Mechanisms for CpG Island Methylation during X Chromosome Inactivation



Komal Yasmin

New College

University of Oxford

Supervised by

Professor Neil Brockdorff

Dr. Tatyana Nesterova

A thesis submitted for the degree of

Doctor of Philosophy

Chromosome and developmental biology

Michaelmas Term 2023

Abstract

During early development in female mammals one of the two X chromosomes is transcriptionally silenced, ensuring an equal dosage of X-linked genes relative to their XY male counterparts. This phenomenon is termed X chromosome inactivation (XCI). XCI is initiated by expression of the long noncoding RNA (lncRNA) Xist from the inactive X chromosome (Xi). Xist orchestrates molecular pathways *in cis* to establish transcriptional silencing and heterochromatinization of the Xi in a multi-stage process. A late stage of XCI is the methylation of CpG-rich regions associated with gene promoters, called CpG Islands (CGIs). Mammals express two highly similar *de novo* DNA methyl transferases, DNMT3A and DNMT3B. Intriguingly, only DNMT3B is crucial for Xi CGI methylation, and it is unclear what factor(s) imparts this specificity.

To address the mechanisms governing DNMT3B specificity, I have developed a sensitive experimental model system, utilising an XX mouse ES cell line with inducible Xist (iXist-ChrX) and quantitative MBD-seq, to conduct high-resolution, allele-specific analysis of XCI CGI methylation dynamics. This model was used to analyse *Dnmt3b*$^{-/-}$ cells and complementation lines expressing specific isoforms or mutants of DNMT3B. A detailed analysis of MBD-seq datasets unveiled that DNMT3B is dispensable for the methylation of a small group of X-linked CGIs and "CGI-like elements", a novel group of CpG-rich regions reported in this work that possess enhancer-like chromatin features and gain methylation on the Xi. I demonstrate that the catalytically active isoform of DNMT3B, DNMT3B1, fully rescues CGI methylation in *Dnmt3b*$^{-/-}$ cells and that the widely expressed catalytically inactive isoform, DNMT3B3, does not contribute to Xi CGI methylation. I also explore the contribution of the less characterized DNMT3B N-terminal domain as a putative specificity imparting region, and report that this is indeed crucial for DNMT3B1 mediated CGI methylation on Xi and autosomes. Intriguingly, both complete or partial deletion of the N-terminal domain of DNMT3B1 abrogates the rescue of CGI methylation observed with full-length DNMT3B1. Furthermore, the N-terminal domain has a context-specific role in the genomic localization of DNMT3B1, being crucial for the localization/residence of DNMT3B1 at CGIs and the Xi.

By studying the effect of a broad range of DNMT3B manipulations on the methylation of X-linked CGIs, and performing an exploratory proteomics analysis to identify proteins associated with DNMT3B before and after XCI, this work presents new insights into Xi CGI methylation and makes progress towards understanding the specificity of DNMT3B on the Xi.

Acknowledgements

Given my humble roots in a remote rural area of Pakistan, submitting a DPhil thesis at the University of Oxford is nothing short of a dream come true and I have a number of people to thank for this great honour.

I would first like to thank my supervisor Professor Neil Brockdorff for giving me the opportunity to work in his lab. Neil has always been kind, encouraging, and generous with his time, and allowed me the freedom to pursue my ideas and perform the experiments I wanted. His excitement to learn about my latest findings kept me motivated during my DPhil; I could not have asked for a better supervisor for my DPhil.

I am grateful to Dr. Tatyana Nesterova for taking me under her wing. Tatyana shared her wealth of knowledge with me, trained me in a wide range of wet-lab techniques from basic tissue culture to next generation sequencing library preparation and equipped me with the confidence and capability to eventually become independent in my project. During my DPhil, Tatyana has been a great support and mentor both inside and outside the lab.

A big thanks to everyone in the Brockdorff lab for being so warm and inviting especially when I first came to the UK and felt out of place. I have learnt and benefited from everyone in one form or another. Thanks to Dr. Heather Coker who has been a huge inspiration, and moral support and has also helped me in the reviewing of this thesis. I have also greatly benefitted from my discussions with Dr. Guifeng Wei and his bioinformatics expertise. Though not directly involved in the project, discussing ideas with him has helped me in the data analysis for this work. I have also enjoyed informal hallway chats (and during COVID, zoom chats) with Dr. Joseph Bowness and Dr.Gabriele Mogorini, which served as a good medium to learn from each other and solve problems.

I am grateful to a past member of the lab, Dr. Michal Gdula, who has provided me the basic training in next generation sequencing data analysis. Although he left shortly after I joined the lab, we kept in touch via emails, and he provided me with useful input on my data analysis. I would also like to thank Dr. Tianyi Zhang, another former member of the lab, for always being available and excited to hear about my research progress and share her insights and protocols.

I would like to thank Professor Rob Klose, who hosted me for my first lab rotation, and served as an examiner in my transfer- and confirmation-of-status exams. His feedback on my project has always been invaluable. Dr. Emilia Dimitrova, Dr. Neil Blackledge in the Klose lab, and Dr. Angelika Feldmann,

a past member of the Klose lab have also provided me with guidance in experiments and bioinformatics. I would also like to extend my gratitude to the people working on the third floor of Oxford Biochemistry for providing a wonderful work environment. During my DPhil, I have taken the advantage of their broad experimental skills and expertise. I am particularly grateful to Dr.Krzysztof Kus and Dr. James Collier for helping me with sucrose gradient fractionation described in chapter 4 and Dr.Gabriele Marciano who was thrilled to train me in protein expression and purification. I am obliged to Amanda Williams at the Zoology Nextseq facility and Marjorie Fournier at the Oxford Biochemistry mass spectrometry facility for processing my samples and Ashvina Sergaran, Kriaucionis lab, for performing LC/MS analysis on my samples.

I would like to express my gratitude to my friends and family whose support has been indispensable during my DPhil. I would like to acknowledge the unwavering support and encouragement of my parents, Muhammad Akram Malik and Bushra Rabbani, who always picked me up when times are hard. I have inherited my enthusiasm for science from my father, who has always been eager to learn about my research every time we saw each other in person and mostly, virtually. I am grateful to my husband Usman Malik, who moved to the UK during the last year of my DPhil. He has been very supportive of my career, and was extremely patient and kind during this critical time. His ability to always make time for family while being fully productive at work is impressive. I am thankful to my brother Arsalan Akram Malik and his wife Haya Fatimah for sharing their PhD experiences with me and always being a source of wisdom and motivation. My time in Oxford was made much more enjoyable because of my friends, Hadiqa Khan, Dr. Hannah Rana, Fatima Zahra, Hira Javaid, Tayyaba Komal and Faseeha Ayaz who started their DPhil journey with me, Dr. Feryal Amir, and Anna Lastuvka, my office neighbour. They were always available to lend an ear, meet for coffee and walks, and feature some of my fondest memories in Oxford of celebrating Ramadan and Eid together. Their passion for science has always been inspiring. I have never experienced a stronger alliance and friendship, which will be carried beyond my years in Oxford.

Finally, I am thankful to my examiners, Dr. Gavin Kelsey, and Professor Skirmantas Kriaucionis for taking the time to read and examine this thesis.

Contents

List of Figures X

List of Tables XIV

List of Abbreviations XV

Chapter 1 Introduction 1

1.1 Chromatin and epigenetics 1

1.1.1 Chromatin structure 4

1.1.2 Histone tail PTMs define types of chromatin 4

1.1.2.1 Histone PTMs associated with active euchromatin 5

1.1.2.2 Histone PTMs associated with repressed heterochromatin 6

1.1.2.3 Polycomb and Trithorax group - exemplary epigenetic regulation system of developmentally regulated loci 7

1.1.3 Chromatin remodelling complexes and histone variants in chromatin structure. 10

1.2 DNA methylation 10

1.2.1 DNA methylation - a prevalent epigenetic modification 10

1.2.2 A brief history of the discovery of DNA methylation 11

1.2.3 Enzymes catalysing DNA methylation 13

1.2.3.1 The maintenance DNA methyltransferase DNMT1 15

1.2.3.2 *de novo* methyltransferases DNMT3A and DNMT3B 16

1.2.4 Enzymes catalysing demethylation of 5mC 19

1.2.5 The enzymology of DNA methylation in plants 21

1.2.6 Methyl-CpG binding proteins 21

1.2.7 Molecular techniques to study CpG methylation 23

1.2.8 The function of CpG methylation 26

1.2.9 The dynamics of CpG methylation during mammalian development 28

1.2.10 Functional importance of developmentally regulated DNA methylation 30

1.2.11 CpG islands (CGIs) 31

1.2.12 Methylation of CGIs 32

1.3 X Chromosome Inactivation (XCI) 34

1.3.1 Initiation of XCI 36
1.3.2 Establishment of XCI 39
1.3.2.1 Xist repeat elements and interacting proteins 39
1.3.2.2 Xist spreading 42
1.3.2.3 Gene silencing during early XCI 42
1.3.3 Maintenance of XCI 44
1.3.3.1 Recruitment of SMCHD1 44
1.3.3.2 Establishment of H3K9me3 46
1.3.3.3 CGI methylation 47
1.4 Specificity among *de novo* DNMTs 50
1.4.1 Current understanding of target specificity for DNMT3A and DNMT3B 50
1.4.2 Significance of understanding specificity among *de novo* methyltransferases 51
1.5 Aims of this thesis 52

Chapter 2 Materials and methods 55

2.1 Experimental procedures 55
2.1.1 Cell culture conditions 55
2.1.2 Cloning of TIGRE homology vector for CRISPR/Cas9 mediated targeting of *Dnmt3b* rescue constructs 55
2.1.3 Cloning of small guide RNAs (sgRNAs) for CRISPR/Cas9 targeting 57
2.1.4 Generation of cell lines 58
2.1.5 Genomic DNA Isolation from ES cells (screening of 96-well plates) 59
2.1.6 Total Protein extraction 59
2.1.7 Western blot 59
2.1.8 Immunofluorescence (IF) 60
2.1.9 Embryoid body (EB) differentiation 61
2.1.10 Neural progenitor cell (NPC) differentiation 62
2.1.11 Genomic DNA isolation for BioCAP and MBD 62
2.1.12 BioCAP 63
2.1.13 Methyl binding domain affinity purification (MBD) 63
2.1.14 qPCR validation of MBD and BioCAP 64
2.1.15 Native Chromatin immunoprecipitation (nChIP) 65
2.1.16 NGS library preparation, validation, and quantification 66

2.1.17 Salt extractions to estimate the chromatin binding strength of DNMT3B and its mutants 67
2.1.18 Sucrose gradient fractionation of chromatin-bound proteins 68
2.1.19 Chromatin extraction under high salt 69
2.1.20 Immunoprecipitation of T7-DNMT3A and T7-DNMT3B from nuclear lysates 69
2.1.21 Chromatin RNA extraction and library preparation for sequencing (chrRNA-seq) 70
2.1.22 Nuclear extract preparation 71
2.1.23 RNA extraction and cDNA preparation for quantitative PCR analysis 71
2.1.24 XX karyotyping for cell lines using immunofluorescence (IF) and allele specific PCR 72
2.2 Data analysis softwares and packages 74
2.2.1 Mapping of fastq files for BioCAP-, ChIP- and MBD-seq and assigning reads to specific alleles 74
2.2.2 Counting reads in regions of interest 75
2.2.3 Normalization of MBD- and ChIP-seq data using spiked-in human DNA 76
2.2.4 Peak-calling using MBD-seq data 76
2.2.5 Finding differentially methylated regions between two conditions 77
2.2.6 Finding genes associated with CGIs 77
2.2.7 Mapping chrRNA-seq fastq files 78
2.2.8 Differential expression analysis 78
2.2.9 Generation of heatmaps and profile plots 78
2.2.10 Defining statistically significant DNMT3A- or DNMT3B-associated proteins from mass spectrometry data after immunoprecipitation 79
2.2.11 Softwares and tools used in routine for data analysis 81
2.2.12 Publicly available datasets 82
2.2.13 Datasets from other lab members analysed 83

Chapter 3 Establishment of CGI methylation during differentiation 84
3.1 Introduction 84
3.2 Tet-On Xist mouse ES model cell line (iXist-ChrX mouse ES cells) 86
3.3 BioCAP-seq - a method to study CGI methylation dynamics on Xi 87
3.4 BioCAP-seq on iXist-ChrX line to study CGI methylation dynamics during EB differentiation timecourse 92
3.5 Testing the sensitivity of BioCAP-seq using MEFs 96
3.6 Analysis of CGI methylation dynamics on Xi during EB differentiation. 100

3.7 Spike-in MBD-seq gives quantitative information on the methylation status of X-linked CGIs 106
3.8 CGI methylation on Autosomes 110
3.9 Gain of methylation at non-CGI CpG-rich elements on the Xi 112
3.10 Impact of *Dnmt3b* deletion on XCI-linked CGI methylation in iXist-ChrX cells 116
3.11 Loss of *Dnmt3b* does not affect XCI or gene expression in general 122
3.12 Role of SMCHD1 in CGI methylation during XCI 124
3.13 Summary and conclusions 126

Chapter 4 Investigating DNMT3B specificity mechanisms 131
4.1 Introduction 131
4.1.1 Isoform model of DNMT3B specificity for CGI methylation on Xi 132
4.1.2 Specificity-imparting feature model of DNMT3B 134
4.2 A complementation strategy to study the role of DNMT3B isoforms in Xi CGI methylation. 137
4.3 The catalytically active isoform DNMT3B1 fully rescues CGI methylation on Xi 138
4.4 The catalytically inactive isoform DNMT3B3 is not required for CGI methylation on Xi. 142
4.5 DNMT3B1 N-terminus is required to complement Xi CGI methylation in *Dnmt3b*$^{-/-}$ cells 146
4.6 Further analysis of the N-terminus of DNMT3B1 to determine the regions required for CGI methylation rescue 152
4.7 CGI methylation in complementation lines expressing DNMT3A with N-terminal domain of DNMT3B1 155
4.8 Role of the N-terminal domain in chromatin binding of DNMT3B1 158
4.9 ChIP-seq analysis of DNMT3B1 after deletion of the N-terminal domain 164
4.10 A point mutation in the DNMT3B1 PWWP domain found in ICF results in complete loss of CGI methylation on the X chromosome 167
4.11 Differences in pericentric localization of DNMT3B1 for Δ1-218 DNMT3B1 and S277P DNMT3B1 mutations 170
4.12 Summary and conclusions 172

Chapter 5 Investigating the role of H3K9me3 and protein interactors in influencing DNMT3A and DNMT3B specificity.....176

5.1 Introduction..... 176

5.1.1 SMCHD1-linked H3K9me3 accumulation. 176

5.1.2 DNMT3B-specific interactors..... 177

5.2 H3K9me3 is enriched on CGIs..... 178

5.3 H3K9me3 correlates with DNMT3B localization on the Xi chromosome 180

5.4 H3K9me3 accumulation at CGI is upstream of their DNA methylation but shows partial loss in *Dnmt3b*$^{-/-}$ cells 183

5.5 H3K9me3 is not enriched on autosomal CGIs dependent upon DNMT3B for methylation .. 185

5.6 A strategy to determine the DNMT3A and DNMT3B interactome in ES cells and EBs..... 186

5.7 DNMT3A and DNMT3B associated proteins in ES cells..... 189

5.8 DNMT3A and DNMT3B associated proteins in EBs 193

5.9 PTMs on DNMT3A and DNMT3B..... 195

5.10 Summary and conclusions..... 197

Chapter 6 Discussion 200

6.1 Differentiation of iXist-ChrX ES cells coupled to calibrated MBD-seq - a versatile model system to study CGI methylation on Xi 200

6.2 Characterization of DNMT3B-dependent CGIs on Xi and genomewide.....203

6.3 Expansion of the current definition of CGIs.....203

6.4 Methylation of “CGI-like elements” on Xi is DNMT3B-independent.....206

6.5 DNMT3B isoforms and specificity.....208

6.6 A role for the DNMT3B N-terminal domain in CGI methylation.....209

6.7 Contribution of N-terminal domain to DNMT3B nuclear localization.....214

6.8 Analysis of the impact of DNMT3B1 PWWP point mutant reported in ICF.....215

6.9 Contribution of H3K9me3 in CGI methylation.....216

6.10 Crosstalk between CGI methylation and H3K9me3 accumulation on Xi.....220

6.11 Proteomic analysis for discovery of DNMT3A and DNMT3B interactome dynamic during differentiation.....221

Bibliography 225

Appendix 255

Appendix 1 255

XX Karyotyping of cell lines 255

Appendix 2 259

Correlation between BioCAP-, MBD-, and ChIP-seq replicates 259

Appendix 3 262

Catalytic assays for DNMT3B rescue constructs 262

Appendix 4 265

Differential methylation analysis of complementation lines expressing DNMT3B1 with smaller deletions in the N-terminus 265

Appendix 5 266

Key information regarding CGI methylation on Xi 266

Appendix 6 271

NGS datasets generated in this work 271

List of Figures



Chapter 1

Figure 1.1: Schematic showing the elements/components of eukaryotic gene 2

Figure 1.2: Structure of nucleosome and chromatin. 3

Figure 1.3: Histone tails are covalently modified forming a complex histone code. 5

Figure 1.4: Reader-writer coupling for establishment and maintenance of H3K9me3 6

Figure 1.5: Polycomb repressive complex components and recruitment mechanisms 9

Figure 1.6: CpG methylation reaction and catalysing enzymes 14

Figure 1.7: UHRF1-mediated recruitment of DNMT1 to replication fork 16

Figure 1.8: DNMT3A and DNMT3L form heterotetramers 17

Figure 1.9: Autoinhibition of DNMT3A and DNMT3B with ADD domain. 18

Figure 1.10: Demethylation of 5mC by TET enzymes 20

Figure 1.11: Dynamics of CpG methylation during mouse development 29

Figure 1.12: Methylation of ICR controls allele-specific expression of imprinted *Igf2/H19* locus. 34

Figure 1.13: Phases of XCI 36

Figure 1.14: Schematic showing X inactivation centre (Xic) in mouse 39

Figure 1.15: Schematic showing Xist repeat elements and RNA binding proteins 41

Figure 1.16: The diverse functions of SMCHD1 in XCI 46

Figure 1.17: Key findings regarding CGI methylation on X. 49

Figure 1.18: Unanswered questions in the establishment of CGI methylation during XCI 54

Chapter 2

Figure 2.1: XX karyotyping of cell lines using allele specific PCR and IF 74

Figure 2.2: Steps of mass spectrometry data analysis to identify statistically significant proteins associated with DNMT3A or DNMT3B after IP from nuclear lysates 81

Chapter 3

Figure 3.1: The iXist-ChrX model cell line. 87

Figure 3.2: BioCAP-seq for studying methylation of CGIs 89

Figure 3.3: Optimization of BioCAP-seq for studying CGIs on Xi. 92

Figure 3.4: BioCAP-seq on EB timecourse in the iXist-ChrX cell line. 94
Figure 3.5: Methylation dynamics of NMIs during EB differentiation. 96
Figure 3.6: Comparison between BioCAP-seq and MBD-seq for studying methylated CGIs..... 100
Figure 3.7: Methylation dynamics of CGIs during EB timecourse of the iXist-ChrX line 103
Figure 3.8: CGI methylation dynamics during NPC differentiation of the iXist-ChrX cell line 105
Figure 3.9: Calibrated MBD-seq reveals the level of DNA methylation gained by CGIs on the Xi. 109
Figure 3.10: Differentiation-coupled methylation at autosomal CGIs. 111
Figure 3.11: Characterization of the "Non-CGI signal" in MBD-seq 113
Figure 3.12: Characterization of the novel "CGI-like elements" on the X chromosome 116
Figure 3.13: CGI methylation in the *Dnmt3b*$^{-/-}$ cell line 119
Figure 3.14: "CGI-like elements" gain methylation independent of DNMT3B. 120
Figure 3.15: CGI methylation at autosomal CGIs in *Dnmt3b*$^{-/-}$ cells 122
Figure 3.16: Interplay between CGI methylation and transcriptional silencing on the Xi 124
Figure 3.17: Role of SMCHD1 in CGI methylation. 125
Figure 3.18: "CGI-like elements" are associated with low-level transcripts. 128

Chapter 4

Figure 4.1: The *Dnmt3b* gene codes for multiple isoforms. 133
Figure 4.2: N-terminal domain of DNMT3A and DNMT3B represents the most diverged regions between the two homologues 136
Figure 4.3: Complementation assay to study the contribution of DNMT3B isoforms and /or the protein features that confer specificity for Xi CGI methylation. 138
Figure 4.4: Complementation of the *Dnmt3b*$^{-/-}$ line with *Dnmt3b1* 141
Figure 4.5: Complementation of *Dnmt3b*$^{-/-}$ line with *Dnmt3b3* (*Dnmt3b3-KI*) 144
Figure 4.6: Testing "accessory" functions of DNMT3B in *Dnmt3b3-KI* complementation line. 145
Figure 4.7: Complementation of *Dnmt3b*$^{-/-}$ line with *Δ1-229 Dnmt3b1*. 149
Figure 4.8: Complementation of *Dnmt3b*$^{-/-}$ line with *Δ1-218 Dnmt3b1* (*Δ1-218 Dnmt3b1-KI*) 152
Figure 4.9: Complementation of *Dnmt3b*$^{-/-}$ line with *Dnmt3b1* bearing small deletions in the N-terminus 155
Figure 4.10: Complementation of *Dnmt3b*$^{-/-}$ cells with chimeric DNMT3 proteins 158
Figure 4.11: Salt extractions to test the chromatin binding strength of DNMT3B1 with N-terminal domain deletion 159
Figure 4.12: High salt chromatin extraction and sucrose gradient fractionation of DNMT3B1 to test DNMT3B-chromatin-binding strength 163

Figure 4.13: DNMT3B1 ChIP reveals differential localization of Δ1-218 DNMT3B1 167
Figure 4.14: MBD-seq on PWWP point mutant *S277P Dnmt3b1* complementation line..... 169
Figure 4.15: Pericentric localization of DNMT3B1, Δ1-218 DNMT3B1 and S277P DNMT3B1..... 172

Chapter 5

Figure 5.1: H3K9me3 accumulation on Xi..... 180
Figure 5.2: H3K9me3 enrichment correlates with DNMT3B localization on the X chromosome 181
Figure 5.2: H3K9me3 enrichment correlates with DNMT3B localization on the X chromosome 182
Figure 5.3: H3K9me3 is more enriched at CGIs in comparison to DNMT3B-independent "CGI-like elements" on the Xi in EBs..... 183
Figure 5.4: Enrichment of H3K9me3 on the Xi is lower in *Dnmt3b*$^{-/-}$ cells as compared to wt..... 184
Figure 5.5: H3K9me3 enrichment on autosomal CGIs. 186
Figure 5.6: Schematic showing the rationale behind performing IP experiments to identify DNMT3A and DNMT3B interactors in undifferentiated and differentiated cells. 187
Figure 5.7: Validation and analysis steps for IP from nuclear lysates 189
Figure 5.8: DNMT3A and DNMT3B interactors..... 191
Figure 5.9: DNMT3A- and DNMT3B-associated proteins immunoprecipitated from nuclear lysates in ES cells 193
Figure 5.10: DNMT3A- and DNMT3B-associated proteins in EBs. 195
Figure 5.11: PTMs on DNMT3A2. 197

Chapter 6

Figure 6.1: The proportion of promoter-associated and inter/intragenic CGIs among those gaining methylation during EB differentiation..... 202
Figure 6.2: Discovery of additional CGIs by MBD-seq. 205
Figure 6.3: "CGI-like elements" gain methylation on Xi independent of DNMT3B..... 208
Figure 6.4: Suggested models for the role of the N-terminus in DNMT3A and DNMT3B. 213
Figure 6.5: Models for maintenance of a hypomethylated state of CGIs in ES cells..... 217
Figure 6.6: TET1-mediated hypomethylation on CGIs in ES cells and proposed model for CGI methylation in EBs..... 222
Figure 6.7: Model for DNMT3B mediated methylation of X-linked CGIs based on findings in this study and preliminary data. 224

Appendix

Figure A.1: Allelic ratio calculation for wt iXist-ChrX (A) and *Dnmt3b*$^{-/-}$ (B) ES and EBs used for MBD-seq. 255
Figure A.2: XX karyotyping for KI lines..... 256
Figure A.3: Allelic ratio calculation for estimating XX karyotype of three clones each of complementary lines. 257
Figure A.4: Estimation of XX-karyotype of wt iXist-ChrX and *Dnmt3b*$^{-/-}$ lines..... 258
Figure A.5: Correlation plots showing reproducibility among replicates.. 259
Figure A.6: LC/MS based assays to test the catalytic activity of DNMT3B rescue constructs. 263
Figure A.7: ELISA based assays to test the catalytic activity of DNMT3B rescue constructs. 264
Figure A.8: Differentially methylated regions found in complementation lines expressing *Dnmt3b* with smaller deletions in the N-terminal domain at day 12 EBs..... 265

List of Tables

Table 1.1: Techniques widely used for studying CpG methylation. 26
Table 2.1: Gibson cloning oligos used to make homology vectors 57
Table 2.2: CRISPR-Cas9 sgRNAs and reverse complement oligos 58
Table 2.3: Antibodies used in this study. 60
Table 2.4: Program used for qPCR validation/estimation of IP efficiency of MBD and BioCAP. 64
Table 2.5: Primers used for qPCR validation/estimation of IP efficiency of MBD and BioCAP. 65
Table 2.6: Program used for quantification of NGS libraries using qPCR 67
Table 2.7: Primers used for quantification of NGS libraries using qPCR 67
Table 2.8: Primers used for estimating gene expression levels. 72
Table 2.9: Primers used for XX karyotyping of the cell lines. 73
Table 2.10: Softwares and packages used in routine NGS data analysis. 82
Table 2.11: Published datasets used in this work. 82
Table 2.12: Unpublished datasets generated by other lab members used in this work. 83
Table A1: Key information, regarding CGIs on Xi, their methylation dynamics, associated genes and dependence on DNMT3B and SMCHD1. 270
Table A2: Summary for the NGS datasets generated in this work. 275

List of Abbreviations

5-Azac	5-Azacytidine
5cac	5 Carboxy Cytosine
5fc	5 Formyl Cytosine
5hmc	5 Hydorxymethyl Cytosine
5mC	C5-Methyl Cytosine
aa	Amino acids
ADD	ATRX-DNMT3-DNMT3L
AGO4/6	Argonaute Proteins 4 Or 6
AP	Affinity Purification
AR	Allelic Ratio, Xi/ (Xi + Xa)
ATAC-Seq	Assay for Transposase-Accessible Chromatin Using Sequencing
ATP	Adenosine Triphosphate
ATRX	Alpha Thalassemia/Mental Retardation, X Linked Syndrome
BER	Base Excision Repair
BioCAP-Seq	Biotinylated-CAP (CXXC Affinity Purification) Sequencing
bp	Base pair
BSA	Bovine Serum Albumin
CAGE-Seq	Cap Analysis Gene Expression Sequencing
CAP-Seq	CXXC Affinity Purification Sequencing
Cas9	CRISPR-Associated Protein 9
CBX	Chromobox
CFP1	CXXC Finger Protein 1
CGI	CpG Island
ChIP	Chromatin Immunoprecipitation
ChrRNA-seq	Chromatin RNA Sequencing
CIZ1	Cip1-Interacting Zinc Finger Protein
CMT3	Chromomethyltransferase 3
COBRA	Combined Bisulphite Restriction Analysis

