (MethylSeekR) to detect PMDs has difficulties in detecting short-shallow PMDs of high methylation profiles, e.g., monocytes and macrophages. On the other hand we also observed the intertwined relationship between DNA methylation

and histone marks on a large scale. Hence, taking the advantage of this knowledge, implementing a HMM-method that account for both marks can not only improve PMDs detection, but also help exploring the relatively large ‘unknown/quiescent’ state that result in usually from histone mark-segmentation tools (e.g., ChromHMM and EpiCSeq). Moreover, the feature patterns coming out of such model will improve our understanding of the dependencies between these two epigenetic marks and maybe capturing new ‘biological’ states that were missed by either of the two methods (i.e., MethylSeekR- and ChromHMM/EpiCSeq-like).

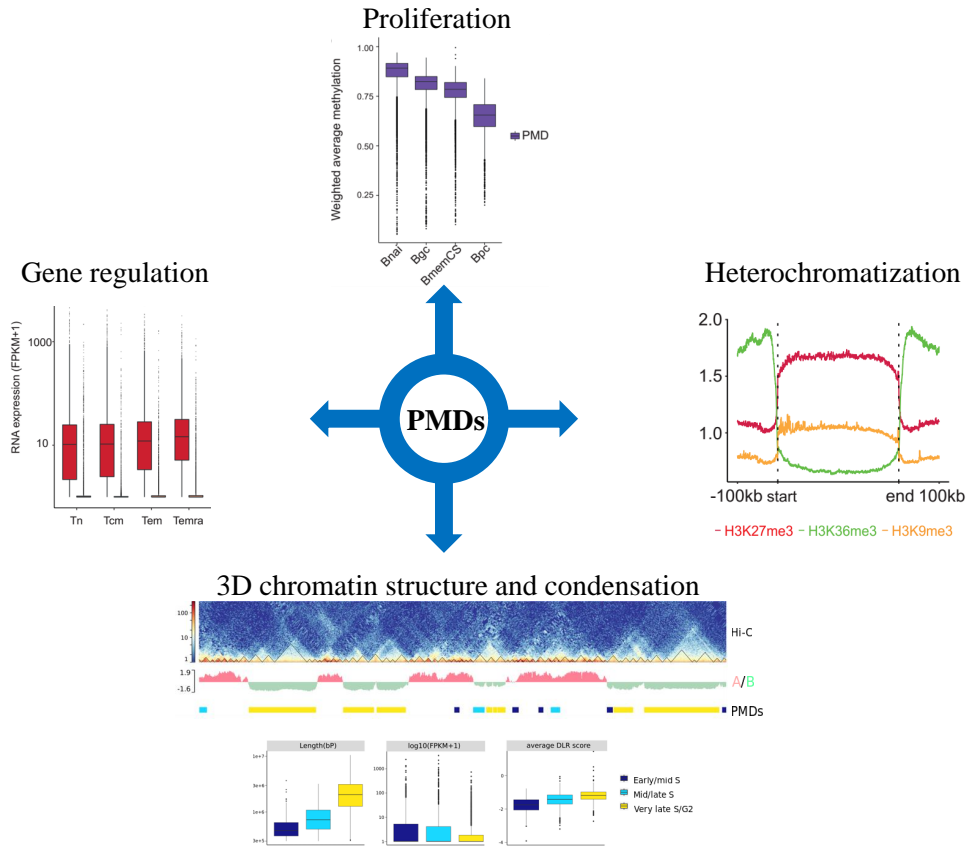
It is of high interest to apply such model on the wealthy data of IHEC. However, unified processing of these data, generated by different members, need to be done first in order to minimize the variance that may arise from the fact that each consortium has its own processing pipelines. Despite of the great effort by IHEC to standardize and unify data processing procedures, it is still facing some obstacles associated to data retrieval and sharing, and lacking of nomenclature-consistency of the metadata. Overcoming these hurdles will ease the access to the epigenetic data through some established databases like DeepBlue [Albrecht et al., 2016]. Hence, helping scientist who are aiming doing integrative analysis to get the most out of the available epigenetic data sets.

Illumina EPIC array becomes an attractive platform for Epigenome-Wide Association Studies (EWAS) because it covers roughly 850,000 CpG positions distributed over many regulatory regions, and it is an efficient and cost-effective approach for studying DNA methylation. It would be of great interest to use methylation arrays for PMD detection taking the advantage of available tool like minfi for A/B compartments segmentation [Fortin and Hansen, 2015] as a proxy for PMDs/HMDs domains, and adapt it to work with individual sample. Similarly, this could be applied for RRBS data.

One Major outcome of this cumulative work is that PMDs are linked to proliferation, DNA replication, differentiation, heterochromatization, 3D chromatin structure and gene regulation (Figure 5.2). These results came as consequences of several integrative analyses, which point us to the importance of data integration in the epigenetics filed. For instance, scientist observed that epigenetic marks are connected to 3D chromatin structure and they started to develop methods to resolve this relationship [Li et al., 2019].

Understanding the mechanism behind the hypomethylation of PMDs in cancer cells may have medical and therapeutic applications through DNA methylation re-programming. However, in-depth examination needs to be carried out to resolve this phenomenon, which was also found to be existed in primary cells. Hence, clear characterization of hypomethylation in primary cells and their cancerous counterparts can help understanding the role of PMDs hypomethylation in cancer.