CpG	5’-C-Phosphate-G-3’ Dinucleotide
CRE	Cis-Regulatory Element
CRISPR `	Clustered Regularly Interspaced Short Palindromic Repeats
CTCF	CCCTC-binding Factor
Ctrl	Control
CXXC	(C represents cysteine and X is any amino acid) domain
DAPI	4’,6-diamidino-2-phenylindole
DCL3	Dicer-Like Protein 3
DHU	Dihydrouracil
DMSO	Dimethyl Sulfoxide
DNA	Deoxyribonucleic Acid
DNase	Deoxyribonuclease
DNM3L	DNA Methyltransferase 3-like
DNMT1	DNA Methyltransferase 1
DNMT3A/3B	DNA Methyltransferase 3A or 3B
dNTP	Deoxyribonucleotide Tri Phosphate
Dox	Doxycycline
DPPA3	Developmental Pluripotency Associated 3
DRM2	Domains Rearranged Methyltransferase 2
DSP	Di-Thiobis-Succinimidyl-Propionate
DTT	Dithiothreitol
EB/EBs	Embryonic Body/ Embryonic Bodies
EDTA	Ethylenediamine Tetra acetic Acid
EED	Embryonic Ectoderm Development
EGF	Epidermal Growth Factor
EMSA	Electrophoretic Mobility Shift Assay
ENCODE	Encyclopedia of DNA Element
Enox	Expressed Neighbour of Xist
EpiSC	Epiblast Stem Cells
ES cells	Embryonic Stem Cells
EZH1/2	Enhancer of Zeste Homolog 1/2

FCS	Foetal Calf Serum
FDR	False Discovery Rate FGF
FGF	Fibroblast Growth Factor
FISH	Fluorescent *In Situ* Hybridization
Ftx	Five Prime to *Xist*
GEO	Gene Expression Omnibus
H2AK119ub1	Monoubiquitination of Lysine 119 of Histone H2A
H3K27ac	Acetylation of Lysine 27 of Histone H3
H3K27me3	Tri-methylation of Lysine 27 of Histone H3
H3K36me3	Trimethylation of Lysine 36 of Histone H3
H3K4me1	Monomethylation of Lysine 4 of Histone H3
H3K4me3	Trimethylation of Lysine 4 of Histone H3
H3K9ac	Acetylation of Lysine 9 of Histone H3
H3K9me2/3	Di/Tri-methylation of Lysine 9 of Histone H3
H4K20me1	Monomethylation of Lysine 20 of Histone H4
HDAC	Hisone Deacetylase Complex
HNRNP	Heterogeneous Nuclear Ribonucleoprotein Complex Protein
Hox	Homeotic
HP1	Heterochromatin Protein 1
HUSH	Human Silencing Hub Complex
IAP	Intracisternal A Particle
ICF	Immunodeficiency, Centromeric instability and Facial anomalies
ICM	Inner Cell Mass
ICR	Imprinting Control Region
IF	Immunofluorescence
IP	Immunoprecipitation
iXist-ChrX	Inducible Xist on Chromosome X
JARID2	Jumonji, AT Rich Interactive Domain 2
Jpx	just proximal to Xist
KI	Knock-in
KO	Knockout

LBR	Lamin B Receptor
LBS	Lamin B Receptor Binding Site
LIF	Leukemia Inhibitory Factor
LINE	Long Interspersed Nuclear Element
Linx	Large Intervening Transcript in the Xic
lncRNA	Long non-coding RNA
LTR	Long Terminal Repeat
m6A	N6-Methyladenosine
MAP-seq	MBD Affinity Purification sequencing
MBD	Methyl-CpG Binding Domain
MBD-seq	Methyl-CpG Binding Domain Sequencing
MCAF1	MBD1-Containing Chromatin-Associated Factor 1
MECP2	Methyl CpG Binding Protein 2
MeDIP	Methylated DNA Immunoprecipitation
MEF	Mouse Embryonic Fibroblast
MET1	Methyltransferase 1
MNase	Micrococcal Nuclease
MRSE	Methylation-Sensitive Restriction Endonucleases
MSP	Methylation-specific PCR
MTase	Methyltransferase
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
NMIs	Non-Methylated Islands
NPC	Neural Progenitor Cell
NuRD	Nucleosome Remodelling and Deacetylase
OCT4	Octamer-binding Transcription Factor 4
PBS	Phosphate Buffered Saline
PCA	Principle Component Analysis
PcG	Polycomb Group
PCGF	Polycomb Group RING Finger protein
PCNA	Proliferating Cell Nuclear Antigen

PCR	Polymerase Chain Reaction
PHD	Plant Homeodomain
PGCs	Primordial Germ Cells
PIC	Protease Inhibitor Cocktail
piRNA	PIWI-interacting RNA
PIWI	P-element Induced WImpy testis
PRC1/2	Polycomb Repressive Complex 1 or 2
PRE	Polycomb Response Elements
PTMs	Post Translational Modifications
PWWP	Pro-Trp-Trp-Pro
RdDM	RNA-Dependent DNA Methylation
RDR2	RNA-dependent RNA polymerase
RFTS	Replication Fork Targeting Sequence
RIME	Rapid Immunoprecipitation Mass Spectrometry of Endogenous Proteins
RISC	RNA-Induced Silencing Complex
RNA	Ribonucleic Acid
RNA Pol II	RNA Polymerase II
RNAi	RNA Interference
RNA-seq	RNA-sequencing
RNF12	Ring Finger protein 12
RNP	Ribonucleoprotein
RRBS	Reduced Representation Bisulphite Sequencing
RRM	RNA Recognition Motif
rRNA	Ribosomal RNA
rt-TA	Tet Trans-Activator
SAF-A	Scaffold Attachment Factor A
SAM	S-Adenosyl Methionine
SET	Su(var)3-9, Enhancer-of-zeste and Trithorax
SETDB1	SET Domain Bifurcated Histone Lysine Methyltransferase 1
sgRNA	Single Guide RNA
SMC	Structural Maintenance of Chromosomes

SMCHD1	Structural Maintenance of Chromosomes Hinge Domain containing protein 1
SMRT	Silencing Mediator for Retinoid or Thyroid-hormone receptors
SNP	Single Nucleotide Polymorphism
SPOC	Spen Paralog and Ortholog C-terminal
SPEN	Split End
SRA	Set and Ring associated domain
SUV	Suppressor of Variegation
SUZ12	Suppressor of Zeste 12
SWI/SNF	Switch/Sucrose Non-Fermentable
TAD	Topologically Associating Domain
TAP-tag	Tandem Affinity Purification Tag
TDG	Thymine DNA Glycosylase
TET	Ten-Eleven Translocation
TRD	Target Recognition Domain
TRE	Trithorax Response Elements
TrxG	Trithorax Group
TSA	Trichostatin A
TSS	Transcription Start Site
Tsx	Testis-Specific X-Linked
TTD	Tandem Tudor Domain
TTS	Transcription Termination Site
UCSC	University of California, Santa Cruz
UHRF1	Ubiquitin Like With PHD And Ring Finger Domains 1
WGBS	Whole Genome Bisulphite Sequencing
Wt	Wildtype
Xa	Active X Chromosome
XCI	X Chromosome Inactivation
Xert	Xist-Enhancing Regulatory Transcript
Xi	Inactive X Chromosome
Xic	X-Inactivation Centre
Xist	X-Inactive Specific Transcript

Xite	X-Inactivation Intergenic Transcription Elements

Chapter 1

Introduction

1.1 Chromatin and epigenetics

The genetic information of living organisms is stored in deoxyribonucleic acid (DNA). DNA is a polymer of four deoxyribonucleotides, the sequence of which forms a code, carrying instructions to make up the organism. DNA is stable and simple to replicate which explains its choice by nature for carrying genetic information despite its apparent simplicity (Watson and Crick, 1953). The genetic information stored in DNA is "read" by cellular machinery. Francis Crick proposed the central dogma of molecular biology suggesting that the DNA sequence is first transcribed into a complementary single-stranded ribonucleic acid (RNA) molecule, a chemically similar polynucleotide, and then translated to proteins which are the main building blocks and regulators of biological processes (Crick, 1970). It is now widely agreed that this central dogma is a simplified concept, as, for instance, many RNAs do not code proteins and act as functional molecules themselves. Alongside genes, the sequence of DNA that codes for functional molecules, a significant proportion of DNA forms gene regulatory sequences, such as promoters and enhancers, that control gene expression quantitatively and temporally. Promoters are present at the start of the genes. In eukaryotes, promoters serve as the site of assembly for proteins called transcription factors (TFs), which help in the initiation of transcription by the enzyme RNA polymerase II (RNA Pol II). Enhancers are usually present far from the transcription start site (TSS) and bind activator proteins. Enhancers and the bound activator proteins can physically interact with the promoters and TFs by "looping" of genomic DNA (Figure 1.1).

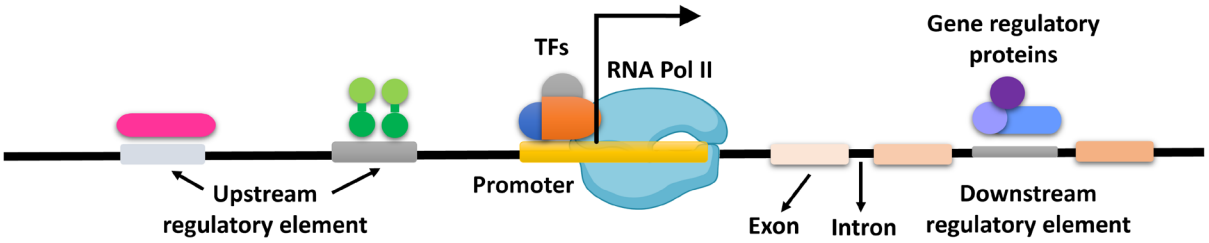


Figure 1.1: Schematic showing the elements/components of eukaryotic gene

Transcription is regulated via gene regulatory elements, which serve as binding sites for regulatory proteins: The promoter acts as a binding site for TFs that aid RNA Pol II to assemble and start transcription (TSS is indicated with an arrow). Enhancers bind activator proteins and promote gene expression, sometimes at a distance from the TSS, requiring chromatin looping to facilitate RNA Pol II initiation. Transcription leads to production of RNA which is processed to remove sequences called introns and retain only exonic sequence. Regulatory elements can be present both upstream and downstream of the TSS including within intronic regions.

The gene regulation mechanisms in living organisms have evolved from relatively simple, in the case of single-celled prokaryotes, to more complex systems in multicellular organisms. One reason for this increase in complexity is the presence of chromatin in which DNA is packaged around nucleosome particles made up by histone proteins. A nucleosome is typically composed of a 147 bp of DNA wrapped around an octamer of 4 types of histones, H2A, H2B, H3, and H4 (Luger et al., 1997) (Figure 1.2 A). Nucleosomes are separated by a short "linker DNA", bound by a fifth type of histone, H1 or H5. This packaging of DNA into chromatin can act as a barrier for DNA-templated processes such as transcription, by altering the access of the DNA to transcriptional machinery (Figure 1.2 B and C). Therefore, transcriptional regulation in eukaryotes is achieved through both, the modifications in chromatin structure (from "open" to "closed" state or vice versa), and genetic elements like promoters and enhancers.

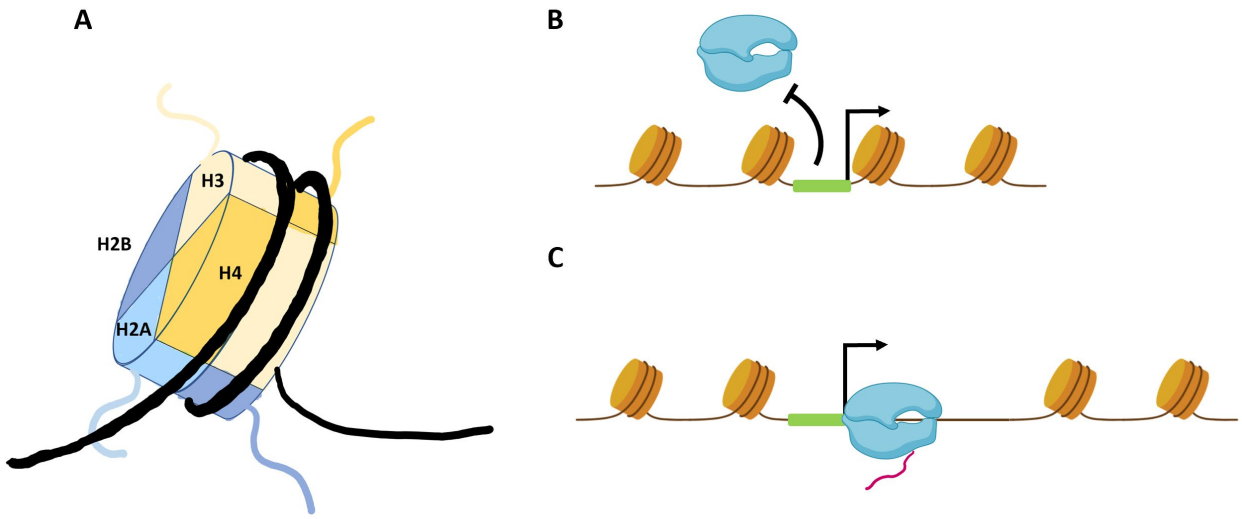


Figure 1.2: Structure of nucleosome and chromatin.

A) Schematic showing the structure of a nucleosome. A nucleosome is an octamer comprising four types of histone proteins; H2A, H2B, H3 and H4. The linker DNA can also be bound by a fifth type of histone, H1/H5. The N-terminal 'tails' of these core histone proteins are unstructured and protrude out of the nucleosome. These N-terminal regions are subject to diverse types of post translational modifications (see text for details).
B) The tight packaging of DNA into chromatin can form a barrier for the processes that require DNA as template such as transcription. Chromatin structure, including histone tails, must therefore be modified to allow access to transcriptional machinery. In the schematic, RNA Pol II is shown in blue, the promoter in green, the TSS as an arrow and newly transcribed nascent RNA in pink.
C) Loosely packed "open" chromatin allows access to transcriptional machinery.

Chromatin structure modification provides a reversible mechanism of transcriptional control, whereby stimuli such as environmental or developmental cues, can lead to change in degree of chromatin packaging and, in turn, gene expression. Once established, the chromatin modifications can be maintained in subsequent cell divisions and as a result, preserve the 'memory' of gene expression pattern. This form of transcriptional regulation is a key feature in the diversity of cell types in multicellular organisms achieved during development, when the organism starts its life from a single cell and, in the case of mammals, gives rise to 200-300 different cell types, all inheriting the same DNA, yet distinct in morphology and function. Thus, chromatin modifications form a layer of information regulating gene expression, in addition to regulatory sequences coded by the DNA, and constitute an

epigenetic (meaning above genetics) mechanism of gene regulation. The epigenetic modifications that shape chromatin include chemical modifications of DNA, post-translational modifications (PTMs) of histones, and replacement of histones with variant histones. These epigenetic modifications can alter the structure of chromatin directly, or by the help of proteins that bind them, referred to as readers.

1.1.1 Chromatin structure

Chromatin can be broadly divided into two categories: (1) open/transcriptionally permissive euchromatin and (2) condensed heterochromatin. The different extent of packaging of chromatin was first observed by Emil Heitz in 1928 (Heitz, 1928). He saw that chromosomal regions were differentially stained indicating differences in their condensed state. The regions that stained strongly, representing highly condensed chromatin, were termed heterochromatin, while the lightly stained less condensed chromatin was termed euchromatin. He later suggested that genes were found in euchromatin, while heterochromatin was mostly silenced and inert. Finally, he coined the term 'facultative heterochromatin' for the regions of chromosomes stained only in certain cells to distinguish it from the constitutively stained heterochromatin (therefore called constitutive heterochromatin).

More recent genome-wide studies using chromatin immunoprecipitation (ChIP) and high throughput sequencing techniques, have now characterized and defined heterochromatin and euchromatin on the basis of epigenetic modifications such as PTMs of histones and chemical modification of DNA.

1.1.2 Histone tail PTMs define types of chromatin

Each of the four histone proteins consist of a core globular domain and protruding unstructured histone tail. The tails of histones (and sometimes globular domains) undergo a remarkable variety of covalent modifications such as acetylation, phosphorylation, ubiquitination, mono-, di- and tri-methylation (Figure 1.3). Histone PTMs can exist alone or in combination, forming a complex "histone code" which defines functional genomic domains. Each of these PTMs is laid down by a specific enzyme(s) which often works on one or few selective sites on the histone tails. Similarly, each histone tail PTM is removed by a different enzyme and bound by different proteins that possess special chromatin reading domains allowing them to bind to a particular PTM on histone. The proteins involved in establishment, maintenance, and removal of histone PTMs are carefully regulated by interconnecting signalling pathways that respond to environmental and developmental cues and have important consequences on gene expression.

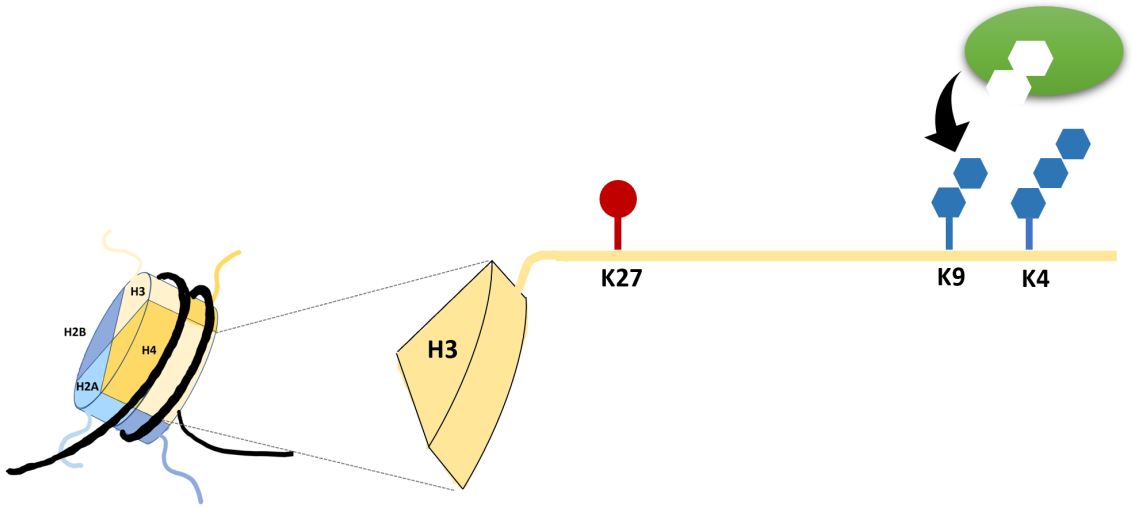


Figure 1.3: Histone tails are covalently modified forming a complex histone code.

The N-terminal tail of histones protrude outside the nucleosome core and can acquire covalent modifications. H3 is shown as an example in the figure. Single histone tail residues can be modified differently, for instance, lysine can be mono-, di- or tri methylated (a single methyl moiety is shown by blue hexagon) or acetylated (shown in red). Furthermore, multiple residues can be modified in combination, creating a complex histone code that plays a role in gene regulation. These modifications either have a direct effect on chromatin structure or they serve as binding sites for proteins that lead to alterations in chromatin structure. For example, negatively charged acetylation (red lollipop), is thought to neutralize the polar interactions between positively charged histone and negatively charged DNA, loosening the DNA wrapping around the nucleosome. Histone 3 and 4 lysine methylation on the other hand influence chromatin structure by binding proteins containing chromodomains (depicted in green).

1.1.2.1 Histone PTMs associated with active euchromatin

Euchromatin is relatively open or uncompacted. It is associated with actively transcribed regions, characterized by high accessibility and reduced nucleosome occupancy. Promoters of the actively transcribed genes are marked by lysine 4 tri-methylation of histone H3 (H3K4me3) and acetylation of histones H3 and H4. H3K4me3 has been implicated in facilitating transcription and shown to recruit transcription initiation factor TFIID on selected genes (Vermeulen et al., 2007). The gene bodies of actively transcribed genes are marked with H3K36me3 which is established co-transcriptionally by the recruitment of H3K36me3 methyltransferase SETD2 to elongating RNA Pol II (Bannister et al., 2005; Baubec et al., 2015; Kizer et al., 2005). Although H3K36me3 is not required for transcription it has been suggested to lead to silencing of cryptic transcription (Carrozza et al., 2005; Xie et al., 2011) and

regulation of splicing (Luco et al., 2010). Ubiquitination of H2B (H2BK120ub) also associates with actively transcribed regions and has been shown to direct activation of some genes (Luco et al., 2010). Enhancers of expressed genes are modified with mono methylation of H3K4 (H3K4me1) and acetylation of H3K27 (H3K27ac), (reviewed in (Millán-Zambrano et al., 2022)).

1.1.2.2 Histone PTMs associated with repressed heterochromatin

Heterochromatin is highly condensed, gene poor and transcriptionally repressed. Heterochromatic regions have been shown to be late replicating and associated with the nuclear lamina or the nucleolus. It is marked by histone modifications correlated with gene silencing, such as di- or tri-methylation of H3K9 (H3K9me2/3) and H3K27me3 (reviewed in (Allshire and Madhani, 2018)). **Constitutive heterochromatin** is marked by H3K9me2/3 and forms on the tandem 'satellite' repeats found in centromeres, the flanking pericentromeric, and telomeric regions of the chromosomes. Additionally, transposons, parasitic DNA elements which possess the ability to "transpose" (multiply by copying their sequence and inserting it at new genetic loci) are repressed at least in part by H3K9me2/3 (reviewed (Allshire and Madhani, 2018)). H3K9me3 is recognized by HP1 via its chromodomain. HP1 can, in turn, interact with H3K9me3 methyltransferase SUV39 and other HP1 molecules creating a positive feedback mechanism for propagating H3K9me3 (Figure 1.4), (Bannister et al., 2001; Lachner et al., 2001).

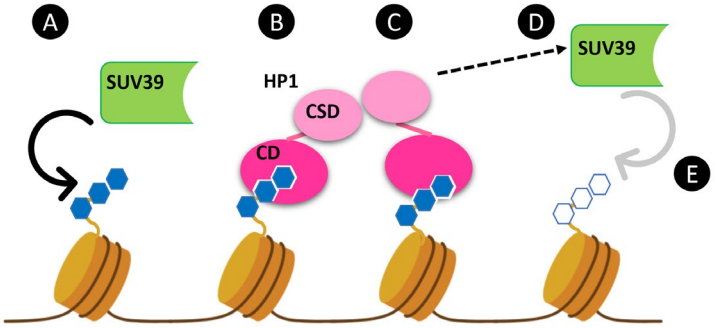


Figure 1.4: Reader-writer coupling for establishment and maintenance of H3K9me3.

(A) SUV39 catalyses methylation of H3K9 leading to H3K9me3 that is then read by the chromodomain (CD) of HP1 (B). HP1 has a chromoshadow domain (CSD) that allows it to interact with more HP1 molecules (C) or SUV39 (D) to further propagate the H3K9me3 modification on chromatin (E). H3K9me3 is displayed as three blue hexagons.

Such reader-writer interactions can allow heterochromatin to spread, a phenomenon first studied as position effect variegation in *Drosophila melanogaster*, where expression of a reporter gene varies due to its relative position near heterochromatin. This was first studied for the *Drosophila* eye colour gene *white* which when juxtaposed to heterochromatin gave red/white mosaic pattern of the facets of *Drosophila* eye. This *Drosophila* eye model was then used for genetic screens to identify the proteins required for establishment of heterochromatin, the mutations of which suppressed variegation, giving SUV39 (suppressor of variegation) its name (Elgin and Reuter, 2013).

Facultative heterochromatin is more dynamic/transient in comparison to constitutive heterochromatin. Facultative heterochromatin is associated with H3K27me3 and ubiquitination of H2A on lysine 119 (H2AK119ub1), established by Polycomb repressive complex 2 and 1 (PRC2 and PRC1), respectively (reviewed in (Millán-Zambrano et al., 2022). Facultative heterochromatin forms on genes that are silenced in response to stimuli, for instance, during development, allowing the genes to be turned off and on at various developmental stages. Examples of facultative heterochromatin include the inactive X chromosome in female mammals, developmental master regulator homeotic (*Hox*) genes and cell-type specific genes. Polycomb group of proteins together with their antagonists Trithorax group of proteins are responsible for creating the optimal balance for the accurate spatial and temporal expression of developmental regulatory loci and present an exemplary system of epigenetic regulation of development (reviewed in (Schuettengruber et al., 2017)).

1.1.2.3 Polycomb and Trithorax group - exemplary epigenetic regulation system of developmentally regulated loci

The Polycomb (*Pc*) gene was discovered by P. Lewis in 1947 as mutation that leads to the appearance of ectopic sex combs on the second and third pairs of legs in male *Drosophila* files, thus named Polycomb (Lewis, 1947). This phenotype is manifested due to failure of silencing of a *Hox* gene *Sex comb reduced* (*Scr*) (Lewis, 1978). Later E.B Lewis showed that loss-of-function mutations in *Pc* lead to the transformation of embryonic body segments of *Drosophila* into posterior most segment (Lewis, 1978). Since *Pc* was not present in *Hox* gene cluster, it was considered as a negative *trans* regulator of the *Hox* genes. Ingham in 1983 observed that mutations in *trithorax* (*trx*) caused the transformation

of embryonic body segments of flies into the anterior segment (Ingham, 1983). Since the combined mutations in *Pc* and *trx* were able to repress homeotic transformation phenotypes, *trx* was also described as a positive regulator of *Hox* genes and antagonist of Polycomb (reviewed in (Grimaud et al., 2006)). Similar genetic screens for mutations causing or reversing homeotic transformations unveiled more protein regulators of *Hox* genes. Many years of research have led to the knowledge that Polycomb and trithorax group proteins function as multi-subunit complexes which are conserved in organisms including plants, mammals and yeast.

Trithorax group (TrxG) is composed of a functionally diverse group of proteins including the ATPase dependent chromatin remodelling SWI/SNF complex which allows access to nucleosomal DNA, and SET domain-containing H3K4 mono-, di- and tri-methyltransferases catalysing the histone modifications hallmark of active chromatin (such as *Drosophila trx* or its mammalian orthologues MLL1-4) (reviewed in (Schuettengruber et al., 2017)). The Polycomb group (PcG) proteins also contribute to several complexes, the best characterized of which are Polycomb repressive complex 1 and 2 (PRC1 and PRC2), comprising several core and accessory subunits (Figure 1.5 A). PRC2-mediated H3K27me3 can lead to recruitment of PRC1 via the chromodomain containing subunit CBX. Additionally, variant PRC1 complexes catalyse a different histone modification, H2AK119ub1 which in turn is recognized by the PRC2 subunit JARID2 (Blackledge et al., 2014; Cooper et al., 2016). The enzymatic activity of PRC2 is stimulated by recognition of its own deposited H3K27me3 suggesting a self-perpetuating feed forward mechanism for its own propagation (Figure 1.5 B) (Jiao and Liu, 2015; Margueron et al., 2009). The canonical PRC1 complexes which recognize H3K27me3 form long range chromatin interactions forming suppressed Polycomb domains (reviewed in (Blackledge and Klose, 2021)).

How TrxG- and PcG-mediated histone modifications alter gene expression and the cause-and-effect relationship between the two is not well understood. H3K4me3 has been implicated in the recruitment of chromatin remodelling complexes (Schnetz et al., 2009; Wysocka et al., 2006), histone acetyltransferase and basic transcriptional machinery (Tang et al., 2013; Vermeulen et al., 2007). H2AK119ub1 is thought to block transcription due to the presence of inhibitory bulky ubiquitin moiety or as recently shown, to reduce the frequency of transcription initiation events by RNA Pol II (Dobrinić et al., 2021).

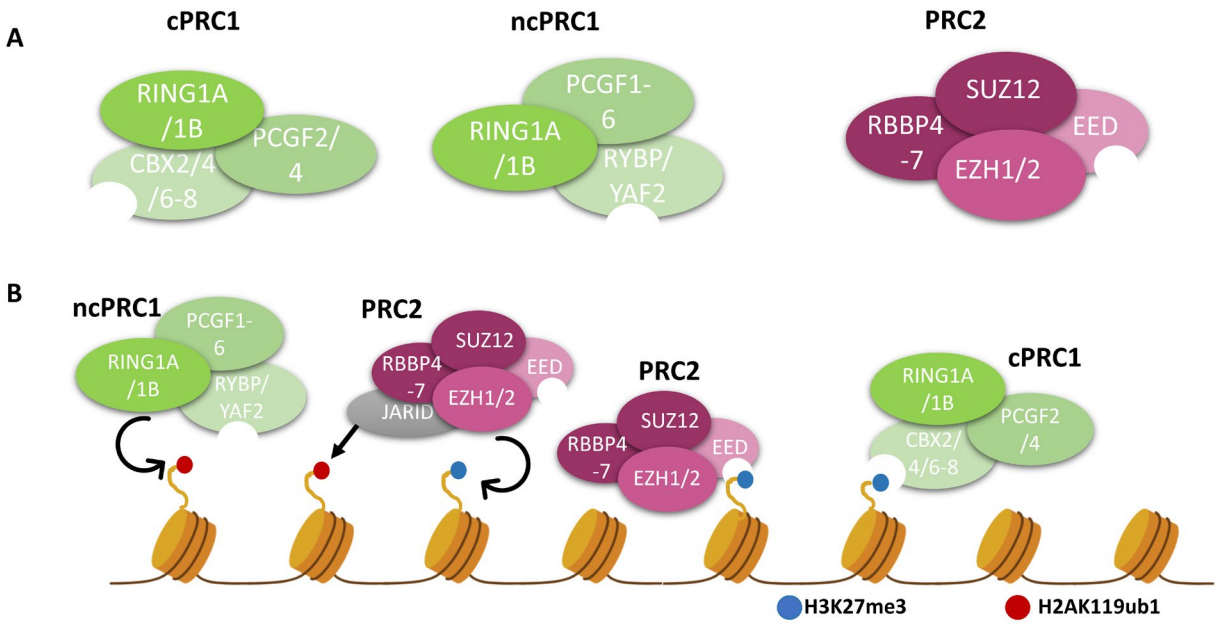


Figure 1.5: Polycomb repressive complex components and recruitment mechanisms

A) Core components of canonical PRC1 (cPRC1), noncanonical PRC1 (ncPRC1) and PRC2. RING1A/B are E3 ubiquitin ligases that monoubiquitinate lysine 119 of H2A. The enzymatic subunit of PRC2 is EZH1/2.

B) ncPRC1 deposits H2AK119ub1 which is recognized by JARID2, an accessory subunit of PRC2. The H3K27me3 deposited by EZH1/2 of PRC2 is recognized by another PRC2 subunit EED. H3K27me3 in turn is recognized by CBX of cPRC1. RYBP/YAF2 of ncPRC1 can also recognize H2AK119ub1 and help to propagate this histone modification.

In *Drosophila* the Polycomb and Trithorax complexes are known to be seeded on DNA elements called Polycomb response elements (PREs) or Trithorax response elements (TREs) identified first in homeotic gene clusters. PREs are CpG-rich and located within 5 kb of gene promoters. A mammalian equivalent of PREs has not been identified, but Polycomb and trithorax complex binding in mammals has been shown to occur at CpG islands (CGIs) which are CpG-rich regions present near TSS. The Polycomb and Trithorax complexes are thought to compete for CGIs and maintain the transcriptional on or off state as first set by early TFs. Many gene promoters in pluripotent cells are identified to carry both H3K27me3 and H3K4me3, termed as bivalent promoters. An accepted model is that these bivalent promoters are present in a "poised" state which is resolved into an active or repressed state during differentiation (reviewed in (Kuroda et al., 2020)).

1.1.3 Chromatin remodelling complexes and histone variants in chromatin structure

In addition to histone PTMs, chromatin structure is also influenced by chromatin remodelling complexes and variant histones. As mentioned before, one class of TrxG is the chromatin remodelling complex SWI/SNF. Positively charged histone makes a contact with phosphate backbone of DNA on 14 points (after every 10.4 bp) which, combined, give nucleosomes positional stability (reviewed in (Clapier and Cairns, 2009)). Chromatin remodelling complexes use the energy of ATP to move, eject nucleosomes or exchange histone dimers and as a result, change the chromatin architecture by making it either more, or less accessible for transcriptional machinery and other DNA templated processes. The activity of these complexes on gene regulatory regions such as promoters and enhancers can therefore impact on gene expression.

Predominantly reported histone variants belong to either H2A or H3 families. The best characterized histone variants include macroH2A, H2A.Z, H2A.X, H3.3 and CENP-A. MacroH2A is generally regarded as a repressive histone variant and has been shown to repress transcription initiation (reviewed in (Sun and Bernstein, 2019)). H2A.X is well studied for its role in DNA damage repair where it becomes phosphorylated at site of DNA damage and leads to the recruitment of DNA repair machinery. H2A.Z forms an extended acidic patch which is important for its interaction with chromatin remodelling complexes explaining its role in higher-order chromatin structure and transcription regulation (reviewed in (Bönisch and Hake, 2012)). CENP-A is an H3 variant that resides in centromeres and functions in chromosome segregation (reviewed in (Jin et al., 2005)). Deposition of H3.3 is thought to produce an unstable nucleosomal conformation and is correlated with transcriptionally active chromatin (Schwartz and Ahmad, 2005).

1.2 DNA methylation

1.2.1 DNA methylation - a prevalent epigenetic modification

Methylation of the 5th carbon of cytosine (5mC) is a primitive epigenetic modification that originated in bacteria. The evolution of CpG DNA methylation is complex as it is known to have been present in the first eukaryotes and is found in many present higher eukaryotes, but has been lost in some organisms, for example *Drosophila melanogaster*, *Caenorhabditis elegans*, bakers and fission yeast (Zemach et al., 2010). In plants 5mC is present in CpG, CpHpH and CpHpG (where H is A , C or T) contexts which has led to the evolution of an elaborate DNA methylation machinery (reviewed in

(Zhang et al., 2018a)). In the plant model *Arabidopsis thaliana* 24%, 6.7% and 1.7% of cytosines in CpG, CpHpG and CpHpH, contexts respectively, are found methylated, occurring predominantly in transposons and tandem repeat elements (Cokus et al., 2008; Zhang et al., 2006). The study of these mechanisms in the plant model *Arabidopsis thaliana* have contributed greatly to our current knowledge of DNA methylation regulation (reviewed in (Zhang et al., 2018a)).

In vertebrates 5mC predominantly is found in the context of CpG dinucleotides. 5mC is prevalent among mammals with 70-80% of all CpGs being methylated (Zemach et al., 2010). The distribution of CpG methylation in the genome is not uniform. In mammals, the regions with lower density of CpGs, which includes tandem repeat elements and heterochromatin, can be found extensively methylated, while the regions with the highest CpG content, such as those near gene promoters, CGIs, are unmethylated. Due to its co-occurrence with constitutive heterochromatin, CpG methylation was classically thought of as a 'silencing' epigenetic modification. However, the later discovery of 5mC existence in euchromatin, such as active gene bodies, changed this perception and brought into light the regulatory functions of CpG methylation, such as in RNA splicing and inhibition of spurious transcription. In this section, I present a brief account of the CpG methylation (with focus on mammals), the enzymes responsible for its establishment and removal, and its distribution in the genome.

1.2.2 A brief history of the discovery of DNA methylation

The timeline of important discoveries in DNA methylation field have been complied in a recent review (Mattei et al., 2022). 5mC was first discovered in 1925 in *Mycobacterium tuberculosis*, long before DNA was shown to be the genetic material in living organisms in the famous Avery-MacLeod-McCarty experiment in 1944. It was also discovered that 5mC usually precedes a G and that during DNA replication, the newly synthesized DNA strand becomes methylated (Sinsheimer et al., 1954), indicating the association of methylation machinery with DNA replication machinery (Billen, 1968). The first mechanistic role of DNA methylation was unveiled by the discovery of restriction-modification system (R-M system) in bacteria. The R-M system is a primitive immune system consisting of a set of DNA methyltransferase and methylation sensitive endonucleases. The methyltransferase methylates the bacterial DNA which is then protected from cleavage by a complementary restriction endonuclease. Unmethylated bacteriophage DNA however is cleaved, protecting the host form this pathogen (Arber, 1965). After the discovery of a function for methylated DNA in bacteria, an exploration of 5mC was begun in other organisms. From 1970 – 1979, 5mC was found to occur in various species of animals like sponges, sea urchin, molluscs, fish, amphibians, reptiles, and mammals.

This widespread evolutionary distribution of 5mC validated its significant role in cellular function. 5mC content was found to differ among animals, with closely related species sharing more similarities in 5mC content (Vanyushin et al., 1970). As noted above, among plants, 5mC was found to occur in contexts other than CpG dinucleotides (Guseinov and Vanyushin, 1975).

Around this time, John Gurdon performed his landmark nuclear transformation experiments showing that transplanting a somatic cell nucleus into enucleated oocytes could reprogram it, thereby proving that the genetic material remains conserved throughout the lifetime of an organism (Gurdon and Uehlinger, 1966). Based on this development, Holliday and Pugh proposed that DNA methylation could provide a mechanism for reversibly turning the genes off and on during development without alterations in the DNA sequence (Holliday and Pugh, 1975). Riggs proposed that 5mC on regulatory sequences could inhibit protein binding there (Riggs, 1975). Sager and Kitchin took inspiration from the bacterial R-M system and suggested that 5mC could lead to "restriction" of modified DNA sequence, leading to the heterochromatinization (and inactivation) of 5mC modified regions (Sager and Kitchin, 1975).

A leap forward in DNA methylation research came from the study of methylation at genes exhibiting tissue-specific expression patterns. This was aided by the discovery of methylation-sensitive and insensitive restriction enzymes, *Hpa*II and *Msp*I (both of which recognize CC|GG sites) (Smith and Welcox, 1970; Waalwijk and Flavell, 1978b). Genes like *β-globin* were variably methylated among different tissues and exhibited lower methylation (at CCGG sites) in tissues where these genes were expressed (Waalwijk and Flavell, 1978a). 5-azacytidine (5-azaC) was another important addition to the tool kit for DNA methylation research. 5-azaC is a cytosine analogue, which incorporates into DNA and RNA, and replaces cytosine and 5mC. Treatment of mouse embryonic fibroblasts with 5-azaC caused them to become tubular and multinucleated, suggesting their differentiation into muscle cells (Constantinides et al., 1977). The mode of action of 5-azaC became clear in an experiment which showed that 5-azaC treatment of *E.coli* caused a reduction in levels of 5mC and not 6 methyl adenosine (6mA, another DNA methylation modification found in bacteria) (Taylor and Jones, 1979). Later experiments showed that 5-azaC treatment also leads to gene activation, establishing a role for CpG methylation in gene silencing (Jones and Taylor, 1980). More contemporary work has shown that 5-azaC functions by irreversibly trapping the enzymes that perform CpG DNA methylation (Shen and Laird, 2012). At this point, it was not clear how the position of 5mC relative to the gene (within or around a gene for example), determined/influenced its expression. Further research showed that loss of 5mC did not always lead to gene activation. For example, a paper showed that loss of 5mC on 5' end of the gene led to its activation while most of the gene was still methylated (Wilks et al., 1982).

Similarly, other studies showed that genes were expressed in some tissues despite high levels of DNA methylation across the gene (Cate et al., 1983; Vedel et al., 1983).

Later experiments revealed that DNA methylation is highly regulated during development. By injecting viral DNA at various stages of mouse embryogenesis, it was shown that in germline and preimplantation embryos, viral DNA became silent and inactivated due to *de novo* DNA methylation activity (Harbers et al., 1981). This finding led Monk and colleagues to study 5mC levels in sperm, oocytes and early-stage mouse embryos where they discovered that mouse embryos undergo shifts in DNA methylation. They showed that sperm DNA has high levels of DNA methylation which is gradually lost in embryos around the onset of implantation. The embryo proper then progressively regains DNA methylation through gastrulation (Monk et al., 1987).

In the late 1980s, the first bacterial 5mC methyltransferase, M.*Hha*I was cloned. This was a major development in the field since the protein sequence also brought to light the evolutionary conservation of the catalytic domain of 5mC methyltransferases (Lauster et al., 1989; Pósfai et al., 1989). In 1993, the crystal structure of M.*Hha*I was solved (Cheng et al., 1993) and Wu and Santi revealed key insights into the chemistry of cytosine methyl-transfer reaction (Wu and Santi, 1987). Importantly, it was revealed that the catalytic domain contained sequence motifs, which were found conserved among all 5mC methyltransferases. With subsequent work on mammalian and plant 5mC methyltransferases, and the discovery of 5mC binding proteins, the field of DNA methylation made fast progress (Bestor et al., 1988; Boyes and Bird, 1991; Cheng et al., 1993). These studies then paved the way for our current understanding of the mediators, mechanisms, and functions of 5mC.

1.2.3 Enzymes catalysing DNA methylation

In vertebrates, cytosine methylation is catalysed by DNA methyl transferases (abbreviated as DNMT) which transfer the methyl group from a donor molecule, S-adenosyl methionine (SAM), to the 5^{th} carbon of the cytosine pyrimidine ring (Figure 1.6 A). DNMTs can be classified into two categories based on their preference for DNA substrate. (1) *de novo* methyltransferases (in mammals two main enzymes are DNMT3A and DNMT3B) which establish methylation of previously unmethylated CpG sites on DNA and (2) maintenance methyltransferase (DNMT1) which copies/reproduces CpG methylation pattern of hemi-methylated DNA on the parent strand onto the daughter strand during DNA replication. DNMT2, originally categorized together with other DNMTs on the basis of its sequence homology, is an tRNA methyltransferase, establishing 5mC on $tRNA^{Asp}$ (Goll et al., 2006). Recently a third *de novo* DNMT, DNMT3C , has been discovered in rodents and characterized for its role in repressing retrotransposons during spermatogenesis (Barau et al., 2016). Another member of

the DNMT family, DNM3L (DNMT3-like), is unusual, as it possesses no catalytic activity. DNMT3L is missing residues in the conserved catalytic domain motifs and lacks a PWWP domain found in other DNMT3 proteins (Figure 1.6 B). Although DNMT3L is not catalytically active itself, it interacts with DNMT3A and DNMT3B, and acts as an accessory factor to stimulate their catalytic activity (Chedin et al., 2002). Below I present a brief account on DNMT1 and then provide details on DNMT3A and DNMT3B which are a major focus of work in this thesis.

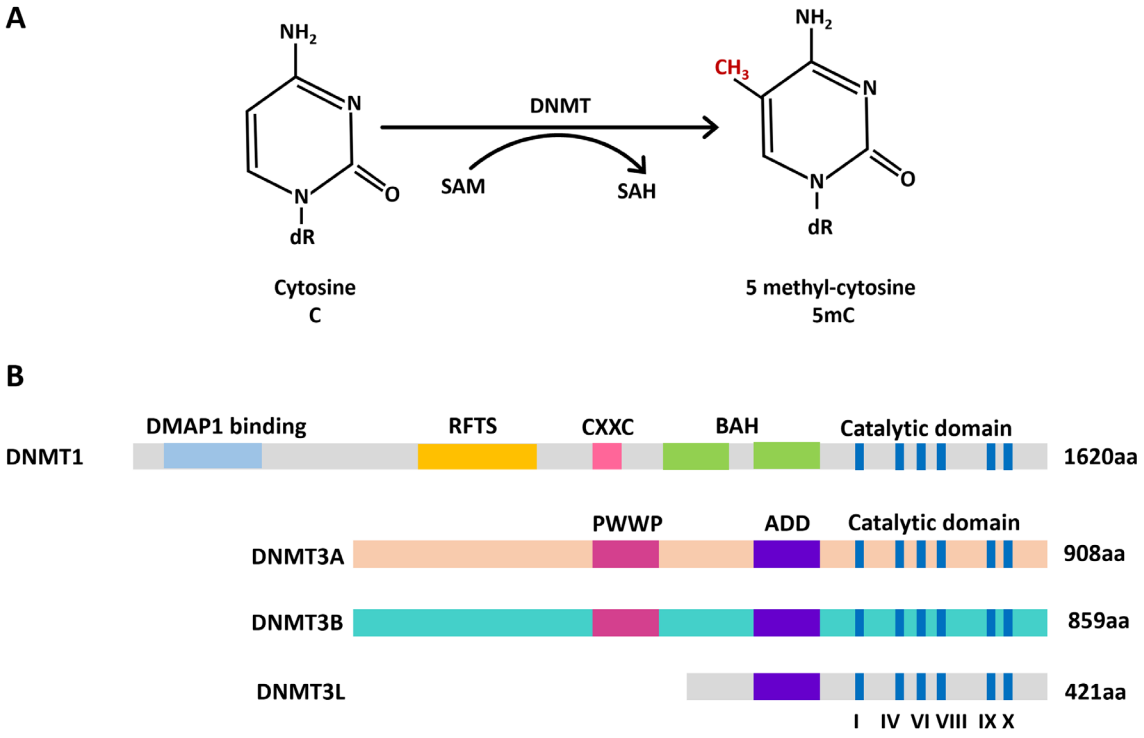


Figure 1.6: CpG methylation reaction and catalysing enzymes

A) Schematic showing the methylation of 5th carbon of cytosine by DNMT. S-adenosyl methionine (SAM) serves as a methyl group donor.

B) Schematic structure of DNMT1, DNMT3A, DNMT3B and DNMT3L. All DNMT proteins carry an N-terminal domain which serves a regulatory role by mediating binding with interacting partners or chromatin, and a C-terminal catalytic domain. The number of amino acids written on the right correspond to the size of the mouse protein. In DNMT1, DMAP1 binding domain mediates binding with a transcriptional repressor DNA methyltransferase-associated protein 1 and RFTS is replication foci targeting sequence which allows targeting of DNMT1 to replication foci. The CXXC is zinc finger domain recognizes unmethylated DNA. The function of bromo-adjacent homology (BAH) domain is unclear but has been proposed to function in protein-protein interactions and efficient targeting to replication fork (Yarychkivska et al., 2018). The blue bars represent the conserved catalytic motifs (labelled in Roman numerals below the figure). In DNMT3A and DNMT3B, the PWWP domain is Pro-Trp-Trp-Pro and ADD is ATRX-DNMT3-DNM3L domain. Both PWWP and ADD domains are chromatin-binding domains and recognize H3K36me3 and unmodified H3K4 respectively.

1.2.3.1 The maintenance DNA methyltransferase DNMT1

DNMT1 was the first member of the vertebrate DNMT family to be studied (Bestor et al., 1988). The maintenance methyltransferase DNMT1 is also the largest member among the DNMT enzymes with a complex regulatory N-terminal domain and a catalytic C-terminal domain (Figure 1.6 B). As mentioned before, the catalytic domain of DNMT1 is composed of conserved sequence motifs which were first identified in bacterial M.*Hha*I, one of the smallest 5mC methyltransferase and the first one to be crystallized in 1993 (Cheng et al., 1993).

DNMT1 works in concert with an E3 ubiquitin ligase UHRF1 which acts as a multifaceted regulator of DNMT1. UHRF1 specifically binds to hemi-methylated CpG at the replication fork via its SET and RING associated (SRA) domain and helps in the recruitment of DNMT1 (Bostick et al., 2007; Sharif et al., 2007). DNMT1 possesses a replication fork targeting sequence (RFTS) which is buried in its catalytic domain, locking the enzyme in an autoinhibited form. In the presence of UHRF1, DNMT1 RFTS is released from the catalytic domain, allowing DNMT1 to carry out the DNA methylation at the replication fork (Song et al., 2011; Takeshita et al., 2011). Ubiquinated H3 lysine 18 and 23 catalysed by the RING domain of UHRF1, are bound by DNMT1 and help to stimulate its catalytic activity (Nishiyama et al., 2013; Qin et al., 2015). UHRF1 interacts with chromatin via its tandem tudor domain (TTD) and plant homeodomain (PHD) which recognize H3K9me3 and H3R2 respectively. Both of these chromatin interactions by UHRF1 have been shown to be required for the faithful recruitment and maintenance of DNA methylation by DNMT1 (Nady et al., 2011; Qin et al., 2015; Rothbart et al., 2012), (Figure 1.7). UHRF1 is also involved in the ubiquitination of DNMT1 itself which causes proteasomal degradation and reduction of the protein levels of DNMT1 in the cell. The interaction of DNMT1 with proliferative cell nuclear antigen (PCNA) which binds to newly replicated DNA, has also been suggested to aid its recruitment to the replication fork (Chuang et al., 1997; Iida et al., 2002).

DNMT1 protein has a CXXC domain that binds unmethylated CpG sites and stabilizes the positioning of an autoinhibitory linker, blocking the interaction of catalytic domain and unmethylated CpG on the DNA and providing the basis for the inhibition of *de novo* methyltransferase activity by DNMT1 (Song et al., 2012; Takeshita et al., 2011). There is some recent evidence of low levels of *de novo* methyltransferase activity of DNMT1 in transposons such as LTRs (long terminal repeats) and IAPs (intracisternal A particle) (Haggerty et al., 2021).

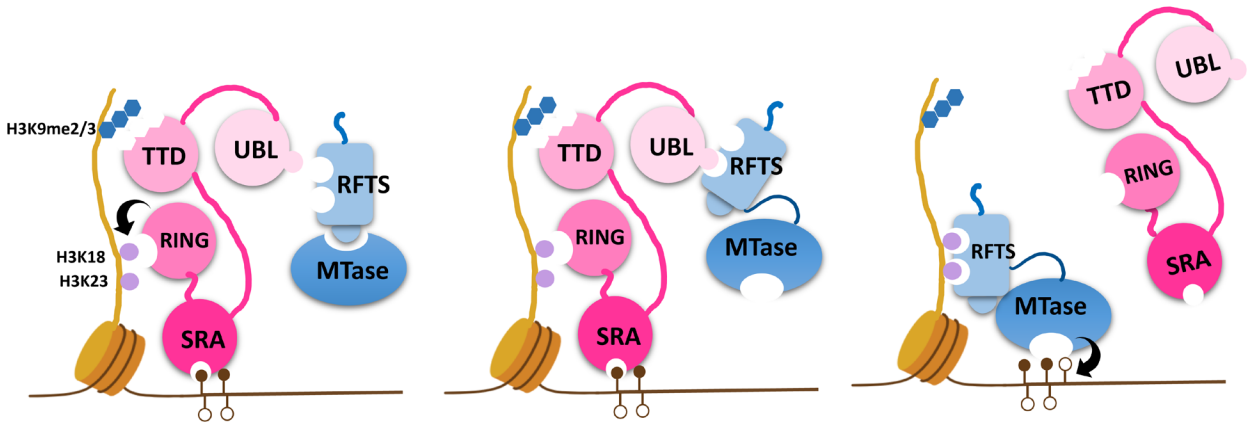


Figure 1.7: UHRF1-mediated recruitment of DNMT1 to replication fork.

UHRF1 (pink) is recruited to the replicating hemi-methylated DNA via its SRA domain. TTD domain associates with H3K9me2/3 (left). DNMT1 (blue) is locked in an autoinhibited form due to the RFTS domain which folds into the methyltransferase domain (MTase) unless encountered by UHRF1. The Ubiquitin-like domain (UBL) of URF1 interacts with DNMT1 RFTS, releasing the later from autoinhibition (middle). The SRA domain of UHRF1 also ubiquitinates the H3 tail at lysine 18 and 23 (shown in purple) which are recognized by the RFTS domain of DNMT1 and have been reported to enhance its catalytic activity and release DNMT1 autoinhibition. H3K9me3 is depicted by blue hexagons and methylated and unmethylated CpG by filled or unfilled lollipops respectively.

1.2.3.2 *de novo* methyltransferases DNMT3A and DNMT3B

After the identification of DNMT1, other 5mC methyltransferases were discovered based on their homology with DNMT1. The mechanism of the transfer of methyl group to cytosine is also conserved and involves flipping and rotating the target base into the catalytic pocket, followed by a nucleophilic attack of the 5mC methyltransferase on the target cytosine, forming a methyltransferase-cytosine intermediate. This is followed by the transfer of methyl group from SAM to C5 of cytosine and deprotonation of C5 by the enzyme to form 5mC (Klimasauskas et al., 1994; Wu and Santi, 1987).

The C-terminal domain of DNMT3A and DNMT3B forms a heterotetramer with DNMT3L, where DNMT3L is on the outer edges of the tetramer, and the two DNMT3A or DNMT3B molecules are in the centre (Figure 1.8) (Gao et al., 2022; Jia et al., 2007). This organization increases the interface of DNMT3A and DNMT3B with DNA and provides an option to catalyse cytosine methylation on both DNA strands. Multimerization of DNMT3A/DNMT3L tetramers has also been suggested to lead to

cooperative DNA binding but the size of these complexes and the exact role of this cooperative binding needs further investigation (Emperle et al., 2014; Rajavelu et al., 2012) and reviewed in (Jurkowska and Jeltsch, 2016).