Altogether, advancing in the epigenetic field stimulates the bioinformatician and computational biologists to develop new methods to deal with the readouts from the new technologies. Nevertheless, there is still a need to develop methods for integrative analysis to deal with the complexity of epigenetic data.



**Figure 5.2: PMDs in different contexts.** A summary scheme for the different roles of PMDs investigated in this thesis.

# Chapter A

## Appendix: List Of Publications

### (Co-)first Author Publications

Durek, P.<sup>§</sup>, Nordström, K.<sup>§</sup>, Gasparoni, G.<sup>§</sup>, **Salhab, A.**<sup>§</sup>, Kressler, C.<sup>§</sup>, De Almeida, M.<sup>§</sup>, ... & Glažar, P. (2016). Epigenomic profiling of human CD4<sup>+</sup> T cells supports a linear differentiation model and highlights molecular regulators of memory development. *Immunity*, 45(5), 1148-1161. doi: <http://doi.org/10.1016/j.immuni.2016.10.022>.

**Salhab, A.**, Nordström, K., Gasparoni, G., Kattler, K., Ebert, P., Ramirez, F., ... & Hengstler, J. G. (2018). A comprehensive analysis of 195 DNA methylomes reveals shared and cell-specific features of partially methylated domains. *Genome biology*, 19(1), 150. doi: <http://doi.org/10.1186/s13059-018-1510-5>.

### Contributing Author Publications

Souren, N. Y., Gerdes, L. A., Lutsik, P., Gasparoni, G., Beltrán, E., **Salhab, A.**, ... & Walter, J. (2019). DNA methylation signatures of monozygotic twins clinically discordant for multiple sclerosis. *Nature communications*, 10(1), 2094. doi: <http://doi.org/10.1038/s41467-019-09984-3>.

Nordström, K., Schmidt, F.<sup>‡</sup>, Gasparoni, N.<sup>‡</sup>, **Salhab, A.**<sup>‡</sup>, Gasparoni, G., Kattler, K., ... & Lengauer, T. (2019). Unique and assay specific features of NOME-, ATAC- and DNase I-seq data. *bioRxiv*, 547596. doi: <http://doi.org/10.1101/547596> (in press in *Nucleic Acids Research*, Accepted: September 11, 2019 )

---

<sup>§</sup>Co-first authors

<sup>‡</sup>These authors contributed equally to this work

## Chapter B

### Appendix: License and Copyright Information

#### B.1 Manuscripts

##### B.1.1 A comprehensive analysis of 195 DNA methylomes reveals shared and cell-specific features of partially methylated domains

The manuscript [Salhab et al., 2018] was originally published in Genome Biology under the terms of the Creative Commons Attribution 4.0 International License which grants the following rights:

“This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated”

##### B.1.2 Epigenomic Profiling of Human CD4<sup>+</sup> T Cells Supports a Linear Differentiation Model and Highlights Molecular Regulators of Memory Development

The manuscript [Durek et al., 2016] was originally published in Immunity (Copyright ©2016 Elsevier Inc.). The author of this paper has the following right:

“Authors can include their articles in full or in part in a thesis or dissertation for non-commercial purposes.”

(see <https://www.elsevier.com/about/policies/copyright/permissions>)

##### B.1.3 DNA methylation signatures of monozygotic twins clinically discordant for multiple sclerosis

The manuscript [Souren et al., 2019] was originally published in Nature Communication under a Creative Common Attribution 4.0 International License which grants the following rights:

“This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.”

## B.2 Figure Reprints

Figure	License	Publisher	Reference
1.1	Open Access (CC BY 4.0)	Taylor & Francis	[Hansen et al., 2018]
1.2	#4593010638458	AAAS	[Lieberman-Aiden et al., 2009]
1.7	#4636550877434	Nature Publishing Group	[Krueger et al., 2012]

**Table B.1:** Licensing information for figure reuse

## Bibliography

- Albrecht, F., List, M., Bock, C., and Lengauer, T. (2016). Deepblue epigenomic data server: programmatic data retrieval and analysis of epigenome region sets. *Nucleic acids research*, 44(W1):W581–W586.
- Allis, C. D., Caparros, M.-L., Jenuwein, T., and Reinberg, D. (2015). *Epigenetics*. Cold Spring Harbor Laboratory Press.
- Aran, D., Toperoff, G., Rosenberg, M., and Hellman, A. (2010). Replication timing-related and gene body-specific methylation of active human genes. *Human molecular genetics*, 20(4):670–680.
- Arrigoni, L., Al-Hasani, H., Ramírez, F., Panzeri, I., Ryan, D. P., Santacruz, D., Kress, N., Pospisilik, J. A., Bönisch, U., and Manke, T. (2018). Relacs nuclei barcoding enables high-throughput chip-seq. *Communications biology*, 1(1):214.
- Bannister, A. J., Schneider, R., Myers, F. A., Thorne, A. W., Crane-Robinson, C., and Kouzarides, T. (2005). Spatial distribution of di-and tri-methyl lysine 36 of histone h3 at active genes. *Journal of Biological Chemistry*, 280(18):17732–17736.
- Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837.
- Beagrie, R. A., Scialdone, A., Schueler, M., Kraemer, D. C., Chotalia, M., Xie, S. Q., Barbieri, M., de Santiago, I., Lavitas, L.-M., Branco, M. R., et al. (2017). Complex multi-enhancer contacts captured by genome architecture mapping. *Nature*, 543(7646):519.
- Bell, A. C. and Felsenfeld, G. (2000). Methylation of a ctcf-dependent boundary controls imprinted expression of the igf2 gene. *Nature*, 405(6785):482.
- Berezney, R. and Jeon, K. W. (1995). *Nuclear matrix: structural and functional organization*. Elsevier.
- Berman, B. P., Weisenberger, D. J., Aman, J. F., Hinoue, T., Ramjan, Z., Liu, Y., Noushmehr, H., Lange, C. P., van Dijk, C. M., Tollenaar, R. A., et al. (2012). Regions of focal dna hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nature Genetics*, 44(1):40–46.
- Bernstein, B. E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D. K., Huebert, D. J., McMahon, S., Karlsson, E. K., Kulbokas III, E. J., Gingeras, T. R., et al. (2005). Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell*, 120(2):169–181.



- Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125(2):315–326.
- Bochtler, M., Kolano, A., and Xu, G.-L. (2017). Dna demethylation pathways: additional players and regulators. *Bioessays*, 39(1):1–13.
- Bock, C. (2012). Analysing and interpreting dna methylation data. *Nature Reviews Genetics*, 13(10):705.
- Bock, C., Tomazou, E. M., Brinkman, A. B., Müller, F., Simmer, F., Gu, H., Jäger, N., Gnirke, A., Stunnenberg, H. G., and Meissner, A. (2010). Quantitative comparison of genome-wide dna methylation mapping technologies. *Nature biotechnology*, 28(10):1106.
- Burger, L., Gaidatzis, D., Schübeler, D., and Stadler, M. B. (2013). Identification of active regulatory regions from dna methylation data. *Nucleic acids research*, 41(16):e155–e155.
- Cavalli, G. and Heard, E. (2019). Advances in epigenetics link genetics to the environment and disease. *Nature*, 571(7766):489.
- Chen, B., Gilbert, L. A., Cimini, B. A., Schnitzbauer, J., Zhang, W., Li, G.-W., Park, J., Blackburn, E. H., Weissman, J. S., Qi, L. S., et al. (2013). Dynamic imaging of genomic loci in living human cells by an optimized crispr/cas system. *Cell*, 155(7):1479–1491.
- Chen, P.-Y., Cokus, S. J., and Pellegrini, M. (2010). Bs seeker: precise mapping for bisulfite sequencing. *BMC bioinformatics*, 11(1):203.
- Chen, Y., Zhang, Y., Wang, Y., Zhang, L., Brinkman, E. K., Adam, S. A., Goldman, R., van Steensel, B., Ma, J., and Belmont, A. S. (2018). Mapping 3d genome organization relative to nuclear compartments using tsa-seq as a cytological ruler. *The Journal of Cell Biology*, 217(11):4025–4048.
- Coarfa, C., Yu, F., Miller, C. A., Chen, Z., Harris, R. A., and Milosavljevic, A. (2010). Pash 3.0: A versatile software package for read mapping and integrative analysis of genomic and epigenomic variation using massively parallel dna sequencing. *BMC bioinformatics*, 11(1):572.
- Cremer, T. and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature reviews genetics*, 2(4):292.
- Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., et al. (2010). Histone h3k27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107(50):21931–21936.
- Dekker, J., Belmont, A. S., Guttman, M., Leshyk, V. O., Lis, J. T., Lomvardas, S., Mirny, L. A., O’shea, C. C., Park, P. J., Ren, B., et al. (2017). The 4d nucleome project. *Nature*, 549(7671):219.
- Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. *Science*, 295(5558):1306–1311.

- Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376.
- Dmitrijeva, M., Ossowski, S., Serrano, L., and Schaefer, M. H. (2018). Tissue-specific dna methylation loss during ageing and carcinogenesis is linked to chromosome structure, replication timing and cell division rates. *Nucleic acids research*, 46(14):7022–7039.
- Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., et al. (2006). Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements. *Genome research*, 16(10):1299–1309.
- Durek, P., Nordström, K., Gasparoni, G., Salhab, A., Kressler, C., De Almeida, M., Bassler, K., Ulas, T., Schmidt, F., Xiong, J., et al. (2016). Epigenomic profiling of human CD4<sup>+</sup> T cells supports a linear differentiation model and highlights molecular regulators of memory development. *Immunity*, 45(5):1148–1161.
- Ebert, P., Müller, F., Nordström, K., Lengauer, T., and Schulz, M. H. (2015). A general concept for consistent documentation of computational analyses. *Database*, 2015.
- ENCODE et al. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57.
- Ernst, J. and Kellis, M. (2012). Chromhmm: automating chromatin-state discovery and characterization. *Nature methods*, 9(3):215.
- Feng, S., Cokus, S. J., Zhang, X., Chen, P.-Y., Bostick, M., Goll, M. G., Hetzel, J., Jain, J., Strauss, S. H., Halpern, M. E., et al. (2010). Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences*, 107(19):8689–8694.
- Forcato, M., Nicoletti, C., Pal, K., Livi, C. M., Ferrari, F., and Bicciato, S. (2017). Comparison of computational methods for hi-c data analysis. *Nature methods*, 14(7):679.
- Forrest, A. R., Kawaji, H., Rehli, M., Baillie, J. K., De Hoon, M. J., Haberle, V., Lassmann, T., Kulakovskiy, I. V., Lizio, M., Itoh, M., et al. (2014). A promoter-level mammalian expression atlas. *Nature*, 507(7493):462.
- Fortin, J.-P. and Hansen, K. D. (2015). Reconstructing a/b compartments as revealed by hi-c using long-range correlations in epigenetic data. *Genome biology*, 16(1):180.
- Fraser, J., Ferrai, C., Chiariello, A. M., Schueler, M., Rito, T., Laudanno, G., Barbieri, M., Moore, B. L., Kraemer, D. C., Aitken, S., Xie, S. Q., Morris, K. J., Itoh, M., Kawaji, H., Jaeger, I., Hayashizaki, Y., Carninci, P., Forrest, A. R., Semple, C. A., Dostie, J., Pombo, A., and Nicodemi, M. (2015). Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Molecular Systems Biology*, 11(12).
- Frith, M. C., Mori, R., and Asai, K. (2012). A mostly traditional approach improves alignment of bisulfite-converted dna. *Nucleic acids research*, 40(13):e100–e100.

- Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L., and Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual dna strands. *Proceedings of the National Academy of Sciences*, 89(5):1827–1831.
- Gibcus, J. and Dekker, J. (2013). The hierarchy of the 3d genome. *Molecular Cell*, 49(5):773 – 782.
- Giehr, P., Kyriakopoulos, C., Ficz, G., Wolf, V., and Walter, J. (2016). The influence of hydroxylation on maintaining cpg methylation patterns: a hidden markov model approach. *PLoS computational biology*, 12(5):e1004905.
- Hajkova, P., Erhardt, S., Lane, N., Haaf, T., El-Maarri, O., Reik, W., Walter, J., and Surani, M. A. (2002). Epigenetic reprogramming in mouse primordial germ cells. *Mechanisms of development*, 117(1-2):15–23.
- Hall, I. M., Shankaranarayana, G. D., Noma, K.-i., Ayoub, N., Cohen, A., and Grewal, S. I. (2002). Establishment and maintenance of a heterochromatin domain. *Science*, 297(5590):2232–2237.
- Han, Z. and Wei, G. (2017). Computational tools for hi-c data analysis. *Quantitative Biology*, 5(3):215–225.
- Hansen, A. S., Cattoglio, C., Darzacq, X., and Tjian, R. (2018). Recent evidence that tads and chromatin loops are dynamic structures. *Nucleus*, 9(1):20–32. PMID: 29077530.
- Hansen, K. D., Timp, W., Bravo, H. C., Sabunciyan, S., Langmead, B., McDonald, O. G., Wen, B., Wu, H., Liu, Y., Diep, D., et al. (2011). Increased methylation variation in epigenetic domains across cancer types. *Nature Genetics*, 43(8):768–775.
- Harris, E. Y., Ponts, N., Le Roch, K. G., and Lonardi, S. (2012). Brat-bw: efficient and accurate mapping of bisulfite-treated reads. *Bioinformatics*, 28(13):1795–1796.
- Harris, E. Y., Ponts, N., Levchuk, A., Roch, K. L., and Lonardi, S. (2009). Brat: bisulfite-treated reads analysis tool. *Bioinformatics*, 26(4):572–573.
- Hashimoto, H., Liu, Y., Upadhyay, A. K., Chang, Y., Howerton, S. B., Vertino, P. M., Zhang, X., and Cheng, X. (2012). Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. *Nucleic acids research*, 40(11):4841–4849.
- He, Q., Johnston, J., and Zeitlinger, J. (2015). Chip-nexus enables improved detection of in vivo transcription factor binding footprints. *Nature biotechnology*, 33(4):395.
- He, Y.-F., Li, B.-Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, L., et al. (2011). Tet-mediated formation of 5-carboxylcytosine and its excision by tdg in mammalian dna. *Science*, 333(6047):1303–1307.
- Heinig, M., Colomé-Tatché, M., Taudt, A., Rintisch, C., Schafer, S., Pravenec, M., Hubner, N., Vingron, M., and Johannes, F. (2015). histonehmm: Differential analysis of histone modifications with broad genomic footprints. *BMC bioinformatics*, 16(1):60.

- Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular cell*, 38(4):576–589.
- Hellman, A. and Chess, A. (2007). Gene body-specific methylation on the active x chromosome. *science*, 315(5815):1141–1143.
- Hess, S. T., Girirajan, T. P., and Mason, M. D. (2006). Ultra-high resolution imaging by fluorescence photoactivation localization microscopy. *Biophysical journal*, 91(11):4258–4272.
- Holliday, R. (1994). Epigenetics: an overview. *Developmental genetics*, 15(6):453–457.
- Holliday, R. and Pugh, J. E. (1975). Dna modification mechanisms and gene activity during development. *Science*, 187(4173):226–232.
- Hon, G. C., Hawkins, R. D., Caballero, O. L., Lo, C., Lister, R., Pelizzola, M., Valsesia, A., Ye, Z., Kuan, S., Edsall, L. E., et al. (2012). Global dna hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Research*, 22(2):246–258.
- Hovestadt, V., Jones, D. T., Picelli, S., Wang, W., Kool, M., Northcott, P. A., Sultan, M., Stachurski, K., Ryzhova, M., Warnatz, H.-J., et al. (2014). Decoding the regulatory landscape of medulloblastoma using dna methylation sequencing. *Nature*, 510(7506):537–541.
- Hughes, J. R., Roberts, N., McGowan, S., Hay, D., Giannoulatou, E., Lynch, M., De Gobbi, M., Taylor, S., Gibbons, R., and Higgs, D. R. (2014). Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nature genetics*, 46(2):205.
- Irizarry, R. A., Ladd-Acosta, C., Carvalho, B., Wu, H., Brandenburg, S. A., Jeddloh, J. A., Wen, B., and Feinberg, A. P. (2008). Comprehensive high-throughput arrays for relative methylation (charm). *Genome research*, 18(5):780–790.
- Irizarry, R. A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., et al. (2009). The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific cpG island shores. *Nature Genetics*, 41(2):178–186.
- Ji, D., Lin, K., Song, J., and Wang, Y. (2014). Effects of tet-induced oxidation products of 5-methylcytosine on dnmt1-and dnmt3a-mediated cytosine methylation. *Molecular bioSystems*, 10(7):1749–1752.
- Ji, H., Ehrlich, L. I., Seita, J., Murakami, P., Doi, A., Lindau, P., Lee, H., Aryee, M. J., Irizarry, R. A., Kim, K., et al. (2010). Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature*, 467(7313):338.
- Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502.

- Jones, P. A. (2012). Functions of dna methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484.
- Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., and Chen, L. (2012). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature biotechnology*, 30(1):90.
- Kaya-Okur, H. S., Wu, S. J., Codomo, C. A., Pledger, E. S., Bryson, T. D., Henikoff, J. G., Ahmad, K., and Henikoff, S. (2019). Cut&tag for efficient epigenomic profiling of small samples and single cells. *bioRxiv*, page 568915.
- Kelly, T. K., Miranda, T. B., Liang, G., Berman, B. P., Lin, J. C., Tanay, A., and Jones, P. A. (2010). H2a. z maintenance during mitosis reveals nucleosome shifting on mitotically silenced genes. *Molecular cell*, 39(6):901–911.
- Khare, S. P., Habib, F., Sharma, R., Gadewal, N., Gupta, S., and Galande, S. (2011). HIstome—a relational knowledgebase of human histone proteins and histone modifying enzymes. *Nucleic Acids Research*, 40(D1):D337–D342.
- Kohli, R. M. and Zhang, Y. (2013). Tet enzymes, tdg and the dynamics of dna demethylation. *Nature*, 502(7472):472.
- Kolovos, P., van de Werken, H. J., Kepper, N., Zuin, J., Brouwer, R. W., Kockx, C. E., Wendt, K. S., van IJcken, W. F., Grosveld, F., and Knoch, T. A. (2014). Targeted chromatin capture (t2c): a novel high resolution high throughput method to detect genomic interactions and regulatory elements. *Epigenetics & Chromatin*, 7(1):10.
- Kouzarides, T. (2007). Chromatin modifications and their function. *Cell*, 128(4):693 – 705.
- Krueger, F. and Andrews, S. R. (2011). Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *bioinformatics*, 27(11):1571–1572.
- Krueger, F., Kreck, B., Franke, A., and Andrews, S. R. (2012). Dna methylome analysis using short bisulfite sequencing data. *Nature methods*, 9(2):145.
- Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C. T., Low, H. M., Sung, K. W. K., Rigoutsos, I., Loring, J., et al. (2010). Dynamic changes in the human methylome during differentiation. *Genome research*, 20(3):320–331.
- Li, G., Liu, Y., Zhang, Y., Kubo, N., Yu, M., Fang, R., Kellis, M., and Ren, B. (2019). Joint profiling of dna methylation and chromatin architecture in single cells. *Nature methods*, pages 1–3.
- Liang, G., Chan, M. F., Tomigahara, Y., Tsai, Y. C., Gonzales, F. A., Li, E., Laird, P. W., and Jones, P. A. (2002). Cooperativity between dna methyltransferases in the maintenance methylation of repetitive elements. *Molecular and cellular biology*, 22(2):480–491.
- Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and

- Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293.
- Lippman, Z., Gendrel, A.-V., Black, M., Vaughn, M. W., Dedhia, N., McCombie, W. R., Lavine, K., Mittal, V., May, B., Kasschau, K. D., et al. (2004). Role of transposable elements in heterochromatin and epigenetic control. *Nature*, 430(6998):471.
- Lister, R., Pelizzola, M., Downen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M., et al. (2009). Human dna methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322.
- Lister, R., Pelizzola, M., Kida, Y. S., Hawkins, R. D., Nery, J. R., Hon, G., Antosiewicz-Bourget, J., O'Malley, R., Castanon, R., Klugman, S., et al. (2011). Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature*, 471(7336):68–73.
- Litt, M. D., Simpson, M., Gaszner, M., Allis, C. D., and Felsenfeld, G. (2001). Correlation between histone lysine methylation and developmental changes at the chicken  $\beta$ -globin locus. *Science*, 293(5539):2453–2455.
- Liu, Y., Siegmund, K. D., Laird, P. W., and Berman, B. P. (2012). Bis-snp: combined dna methylation and snp calling for bisulfite-seq data. *Genome biology*, 13(7):R61.
- Lock, L. F., Takagi, N., and Martin, G. R. (1987). Methylation of the hprt gene on the inactive x occurs after chromosome inactivation. *Cell*, 48(1):39–46.
- López-Moyado, I. F., Tsagaratou, A., Yuita, H., Seo, H., Delatte, B., Heinz, S., Benner, C., and Rao, A. (2019). Paradoxical association of tet loss of function with genome-wide dna hypomethylation. *Proceedings of the National Academy of Sciences*, page 201903059.
- Luc, S., Luis, T. C., Boukarabila, H., Macaulay, I. C., Buza-Vidas, N., Bouriez-Jones, T., Lutteropp, M., Woll, P. S., Loughran, S. J., Mead, A. J., et al. (2012). The earliest thymic t cell progenitors sustain b cell and myeloid lineage potential. *Nature immunology*, 13(4):412.
- Maiti, A. and Drohat, A. C. (2011). Thymine dna glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine potential implications for active demethylation of cpg sites. *Journal of Biological Chemistry*, 286(41):35334–35338.
- Mammana, A. and Chung, H.-R. (2015). Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. *Genome biology*, 16(1):151.
- Manuelidis, L. (1985). Individual interphase chromosome domains revealed by in situ hybridization. *Human genetics*, 71(4):288–293.
- Martens, J. H., O'Sullivan, R. J., Braunschweig, U., Opravil, S., Radolf, M., Steinlein, P., and Jenuwein, T. (2005). The profile of repeat-associated histone lysine methylation states in the mouse epigenome. *The EMBO journal*, 24(4):800–812.
- Mayer, W., Niveleau, A., Walter, J., Fundele, R., and Haaf, T. (2000). Embryogenesis: demethylation of the zygotic paternal genome. *Nature*, 403(6769):501.

- Meissner, A., Mikkelsen, T. S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B. E., Nusbaum, C., Jaffe, D. B., et al. (2008). Genome-scale dna methylation maps of pluripotent and differentiated cells. *Nature*, 454(7205):766–770.
- Moarefi, A. H. and Chédin, F. (2011). Icf syndrome mutations cause a broad spectrum of biochemical defects in dnmt3b-mediated de novo dna methylation. *Journal of molecular biology*, 409(5):758–772.
- Morgan, H. D., Santos, F., Green, K., Dean, W., and Reik, W. (2005). Epigenetic reprogramming in mammals. *Human molecular genetics*, 14(suppl\_1):R47–R58.
- Nir, G., Farabella, I., Pérez Estrada, C., Ebeling, C. G., Beliveau, B. J., Sasaki, H. M., Lee, S. D., Nguyen, S. C., McCole, R. B., Chatteraj, S., Erceg, J., AlHaj Abed, J., Martins, N. M. C., Nguyen, H. Q., Hannan, M. A., Russell, S., Durand, N. C., Rao, S. S. P., Kishi, J. Y., Soler-Vila, P., Di Pierro, M., Onuchic, J. N., Callahan, S. P., Schreiner, J. M., Stuckey, J. A., Yin, P., Aiden, E. L., Marti-Renom, M. A., and Wu, C.-t. (2018). Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. *PLOS Genetics*, 14(12):1–35.
- Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N. L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, 485(7398):381.
- Nordström, K., Schmidt, F., Gasparoni, N., Salhab, A., Gasparoni, G., Kattler, K., Müller, F., Ebert, P., Costa, I. G., Pfeifer, N., et al. (2019). Unique and assay specific features of NOME-, ATAC-and DNase I-seq data. *bioRxiv*, page 547596.
- Nothjunge, S., Nührenberg, T. G., Grüning, B. A., Doppler, S. A., Preissl, S., Schwaderer, M., Rommel, C., Krane, M., Hein, L., and Gilsbach, R. (2017). Dna methylation signatures follow preformed chromatin compartments in cardiac myocytes. *Nature communications*, 8(1):1667.
- Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N., and Mirny, L. A. (2018). Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proceedings of the National Academy of Sciences*, 115(29):E6697–E6706.
- Oswald, J., Engemann, S., Lane, N., Mayer, W., Olek, A., Fundele, R., Dean, W., Reik, W., and Walter, J. (2000). Active demethylation of the paternal genome in the mouse zygote. *Current Biology*, 10(8):475–478.
- Otani, J., Kimura, H., Sharif, J., Endo, T. A., Mishima, Y., Kawakami, T., Koseki, H., Shirakawa, M., Suetake, I., and Tajima, S. (2013). Cell cycle-dependent turnover of 5-hydroxymethyl cytosine in mouse embryonic stem cells. *PLoS one*, 8(12):e82961.
- Otto, C., Stadler, P. F., and Hoffmann, S. (2012). Fast and sensitive mapping of bisulfite-treated sequencing data. *Bioinformatics*, 28(13):1698–1704.
- Pedersen, B., Hsieh, T.-F., Ibarra, C., and Fischer, R. L. (2011). Methylcoder: software pipeline for bisulfite-treated sequences. *Bioinformatics*, 27(17):2435–2436.