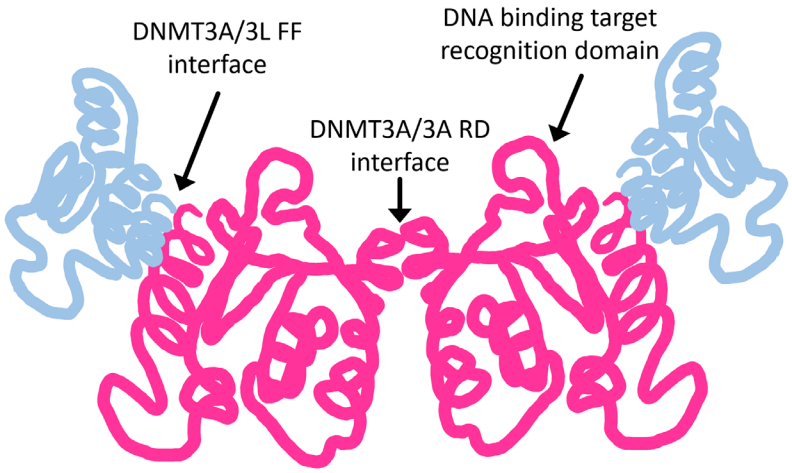


Figure 1.8: DNMT3A and DNMT3L form heterotetramers.

Schematic depicting the catalytic domain of DNMT3A and catalytic-like domain of DNMT3L. Both DNMT3A and DNMT3B associate with DNMT3L to form heterotetramers, with DNMT3L on the outer edges and DNMT3A or DNMT3B in the centre. The DNMT3A/3L interface is called the FF interface, due to the crucial role of two phenylalanine residues that form a hydrophobic interaction. The DNMT3A/3A interface called the RD interface, is formed by polar hydrogen bonding between arginine and aspartate residues from both binding partners (DNMT3A, as shown here, or DNMT3B). This positioning of the two DNMT3A molecules enables methylation of CpG sites on two strands of DNA if located ~10 bp apart. Due to the absence of residues for RD interface, DNMT3L caps this tetramer from further oligomerization and hence the models presented for oligomerization of DNMT3A/3L involve multiple heterotetramers arranged parallel to each other. DNMT3A has also been shown to form heterodimers with DNMT3B using the same RD interface as for DNMT3A/3A interface (Li et al., 2007), also reviewed in (Jurkowska & Jeltsch, 2016)

The N-terminal region of DNMT3A and DNMT3B includes PWWP and ADD domains, both of which mediate chromatin interactions (Figure 1.6 B). The PWWP domain (named after the conserved Pro-Trp-Trp-Pro motif) forms an aromatic cage structure to bind methylated histone tails. The PWWP domain of DNMT3 proteins has an affinity for H3K36me3-modified chromatin which allows them to be recruited to transcriptionally active gene coding regions. The structure of the DNMT3B PWWP domain revealed the presence of two large basic patches which were proposed to bind the DNA helix (Qiu et al., 2002). A recently solved structure of the PWWP domain of a transcriptional coactivator LEDGF in complex with nucleosome indeed confirmed the simultaneous interaction of PWWP domain

with both H3K36me3 and DNA (Wang et al., 2020). Wang et al., proposed that this multivalent binding mode of the PWWP domain explains why PWWP-H3K36me3 interaction in a nucleosome context strengthens by 10,000-fold in comparison to H3K36me3 peptide, observed by (Eidahl et al., 2013; van Nuland et al., 2013; Wang et al., 2020).

The second chromatin interacting domain of DNMT3A and DNMT3B is the ADD (ATRX-DNMT3-DNMT3L) domain. The ADD domain shows an affinity for unmodified H3K4 and causes the exclusion of DNMT3 proteins from genetic loci modified with H3K4me3 (Otani et al., 2009). When not engaged with H3K4, the ADD domain binds the catalytic domain of DNMT3, locking it in an autoinhibited state (Guo et al., 2015b). This mechanism ensures that promoters of actively transcribed genes, which are enriched with H3K4me3, stay unmethylated (Guo et al., 2015b; Piunti and Shilatifard, 2016). In the presence of unmodified H3K4, the catalytic domain is released from the ADD domain, freeing DNMT3 proteins to methylate DNA (Figure 1.9). Since active transcription causes H3K4me3 deposition on gene promoters, it can antagonize the CpG methylation of these loci (Clouaire et al., 2014). DNA methylation can accordingly be thought of as a locking mechanism that maintains gene silencing once transcriptional repression has already taken place.



Figure 1.9: Autoinhibition of DNMT3A and DNMT3B with ADD domain.

Schematic showing the autoinhibition of DNMT3A or DNMT3B by the ADD domain. The ADD domain blocks the catalytic domain (denoted as MTase for methyltransferase) at the H3K4me3 modified regions (shown by three blue hexagons) such as promoters of actively transcribing genes. Upon binding unmodified H3K4, ADD domain releases MTase, allowing it then to catalyse CpG methylation (teal lollipops).

The ADD domain of the protein ATRX (protein product of the gene mutated in alpha thalassemia/mental retardation, X-linked syndrome) was shown to have an affinity for H3K9me3. The residues shown to engage in the ATRX-H3K9me3 binding are not conserved in DNMT3 proteins (Iwase et al., 2011). However, DNMT3A and DNMT3B have been shown to carry out DNA methylation on

H3K9me3-modified chromatin. An *in vitro* study showed that while DNMT3A poorly methylated the H3K4me3-modified nucleosomes, the levels of CpG methylation on H3K9me3-modified or unmodified H3 nucleosomes were comparable and 5 times higher than on H3K4me3-modified nucleosomes (Zhang et al., 2010). This could indicate that H3K9me3 modification enables catalysis and that DNMT3A is no longer autoinhibited in the presence of H3K9me3. This difference in the catalysis between H3K4me3- and H3K9me3-modified chromatin was not seen with DNMT3A catalytic domain alone, therefore suggesting that ADD domain-mediated autoinhibition hindered methylation on H3K4me3 and by extension, not at H3K9me3 (Zhang et al., 2010). The ADD domain structure suggests that the conformation of key residues in the binding pocket may allow DNMT3B to have a higher tolerance to bind H3K9me3 than DNMT3A but this has not been experimentally validated (Boyko et al., 2022).

The role of chromatin remodelling factors such as LSH (also known as HELLS), - a protein from the SNF2 family of chromatin remodelling complexes, has been shown to be crucial for DNA methylation, especially at pericentromeric regions (Dennis et al., 2001; Myant and Stancheva, 2008). Recent studies have shown that *de novo* methyltransferases preferentially methylate CpG sites on linker DNA, providing support for the importance of chromatin remodelling factors in the function of *de novo* methyltransferases (Xu et al., 2020). A good example of the close link between *de novo* methyltransferases and chromatin remodelling factors comes from the disease ICF (immunodeficiency, centromeric instability and facial anomalies) which predominantly results from mutations in *DNMT3B* (subtype 1), but in some cases by mutations in *LSH* (the resulting disease called ICF subtype 4) (reviewed in (Kiaee et al., 2021)).

1.2.4 Enzymes catalysing demethylation of 5mC

The CpG methylation patterns in the genome are formed by the dynamic action of enzymes catalysing the methylation and demethylation of 5mC. The first observations of a potential demethylating activity were made in mouse development where the paternal pronucleus was seen to lose 5mC before the first cleavage division (Mayer et al., 2000).

In 2009 it was described that DNA methylation can be removed by Ten-eleven translocation (TET) enzymes which hydrolyse 5mC step-by-step to 5 hydorxymethyl cytosine (5hmC), formyl cytosine (5fC) and carboxy cytosine (5caC) (Ito et al., 2010; Ito et al., 2011; Kriaucionis and Heintz, 2009; Tahiliani et al., 2009) (Figure 1.10). 5fC and 5caC are recognized and removed by thymine DNA glycosylase (TDG) creating an abasic site. These abasic sites are in turn, targeted by DNA base excision repair (BER) machinery to an unmodified cytosine (Hashimoto et al., 2012; Weber et al., 2016).

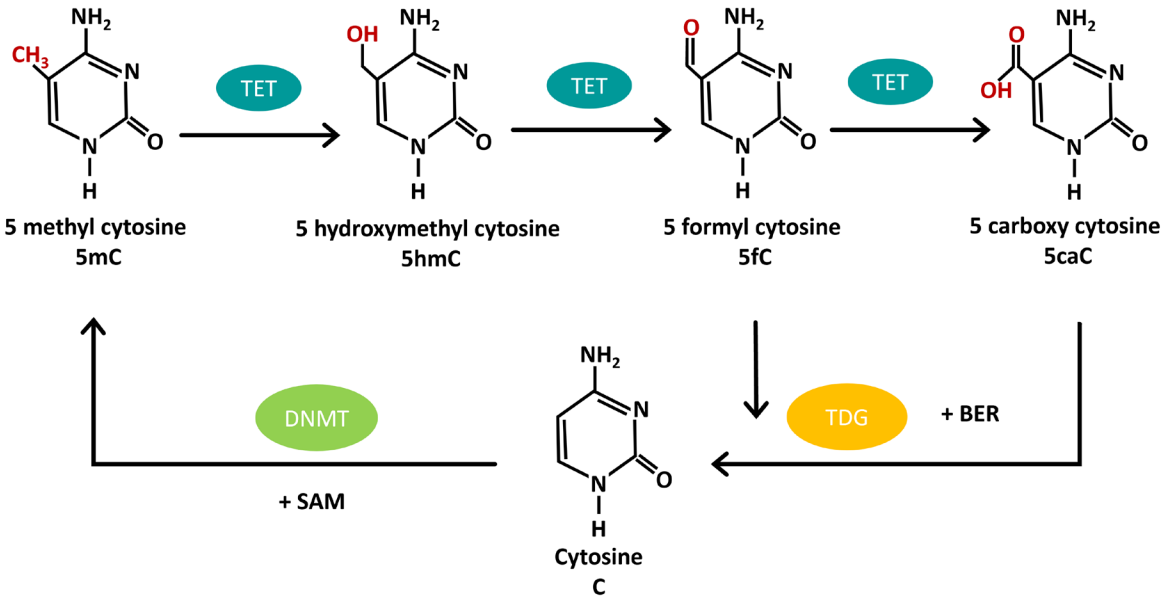


Figure 1.10: Demethylation of 5mC by TET enzymes

TET enzymes demethylate 5mC in a step-wise reaction converting 5mC to 5hmC and then to 5fC and 5caC. Both 5fC and 5caC can be removed by thymine DNA glycosylase (TDG) to create an abasic site which is removed by Base Excision Repair (BER) machinery and replaced by cytosine.

The name ten eleven translocation comes from the translocation observed in acute myeloid leukaemia (AML) leading to TET-MLL fusion protein (Ono et al., 2002). There are three TET proteins in mammals (TET1-3) and they differ in the binding affinities and choice/preference for substrates. A number of studies have indicated that despite their low abundance, 5hmC and 5fC perform independent roles in cells and are thus, more than just transitionary products. TET-mediated removal of DNA methylation is referred to as “active demethylation” to distinguish it from “passive demethylation” which occurs as a result of deficient maintenance of CpG methylation after DNA replication. Failure to efficiently maintain methylation results in dilution and subsequently loss of DNA methylation upon successive cell divisions. Both active and passive mechanisms are employed to regulate DNA methylation during mammalian development (reviewed in (Rasmussen and Helin, 2016)).

1.2.5 The enzymology of DNA methylation in plants

In *Arabidopsis* DNA methylation is concentrated at centromeres and repetitive sequences (Zhang et al., 2006). The DNA methylation machinery in plants is more elaborate as compared to mammals, owing to the methylation of cytosines in CpHpG and CpHpH context in addition to CpG dinucleotide.

In plants, DNA methylation is catalysed by *de novo* methyltransferase Domains rearranged methyltransferase 2 (DRM2), a homologue of mammalian DNMT3 methyltransferases. Maintenance of DNA methylation is conducted by three different enzymes; maintenance of CpG methylation is carried out methyltransferase 1 (MET1, homologue of mammalian DNMT1), maintenance of CpHpG methylation is done by chromomethyltransferase 3 (CMT3) and CpHpH methylation maintenance is performed by DRM2 (reviewed in (Matzke et al., 2009)). An RNA-dependent DNA methylation (RdDM) pathway is responsible for the establishment of *de novo* methylation in plants (Law and Jacobsen, 2010; Matzke et al., 2009; Wassenegger et al., 1994). A plant-specific RNA polymerase IV (Pol IV) transcribes RNA from transposons and repeats (Huettel et al., 2007; Pikaard et al., 2008), which is used as template by RNA-dependent RNA polymerase (RDR2) to generate double stranded RNA (dsRNA) (reviewed in (Matzke et al., 2009)). A dicer-like protein DCL3 then generates 24-nucleotide small interfering RNAs (siRNA) from the dsRNA which are bound by argonaute proteins AGO4 and AGO6 (Li et al., 2008). AGO4 interacts with a subunit of Pol V (called NRPE1), through the interaction between the AGO4-bound siRNA and the nascent transcript from Pol V (Wierzbicki et al., 2008), ultimately leading to the recruitment of DRM2. The recruitment and retention on DRM2 is thought to be aided by suppressor of variegation 3-9 homologue 9 and 2 (SUVH9 and SUVH2) (Johnson et al., 2008).

1.2.6 Methyl-CpG binding proteins

A well-characterized role of DNA methylation in gene silencing is linked to proteins that specifically recognise and bind 5mC, often recruiting histone-modifying complexes. The first discovery in this regard was made by Bird and colleagues. They noted that in an *in vitro* assay, when the methyl DNA binding protein MECP1 was diluted by addition of methylated competitor DNA, the transcription of four gene promoters was equivalent regardless of whether they were methylated or unmethylated (Boyes and Bird, 1991). Similar assays using the extracts of cells deficient in MECP1 showed an inability to efficiently repress methylated DNA. Based on these observations, Bird and colleagues proposed that the role of DNA methylation in transcriptional silencing is indirect and mediated by the methyl DNA binding proteins (Boyes and Bird, 1991). Soon afterwards another methyl CpG binding protein, MECP2, was found to be able to bind single symmetrical methylated CpG and inhibit transcription from methylated DNA *in vitro* (Meehan et al., 1992). Pulldown experiments revealed that MECP2 interacts with histone deacetylase complexes (HDACs) and transcriptional repressor SIN3A, suggesting that histone deacetylation may mediate transcriptional silencing at MECP2-bound sites. The treatment of cells with the HDAC inhibitor trichostatin A (TSA) was shown to relieve the MECP2-mediated transcriptional repression *in vivo* validating the functional association of HDACs and MECP2

(Nan et al., 1998). Subsequently, more proteins were identified based on their homology with MECP2 methyl CpG binding domain (MBD).

In mammals, the MBD protein family has 7 members, MBD 1-6 and MECP2 (reviewed in (Du et al., 2015)). Most of the MBD proteins have been found to bind to 5mC. The exceptions are MBD3, which bears mutations preventing it from binding methylated DNA (Hendrich and Bird, 1998). MBD5 and MBD6 have also been shown not to bind methylated CpGs *in vitro* (Laget et al., 2010). Association of MBD proteins with HDACs is common and has been shown for MECP2 and MBD1, while MBD2/MBD3 are a part of HDAC complex called Mi-2/Nucleosome remodeling and deacetylase (NuRD/Mi-2) (Lyst et al., 2013; Ng et al., 1999). Although histone deacetylation was considered the main mechanism behind transcriptional silencing facilitated by MBD proteins, they have also been implicated in additional chromatin modifications. MBD1 for instance, has been shown to associate with HP1 and H3K9me3 methyltransferase SUV39H1, which enhance MBD1-mediated transcriptional silencing (Fujita et al., 2003). MBD1 engages with MBD1-containing chromatin-associated factor 1 (MCAF1, also known as ATF7IP) and this leads to the recruitment of another H3K9me3 methyltransferase SETDB1 (Ichimura et al., 2005). Links between MBD proteins and chromatin remodelling complexes have also been proposed. For example, Mi-2 subunits of NuRD/Mi-2 are components of ATPase-dependent SWI/SNF complex. Disruption of MBD2-Mi-2 interaction has been shown to increase chromatin accessibility and prevent the silencing of *mb-1* gene in B cells (Ramírez et al., 2012). Further evidence comes from the interaction of MECP2 with another SWI/SNF family protein called BRM (KN et al., 2005). Finally, MBD proteins facilitate the maintenance of CpG methylation creating a feed forward loop. By associating with DNMT1 and DNMT3B, MBD3 and NuRD/Mi-2 complex has also been proposed to maintain CpG methylation at tumour suppressor genes in colon cancer cells (Cai et al., 2014).

MBD is not the only protein domain/fold with the ability to bind methylated DNA. KAISO is a DNA binding protein with dual specificity. It can bind two methylated CpGs or an unmethylated consensus sequence motif via its zinc fingers (Daniel et al., 2002; Shumskaya et al., 2015). KAISO brings about gene repression of its target loci by associating with histone deacetylase containing Nuclear receptor corepressor (NCoR) complex (Yoon et al., 2003).

1.2.7 Molecular techniques to study CpG methylation

Molecular techniques to study CpG methylation have evolved rapidly over several years. The advent of more sensitive and precise experimental procedures provided an understanding of the distribution and function of CpG methylation in the genome and played a pivotal role in the field. The techniques

to study DNA methylation can be broadly divided into three categories based on the underlying principle: (1) restriction digestion-based methods, (2) methods using sodium bisulphite treatment and (3) affinity enrichment methods. Among the pioneering methods to study CpG methylation, is the use of methylation-sensitive restriction endonucleases (MRSEs). MRSEs such as *Hpa*II or *Sma*I, which recognize CpG-rich sequences, were used in conjunction with their methylation insensitive isochizomers (*Msp*I for *Hpa*II and *Xma*I for *Sma*I) to digest genomic DNA. The pattern of the restriction fragments for both the digests was compared for the regions of interest and visualized using Southern blot. However, this method only allowed the study of regions containing the target sequence for the MRSEs and was prone to inaccuracies which could arise due to incomplete digestion by MRSEs (reviewed (Dahl and Guldberg, 2003)).

In the 1990s the field of DNA methylation research was revolutionized by the discovery that unmethylated cytosines are deaminated rapidly upon the treatment with sodium bisulphite unlike 5mC (Frommer et al., 1992). Thus, bisulphite treatment converts unmethylated C to T, while 5mC is retained as cytosine. The use of sodium bisulphite treatment of DNA gave way to the development of a wide variety of techniques (summarized in Table 1.1). Initially bisulphite treated DNA was analysed using Sanger sequencing of cloned PCR products from a region of interest. Illumina advanced the DNA methylation analysis of bisulphite converted DNA by introducing hybridization array technology (such as GoldenGate and Infinium methylation, described in Table 1.1). In these methods, bisulphite converted DNA was amplified, fragmented, and hybridized to site-specific oligos/probes immobilized on bead chips. Each CpG locus is represented by two oligos that are complementary to either methylated or unmethylated state of the locus. The genomic DNA fragments would hybridize to one of the oligos depending on the methylation status of the CpGs in the locus. The hybridized site-specific oligos were extended by labelled deoxy nucleotides (dNTPs) and the fluorescent signal could then be read and analysed (Bibikova et al., 2009). A further improvement was the processing of bisulphite-converted DNA using Next Generation Sequencing (NGS) technology. Whole genome bisulphite sequencing (WGBS) is considered a gold standard method and provides single base pair resolution for studying 5mC methylation across the genome. Importantly, WGBS provides a quantitative measure of the absolute methylation level at the given cytosine in the sample. WGBS, however, requires high sequencing depth which increases the cost of the assay. Therefore, approaches that enrich a proportion of the genome, such as reduced representation bisulphite sequencing (RRBS), are commonly used to make bisulphite sequencing more cost-effective. In RRBS, genomic fragments obtained after digestion with restriction enzymes like *Msp*I are sequenced, which concentrate/enrich for CpG-rich genomic loci. RRBS is claimed to capture 70% of gene promoters in mammalian cells (Guo et al., 2015a). A main disadvantage of the bisulphite treatment of DNA is the harsh nature of chemical

reaction which results in degradation of DNA. Furthermore, since majority of cytosines are unmethylated, the sodium bisulphite treatment cause reduction in sequence complexity of input DNA, leading to its poor mappability. Recently a method called TAPS (TET-assisted pyridine borane sequencing) has been developed which overcomes these limitations. TAPS employees TET mediated oxidation of 5mC and 5hmC to 5caC, with subsequent conversion of 5caC to dihydrouracil (DHU) by pyridine borane reduction. In downstream PCR, DHU is converted to T (Liu et al., 2019).

Another category of techniques used for the profiling CpG methylation is based on affinity enrichment. These approaches work on the principle of pulldown to enrich for genomic regions based on their methylation status. Methylated DNA immunoprecipitation (MeDIP) makes use of antibodies that recognize 5mC (Weber et al., 2005). Other methods use the affinity of protein domain to bind CpG-rich DNA. MBD affinity purification (MAP) uses the affinity of MBD for methylated CpG-rich DNA (Rauch and Pfeifer, 2005). A complementary technique called CAP (CXXC affinity purification) uses the enrichment of unmethylated CpG-rich DNA by the CXXC protein domain (Illingworth et al., 2008). These methods are now frequently used in combination with NGS. MAP/MBD- and BioCAP-seq (an improved CAP-seq technique), owing to their selective enrichment of CpG-rich regions are good methods to study gene promoters which tend to be rich in CpGs. Unlike, WGBS, however, they do not allow a quantitative measure of methylation levels at a given locus. Their higher affordability (due to the need for less sequencing depth) in comparison to WGBS has made them popular methods. Table 1.1 summarizes some of the widely used techniques for studying CpG methylation.

Technique	Description	Genomic coverage
MRSE analysis	Digestion of genomic DNA with methylation sensitive and methylation insensitive enzymes, followed by analysis of digestion pattern of the regions of interest by southern blot.	Low. Limited to the region of interest
COBRA (Combined bisulphite restriction analysis) (Xiong and Laird, 1997)	Bisulphite treatment of DNA and subsequent analysis of methylation dependent retention of pre-existing restriction sites or creation of new restriction sites in the region of interest.	Low. Limited to the region of interest
MSP (Methylation-specific PCR) (Ku et al., 2011)	Bisulphite treatment of DNA followed by PCR analysis with primers specially designed to discriminate the methylated or unmethylated state of target DNA.	Low. Limited to the region of interest
EpiTyper (Sequenom MassARRAY platform)	Bisulphite conversion of DNA, amplification of target DNA*, in vitro* transcription, cleavage of resulting RNA at uracil residues followed by mass spectrometery analysis (called MassARRAY analysis).	Limited to target regions but allows simultaneous analysis of few-hundred genomic regions.

Technique	Description	Genomic coverage
GoldenGate (Bibikova and Fan, 2009)	Bisulphite treated DNA is hybridized with bead array containing region-specific probes. Each CpG locus is represented by two oligos, complementary to either methylated or unmethylated state of the locus.	A microarray chip allows analysis of 1536 CpG sites corresponding to 371 genes in human genome.
Infinium methylation (Dedeurwaerder et al., 2011)	The principle is similar to GoldenGate technology.	Infinium Methylation 450K allows evaluation of 480,000 chosen cytosines genomewide.
EPIC array (Pidsley et al., 2016)	The principle is similar to GoldenGate technology.	Covers 850,000 CpG sites genomewide.
MeDIP-seq (Methylated DNA immunoprecipitation) (Down et al., 2008; Weber et al., 2005)	Fragmented genomic DNA is incubated with antibody directed against 5mC to perform an IP. Enriched DNA is sequenced using NGS.	Full genomic coverage (with repeat masking) with bias towards CpG-rich loci.
Enzymatic methyl-seq (Vaisvila et al., 2021)	Enzyme treatment of genomic DNA where 5mC is converted to 5caC by TET2 while unmethylated cytosine is deaminated to uracil by APOBEC3A. Enzyme treated DNA is sequenced using NGS.	Good genomic coverage (with repeat masking).
CAP-seq and BioCAP-seq (Blackledge et al., 2012; Illingworth et al., 2008)	Enrichment of unmethylated CpG rich regions using an affinity pulldown with CXXC domain.	Limited genomic coverage (1-5%) but cost-effective method of studying CpG-rich loci genomewide.
MBD-seq (Serre et al., 2010)	Enrichment of methylated CpG rich regions in the genome by employing affinity of MBD domain.	Low (1-5%). Cost-effective method to study methylation status of CpG-rich loci such as gene promoters.
WGBS (Urich et al., 2015)	NGS of bisulphite treated genomic DNA.	Full genomic coverage (with repeat masking).
RRBS (Meissner et al., 2005; Meissner et al., 2008)	Restriction digestion of genomic DNA with *Msp*I or *Bgl*II , followed by bisulphite conversion and NGS.	Coverage of CpG-rich genomic sites
TAPS (TET-assisted pyridine borane sequencing) (Liu et al., 2019)	TET mediated oxidation of 5mC and 5hmC to 5caC, with subsequent conversion of 5caC to dihydrouracil (DHU).	Full genomic coverage (with repeat masking).

Technique	Description	Genomic coverage
Oxford Nanopore (Rand et al., 2017)	Nucleotides (included modified) sequence is read by passing DNA through a pore that detects changes in ionic charges specific to each type of nucleotide.	Full genomic coverage.

Table 1.1: Techniques widely used for studying CpG methylation.

1.2.8 The function of CpG methylation

In the mammalian genome, CpG methylation shows non-uniform distribution where the low CpG density regions are found methylated while regions with high CpG density are hypomethylated. In plants and mammals, 5mC accumulates predominantly over transposons and repeats. Active transposons are detrimental/parasitic as they create double stranded breaks and result in insertional mutagenesis. DNA methylation is considered to play a pivotal role in the defence against transposons and this is thought to be a driving force for its evolution in eukaryotes. Human and mouse genomes contain millions of copies of retrotransposons (transposons that copy themselves via an RNA intermediate) which serve as the main target for DNA methylation. Active retrotransposons contain CpG-rich promoters, the methylation of which represses their transcription (reviewed in (Jansz, 2019)). *Dnmt1* null embryos exhibit upregulation of transcripts originating from IAP elements, the most aggressive retrotransposons in the mouse genome (Walsh et al., 1998).

DNA methylation can also render transposons permanently silenced due to its inherent mutability. Specifically, DNA methylation of transposon-coding sequences can lead to the accumulation of mutations owing to the spontaneous C ->T conversions (Greenberg and Bourc'his, 2019). "Old" inactive transposons are therefore thought to have lost their ability to move/jump in the genome due to mutations, especially in promoters, gathered over time. Derepression of transposons causes them to acquire H3K4me3, which in turn recruits meiotic recombination machinery. Thus, the methylation of transposable elements also prevents homologous recombination which could lead to improper synapsis and aberrant meiosis (Zamudio et al., 2015).

DNA methylation also plays a crucial role in maintaining the integrity of centromeres. Major and minor satellite repeats at the pericentromeric regions and centromeres are known to have transcriptional potential (Chen et al., 2004). Methylation of satellite repeats is crucial for ensuring proper centromere

structure and chromosome alignment for segregation during mitosis. Several mechanisms for the methylation of centromeric repeats have been suggested. A direct interaction between DNMT3B and centromeric protein CENP-C has been reported which leads to recruitment of DNMT3B to centromeric repeats (Gopalakrishnan et al., 2009a). At the pericentromeric regions H3K9 methyltransferases SUV39H1 and H2 are known to recruit DNMT3B to these sites via HP1 (Lehnertz et al., 2003). H4K20me3 is also known to occur at the pericentric heterochromatin which could provide an alternative mechanism for the recruitment of DNMT3B via its PWWP domain which has been shown to have an affinity for H4K20me3 (Karachentsev et al., 2005; Kidder et al., 2017).

Both mammals and plants have high levels of CpG methylation present over gene bodies of actively transcribing genes, the functions of gene body methylation however have only been made clear in mammals (Schübeler, 2015; Zhang et al., 2006). During transcription, RNA Pol II leads to the recruitment of an H3K36 histone methyltransferase called SETD2 which establishes H3K36me3 on gene bodies. DNMT3A and DNMT3B get recruited to these gene bodies via their PWWP domain, which has an affinity for H3K36me3, and establish DNA methylation (Baubec et al., 2015). CpG methylation in gene bodies is known to silence cryptic transcripts. High-resolution RNA-seq revealed that loss of CpG methylation leads to higher expression of transcripts originating from intragenic regions compared to wildtype (Neri et al., 2017). Gene body methylation has also been shown to encourage the inclusion of alternatively spliced exons (Maunakea et al., 2013; Shayevitch et al., 2018).

Gene promoters are generally CpG-rich and known to be devoid of CpG methylation. 5mC at gene promoters is correlated with gene silencing. A few mechanisms have been proposed via which CpG methylation can be inhibitory to transcription (1) by physically obstructing the binding of TFs (Rausch et al., 2020; Yin et al., 2017), (2) by changing the chemical properties of DNA such as the increase in the rigidity and melting temperature of DNA or causing tighter packaging of DNA on nucleosomes (Choy et al., 2010; Lee and Lee, 2012) and (3) by recruitment of MBD proteins (Nan et al., 1998; Ng et al., 1999; Yoon et al., 2003). A recent study has shown that the direct inhibition of methylation-sensitive TFs is the prevalent mechanism for repression of gene regulatory regions by CpG methylation in mammals (Kaluscha et al., 2022). Only a few TFs prefer to bind to their recognition motifs when they are 5mC modified. Interestingly, this ability to bind 5mC modified sequence motifs was shown for TFs KLF4 and POUF, two of the four TFs required to reprogramme somatic cells to a pluripotent state (Takahashi and Yamanaka, 2006). This rare ability could allow them to bind to their target loci when they are methylated, for instance, in differentiated cells. Many target sites of KLF4 are hypomethylated in embryonic stem cells (ES cells) and become methylated upon differentiation and thus refractory to binding for most TFs. A recent study showed that TFs like KLF4 and C/EBPα recruit

TET2 and lead to the demethylation of promoters and enhancers during cell reprogramming, which, in turn, is correlated with a gain in chromatin accessibility (Sardina et al., 2018).

CpG methylation can influence long-range chromatin interactions by inhibiting the binding of a zinc finger protein CCCTC-binding factor (CTCF) (Kanduri et al., 2000; Pant et al., 2003). Gene regulatory regions such as promoters and enhancers which are separated by large linear distances on the DNA, frequently interact with each other forming topologically associated domains (TADs). The widely accepted model for the formation of TADs is loop extrusion, where DNA is extruded out of a structural maintenance of chromosomes (SMC) complex called cohesin. SMC complexes such as cohesin form ring-like structures, through which DNA is extruded until it is halted at boundaries which are usually CTCF binding sites (reviewed in (Banigan and Mirny, 2020)). The binding of CTCF is sensitive to the methylation of CpGs within its binding motif. However, one study revealed that CTCF shows sensitivity to DNA methylation at only a subset of binding sites. CTCF profiling in 19 cell types showed that while CTCF showed a variable binding pattern in cell-type dependent manner, 41% of these differential sites were associated with DNA methylation (Wang et al., 2012).

CpG methylation patterns in mammals are largely set early during mammalian development. A brief description of the dynamics of DNA methylation during mammalian development is mentioned in the following section.

1.2.9 The dynamics of CpG methylation during mammalian development

Mammalian embryos undergo two phases or “waves” of DNA methylation reprogramming (Figure 1.11). The first phase of reprogramming ensues after fertilisation and removes the maternal and paternal DNA methylation patterns inherited from oocyte and sperm. This is assumed to contribute to creating a permissive/plastic epigenetic landscape to acquire plasticity during development. There is evidence that maternal and paternal genomes are demethylated via different mechanisms. The paternal genome undergoes an initial phase of active demethylation, for which the action of the TET3 enzyme was originally proposed (Shen et al., 2014). However, more recent evidence suggests an unknown mechanism for active paternal genome demethylation independent of TET3 (Amouroux et al., 2016). Later both the paternal and maternal genomes undergo passive demethylation. Passive demethylation involves the exclusion of DNMT1 from the nucleus and the dilution of DNA methylation during successive cellular division (Howell et al., 2001). The role of DPPA3/STELLA is well-studied for causing the exclusion of DNMT1. A recent study shows that DPPA3 binds to UHRF1, which is required for DNMT1 targeting, and displaces it from chromatin and then the nucleus (Mulholland et al., 2020). Some studies have reported exceptions to this largely agreed model of paternal and maternal genome

demethylation. 5hmC, a product of 5mC oxidation, was also found in the maternal genome indicating active demethylation (Wang et al., 2014). Similarly, passive demethylation of the paternal genome was pointed to in other studies (Guo et al., 2014; Shen et al., 2014).

Upon reaching the blastocyst stage, embryos attain a low point of DNA methylation. It is estimated that 20% of CpGs resist this demethylation. These CpGs comprise regulatory regions of imprinted genes called imprinting control regions (ICRs) and young transposons (Lane et al., 2003; Wang et al., 2014). Both of these categories for demethylation resisting loci are thought to be protected by KRAB-ZFP and then KAP1 binding mechanisms (Messerschmidt et al., 2012; Quenneville et al., 2011).

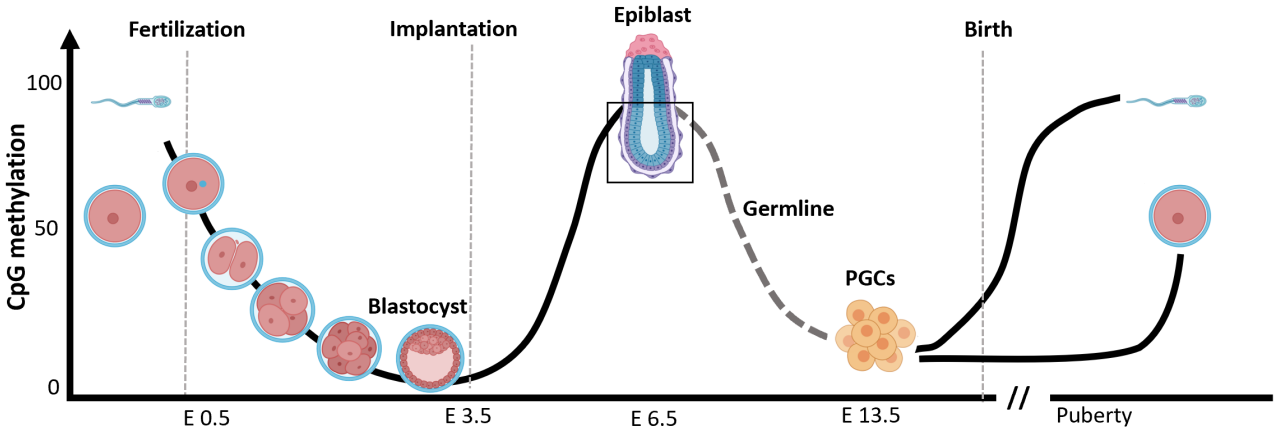


Figure 1.11: Dynamics of CpG methylation during mouse development.

CpG methylation changes dynamically through embryonic development from fertilisation to birth (see text for details).

After implantation, *de novo* methyltransferases DNMT3A, DNMT3B and DNMT3L are expressed, re-establishing global DNA methylation. At E4.5 the blastocyst is composed of inner cell mass (ICM, which subsequently forms epiblast), and trophectoderm and primitive endoderm which form extraembryonic tissues. Extraembryonic tissues are hypomethylated in comparison to epiblast likely due to lower expression of DNMT3 proteins (Zhang et al., 2018b). The re-methylation of epiblast is rapid and by E6.5, methylation levels are comparable to somatic cells (Seisenberger et al., 2012; Wang et al., 2014).

Primordial germ cells (PGCs) that give rise to germ cells in the embryo, undergo a second wave of demethylation. Here too, young retrotransposons, such as IAP and LTR elements, resist demethylation (Guibert et al., 2012; Hajkova et al., 2002). In contrast to preimplantation demethylation, this phase

ensures the removal of inherited imprints so that gamete-specific DNA methylation patterns can be established. The germ line demethylation occurs in two stages: the first passive demethylation which leads to global demethylation, sparing imprinted loci, and then a second phase of TET1- and TET2-dependent demethylation, mainly targeting imprinted loci and germ-line-specific genes (Hackett et al., 2013; Vincent et al., 2013; Yamaguchi et al., 2013a; Yamaguchi et al., 2012; Yamaguchi et al., 2013b). Male gametes gain methylation by the action of DNMT3L and DNMT3A, and DNMT3C in rodents. Before birth, male gametes have already accumulated high levels of DNA methylation. The oocyte gains methylation before ovulation also by DNMT3A and DNMT3L. At the end of gametogenesis, both male and female germ cells have distinct methylation patterns; spermatogonia are 80% methylated while oocytes are 50% methylated (Smallwood et al., 2011). In oocytes the methylated regions reside within transcribed regions (Chotalia et al., 2009; Veselovska et al., 2015) and depend upon H3K36me3 depositing enzyme, SETD2, that is thought to recruit DNMT3A and DNMT3L (Xu et al., 2019). Recent evidence suggests that the transcription required for female germline methylation pattern establishment, occurs from alternative transcription start sites residing in the LTR elements located upstream of the methylated regions (Bogutz et al., 2019; Brind'Amour et al., 2018). In contrast to oocytes, most of the sperm genome is methylated except for CpG-rich/H3K4me3 modified regions. Recently, it has been shown that the sperm genome methylation is assisted by the activity of the histone methyltransferase NSD1, which deposits H3K36me2 and mediates DNMT3A/3L recruitment (Shirane et al., 2020).

Germline DNA methylation patterns are of importance as the DNA methylation patterns in gametes regulate the gene expression in the next generation. Failure in the establishment of proper DNA methylation patterns in PGCs have profound consequences on the post-fertilization development and survival. As development proceeds, different somatic cell types acquire distinct methylation patterns. It has been suggested that during cell fate determination in later development, cells can undergo smaller waves of methylation and demethylation, mostly targeting tissue-specific enhancers and other gene regulatory regions (Rasmussen et al., 2019; Sardina et al., 2018).

1.2.10 Functional importance of developmentally regulated DNA methylation

The establishment of the germline DNA methylation pattern serves the crucial role of silencing transposons and ensuring the correct expression of imprinted genes. Defence against retrotransposons during gametogenesis is critical as their activity can adversely impact the fitness of

the next generation. Male germ cells undergo heavy methylation to minimize the transposition of IAPs. The loss of *Dnmt3l* is known to cause sterility in male mice, indicating the significance of DNA methylation and transposon silencing in reproduction (Bourc'his et al., 2001). In rodents, DNMT3C functions specifically to silence transposable elements during spermatogenesis. DNMT3C uses a small RNA- (piwi RNA) directed pathway to silence young transposable elements (Barau et al., 2016). Homozygous mutation in *Dnmt3c* gene in mice is viable but causes reactivation of retrotransposons, small testes and infertility in males (Barau et al., 2016).

Imprinting is the parent-of-origin-dependent monoallelic expression of a gene, which is largely regulated by epigenetic chromatin modifications, including DNA methylation (reviewed in (Tucci et al., 2019)). As described above, these epigenetic patterns are established in, and then inherited from the parental germline. Although DNA methylation was originally considered the main epigenetic modification for maintaining imprints, recently H3K27me3-silenced maternal imprints have also been reported, termed as non-canonical imprinting (reviewed in (Chen and Zhang, 2020)). Imprinted genes tend to occur in clusters and are usually involved in growth and development. Misregulation of imprinted genes in humans causes specific developmental disorders such as Prader-Willi syndrome, Beckwith-Wiedemann syndrome, and Silver-Russell syndrome (reviewed in (Butler, 2009)).

1.2.11 CpG islands (CGIs)

CGIs are ~1-2 kb long CpG-rich DNA sequences associated with approximately 70% of gene promoters in vertebrates (Bird et al., 1985). CGIs constitute the most common promoter type in vertebrates, encompassing housekeeping genes, tissue-specific genes and developmental regulator genes (Larsen et al., 1992; Saxonov et al., 2006). In humans and mice, a class of CGIs is also present inter- and intragenically, and these have been proposed to possess features of promoters such as H3K4me3 enrichment, low-level transcription and co-localization with RNA Pol II at transcription initiation sites (Illingworth et al., 2010). Bird and colleagues described that CpG islands are mostly devoid of DNA methylation and suggested that this is the main reason for the preservation of their high GC content, unlike the rest of the genome which has reduced GC content owing to spontaneous C ->T deamination (Bird et al., 1985).

The early model presented to explain the unmethylated status of CGIs proposed that they are bound by proteins, such as TFs, providing protection against methylation (Bird, 1986). Over the years, CGIs have indeed been shown to serve as binding sites for myriad of proteins. A study, using transient

transfection reporter gene assays for examining activity of human TFs, showed that CGIs are enriched for recognition motifs for SP1, NRF-1, E2F, and ETS (Landolin et al., 2010). As mentioned above, CGIs are also the binding sites for the Polycomb and Trithorax complexes. MLL1 and MLL2, the mammalian orthologues of *Drosophila trithorax*, bind CGIs via their CXXC domains (Cierpicki et al., 2010; Xu et al., 2018). CGIs are also recognized by CFP1 (CXXC finger protein 1), a subunit of the H3K4me3 methyltransferase SETD1, via its CXXC domain. CFP1 is thought to play a key role in the recruitment of SETD1 complex to CGIs in mouse. Polycomb complex PRC1 is targeted to CGIs via its interaction with CXXC domain-containing protein KDM2B (Farcas et al., 2012). KDM2B is known to bind most of the CGIs and does not always recruit PRC1. Insertion of synthetic CGI-like elements in mouse cells also results in the recruitment of PRC2 and CFP1 (Mendenhall et al., 2010; Thomson et al., 2010). Thus, CGIs form important regulatory elements, controlling gene expression by the recruitment of PcG and TrG proteins. This type of regulation is more reversible and plastic in comparison to DNA methylation.

The predominant hypomethylated status of CGIs is accredited to the active transcription and accumulation of H3K4me3. Low nucleosome occupancy (Ramirez-Carrozzi et al., 2009) and acetylation of H3 and H4 histones make CGIs permissive to active transcription. The binding of DNMT enzymes has been proposed to be sterically inhibited due to the occupancy of TFs on CGIs. Ablation of binding sites for TFs like SP1 have been shown to lead to methylation of the *Aprt* gene promoter CGI (Brandeis et al., 1994; Macleod et al., 1994). Similarly, binding of TFs like SP1, NRF-1, and YY1 to CGIs in monocytes correlates with the protection of their target CGIs from methylation (Gebhard et al., 2010). Accumulation of H3K4me3 on CGIs is also thought to occur as a result of active transcription (Brown et al., 2017) which can inhibit DNA methylation due to the autoinhibition of DNMTs by ADD domains. Active transcription however does not seem to be the only determinant for the maintenance of CGIs in a hypomethylated state since CGI methylation does not occur on a number of promoters of silent genes. Hypomethylation of CGIs, therefore, has also been proposed to exist due to continuous demethylation by the action of TET enzymes and there is some evidence indicating that TET1, which has a CXXC domain and localizes to CGIs (Wu et al., 2011).

1.2.12 Methylation of CGIs

Despite the prevalent hypomethylated status of CGIs, there are some exceptions. CGIs of the germline-specific genes, which function in the differentiation of germ cells, constitute 5% of total CGIs in the mouse genome. These CGIs resist complete demethylation in the preimplantation embryo and later gain methylation in a DNMT3B-dependent manner at the onset of somatic differentiation in post-

implantation embryos (Auclair et al., 2014; Borgel et al., 2010). The loss of *Dnmt3b* has also been shown to cause a modest derepression of these germline-specific genes in mouse embryos (Auclair et al., 2014). These CGIs are known to be targeted by a subtype of PRC1 complex, PRC1.6. PRC1.6 contains sequence-specific DNA-binding proteins MGA, MAX and E2F6, which have been implicated in the targeting of PRC1.6 to germline-specific genes (Endoh et al., 2017). E2F6 has been shown to interact with DNMT3B and proposed to target it to CGIs of germline-specific genes (Velasco et al., 2010). PRC1.6 is also known to associate with L3MBT2 which in turn interacts with the H3K9me2 methyltransferase G9A. G9A has been shown to be specifically targeted to CGIs of germline-specific genes and the loss of *G9a* leads to a reduction in CGI methylation at these loci in mouse embryos (Auclair et al., 2016).

Analysis of the CGI methylation in mouse embryos has shown that CGIs of genes important in development, morphogenesis and neural development are also methylated in some cases. Examples of these loci include key TFs like *Klf3, Foxo3, Cux1, Zfp64, Daxx* and several genes in Wnt/β catenin pathway (Auclair et al., 2014). Furthermore, tissue-specific methylation of CGIs has also been reported. For instance, CGIs of *Sec31b* that is involved in vesicular trafficking, is methylated in blood and spleen. Similarly, CGIs of genes determining early development, such as *Hox*C cluster and *Pax6,* are also found methylated in somatic cells (Illingworth et al., 2008).

A further example of CGIs that gain methylation is the ICRs that regulate the expression of imprinted genes. ICRs contain CGI-like elements and work in-*cis* to cause monoallelic expression. ICRs can efficiently regulate the monoallelic expression over large genomic areas comprising tens of kilobases, by functioning via two main mechanisms: (1) expression of long noncoding RNA (lncRNA) and (2) through chromatin insulation. An example of the former is the *Airn/Igf2r* locus, which spans 10 Mb and is regulated by the expression of the lncRNA Airn from the paternal allele. It was proposed that transcription of lncRNAs, rather than the RNA itself, causes silencing over such long genomic loci. The expression of *Igf2r*, for example, is interrupted due to the transcription of Airn from the opposite strand, rather than Airn RNA itself (Latos et al., 2012). It was also proposed that the transcription of Airn could ablate promoter-enhancer interactions which could repress genes located farther from the ICR. However, deletion of *Airn* gene on the maternal allele did not affect the expression of genes in *Airn/Igf2r* locus like *Slc22a3* and *Tcp1*, arguing against the presence of crucial promoter-enhancer interactions. On the other hand, the inhibition of paternal expression of Airn RNA led to the derepression of these genes, indicating that the Airn RNA is required to repress the genes in the locus (Andergassen et al., 2019a). Airn and potentially other lncRNAs have been proposed to recruit PRC1/2 and G9A which catalyse H3K9me2. The physical interaction between the RNA and these complexes

has been shown in pulldown experiments (Nagano et al., 2008; Schertzer et al., 2019; Zhao et al., 2010).

Chromatin insulation among imprinted loci is best exemplified by the *Igf2/H19* locus. The hypomethylated maternal allele binds CTCF, which insulates the interaction between an upstream enhancer and *Igf2* promoter. The enhancer instead interacts with *H19* promoter and allows the expression of H19. On the paternal allele, methylation prevents CTCF binding, allowing for expression of *Igf2* (Bell and Felsenfeld, 2000) (Figure 1.12).

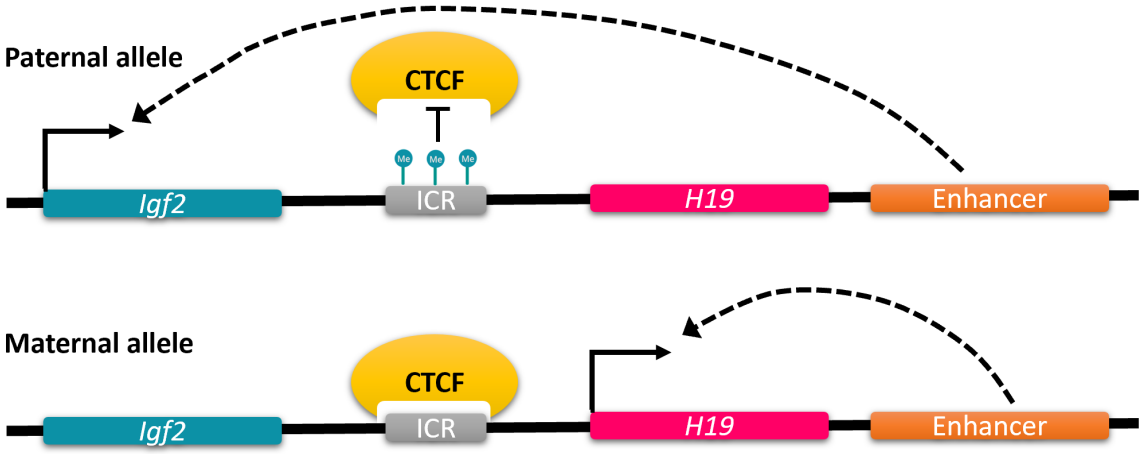


Figure 1.12: Methylation of ICR controls allele-specific expression of imprinted *Igf2/H19* locus.

Methylation of the *Igf2/H19* locus ICR inhibits the binding of the insulator protein CTCF on the paternal allele, allowing the upstream enhancer to activate the transcription of *Igf2*.

A final example of CGIs that become methylated is those present on the inactive X (Xi) chromosome in female mammals. This process is the central focus of my thesis and is discussed in more detail below.

1.3 X Chromosome Inactivation (XCI)

Barr and Bertram (1949) first observed a difference between the nuclei of neurons from male and female cats; upon the staining of nuclei with basic dyes, a dense chromatin structure was only visible in neurons of female cats (Barr and Bertram, 1949). It was later named after its discoverer as the "Barr body" and was shown to consist of a condensed allocyclic X chromosome (Ohno & Hauschka, 1960). In 1961, Mary Lyon first highlighted the functional relevance of the Barr body and proposed that one X chromosome in female mammals is transcriptionally silenced, termed X chromosome inactivation (XCI) (Lyon, 1961).

Subsequent work has shown that XCI occurs early in mammalian embryogenesis and then is maintained during the entire life span. This process ensures an equal dosage of X-linked genes between male and female mammals, which inherit one and two X chromosomes, respectively. Dosage compensation mechanisms originated among organisms in which sex determination occurs via specialized sex chromosomes. The exact mechanisms to achieve equal dosage vary among organisms. For instance, in *Drosophila* the male X chromosome is hyperacetylated to allow two-fold expression of X-linked genes in XY males and equate the dosage to XX female counterparts (reviewed in (Conrad and Akhtar, 2012)). Even among mammals the strategies to silence the X chromosomes differ. For instance, in marsupials, imprinted XCI occurs where the paternal X chromosome is inactivated (reviewed in (Deakin et al., 2009)). In mice, XCI occurs in two phases during development. At 2-4 cell stage embryo (E1.5-2.5), the paternal X chromosome is silenced in all cells of the embryo (termed imprinted XCI or iXCI), and this silenced state is maintained in extraembryonic tissues. However, in the early blastocyst's (E3.5) ICM, that gives rise to the embryo proper, the paternal X chromosome is reactivated, followed by a second phase of random X inactivation (rXCI) of either paternal or maternal X chromosomes shortly after implantation (E5.5) (reviewed in (Brockdorff et al., 2020)). rXCI can be recapitulated in mouse ES cells derived from the blastocytes, using *in vitro* differentiation methods, making them a widely used model system to study XCI (Navarro et al., 2008).

The master regulator of XCI is a 17 kb lncRNA called Xist (X-inactive specific transcript) (Brockdorff et al., 1991; Brown et al., 1991). Xist acts in *cis* and accumulates/spreads along the X chromosome from where it is expressed. Xist RNA orchestrates a number of downstream processes such as the exclusion of RNA Pol II, recruitment of silencing complexes such as PRCs, and eventually causing long-range chromatin conformation changes. All of these processes culminate in changing the chromatin

landscape of X chromosome to repressed heterochromatin forming the “Barr body”, - the observation of which led to the discovery of XCI (reviewed in (Loda et al., 2022)). XCI can be seen as a stepwise process which can be divided into three main stages: (1) Initiation – marked by the expression of Xist, (2) establishment – involving recruitment and spread of complexes to bring about transcriptional silencing and (3) maintenance – mechanisms that lock transcriptional silencing in mitotically stable state (Figure 1.13). A brief account of these steps is presented below.

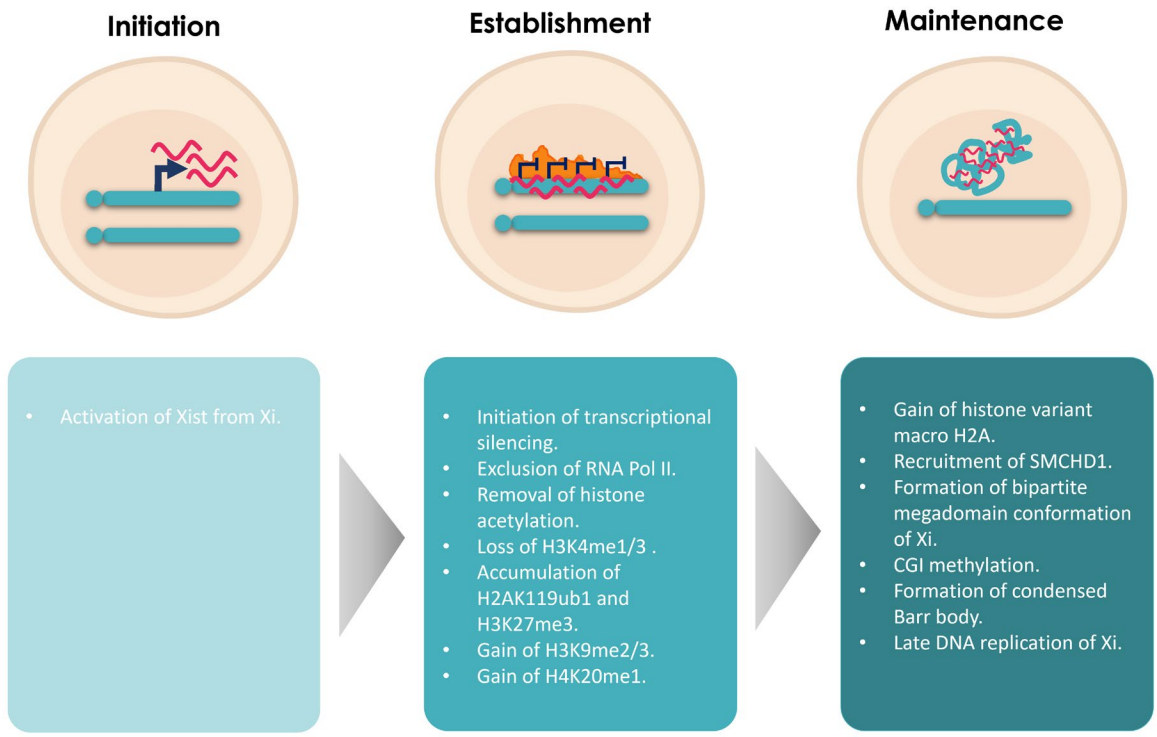


Figure 1.13: Phases of XCI

XCI can be divided into three phases based on the timing of the appearance of hallmark features. The process initiates with the expression of lncRNA Xist which spreads along the Xi. The processes / chromatin modifications occurring during the establishment phase are regarded as early features of XCI and occur within 24 hours. The maintenance phase ensues several days after initiation of XCI and includes SMCHD1 recruitment and CGI methylation which require cell differentiation (Bowness et al.,2022).