- Pombo, A. and Dillon, N. (2015). Three-dimensional genome architecture: players and mechanisms. *Nature Reviews Molecular Cell Biology*, 16(4):245–257.
- Quinlan, A. R. and Hall, I. M. (2010). Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.
- Ramírez, F., Bhardwaj, V., Arrigoni, L., Lam, K. C., Grüning, B. A., Villaveces, J., Habermann, B., Akhtar, A., and Manke, T. (2018). High-resolution tads reveal dna sequences underlying genome organization in flies. *Nature communications*, 9(1):189.
- Rhee, H. S. and Pugh, B. F. (2011). Comprehensive genome-wide protein-dna interactions detected at single-nucleotide resolution. *Cell*, 147(6):1408–1419.
- Riggs, A. D. (1975). X inactivation, differentiation, and dna methylation. *Cytogenetic and Genome Research*, 14(1):9–25.
- Ringrose, L. and Paro, R. (2004). Epigenetic regulation of cellular memory by the polycomb and trithorax group proteins. *Annu. Rev. Genet.*, 38:413–443.
- Rodley, C., Bertels, F., Jones, B., and O’Sullivan, J. (2009). Global identification of yeast chromosome interactions using genome conformation capture. *Fungal Genetics and Biology*, 46(11):879 – 886.
- Rosenbloom, K. R., Armstrong, J., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T. R., Fujita, P. A., Guruvadoo, L., Haeussler, M., et al. (2014). The ucsc genome browser database: 2015 update. *Nucleic acids research*, 43(D1):D670–D681.
- Salhab, A. (2014). Correlation of dna-methylation topography with chromatin and transcriptional features in epigenomes of human cells. Master’s thesis, Saarland university, Germany.
- Salhab, A., Nordström, K., Gasparoni, G., Kattler, K., Ebert, P., Ramirez, F., Arrigoni, L., Müller, F., Polansky, J. K., Cadenas, C., G.Hengstler, J., Lengauer, T., Manke, T., DEEP Consortium, and Walter, J. (2018). A comprehensive analysis of 195 dna methylomes reveals shared and cell-specific features of partially methylated domains. *Genome Biology*, 19(1):150.
- Santos-Rosa, H., Schneider, R., Bannister, A. J., Sherriff, J., Bernstein, B. E., Emre, N. T., Schreiber, S. L., Mellor, J., and Kouzarides, T. (2002). Active genes are tri-methylated at k4 of histone h3. *Nature*, 419(6905):407.
- Schmidl, C., Klug, M., Boeld, T. J., Andreesen, R., Hoffmann, P., Edinger, M., and Rehli, M. (2009). Lineage-specific dna methylation in t cells correlates with histone methylation and enhancer activity. *Genome research*, 19(7):1165–1174.
- Schroeder, D. I., Blair, J. D., Lott, P., Yu, H. O. K., Hong, D., Crary, F., Ashwood, P., Walker, C., Korf, I., Robinson, W. P., and LaSalle, J. M. (2013). The human placenta methylome. *Proceedings of the National Academy of Sciences*, 110(15):6037–6042.



- Schroeder, D. I., Lott, P., Korf, I., and LaSalle, J. M. (2011). Large-scale methylation domains mark a functional subset of neuronally expressed genes. *Genome Research*, 21(10):1583–1591.
- Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the drosophila genome. *Cell*, 148(3):458 – 472.
- Shipony, Z., Mukamel, Z., Cohen, N. M., Landan, G., Chomsky, E., Zeligler, S. R., Fried, Y. C., Ainbinder, E., Friedman, N., and Tanay, A. (2014). Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. *Nature*, 513(7516):115.
- Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R., and Oberdoerffer, S. (2011). Ctf-promoted rna polymerase ii pausing links dna methylation to splicing. *Nature*, 479(7371):74.
- Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., De Wit, E., Van Steensel, B., and De Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4c). *Nature genetics*, 38(11):1348.
- Skene, P. J., Henikoff, J. G., and Henikoff, S. (2018). Targeted in situ genome-wide profiling with high efficiency for low cell numbers. *Nature protocols*, 13(5):1006.
- Skene, P. J., Hernandez, A. E., Groudine, M., and Henikoff, S. (2014). The nucleosomal barrier to promoter escape by rna polymerase ii is overcome by the chromatin remodeler chd1. *Elife*, 3:e02042.
- Smith, A. D., Chung, W.-Y., Hodges, E., Kendall, J., Hannon, G., Hicks, J., Xuan, Z., and Zhang, M. Q. (2009). Updates to the rmap short-read mapping software. *Bioinformatics*, 25(21):2841–2842.
- Song, Q. and Smith, A. D. (2011). Identifying dispersed epigenomic domains from chip-seq data. *Bioinformatics*, 27(6):870–871.
- Souren, N. Y., Gerdes, L. A., Lutsik, P., Gasparoni, G., Beltrán, E., Salhab, A., Kümpfel, T., Weichenhan, D., Plass, C., Hohlfeld, R., and Walter, J. (2019). DNA methylation signatures of monozygotic twins clinically discordant for multiple sclerosis. *Nature Communications*, 10(1):2094.
- Stadhouders, R., Kolovos, P., Brouwer, R., Zuin, J., Van Den Heuvel, A., Kockx, C., Palstra, R.-J., Wendt, K. S., Grosveld, F., Van Ijcken, W., et al. (2013). Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. *Nature protocols*, 8(3):509.
- Stadler, M. B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E. J., Gaidatzis, D., et al. (2011). Dna-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, 480(7378):490.