1.3.1 Initiation of XCI

Initiation of XCI occurs with the expression of Xist, controlled by a large regulatory region, that spans several hundred kilobases on the X chromosome, termed the X inactivation centre (Xic). The Xic has been identified as the minimum region which is sufficient to trigger XCI in mice and humans (reviewed in (Galupa and Heard, 2018)). The region spanning Xic was narrowed down from a larger candidate region defined by the study of balanced X-autosomal translocations and series of X chromosome truncations in female ES cells and transgenic mice (Cattanach, 1983; Cattanach et al., 1989; Keer et al., 1990; Rastan, 1983; Rastan and Robertson, 1985; Russell, 1983). The Xic harbours the gene for lncRNA Xist, and several other lncRNAs that regulate the expression of Xist. Expression of Xist is crucial for both iXCI and rXCI. In mice Xist is 17 kb long, undergoes capping, splicing and is polyadenylated, but nevertheless bypasses export out of the nucleus. Upon expression, Xist spreads along the X chromosome forming an accumulation that can be visualised by RNA FISH (Fluorescence in *situ* hybridization). Accumulated Xist is sometimes referred to as an Xist "domain" or "cloud".

Xist expression and XCI are linked with pluripotency. Pluripotency factors such as KLF4, NANOG, OCT4, SOX2 and PRDM14 are known to act as negative regulators of Xist expression directly or indirectly by repressing the activators of XCI (Ma et al., 2011; Navarro et al., 2008). Direct repression by NANOG, SOX2 and OCT3/4 is caused by their binding to the genomic region corresponding to first intron of *Xist* gene which is considered to act as an enhancer of Xist expression. *Xist* intron 1 has also been shown to bind other pluripotency regulators like TCF3, PRDM14, and developmental regulators like CDX2 (Erwin et al., 2012; Ma et al., 2011). Repression of OCT3/4 was observed to cause upregulation of Xist in a proportion of male ES cells (Navarro et al., 2008). On the other hand, deletion of *Xist* intron 1 in mouse ES cells showed that it is dispensable for the repression of Xist, upregulation of Xist during differentiation or silencing of Xist during reprogramming to pluripotent state (Minkovsky et al., 2013).

Xist expression is also regulated, both positively and negatively by lncRNAs residing within Xic (Figure 1.14). The most well-known of these is Tsix, which is expressed from the overlapping region on the opposite strand of Xist and acts as an antisense negative regulator of Xist. Tsix is expressed from both the X chromosomes in ES cells but its expression from the Xi is downregulated upon differentiation (reviewed in (Maclary et al., 2013)). In mouse embryos and ES cells, heterozygous mutations in *Tsix* cause a bias in the choice of Xi, with the allele bearing *Tsix* mutation being chosen for inactivation (Sado et al., 2001). Expression of Tsix has also been proposed to help protect the maternal X chromosome from undergoing inactivation during iXCI. Mutations in the maternal *Tsix* gene lead to mouse embryo lethality due to poor development of the extraembryonic tissues where paternal X chromosome is silenced due to iXCI (Lee, 2000; Sado et al., 2001). However, expression of Tsix alone

is not sufficient to inhibit Xist expression, as the mutations in *Tsix* do not always lead to Xist induction. In undifferentiated ES cells, mutations in the *Tsix* promoter lead to only weak Xist upregulation in a proportion of cells (Sado et al., 2002).

Jpx (just proximal to Xist; also called Enox, expressed neighbour of Xist) is another lncRNA located 10 kb upstream of Xist TSS and is known to function as an activator of Xist. Jpx is upregulated with Xist induction in differentiating female ES cells. Deletion of *Jpx* in female ES cells compromised Xist upregulation with fewer cells showing the Xist clouds/domains, increased cell death and abnormal phenotype upon differentiation (Tian et al., 2010). Jpx has been suggested to antagonize Tsix. Mutation in *Tsix* restores Xist expression and rescues cell death observed in differentiated female cells bearing deletion in *Jpx* (Tian et al., 2010). A recent report has shown that *Jpx* knockout mice are viable and undergo normal XCI, implying that the role of Jpx is limited (Collombet et al., 2020). The lncRNA Ftx (five prime to *Xist*) is located further upstream of *Xist* and is also a positive regulator of Xist (Chureau et al., 2011). Mutation of *Ftx* in ES cells results in downregulation of Xist and this is correlated with hypermethylation of the *Xist* CGI (Chureau et al., 2011). Deletion of *Ftx* alone has been shown to be dispensable for iXCI, but the deletion of both *Jpx* and *Ftx* on the maternal allele were shown to compromise female viability (Collombet et al., 2020). Recently another lncRNA called Xert (Xist-enhancing regulatory transcript) has been described to be expressed from 200 kb upstream of *Xist* (Gjaltema et al., 2022). Xert has been shown to activate Xist transcription by forming an in *cis* regulatory hub that interacts with the Xist promoter in 3D space (Gjaltema et al., 2022).

The Xic also harbours negative regulators of Xist. Xite (X-inactivation intergenic transcription elements) is a positive regulator of Tsix. Deletion of *Xite* was also shown to increase probability of the allele bearing the mutation, to get inactivated (Ogawa and Lee, 2003). Another positive regulator of Tsix is Tsx (testis-specific X-linked). Deletion of *Tsx* was shown to cause downregulation of Tsix levels and partial derepression of Xist in male and female ES cells. *Tsx* null mice however, are viable (Anguera et al., 2011)

Apart from the lncRNAs, an E3 ubiquitin ligase, RING finger protein 12 (RNF12) is a key regulator of XCI. RNF12 is expressed from 500 kb upstream of *Xist* and acts in *trans* to upregulate Xist in a dose dependent manner (Jonkers et al., 2009). Additional copies of *Rnf12* cause X inactivation in male mouse ES cells, and inactivation of both X chromosomes in a large proportion of female mouse ES cells (Jonkers et al., 2009). Heterozygous deletion of *Rnf12* markedly reduces initiation of XCI in differentiating XX mouse ES cells (Jonkers et al., 2009). RNF12 causes proteasomal degradation of a pluripotency factor REX1. REX1 binds to the *Xist* and *Tsix* regulatory regions and serves as a negative

regulator of Xist. Overexpression of *Rex1* leads to reduced Xist expression while heterozygous mutations of *Rex1* cause ectopic XCI in female mouse ES cells (Gontan et al., 2012).

Recent studies have shown that the Xic in mouse forms two TADs. *Xist* lies at the border of the two TADs, with the promoter and activators of Xist such as *Jpx, Ftx* and *Rnf12* residing in one TAD, all of which are upregulated during differentiation. The genes encoding negative regulators of Xist such as *Xite* and *Tsix*, which get downregulated upon differentiation, are present in another TAD (Figure 1.14). It has been suggested that the organization of Xist and Tsix in TADs allows them to efficiently interact with their regulators and segregates them from the antagonistic effect of neighbouring genes (Nora et al., 2012). Deletion of 58 kb region on the TAD boundary led to aberrant expression from both TADs, indicating that the 'insulation' segregating the two TADs was lost (Nora et al., 2012). Similarly, inversion of the TAD boundary lead to ectopic expression of Xist in ES cells and prolonged expression of Tsix during differentiation ((Van Bemmel et al., 2019) and reviewed in (Loda et al., 2022)). A recent study described that the promoter of lncRNA large intervening transcript in the Xic (Linx, promoter denoted as *LinxP*) that resides within the Tsix TAD, acts as a long range negatitive regulator of Xist. Interestingly however, when introduced in the Xist TAD, *LinxP* acts as an enhancer of Xist indicating the significance of TAD partitioning of Xic in XCI (Galupa et al., 2020).

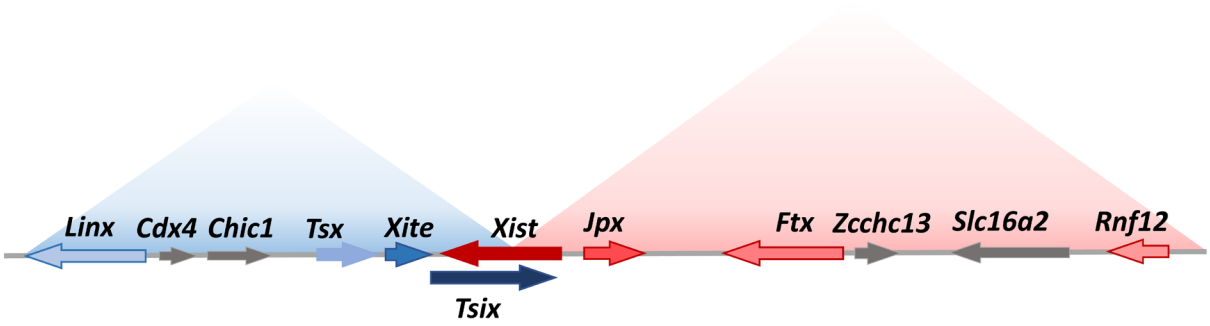


Figure 1.14: Schematic showing X inactivation centre (Xic) in mouse.

Xic harbours several lncRNAs, which both positively and negatively regulate Xist expression. The positive regulators of Xist are shown in shades of red and reside in 500 kb large 'Xist-TAD' which is insulated from the 300 kb 'Tsix-TAD' which contains *Tsix* and other negative regulators or Xist. RNF12 encodes an E3 ubiquitin ligase which causes proteasomal degradation of the pluripotency factor REX1, implicated in Xist repression. Other genes that exist in the Xic but are not involved in Xist regulation are shown in grey.

1.3.2 Establishment of XCI

1.3.2.1 Xist repeat elements and interacting proteins

Xist RNA sequence is composed of 6 structured repeat elements (A-F), which are known to serve as a scaffold, allowing docking, assembly, and concentration of protein complexes that then aid in the coating of Xist on Xi and establishment of transcriptional silencing (Figure 1.15, reviewed in (Pintacuda et al., 2017b)). Xist tandem repeats are conserved in mammalian vertebrates (Brockdorff et al., 1992; Nesterova et al., 2001). Wutz et al, assigned function to specific repeat elements by creating a series of deletions in *Xist,* in XY ES cells with inducible *Xist* transgene (Wutz et al., 2002). A key finding was that the A-repeat is essential for Xist-mediated chromosome silencing. Protein interactors that recognise Xist were unveiled in studies employing RNA-protein cross-linking approaches and genetic screens (Chu et al., 2015; Lu et al., 2016; Moindrot et al., 2015; Monfort et al., 2015). A brief account of these proteins is presented below.

A-repeat element of Xist, located near the 5' end, is the most conserved region of Xist RNA in terms of sequence and copy number (Nesterova et al., 2001). A-repeat has been shown to bind SPEN (Split End, also known as MINT and SHARP) (Cirillo et al., 2017; Lu et al., 2016; Monfort et al., 2015). SPEN has N-terminal RNA recognition motifs (RRM) and a C-terminal protein interacting SPEN Paralog and Ortholog C-terminal (SPOC) domain, both of which have distinct contributions in XCI. RRMs allow direct interaction between SPEN and Xist A-repeat. Recently a study from our lab has shown that deletion of the SPEN RRM domain leads to significant defects in the localization of Xist on Xi resulting in failure to form a proper Xist cloud (Rodermund et al., 2021). SPEN has also been implicated in local accumulation of Xist via its ability for homotypic assemblies (Jachowicz et al., 2022; Markaki et al., 2021). The SPOC domain of SPEN contributes to gene silencing via its interaction with SMRT/NCoR-HDAC3 histone deacetylase complex (Ariyoshi and Schwabe, 2003). Defective silencing in SPEN mutants can be observed within the first 24 hours of Xist expression, demonstrating that SPEN as an early effector protein in XCI (Bowness et al., 2022). In agreement with this observation, the deacetylation of H4 and H3K27 have been reported to be among the earliest chromatin modifications to be established upon the expression of Xist (Żylicz et al., 2019).

Xist B- and C-repeats are bound by HNRNPK, which then recruits Polycomb repressive complexes PRC1 and PRC2 to establish H2AK119ub1 and H3K27me3 histone modifications. The resulting histone modifications can be visualised on Xi by fluorescent antibody staining and imaging and can serve as a proxy for Xist RNA. Depletion of HNRNPK leads to loss of H2AK119ub1, and artificial tethering of

HNRNPK to mutated Xist lacking B- and C-repeats restores Polycomb recruitment. PCGF3/5-PRC1 functions as the initiating complex for Polycomb recruitment to Xi (Almeida et al 2016), and HNRNPK directly interacts with the PCGF3/PCGF5 subunit (Pintacuda et al., 2017a). Similar to SPEN, PRC1/2 recruitment and the resulting chromatin modifications occur very early in XCI and can be observed as soon as 6 hours after the onset of Xist expression. Polycomb and SPEN pathways together work additively to silence genes on the X chromosome (Bowness et al., 2022).

Other interactors of Xist A-repeat include components of N6-methyladenosine (m6A) machinery such as WTAP, and RBM15 which is an accessory factor of m6A-establishing METTL3/METTL14 complex (Chu et al., 2015). Studies conducted at the Brockdorff lab have shown that disruption of m6A machinery (with deletions of RBM15 , METTL3/ METTL14) or deletions of m6A sites on Xist do not lead to significant defects in X chromosome silencing (Coker et al., 2019). Since m6A has been shown to regulate various aspects of RNA regulation and processing, it is plausible that m6A modifications participate in XCI by regulating Xist stability (Wei et al., 2021).

Nuclear laminar protein LBR (Lamin B Receptor) is an inner nuclear membrane protein that has been reported to mediate the binding of Xist to nuclear lamina and localize Xist domain with heterochromatin regions in the nuclear periphery (Chen et al., 2016). LBR binds predominantly to a site located 0.2 kb downstream of A-repeat in addition to several other sites across Xist (Chen et al., 2016). Mutations in *Lbr* show only minor defect in gene silencing during XCI in mouse ES cell system (Nesterova et al., 2019). Similarly, mutations of *Lbr* have been shown to lead to no defects in XCI in mice (Young et al., 2021).

CIZ1 (CDKN1A-interacting zinc-finger protein 1) is another well-studied Xist interactor (Sunwoo et al., 2017). CIZ1 is highly enriched on the Xist domain (as viewed by microscopy) and binds to Xist E-repeat (Sunwoo et al., 2017). The interaction between CIZ1 and Xist has been proposed to occur via zinc finger domains located at the C-terminus (Ainscough et al., 2007; Warder and Keherly, 2003). A recent publication has shown that E-repeat binds other proteins that work in RNA processing such as PTBP1, MATR3, TDP-43 and CELF1, which via self-aggregation form condensates and contribute to gene silencing during the transition of XCI from Xist-dependent to Xist-independent stage (Pandya-Jones et al., 2020).

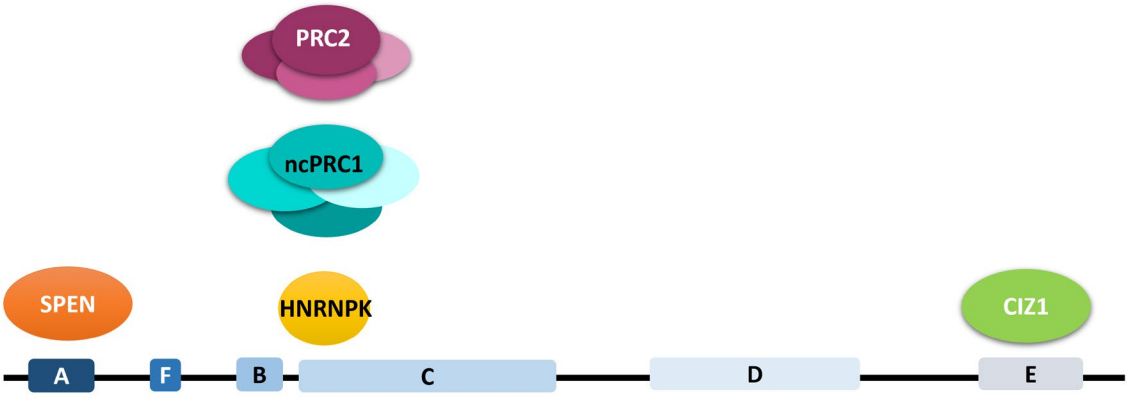


Figure 1.15: Schematic showing Xist repeat elements and RNA binding proteins

Xist is a 17 kb long lncRNA containing conserved repeat elements A-F. These repeat elements are bound by RNA binding proteins which have a critical role in the localization of Xist and / or establishment of XCI. HNRNPK leads to the recruitment of ncPRC1 and PRC2. See text for details.

1.3.2.2 Xist spreading

Mapping of chromosomal sites with which Xist associates using RNA cross-linking has shown that Xist first accumulates at the site of its own transcription and subsequently spreads to "entry sites" distributed along the Xi. These entry sites correlate with regions depicting relatively high 3D contact frequency with the *Xist* locus (Engreitz et al., 2013). They are generally gene-rich regions, although spread to gene-poor regions appears to occur subsequently (Simon et al., 2013). Following Xist RNA dynamics over time using super-resolution microscopy has revealed two phases of Xist spreading. (1) An early expansion phase marked by a high number of Xist molecules and enlargement of Xist domain, followed by a (2) steady-state phase with greater stability/lower turnover of Xist RNA and no further increase in the radial Xist domain (Rodermund et al., 2021).

Some Xist binding proteins have been implicated in mediating Xist localization to Xi. A nuclear matrix protein HNRNPU (also known as SAF-A for scaffold attachment factor A) is one of the earliest characterized Xist interactors known for its role in anchoring Xist to Xi (Hasegawa et al., 2010). HNRNPU shows a broad and nonspecific binding across Xist (Cirillo et al., 2017). HNRNPU is crucial for Xist localization on Xi as the deletion of *Hnrnpu* or its RNA-binding domains, causes Xist to diffuse throughout the nucleoplasm (Hasegawa et al., 2010). CIZ1 is another protein factor well-studied for its role in Xist localization. The role of CIZ1 and E-repeat is thought to become significant in later stages of XCI. Deletions in *Xist* E-repeat were shown to result in reduced efficiency to form clouds only in late differentiation stage of XX cells (Yamada et al., 2015). The deletion of *Ciz1* in mouse embryonic

fibroblasts (MEFs) leads to the dispersion of the Xist cloud (Sunwoo et al., 2017). An imaging study from our lab has also shown the role of CIZ1 in Xist localization in differentiated cells (Rodermund et al., 2021). Other protein factors binding the E-repeat such as PTBP1, MATR3, TDP-43 and CELF1, have also been suggested to help retain Xist on Xi during late XCI (Pandya-Jones et al., 2020).

1.3.2.3 Gene silencing during early XCI

Xist-mediated silencing is largely attributable to the SPEN and Polycomb pathways (Bowness et al., 2022). The silencing defect observed Wutz et al., upon deletion of the A-repeat deletion is recapitulated/comparable to SPEN loss of function, indicating SPEN recruitment to be the principle pathway working via Xist A-repeat (Nesterova et al., 2019). SPEN has been shown to enrich/localize on Xi soon after its induction and becomes dispensable to gene silencing in late stages specifying its role as an early effector in XCI (Dossin et al., 2020). As mentioned above, SPEN ensures efficient localization of Xist on Xi by RRM domain (Rodermund et al., 2021) and contributes to gene silencing via its interaction with SMRT/NCoR-HDAC3 through the SPOC domain (Ariyoshi and Schwabe, 2003). In agreement with this observation, knockout or inhibition of HDAC3 has been shown to lead to a defect in Xist-mediated gene silencing (Żylicz et al., 2019). However the genes showing silencing defects upon the mutation in HDAC3 represent only a subset of those affected by SPEN loss-of-function or SPOC domain mutation, suggesting additional pathways for SPEN-mediated gene silencing during XCI (Żylicz et al., 2019). A recent study by our lab has investigated the relative contribution of the two SPEN domains in X chromosome silencing and shown that mutation of the SPOC domain shows an incomplete silencing deficiency relative to the loss of function mutations or deletion of the A-repeat. On Xi the SPOC-dependent silencing is largely limited to genes showing high expression levels in ES cells (Bowness et al., 2022). The SPEN SPOC domain has also been shown to interact with WTAP, members of NuRD complex, and RNA Pol II (Dossin et al., 2020). While an overlap between the binding sites for SPEN, RNA Pol II associated with transcription initiation (phosphorylated at serine 5) and members of NuRD complex (CHD4 and MBD3) on Xi has also been reported (Dossin et al., 2020), the underlying mechanisms via which SPEN mediates gene silencing through these interactors awaits further investigation.

A recent study from our lab has described a complementary/synergistic role for SPEN and Polycomb pathways in gene silencing during XCI. SPEN brings about the silencing of genes expressed at high levels in the ES cells whereas Polycomb silenced genes, tend to be lowly expressed (Bowness et al., 2022). The mechanisms by which PRC1/2 and associated histone modifications cause silencing are still being researched, but several models have been presented (described in review article (Blackledge and Klose, 2021)). Due to the ability of protein components of PRC1/2 to bind the histone modifications established by the complex itself, PRC1/2 were suggested to create large, repressed domains by spreading the histone modification via a self-propagating mechanism. These Polycomb domains can lead to chromatin compaction and reduced accessibility for transcription. In support of this model, the deletion of the B-repeat was shown to lead to reduced compaction of Xi domain as compared to wildtype in ES cells (Markaki et al., 2021). The addition of the large ubiquitin protein on nucleosomes at gene regulatory elements might hinder the function of transcription machinery directly by antagonizing transcription permissive chromatin states like H3K4me3 (Nakagawa et al., 2008) or alternatively, recruit reader proteins that then silence the target genes (Jansz et al., 2018b). A recent study shows that loss of PRC1 leads to enhanced RNA Pol II initiation, suggesting that PRC1 and H2AK119ub1 counteract this step of transcription (Dobrinić et al., 2021). More studies are needed to give a clearer picture of the underlying mechanistic details of Polycomb-mediated gene repression.

1.3.3 Maintenance of XCI

Once established, XCI and gene silencing is stable/robust such that in somatic cells the silenced state of Xi is maintained even without ongoing expression of Xist (Csankovszki et al., 2001). Upon achieving transcriptional silencing, an XCI maintenance phase ensures that Xi is stably retained over mitotic divisions during the lifespan of the organism. These maintenance pathways include deposition of histone variant macroH2A, recruitment of SMCHD1 (structural maintenance of chromosomes hinge domain containing 1), deposition of H3K9me3, CGI methylation, and large-scale chromatin structure reorganization.

1.3.3.1 Recruitment of SMCHD1

The silencing dynamics of individual genes on Xi are highly variable, ranging from several hours to days after the initiation of XCI (Bowness et al., 2022). Although expression of Xist achieves gene silencing

in ES cell systems, the complete silencing of such "slow silencing/slow dynamic" genes has been shown to require cell differentiation, thus indicating the involvement of some differentiation-dependent effectors. SMCHD1 is one such differentiation-dependent "late effector" of XCI. There is a good overlap between the genes dependent on SMCHD1 for their silencing and the genes exhibiting failure to silence even upon long term Xist expression in ES cells (Bowness et al., 2022). Genes showing a silencing defect upon deletion of *Smchd1* depict slow silencing dynamics (Bowness et al., 2022).

The recruitment of SMCHD1 to the Xi is dependent on PRC1-established H2AK119ub1 (Jansz et al., 2018b). Although expressed in ES cells, SMCHD1 recruitment to Xi is differentiation-dependent and the protein can be detected co-localising with Xi domains 3-4 days after differentiation (Bowness et al., 2022). The reason behind the lapse/delay between SMCHD1 recruitment on Xi and the establishment of H2AK119ub1 on Xi, which occurs within hours of Xist induction, is not clear. It has been proposed that a differentiation-linked factor mediates SMCHD1 recruitment to the Xi (Bowness et al., 2022).

SMCHD1 has a key role in the establishment of CGI methylation on Xi. MEFs obtained from *Smchd1* mutant XX embryos show loss of DNA methylation at the majority of CGIs on Xi. Additionally, transcriptional activation was observed for a substantial number of genes on Xi in these MEFs derived from *Smchd1* mutant embryos, suggesting their dependence on SMCHD1 for their complete silencing (Gdula et al., 2019). The link between CGI methylation on Xi and SMCHD1 was first observed in an earlier study, which established that SMCHD1 is required for the methylation of CGIs that display slow methylation dynamics in mouse ES cells differentiation timecourse experiment (Gendrel et al., 2012). The mechanism via which SMCHD1 regulates CGI methylation on Xi has not been determined. Conversely, loss of SMCHD1 leads to a rise in DNA methylation on Xi regions spanning derepressed genes. These regions also show reduced levels of Xi-specific H3K27me3. This contrasts with overall levels of H3K27me3 on Xi which are slightly increased in the *Smchd1* mutant MEFs (Gdula et al., 2019).

Our understanding of the mechanisms of SMCHD1-mediated gene silencing and CGI methylation on Xi is incomplete. A study from the lab showed that while CGI methylation is lost in both *Smchd1* and *Dnmt3b* null embryos, the derepression of genes is much more profound in *Smchd1* knockout embryos, suggesting that loss of CGI methylation on Xi does not account for the loss of gene silencing upon loss of SMCHD1 (Gendrel et al., 2013). Thus, SMCHD1 appears to operate via independent/distinct pathways to bring about gene silencing and CGI methylation.

With the advent of methods to study chromatin conformation, it is now known that inactive and active X chromosomes (Xa) engage in distinct long-range interactions. In contrast to Xa, Xi is depleted of TADs (Gdula et al., 2019); instead, it attains a bipartite chromatin structure with two megadomains separated by a hinge/ boundary at repetitive *Dxz4* locus in the late blastocyst stage (Gdula et al., 2019). Several independent studies have demonstrated that SMCHD1 is crucial for the formation of this megadomain structure and long-range interactions on Xi (Gdula et al., 2019; Jansz et al., 2018a; Wang et al., 2018). Mutation of *Smchd1* in MEFs leads to the strengthening of TADs and conversion of Xi chromatin conformation into that resembling Xa. The appearance of TADs on Xi in *Smchd1* mutant MEFs was correlated with the occupancy of the insulator protein CTCF that binds the TAD boundaries, suggesting that SMCHD1 interferes with the binding of CTCF (Gdula et al., 2019). The significance of this megadomain conformation of Xi is not clear, as the XCI seems to be largely unaffected upon the deletion of *Dxz4* which abolishes the megadomains (Darrow et al., 2016). *In vivo* study has shown that mice with deletion of *Dxz4* are also viable and fertile (Andergassen et al., 2019b). Figure 1.16 summarizes the diverse functions performed by SMCHD1 in XCI.

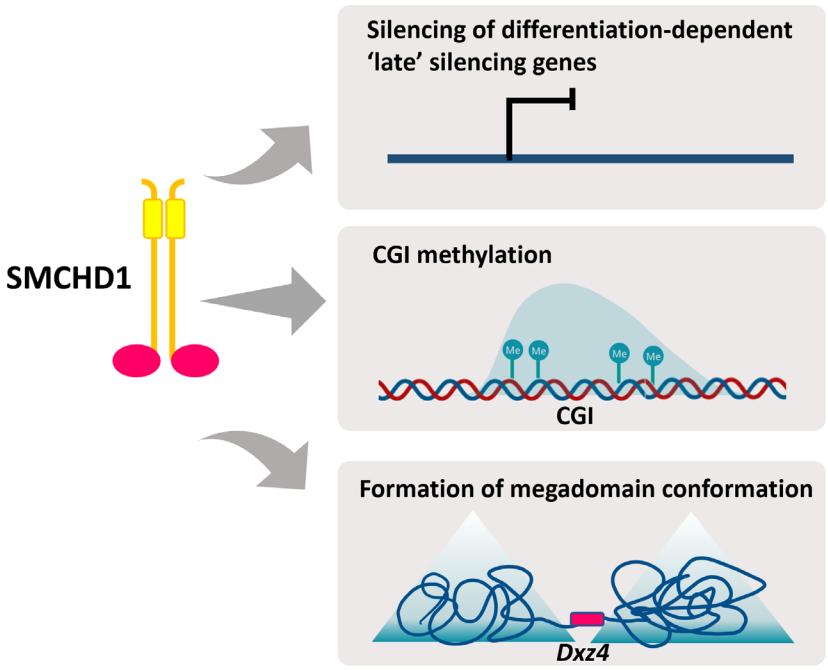


Figure 1.16: The diverse functions of SMCHD1 in XCI

SMCHD1 exists as a homodimer via association between the C-terminus Smc hinge domain (shown in yellow) and N-terminal GHKL ATPase domains (shown in pink). SMCHD1 performs three main functions in XCI; silencing of late silencing genes (Bowness et al, 2022), methylation of CGIs (Gendrel et al, 2012), and formation of the bipartite megadomain chromatin conformation of the Xi (Gdula et al, 2019).

1.3.3.2 Establishment of H3K9me3

The accumulation of H3K9me2 and H3K9me3 on Xi varies among species. In humans, H3K9me3 is highly enriched on Xi and its accumulation can be visualized using microscopy similar to the H2AK119ub1 or H3K27me3 chromatin modifications deposited by Polycomb complexes (Nozawa et al., 2013). In mouse on the contrary, H3K9me3 cannot be seen enriched on the Xi domain via imaging which is why it is relatively understudied. ChIP-seq has revealed that H3K9me2/3 exhibits a modest accumulation on Xi in comparison PRC1/2-mediated histone modification in mouse, and that it is enriched mostly on intergenic, gene-poor regions and predominantly on repeats (Keniry et al., 2016).

SETDB1 was identified as the main H3K9me3 methyltransferase in XCI in an *in vitro* reporter screen. Depletion of SETDB1 was shown to cause a reduction of H3K9me3 in female MEFs. Knockdown of SETDB1 in differentiating mouse ESCs also showed a modest gene silencing defect (Keniry et al., 2016). The contribution of SETDB1 in the establishment of H3K9me3 on Xi in differentiating ES cells was not determined by Keniry et al, but they did report that depletion of SETDB1 leads to a reduction in CGI methylation on Xi, suggesting partial dependence of CGI methylation on Xi on SETDB1 or its established H3K9me3 modification. ATF7IP, a protein interactor of SETDB1, has been implicated in serving as a bridge between CGI methylation and H3K9me3 via its interaction with MBD1 (Minkovsky et al., 2014). However, the impact of ATF7IP depletion or loss of CGI methylation on H3K9me3 establishment was not investigated in that study and therefore there is no clear evidence for CGI methylation-MBD1-ATF7IP-SETDB1 pathway on Xi. A recent study has also implicated the involvement of SMCHD1 in the formation of H3K9me3 domains (or blocks, as described by the authors) on Xi in MEFs. Mutation of *Smchd1* in MEFs resulted in the loss of H3K9me3 levels on Xi and a consequential gain in H3K27me3, likely as a compensatory mechanism. The underlying mechanism via which SMCHD1 influences H3K9me3 establishment on Xi has not been determined (Ichihara et al., 2022).

1.3.3.3 CGI methylation

The analysis of *Dnmt3a* and *Dnmt3b* null embryos has shown that CGI methylation is dispensable for the establishment of XCI and transcription silencing (Sado et al., 2004). However, once transcriptional silencing has been fully established by upstream players of XCI, CGI methylation is believed to act as

the final locking mechanism to maintain XCI. Evidence supporting this idea comes from studies on X chromosome reactivation. The treatment of XX somatic cells with 5-azaC was shown to lead to the reactivation of X-linked *G6pd* or *Pgk1* in a small proportion of cells (mouse-human hybrid fibroblasts, (Mohandas et al., 1981). A follow up study on MEFs also showed reactivation of *Hprt*, *G6pd* or *Pgk1* in 5% of the cells upon 5-azaC treatment (Graves, 1982). In both studies the number of cells that reactivated Xi were nominal (0.3% cell reactivated Xi in Graves et al.,1982 study) and not all tested genes were reactivated in the same cells i.e., full reactivation of Xi was not achieved. A subsequent study showed that treatment of MEFs with 5-azaC or deletion of *Dnmt1* results in activation of a GFP reporter gene on Xi in more cells as compared to deletion of *Xist* or treatment with histone acetyltransferase inhibitor, TSA. 5-azaC, TSA treatment and deletion of *Xist* synergistically improved the rate of reactivation of GFP reporter gene in comparison to 5-azaC treatment alone (Csankovszki et al., 2001). Consistent with the idea that CGI methylation is important for XCI maintenance, studies on the reprogramming of somatic cells into induced pluripotent stem cells have shown that CGI methylation on Xi is among the most persistent epigenetic modifications (Pasque et al., 2014).

Some key aspects of CGI methylation during XCI were unveiled in earlier studies (Lock et al., 1987) Keohane et al, 1996; Gendrel et al., 2012) that demonstrated that CGI methylation occurs late in XCI. Analysis of CGI methylation in XX ES cells undergoing differentiation showed that CGI methylation on Xi occurs in two phases, fast and slow (Gendrel et al., 2012). The majority of the CGIs on Xi show slow methylation dynamics and it is these CGIs that are dependent on SMCHD1 for their methylation (Figure 1.17), (Gendrel et al., 2012).

A systematic study of CGI methylation in *Dnmt3a, Dnmt3b* and *Dnmt3l* null female embryos showed that DNMT3B is specifically required for the CGI methylation on Xi. Loss of *Dnmt3b* resulted in hypomethylation of CGIs belonging to both the fast and slow methylating dynamic classes showing that variable dynamics of CGI methylation on Xi are not determined by distinct *de novo* methyltransferases. Hypomethylation of CGIs on Xi was also observed in embryos heterozygous for *Dnmt3b* deletion indicating a dose-dependent role of DNMT3B in establishing CGI methylation (Gendrel et al., 2012). The specificity of DNMT3B for CGI methylation on Xi was also shown by (Auclair et al., 2014). This specificity of DNMT3B for CGI methylation on Xi is a fascinating observation given the two *de novo* methyltransferases are highly similar.

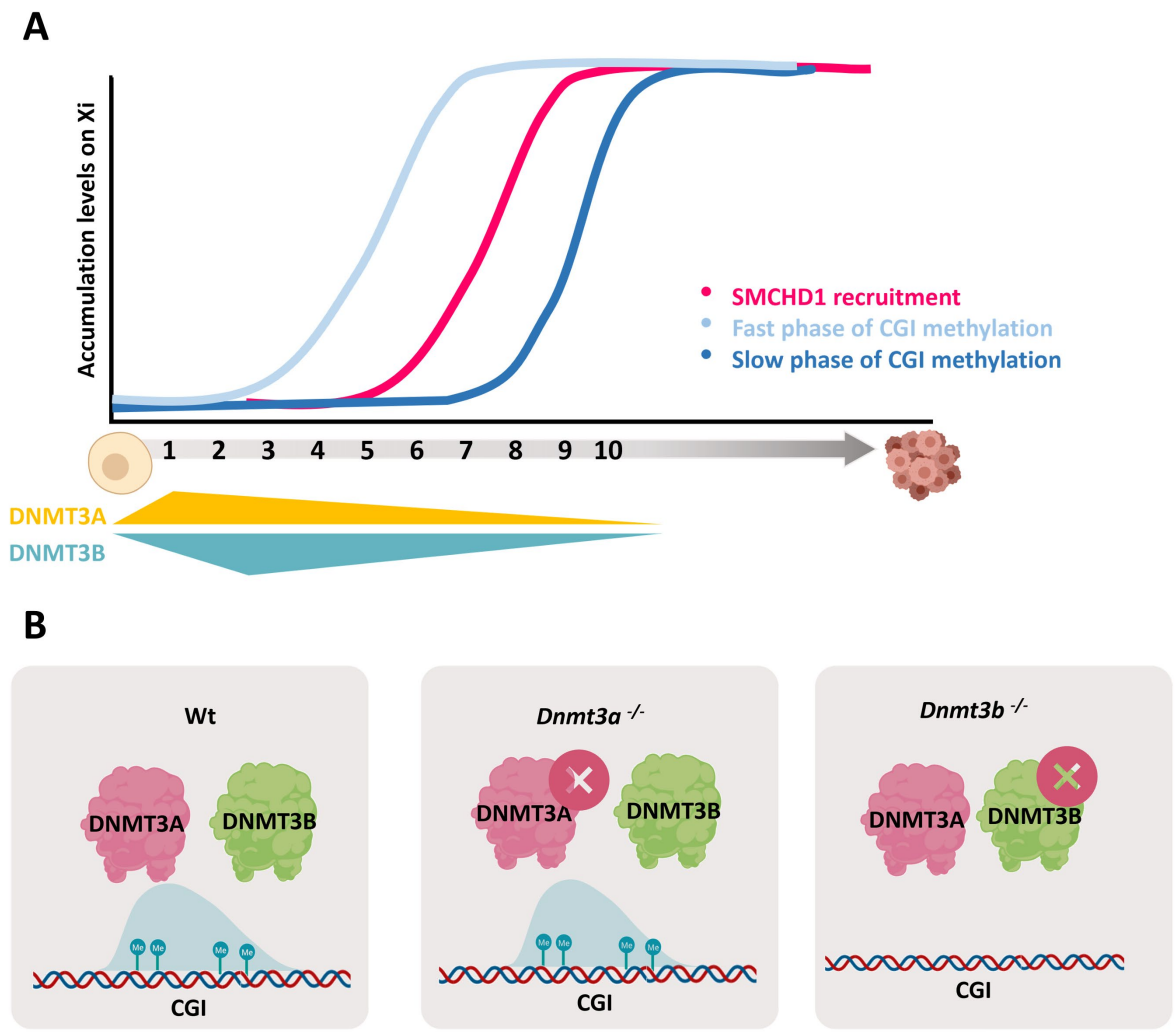


Figure 1.17: Key findings regarding CGI methylation on X.

A) Schematic representation of the key findings from Gendrel et al., 2012 publication: CGI methylation occurs in two phases. Fast methylating CGIs were methylated at day 7 of differentiation while slow methylating CGIs were methylated at day 10 or later in differentiation. Late methylating CGIs form the largest group and show dependency on SMCHD1 for methylation. Timing of SMCHD1 recruitment and protein levels of DNMT3A and DNMT3B during differentiation were also measured in this study. DNMT3A protein levels peak around day 2-3, while DNMT3B peaks between day 1-5 of differentiation. SMCHD1 could be seen enriched by immunofluorescence on the Xi between day 7-9 of differentiation.

B) Another key finding in this study was the specificity of DNMT3B for CGI methylation on the Xi. This was revealed by methylation analysis of selected CGIs in *Dnmt3a* or *Dnmt3b* null E9.5 mouse embryos. (wt = wildtype, *Dnmt3a*$^{-/-}$ = *Dnmt3a* knockout and *Dnmt3b*$^{-/-}$ = *Dnmt3b* knockout)

1.4 Specificity among *de novo* DNMTs

1.4.1 Current understanding of target specificity for DNMT3A and DNMT3B

DNMT3A and DNMT3B are highly related with around 85% amino acid similarity. Conservation is especially high for the region spanning the catalytic domain. There is however a division of labour between the two *de novo* methyltransferases during developmental processes. As mentioned above, the establishment of germline DNA methylation patterns in oocytes and spermatocytes is dependent upon DNMT3A and DNMT3L. The CGIs of germline-specific genes and those on Xi, on the contrary, depend on DNMT3B for their methylation. Deletion of *Dnmt3b* leads to a greater loss of CpG methylation in E8.5 mouse embryos in comparison to the deletion of *Dnmt3a* (Auclair et al., 2014). The greater contribution of DNMT3B in the establishment of CpG methylation during mouse embryogenesis could explain a prior observation by Okano and colleagues that deletion of *Dnmt3b* causes earlier lethality (at E11.5, with abnormal phenotype becoming apparent as early as E9.5) as compared to *Dnmt3a* deletion, where death occurs after birth (Okano et al., 1999).

This specificity of DNMT3A and DNMT3B during development cannot be explained by the expression pattern of *de novo* methyltransferases during development. For instance, Gendrel and colleagues (2012) have shown that both DNMT3A and DNMT3B are expressed during the differentiation of mouse ES cells, specifically in the time window when CGI methylation is established. Similarly, both DNMT3A and DNMT3B are expressed in germinal vesicle oocytes and in spermatogonia (La Salle and Trasler, 2006; Uysal et al., 2017). Offspring of mice bearing a conditional deletion of *Dnmt3b* in oocytes has been shown to have an unaffected phenotype. Contrastingly, the progeny of mice with conditional knockout of *Dnmt3a* and *Dnmt3l* in the germline leads to hypomethylation of differentially methylated regions, aberrant expression of maternal ICRs and lethality at around E9.5 (Kaneda et al., 2004).

Some studies have strived to understand the genomic features determining the differential preference of DNMT3A and DNMT3B for their target sites. *In vitro* enzymatic assays have revealed that DNMT3A prefers to methylate CpG with pyrimidines at +2 position while DNMT3B prefers purines (Handa and Jeltsch, 2005; Lin et al., 2002). Similarly, it has been shown that DNMT3B shows higher preference for CpA sequences in comparison to DNMT3A. These differences in target sequence preferences between DNMT3A and DNMT3B have been explained by subtle differences in the structure of their catalytic domains which lead to the different dynamic behaviour of the catalytic loops and flexibility of DNMT3B for CpA and CpG methylation (Gao et al., 2020). A recent study tested the sequence preference of DNMT3A and DNMT3B using triple knockout (*Dnmt1,3a,3b* null) mouse ES cells, complemented with either *Dnmt3a* or *Dnmt3b*. This study validated the observations first made in the *in vitro* study earlier

and showed that DNMT3A and DNMT3B recognized TACGYC and TACGRC target sequence motifs, respectively. This sequence preference was proposed to be due to one arginine residue (R836 in human DNMT3A) in the catalytic domain of DNMT3A which is substituted by lysine in DNMT3B. Interestingly, the substitution of the arginine with lysine (R836K) in DNMT3A, caused DNMT3A to carry out methylation on DNMT3B-sequence motifs (Mallona et al., 2021). Another *in vitro* study reported that R882H substitution in DNMT3A, observed in AML patients, causes the preference for residue 3' to methylated CpG site to change from pyrimidine to purine (Emperle et al., 2019). This sequence preference was later validated in differentially hypermethylated regions in AML model cell line (HCT116 line expressing wildtype or R882H DNMT3A) (Emperle et al., 2019).

The *in vivo* relevance of the target sequence preference of DNMT3A and DNMT3B is questionable. It has been shown that the contribution of these DNMT3A and DNMT3B target site sequence preferences are secondary to the role of histone modifications in determining DNA methylation (Mallona et al., 2021). An increase in DNA methylation levels at H3K36me3-enriched sites in the genome was observed for both highly and lowly preferred target sequences, suggesting that H3K36me3-mediated localization of both DNMT3A and DNMT3B was a predominant factor in determining DNA methylation (Mallona et al., 2021). Furthermore, the developmental processes such as XCI and imprinting which involve allele-specific methylation patterns, occur in inbred mouse species despite the same DNA sequence on both paternal and maternal alleles, indicating that target-site DNA sequence is not the main determinant/predictor for DNA methylation. The basis of DNMT3A and DNMT3B specificity for target loci becomes more complicated as literature also suggests variable affinities of the PWWP and ADD domains of the two *de novo* DNA methyltransferases for their target sites (Baubec et al., 2015; Boyko et al., 2022; Weinberg et al., 2019). Given the significant role of DNMT3A and DNMT3B in development and disease (briefly described in the next section), it is crucial to understand the specificity imparting factor in *de novo* DNA methyltransferases.

1.4.2 Significance of understanding specificity among *de novo* methyltransferases

Complete loss of *de novo* methyltransferases in mammals is lethal (Okano et al., 1999) and mutations in *DNMT3A* and *DNMT3B* lead to diseases which range from congenital syndromes to cancers. ICF-1 (ICF subtype 1) is a rare recessive autosomal disease with only 70 cases reported worldwide (Kiaee et al., 2021) with the majority of the causal mutations being single amino acid substitutions or deletions in the catalytic domain of DNMT3B (Shirohzu et al., 2002). ICF is marked by immunoglobin deficiency

and hypomethylation of pericentric repeats. The immunoglobulin deficiency causes high vulnerability to infectious diseases among the patients of ICF. The diagnostic feature of ICF is the karyotype with multiradial chromosomes 1, 9 and 16 due to the hypomethylation of pericentric repeats (Ehrlich, 2003). Given the lethality associated with complete loss of *DNMT3B,* ICF-1 mutations are considered to reduce, rather than abrogate DNMT3B activity. ICF-1 has been also shown to cause hypomethylation of some CGIs in Xi (Hansen et al., 2000).

Ectopic DNA hypo and hyper methylation is a hallmark of many cancers, often leading to methylation, and subsequent silencing of tumour suppressor genes, or hypomethylation and upregulation of oncogenes. A frequently observed phenomenon in cancers is the upregulation of/mutations in *de novo* methyltransferases, eventually causing their abnormal targeting. For instance, *DNMT3B* is highly expressed in early embryogenesis and is downregulated in somatic cells. Contrastingly, *DNMT3B* is upregulated in 30% of breast cancers (Bishop and Ferguson, 2015). Aberrant splicing of *DNMT3B* locus has also been observed in cancers. About 30 isoforms of human DNMT3B have been reported which result in the production of catalytically inactive or truncated forms of DNMT3B, the roles of which have not been fully elucidated (Gopalakrishnan et al., 2009b; Ostler et al., 2007; Shah et al., 2010). For instance, the delta variants of DNMT3B, bearing truncations in the N-terminal domain, were observed in non-small lung cancer which are responsible for the methylation of tumour suppressor *p16* and *RASSF1A* (Wang et al., 2006). A catalytically inactive isoform DNMT3B7 is thought to act as a dominant negative (Vasanthakumar et al., 2013) and is found associated with a wide range of cancers ranging from hematopoietic malignancies, neuroblastoma to breast cancer (Brambert et al., 2015; Ostler et al., 2012; Shah et al., 2010).

1.5 Aims of this thesis

The study of epigenetic gene regulation is fundamental to our understanding of development and disease. CpG methylation is carefully regulated during mammalian development and acts as a stable locking mechanism to maintain the cell-type specific gene expression pattern, once established. Aberrations in CpG methylation are frequently observed in congenital syndromes and cancers, the latter of which is thought to reverse differentiation and cell-fate decision, resulting in loss of cell memory (marked by lineage transformation such as the well-studied endoderm to mesoderm transformation in cancers) and stem cell-like physiology. Although many studies have indicated the specificity of DNMT3A or DNMT3B in various developmental contexts, we do not fully understand the

factors that govern/determine the participation of one *de novo* DNMT in a given context and not the other. Study of the mechanisms underpinning this specificity is crucial to our understanding of developmental processes and their misregulation in diseases like cancers.

In this work, I aim to address this question. I have used XCI as a model system which presents a well-characterized developmental process, and where methylation of CGIs on Xi has been shown to be specifically mediated by DNMT3B. The amenability to study XCI in easily cultured ES cells, which can be differentiated by the well-established lab protocols like embryoid body differentiation, make it an excellent experimental system and overcome the scarcity of material observed in research conducted with mouse embryos. Recently, the generation of engineered ES cell lines with inducible expression of Xist (discussed in detail in Chapter 3) has facilitated the research of processes occurring during the establishment of XCI. Studying CGI methylation on Xi can tell us if DNMT3B specificity could be driven by a unique domain or isoform of DNMT3B that distinguishes it from DNMT3A. This DNMT3B specificity imparting domain, if its existence is proved, could target DNMT3B to CGIs on Xi, either directly by reading a chromatin modification accumulated on CGIs on Xi, or indirectly by mediating the interaction of DNMT3B with DNMT3B-specific protein interactors that then target it to Xi. Figure 1.18 summarizes our current understanding of the processes that influence CGI methylation on Xi and the questions that remain unanswered.

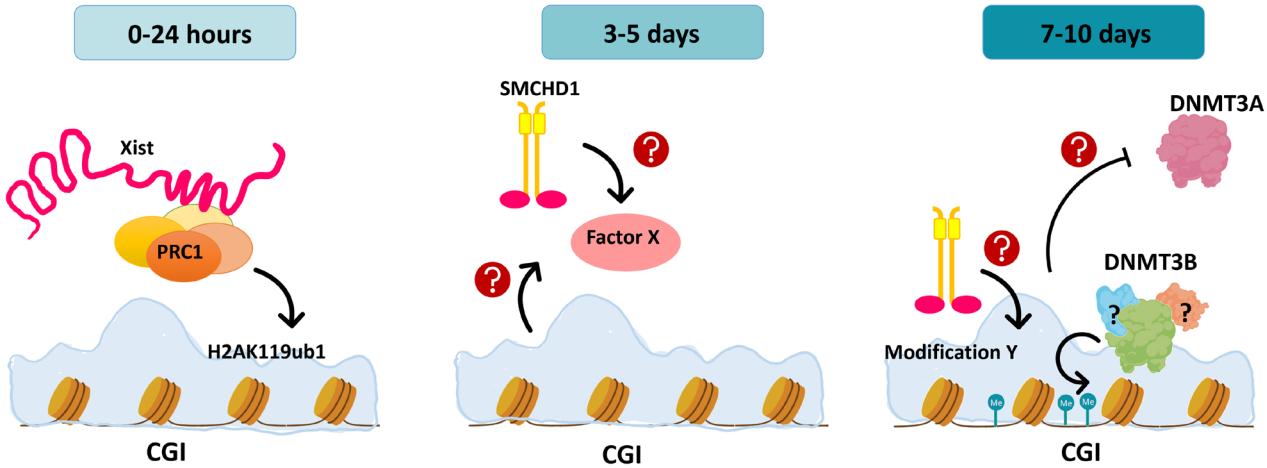


Figure 1.18: Unanswered questions in the establishment of CGI methylation during XCI (shown on the previous page).

Within the first 24 hours of Xist expression, PRC1-mediated H2AK119ub1 establishes on Xi. As XCI and differentiation progresses, SMCHD1 is recruited to the Xi (Bowness et al., 2022). The recruitment of SMCHD1 is dependent on H2AK119ub1 accumulation (Jansz et al., 2018b). This recruitment of SMCHD1 to H2AK119ub1 has been proposed to be dependent upon a differentiation-specific and yet undiscovered "factor X" (Bowness et al., 2022). Gendrel et al., demonstrated that the majority of CGI methylation is dependent upon SMCHD1 and carried out specifically by DNMT3B. It is unclear how SMCHD1 impacts this process. SMCHD1 could bring about a "chromatin modification Y" that specifically targets DNMT3B to CGIs on Xi. This would imply that DNMT3B (unlike DNMT3A) could recognize this modification directly (via a specificity imparting domain) or specific interactors.

Taking into consideration all the prior literature about XCI and CGI methylation on Xi, this study is aimed at addressing the following:

- To establish a robust experimental system to study CGI methylation during XCI and analyse the dynamics of CGI methylation during XCI. Test if the new experimental system validates the findings reported in the literature in mouse embryos (for example, the requirement of DNMT3B for the establishment of CGI methylation on Xi).
- To explore the role of DNMT3B isoforms in CGI methylation on Xi and to dissect the catalytic and accessory roles of DNMT3B in XCI.
- To characterize the specificity imparting domain/region of DNMT3B.
- To investigate the chromatin features that may impact localization of DNMT3B on Xi.
- To investigate DNMT3B specific interacting partners and the dynamics of DNMT3B interactome during differentiation.

In chapter 3, I describe the development of an ES cell-based model system to study CGI methylation, in chapter 4, I discuss experiments designed to search for the specificity imparting isoform or domain of DNMT3B and in chapter 5, I have explored the chromatin modifications and DNMT3B interactors that could play a role in its specificity for CGIs on Xi.

Chapter 2

Materials and methods



2.1 Experimental procedures

2.1.1 Cell culture conditions

Mouse ES cells (iXist-ChrX cells (Nesterova et al., 2019), and the derived cell lines) were cultured in Dulbecco's Modified Eagle Medium (DMEM, Gibco) supplemented with 10% foetal calf serum, 2 mM L- glutamine, 0.1 mM non-essential amino acids, 50 µM 2-mercaptoethanol, 100 U/ml penicillin, 100 µg/ml streptomycin (all from Gibco) and leukaemia inhibitory factor (LIF, made in-house), under standard culture conditions (37^{o}C, 5% CO_2, humid). ES cells were seeded atop a monolayer culture of inactivated mouse SNL fibroblast cells (previously treated with Mitomycin C, referred to as "feeder cells"), on gelatinised tissue culture plates. Gelatine coating of plates was achieved by treatment with 0.1% gelatine in PBS (Gibco) for 2-3 minutes at room temperature. Upon attaining 80% confluency, cells were detached using TrypLE Express (Gibco) and passaged. While in culture, cells were fed with LIF supplemented ES medium every day.