- Strahl, B. D. and Allis, C. D. (2000). The language of covalent histone modifications. *Nature*, 403(6765):41.
- Strahl, B. D., Ohba, R., Cook, R. G., and Allis, C. D. (1999). Methylation of histone h3 at lysine 4 is highly conserved and correlates with transcriptionally active nuclei in tetrahymena. *Proceedings of the National Academy of Sciences*, 96(26):14967–14972.
- Stunnenberg, H. G., Abrignani, S., Adams, D., de Almeida, M., Altucci, L., Amin, V., Amit, I., Antonarakis, S. E., Aparicio, S., Arima, T., et al. (2016). The international human epigenome consortium: a blueprint for scientific collaboration and discovery. *Cell*, 167(5):1145–1149.
- Timp, W., Bravo, H. C., McDonald, O. G., Goggins, M., Umbricht, C., Zeiger, M., Feinberg, A. P., and Irizarry, R. A. (2014). Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors. *Genome Medicine*, 6(8):61.
- van Steensel, B., Delrow, J., and Henikoff, S. (2001). Chromatin profiling using targeted dna adenine methyltransferase. *Nature genetics*, 27(3):304.
- von Meyenn, F., Iurlaro, M., Habibi, E., Liu, N. Q., Salehzadeh-Yazdi, A., Santos, F., Petrini, E., Milagre, I., Yu, M., Xie, Z., et al. (2016). Impairment of dna methylation maintenance is the main cause of global demethylation in naive embryonic stem cells. *Molecular cell*, 62(6):848–861.
- Wang, Z., Zang, C., Rosenfeld, J. A., Schones, D. E., Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Peng, W., Zhang, M. Q., et al. (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature genetics*, 40(7):897.
- Weber, A. R., Krawczyk, C., Robertson, A. B., Kuśnierczyk, A., Vågbø, C. B., Schuermann, D., Klungland, A., and Schär, P. (2016). Biochemical reconstitution of tet1–tdg–ber-dependent active dna demethylation reveals a highly coordinated mechanism. *Nature communications*, 7:10806.
- Weng, Y.-I., Huang, T. H.-M., and Yan, P. S. (2009). Methylated dna immunoprecipitation and microarray-based analysis: detection of dna methylation in breast cancer cell lines. In *Molecular Endocrinology*, pages 165–176. Springer.
- Wolf, S. F., Jolly, D. J., Lunnen, K. D., Friedmann, T., and Migeon, B. R. (1984). Methylation of the hypoxanthine phosphoribosyltransferase locus on the human x chromosome: implications for x-chromosome inactivation. *Proceedings of the National Academy of Sciences*, 81(9):2806–2810.
- Wolff, J., Bhardwaj, V., Nothjunge, S., Richard, G., Renschler, G., Gilsbach, R., Manke, T., Backofen, R., Ramírez, F., and Grüning, B. A. (2018). Galaxy HiCExplorer: a web server for reproducible Hi-C data analysis, quality control and visualization. *Nucleic Acids Research*, 46(W1):W11–W16.
- Wolffe, A. (1998). *Chromatin: structure and function*. Academic press.

- Wu, H. and Zhang, Y. (2014). Reversing dna methylation: mechanisms, genomics, and biological functions. *Cell*, 156(1-2):45–68.
- Wu, S. C. and Zhang, Y. (2010). Active dna demethylation: many roads lead to rome. *Nature reviews Molecular cell biology*, 11(9):607.
- Wu, T. D. and Nacu, S. (2010). Fast and snp-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–881.
- Wu, X. and Zhang, Y. (2017). Tet-mediated active dna demethylation: mechanism, function and beyond. *Nature Reviews Genetics*, 18(9):517.
- Xi, Y., Bock, C., Müller, F., Sun, D., Meissner, A., and Li, W. (2011). Rrbsmap: a fast, accurate and user-friendly alignment tool for reduced representation bisulfite sequencing. *Bioinformatics*, 28(3):430–432.
- Xi, Y. and Li, W. (2009). Bsmep: whole genome bisulfite sequence mapping program. *BMC bioinformatics*, 10(1):232.
- Yong, W.-S., Hsu, F.-M., and Chen, P.-Y. (2016). Profiling genome-wide dna methylation. *Epigenetics & chromatin*, 9(1):26.
- Zemach, A., McDaniel, I. E., Silva, P., and Zilberman, D. (2010). Genome-wide evolutionary analysis of eukaryotic dna methylation. *Science*, 328(5980):916–919.
- Zentner, G. E., Kasinathan, S., Xin, B., Rohs, R., and Henikoff, S. (2015). Chec-seq kinetics discriminates transcription factor binding sites by dna sequence and shape in vivo. *Nature communications*, 6:8733.
- Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., et al. (2008). Model-based analysis of chip-seq (macs). *Genome biology*, 9(9):R137.
- Zhao, M.-T., Whyte, J. J., Hopkins, G. M., Kirk, M. D., and Prather, R. S. (2014). Methylated dna immunoprecipitation and high-throughput sequencing (medip-seq) using low amounts of genomic dna. *Cellular Reprogramming (Formerly Cloning and Stem Cells)*, 16(3):175–184.
- Zhou, W., Dinh, H. Q., Ramjan, Z., Weisenberger, D. J., Nicolet, C. M., Shen, H., Laird, P. W., and Berman, B. P. (2018). Dna methylation loss in late-replicating domains is linked to mitotic cell division. *Nature genetics*, 50(4):591.