In order to freeze cells, pelleted cells were resuspended in 1 ml "freeze medium" (FCS + 10% DMSO, (Gibco)), and stored in 1 ml cryovials in a Mr Frosty freezing container (Nalgene) and placed at -80^{o}C for 24-48 hrs, before long term storage in liquid nitrogen.

2.1.2 Cloning of TIGRE homology vector for CRISPR/Cas9 mediated targeting of *Dnmt3b* rescue constructs

A plasmid previously gifted to the lab (Nora et al., 2017) contained a CAG element to allow for constitutive expression of a cloned gene knocked into the TIGRE locus using 1000 bp long homology arms and a gene for puromycin resistance to allow downstream selection of successfully targeted cells. Full length *Dnmt3b* or truncated variants of *Dnmt3b* were PCR amplified from cDNA (cloned previously

by Dr. Tatyana Nesterova) using a high fidelity Phusion DNA polymerase (Invitrogen) as per manufacturer's guidelines (i.e., 200 µM dNTPs, 500 µM primers, 0.02 U/µl Polymerase, 50 ng template DNA, Phusion buffer, visualized on 1-2% agarose gel run in 1X TAE buffer (40 mM Tris pH 8, 20 mM glacial acetic acid, 0.5 M EDTA) at 200 V). The PCR amplified DNA was extracted from the agarose gel using Zymoclean gel recovery kit (Zymo Research) and cloned into the TIGRE targeting plasmid using Gibson assembly to create a 'donor homology construct' for generation of stable cell lines. Gibson assembly was performed using NEB master mix, following the protocol recommended by the manufacturer. The primers used for generation of donor constructs for knock-in lines are detailed in Table 2.1. The plasmids were transformed into Top10 XL Gold cells using protocol recommended by Stratagene. After cloning, the sequence of the *Dnmt3b* donor homolgy construct plasmid was verified using Sanger sequencing.

Primer	Sequence	Purpose
Dnmt3b_FL_F_KY5	GTGCTGTCTCATCATTTTGGCAAAGAATTCCTCGAGACCATGCCAAAGAGACCCAGACCCTCTAGAATGAAGGGAGACAGCAGACATC	Cloning of *Dnmt3b1* into TIGRE targeting vector by Gibson assembly (also used *Dnmt3b3* and *Dnmt3b(N)3a* cloning
Dnmt3b_FL_R_KY6	CACAGTCGAGGCTGATCAGCGGATCCTATTCACAGGCAAAGTAGTCCTTCAAG	Cloning of *Dnmt3b1* into TIGRE targeting vector by Gibson assembly (also used for *Δ1-229, Δ1-218 Dnmt3b1*)
Dnmt3b3_Frag1_R_KY54	CGTCTTCGACTTTTGTTCTCTGGTTGCTTCTTGTTGGGTTTGAGGCC	Cloning of *Dnmt3b3* into TIGRE targeting vector by Gibson assembly
Dnmt3b3_Frag2_F_KY55	AACCCAACAAGAAGCAACCAGAGAACAAAAGTCGAAGACGCACAACCAATGACT	Cloning of *Dnmt3b1* into TIGRE targeting vector by Gibson assembly
Dnmt3b3_Frag2_R_KY56	GCAGGGAAGCCGAAGATCCTGTTCATTCCGGGTAGGTTACCCCA	Cloning of *Dnmt3b1* into TIGRE targeting vector by Gibson assembly
Dnmt3b3_Frag3_F_KY57	GTAACCTACCCGGAATGAACAGGATCTTCGGCTTCCCTG	Cloning of *Dnmt3b1* into TIGRE targeting vector by Gibson assembly
Delta_N_Dnmt3b_F_KY10	ATTTTGGCAAAGAATTCCTCGAGACCATGCCAAAGAGACCCAGACCCTCTAGAATAGGTGACCTCGTGTGGGGA	Cloning of *Δ1-229 Dnmt3b1* into TIGRE targeting vector by Gibson assembly
Del218aa_Dnmt3b1_F_KY131	CTGGTTATTGTGCTGTCTCATCATTTTGGCAAAGAATTCCATGCCAAAGAGACCCAGACCCTCTAGAGGAGACAGCACAGAGTATCAGGAT	Cloning of *Δ1-218 Dnmt3b1* into TIGRE targeting vector by Gibson assembly
Del1_50_Dnmt3b_F_KY132	GTGCTGTCTCATCATTTTGGCAAAGAATTCCTCGAGACCATGCCAAAGAGACCCAGACCCTCTAGAGTCTGCACACCAGAGACCAGAGG	Cloning of *Δ1-50 Dnmt3b1* into TIGRE targeting vector by Gibson assembly
Del_51_98_Dnmt3b_F2_F_KY133	AGGCAATCTGCACAGAGCCACTAATGCCAAAGCTCACCCG	Cloning of *Δ51-98 Dnmt3b1* into TIGRE targeting vector by Gibson assembly
Del_51_98_Dnmt3b_F1_R_KY134	CGGGTGAGCTTTGGCATTAGTGGCTCTGTGCAGATTGCCTCC	Cloning of *Δ51-98 Dnmt3b1* into TIGRE targeting vector by Gibson assembly
Del_98_145_Dnmt3b_F2_F_KY135	AAGTAGATGATGGGAATGGCTCTGATATTCGCCACCATGTGCAGGAG	Cloning of *Δ98-145 Dnmt3b1* into TIGRE targeting vector by Gibson assembly
Del_98_145_Dnmt3b_F1_R_KY136	TACTCCTGCACATGGTGGCGAATATCAGAGCCATTCCCATCATCTACTT	Cloning of *Δ98-145 Dnmt3b1* into TIGRE targeting vector by Gibson assembly
Del_146_177_Dnmt3b_F2_F_KY137	TCACCCGAGGTCGGCAGGGCGCCAGCGTCGACTTCATGG	Cloning of *Δ146-177 Dnmt3b1* into TIGRE targeting vector by Gibson assembly
Del_146_177_Dnmt3b_F1_R_KY138	TCCATGAAGTCGACGCTGGCGCCCTGCCGACCTCGG	Cloning of *Δ146-177 Dnmt3b1* into TIGRE targeting vector by Gibson assembly
Del_178_229_Dnmt3b_F2_F_KY139	GCACGCCATGGTCATCCCCTTTTGGAATAGGTGACCTCGTGTGG	Cloning of *Δ178-229 Dnmt3b1* into TIGRE targeting vector by Gibson assembly
Del_178_229_Dnmt3b_F1_R_KY140	CCCCACACGAGGTCACCTATTCCAAAAGGGGATGACCATGGCG	Cloning of *Δ178-229 Dnmt3b1* into TIGRE targeting vector by Gibson assembly
Dnmt3a_NLS_F_KY114	AATTCCTCGAGATGCCAAAGAGACCCAGACCCTCTAGAATGAATGCTGTGGAAGAGAACCAGG	Cloning of *Dnmt3a(N)3b* into TIGRE targeting vector by Gibson assembly
PWWP_Dnmt3b_F_KY115	AGTATGAGGATGGCCGGGGCATAGGTGACCTCGTGTGGGGAAAG	Cloning of *Dnmt3a(N)3b* into TIGRE targeting vector by Gibson assembly
Dnmt3a_Nter_R_KY116	CCCCACACGAGGTCACCTATGCCCCGGCCATCCTCATACT	Cloning of *Dnmt3a(N)3b* into TIGRE targeting vector by Gibson assembly

Primer	Sequence	Purpose
Dnmt3b_3a_fusion_frag2_F_KY12	TCAGGATGATAAAGAGTTTGGATTTGGCATTGGAGAGCTGGT	Cloning of *Dnmt3b(N)3a* into TIGRE targeting vector by Gibson assembly
Dnmt3b_3a_fusion_frag1_R_KY13	AGCTCTCCAATGCCAAATCCAAACTCTTTATCATCCTGATACTCTGT	Cloning of *Dnmt3b(N)3a* into TIGRE targeting vector by Gibson assembly
Dnmt3a_FL_R_KY14	GAAGGCACAGTCGAGGCTGATCAGCGCACACAAGCAAAATATTCCTTCAGCG	Cloning of *Dnmt3b(N)3a* into TIGRE targeting vector by Gibson assembly
ICF_F1_F_KY147	CAGCACAGAGTATCAGGATGATAAAGAGTTTGGAATAGGTGACCTCGTGTGGGGAAAGATCAAGG	Cloning of *S277P Dnmt3b1* into TIGRE targeting vector by Gibson assembly
ICF_F1_R_KY148	CAGTTTGTCAGCAGAGATCTCGGGAAACTTGCCATCACCAAACCAC	Cloning of *S277P Dnmt3b1* into TIGRE targeting vector by Gibson assembly
ICF_F2_F_KY149	CCATGCCCGGAATGCGCTGGGTACAGTGGTTTGGTGATGGCAAGTTTCCCGAGATCTCTGCTGACAAACTG	Cloning of *S277P Dnmt3b1* into TIGRE targeting vector by Gibson assembly
ICF_F2_R_KY150	GCCATCACGGGCCTGTTCATTCCGGGTAGGTTACCCCAGAAGTACCGGGCCCTG	Cloning of *S277P Dnmt3b1* into TIGRE targeting vector by Gibson assembly
Nter_T7_Dnmt3b_Frag1_F_KY124	CTGGTTATTGTGCTGTCTCATCATTTTGGCAAAGAATTCCCCTGGCCCACTCGAGGCCACCATG	Cloning of *T7-Dnmt3b1* into TIGRE targeting vector by Gibson assembly
Nter_T7_Dnmt3b_Frag1_R_KY125	CACCCTCTTCTTCATTCAGATGTCTGCTGTCTCCCTTCATGCCTCCACTTCCACCTGATCT	Cloning of *T7-Dnmt3b1* into TIGRE targeting vector by Gibson assembly
Nter_T7_Dnmt3b_Frag2_F_KY126	GAGAAAGGAGCTAGATCAGGTGGAAGTGGAGGCATGAAGGGAGACAGCAGACATCT	Cloning of *T7-Dnmt3b1* into TIGRE targeting vector by Gibson assembly
Dnmt3b_FL_R_KY130	ATGGCTGGCAACTAGAAGGCACAGTCGAGGCTGATCAGCGCTATTCACAGGCAAAGTAGTCCTTC	Cloning of *T7-Dnmt3b1* into TIGRE targeting vector by Gibson assembly
Nter_T7_Dnmt3a_Frag1_R_KY127	AGACTCTCCAGAGGCCTGGTTCTCTTCCACAGCATTCATGCCTCCACTTCCACCTGATCT	Cloning of *T7-Dnmt3a2* into TIGRE targeting vector by Gibson assembly
Nter_T7_Dnmt3a_Frag2_F_KY128	GCAGTTCGAGAAAGGAGCTAGATCAGGTGGAAGTGGAGGCATGAATGCTGTGGAAGAGAACCAGGCC	Cloning of *T7-Dnmt3a2* into TIGRE targeting vector by Gibson assembly
Dnmt3A_R_FL_KY129	ATGGCTGGCAACTAGAAGGCACAGTCGAGGCTGATCAGCGTTACACACAAGCAAAATATTCCTTCAGCG	Cloning of *T7-Dnmt3a2* into TIGRE targeting vector by Gibson assembly
Nter_T7_Del218Dnmt3b_F1_R_KY145	CTCCTCTAGAGGGTCTGGGTCTCTTTGGCATGCCTCCACTTCCACCTGATCT	Cloning of *T7-Δ1-218-Dnmt3b1* into TIGRE targeting vector by Gibson assembly
Nter_T7_Del218_Dnmt3b_F2_F_KY146	GGAGCTAGATCAGGTGGAAGTGGAGGCATGCCAAAGAGACCCAGACCCTCTAGAGGAGACAGC	Cloning of *T7-Δ1-218-Dnmt3b1* into TIGRE targeting vector by Gibson assembly

Table 2.1: Gibson cloning oligos used to make homology vectors

2.1.3 Cloning of small guide RNAs (sgRNAs) for CRISPR/Cas9 targeting

sgRNAs for the CRISPR/Cas9 mediated double stranded breaks at the target loci were cloned into pX459 plasmid (Addgene plasmid #62988) using a single step digestion-ligation protocol developed by the Zhang lab (Ran et al., 2013). 2 µl of the ligated product was transformed into Top10 XL Gold cells (Stratgene) according to the manufacturer's guidelines. sgRNAs for the generation of *Dnmt3b*$^{-/-}$ lines were designed by a previous post-doctorate researcher in the lab, Dr. Mankgopo Kgatle using the online tool called CRISPOR. sgRNAs to target *Dnmt3b* rescue constructs to TIGRE were designed and cloned by Dr. Tatyana Nesterova. All DNA oligonucleotide sequences used to generate sgRNAs are given in Table 2.2.

Oligo	Sequence	Purpose
Dnmt3b_MK1_F	CACCGAGGAGTGCATTATCGTTAAC	Generation of *Dnmt3b*$^{-/-}$ line (Targets Dnmt3b exon3)
Dnmt3b_MK1_R	AAACTTAACGATAATGCACTCCTC	Generation of *Dnmt3b*$^{-/-}$ line (Targets Dnmt3b exon3)
Dnmt3b_MK9_F	AAACGCGCGCTTGAGGTGTCTGCAC	Generation of *Dnmt3b*$^{-/-}$ line (Targets Dnmt3b intron 23)
Dnmt3b_MK9_R	CACCGTGCAGACACCTCAAGCGCGC	Generation of *Dnmt3b*$^{-/-}$ line (Targets Dnmt3b intron 23)
TNT476_sgTigre_C7H8t	CACCGTTACTGCCATAACACCTTT	sgRNA for TIGRE targeting
TNT476_sgTigre_C7H8b	AAACAAAGGTGTTATGGCAGTAAC	sgRNA for TIGRE targeting

Table 2.2: CRISPR-Cas9 sgRNAs and reverse complement oligos

2.1.4 Generation of cell lines

Prior to transfection, cells were plated in 6 well plates with feeders for 24 hours. An hour before transfection, the medium was replaced by an antibiotic-free medium (i.e., without penicillin and streptomycin). Lipofectamine 2000 (Life Technologies, Invitrogen) was used for transfection. Lipofectamine complexes were allowed to form in serum-free Opti-MEM (Gibco), as per manufacturer's instructions. Plasmid constructs for transfection were mixed with 100 µl of Opti-MEM and were subsequently added to lipofectamine solution (containing lipofectamine equal to 3X volume of plasmid DNA in 100 µl of Opti-MEM) to form DNA-lipid complexes, incubated at room temperature for 20 minutes, then applied drop-wise to the cultured cells. 24 hours after transfection, the cells were trypsinized and split into 90 mm gelatinized dishes with feeders in ratio 1/3, 1/10 and 1/15, and fed with 2.5 µg/ml puromycin-supplemented medium to allow for selection of transfected cells. Cells were maintained in the selection medium for 2 days and then later were switched to regular cell culture medium, till the colonies appeared (around 10 days later). Single colonies were picked in 96-well plates and split 1/3 upon reaching 80% confluency to give 3 replica plates which were used for expansion, screening and freezing of the clones.

For generation of the knockout lines, 1 µg each of the pX459 plasmids with CRISPR guides cloned for targeting the 5' and 3' cleavage sites were added. For the knock-in lines the "donor homology construct" plasmid and pX459 plasmid with cloned CRISPR sgRNA were co-transfected in a molar ration of 6:1 (3.5 µg of the plasmid with the donor homology construct and 1 µg of the pX459 with cloned guide RNA).

2.1.5 Genomic DNA Isolation from ES cells (screening of 96-well plates)

When cells were confluent in the 96-well plate, they were incubated overnight with 50 µl of lysis buffer (10 mM NaCl, 10 mM Tris-HCl pH 7.5, 10 mM EDTA-NaOH pH 8, 0.5% sodium lauroyl sarcosinate and proteinase-K 200 µg/ml) at 55°C. Afterwards, the DNA was allowed to precipitate by incubation with absolute ethanol first at -20°C for 30 minutes, followed by an incubation at room temperature for 30 minutes. The liquid fraction was then discarded, and the pellets were washed thrice with 100 µl of 70% ethanol. Extracted DNA (attached to bottom of plate well) was allowed to dry and then resuspended in 30 µl water.

2.1.6 Total Protein extraction

Cells growing in a 6-well plate were trypsinized and "pre-plated". For pre-plating, the cell suspension was plated on a non-gelatinized 6-well plate and incubated at 37°C for 30 minutes; this allows the feeder cells to settle and adhere to the plate (forming a translucent layer on the bottom of the plate), and separate from the ES cells which stay in suspension. The ES cells were then collected and centrifuged (1000 rpm for 4 minutes at room temperature). The cell pellet was then washed with PBS. Subsequently cell pellets were resuspended with TB+ buffer (9.8 ml of Transport buffer (20 mM HEPES (pH 7.3) 10 ml, 110 mM KOAc, 5 mM NaOAc, 2 mM MgOAc, 1 mM EGTA in H_20) supplemented with 2 mM DTT and 1X complete mini EDTA-free proteinase inhibitor, Roche). Cells were centrifuged at 4000 rpm for 3 minutes at 4°C. Cell pellets were resuspended in 50 µl TB+/NP-40/DNase I buffer (2.88 ml TB+, 30 µl NP-40, 30 µl 1 M $MnCl_2$, 60 µl DNAse I (10 U/µl Roche)) and incubated at 37°C for 30 minutes in the thermomixer at 1100 rpm.

Quantification of the protein preparation was conducted using Bradford assay. Bradford Microassay was conducted using Bio-Rad Protein assay dye concentrate (500-0006) as per the manufacturer's recommended protocol.

2.1.7 Western blot

20 µg of the total protein extract was boiled (95°C for 10 minutes) with Laemmli buffer (31.5 mM Tris-HCl pH 6.8, 10% glycerol, 1% SDS, and 0.005% Bromophenol Blue), loaded on the sodium dodecyl sulphate (SDS) PAGE gel (10% resolving gel and 5% stacking gel) and allowed to run for 45 minutes at 200 V. Proteins were then transferred to nitrocellulose membrane using Bio-Rad trans-blot turbo semi-dry transfer protocol recommended by the manufacturer.

After transfer, the membrane was washed with TBS (150 mM NaCl, 20 mM Tris base (pH 7.5) for 5 minutes. The membrane was then blocked in blocking buffer (5% skimmed milk powder (Marvel) in TBST (TBS containing 0.1% Tween 20)), at room temperature for 1 hour. The membrane was washed 3 times with TBST for 5 minutes and incubated with primary antibody (diluted as per vendor's guide, in 5% BSA in TBST) overnight at 4^{o}C. After three washes in TBST, the membrane was incubated with secondary antibody (diluted as per vendors recommendation in blocking buffer) for 1 hour at room temperature with shaking. Excess secondary antibody was removed by washing the membrane three times with TBST. For detection, membrane was incubated with ECL solution from Bio-Rad, following the manufacturers recommendation. The membrane was then blotted dry, placed in cassette and exposed to photographic films in dark. The antibodies used are mentioned in Table 2.3.

Antibody	Host	Experiment	Company	Dilution	Catalogue #
Anti-DNMT3A	Rabbit	Western blot	Cell Signalling Tech.	1:1000	E9P2F
Anti-DNMT3B	Rabbit	Western blot	Cell Signalling Tech.	1:1000	E9X7R
Anti-T7 tag	Rabbit	Western blot/IP/ChIP-seq/IF	Cell Signalling Tech.	1:1500, 5 µl/ChIP, 1:200 IF	D9E1X
Anti-H3	Rabbit	Western blot	Abcam	1:40000	ab199712
Anti-H3K9me3	Rabbit	Western blot/ChIP-seq	Abcam	1:1500, 5 µl/ChIP	ab8898
Anti-H3K36me3	Rabbit	Western blot	Abcam	1:1500	ab 9050
Anti-H3K4me3	Rabbit	Western blot	Millipore	1:1500	17-614
Anti-H2AK119ub1	Rabbit	IF	Cell Signalling Tech.	1:1000	8240S
Anti-H3K27me3	Mouse	IF	Active Motif	1:1000	ab61017

Table 2.3: Antibodies used in this study.

2.1.8 Immunofluorescence (IF)

In order to seed cells for immunofluorescence, a microscopy slide was placed in 90 mm petri dish and gelatinized for 2-3 minutes. Cells were then seeded on the gelatinized dishes (without feeder cells) and cultured in medium (supplemented with 1.5 µg/ml doxycycline if aiming to image chromatin modifications on Xi) in order to allow for formation of flat colonies 24 hours prior to the experiment. The next day, slides were washed with PBS and fixed with 2% paraformaldehyde (Sigma-Aldrich) for 15 minutes at room temperature. A second wash with PBS was performed to remove paraformaldehyde. The cells were then permeabilized using 0.4% Triton X-100 and incubated for 5 minutes at room temperature. The slides were then given another wash in PBS and blocked by performing 3 washes of 2 minutes each, in 0.5% cold sea fish gelatin (Sigma) prepared in PBS. Blocked

slides were incubated with primary antibody (prepared in 0.5% fish gelatin and supplemented with 5% normal goat serum (Sigma)) for 2 hours at room temperature.

Slides were again washed 3 times with 0.5% cold sea fish gelatin, followed by secondary antibody incubation. Secondary antibody was diluted in 0.5% cold sea fish gelatin and allowed to incubate with slides for 45 minutes at 37^{o}C. Slides were again washed twice with fish gelatin and once with PBS. 30 µl of DAPI supplemented Vectashield mounting medium (Vectalabs) was then added to the slides, a coverslip was placed on the slide and the two were sealed together using clear nail polish.

For performing DNMT3B IF, the cells were first permeabilized to allow the nuclear soluble proteins to escape the nucleus (using 0.4% Triton X-100 in PBS incubation for 5 minutes at room temperature). This strategy offered lower background. The cells were then fixed under the same conditions as mentioned above and the remaining protocol was followed as usual without further permeabilization.

2.1.9 Embryoid body (EB) differentiation

EB differentiation had previously been optimized by Dr.Tatyana Nesterova in the lab for iXist-ChrX lines: ES cells were pre-plated to separate and remove the feeder cells form ES cells in EC10 medium. EC10 medium had the same composition as the medium used for culturing ES (described in section 2.1.1) but does not contain LIF. After pre-plating, the cell suspension was moved to a falcon tube and the cells were counted using LUNA-II automated counter (Logos Biosystems). For setting up EB differentiation, 1 x 10^6 cells were plated in 10 cm, ungelatinized petri dish containing 10 ml of EC10 medium supplemented with 1.5 µg/ml doxycycline. For the next two days (day 1 and day 2), cells were fed with EC10 containing 1.5 µg/ml doxycycline; at this stage cells made small sparse colonies. On day 3, the small colonies were gently blown using a pipette to dislodge the weakly attached colonies from the plate. The suspended colonies were then moved to 10 cm bacterial culture dishes, which owing to their different material, provides a less sticky surface. Till day 7, these colonies stay suspended and form spherical masses called embryoid bodies (EBs). From day 3-day 7 EBs were fed with EC10 + 1.5 µg/ml doxycycline by taking the contents of the plate (medium with suspended EBs) out in a 50 ml falcon tube, allowing the EBs to settle (usually within 5 minutes, the exact time depends on the size of EBs; when settled EBs form a loose pellet), aspirating the old medium, gently resuspending EBs in fresh medium, and placing them back in the culture dish. Since some cells attach or remain on the dish during this process, I placed 5 ml of fresh medium on the culture dish, to save these cells from dying and potentially making the dish sticky which would cause the EBs to attach and then differentiate poorly. On day 7, the EBs start to attach to the dish. The weakly attached EBs were carefully detached by pipetting and moved back to 10 cm cell culture dishes. From day 7-day 11 these, attached EBs were

fed every day. On day 12, the EBs were collected, counted, pelleted, and used for further experiments. For certain experiments, I collected cells at timepoints within this full timecourse such as day 4, day 8 and day 10.

2.1.10 Neural progenitor cell (NPC) differentiation

The NPC differentiation protocol was optimized by Dr. Tatyana Nesterova in the lab for iXist-ChrX cells, adapted from previous work (Conti et al., 2005; Splinter et al., 2011). ES cells were trypsinized and pre-plated 4 times for 40 minutes each on gelatinized petri dishes to ensure the complete removal of feeder cells. Cells were then centrifuged (1000 rpm, 3 minutes), counted and 0.45 x 10^6 were plated in gelatinized T25 flasks in N2B27 media (50:50 DMEM/F-12: Neurobasal (Gibco) supplemented with 1X N2 and 1X B27 (ThermoFisher)) and 1.5 µg/ml doxyxycline. The cells were fed every day with fresh N2B27 medium + doxycycline for the next three days and then every other day till day 7. On day 7, cells were treated with Accutase (Millipore) to detach them from the flask, and 3 x 10^6 cells were plated on 10 cm bacterial dish in N2B27 + doxycycline medium supplemented with EGF and FGF (10 mg/ml, Peprotech). Cells stay in suspension and form EB-like aggregates. On day 10, the EB-like aggregates were mildly centrifuged (2 min, 800 rpm, 100 g) and plated on gelatinized 90 mm dishes in N2B27 + doxycycline medium supplemented with FGF and EGF. After 3-7 days cell aggregates attached on the dish and then later developed neural stem cell like morphology. For collecting, cells were treated with Accutase for 5 minutes, washed with PBS, counted, and pelleted by centrifugation.

2.1.11 Genomic DNA isolation for BioCAP and MBD

2-3 x 10^6 cells were pelleted and washed with PBS (Gibco). Cell pellet was resuspended in 200 µl of lysis buffer (10 mM NaCl, 10 mM Tris-HCl pH 7.5, 10 mM EDTA-NaOH pH 8, 0.5% sodium lauroyl sarcosinate and proteinase K 200 µg/ml) and incubated overnight at 55^{o}C. Lysed cells were then treated with RNase A (Sigma) at 37^{o}C for 1 hour. Equal volume of phenol: chloroform: isoamylalcohol (25:24:1, saturated with 10 mM Tris pH 8.0 and 1 mM EDTA, Thermoscientific) was added to the lysed cells and mixed well by inverting 8-10 times. The aqueous (containing nucleic acids) and organic phase were then separated by centrifugation at 10,000 x *g* for 5 minutes. Another extraction with phenol: chloroform: isoamylalcohol (25:24:1) was performed. A final extraction in chloroform: isoamylalcholol (24:1, Thermoscientific) was then carried out. The aqueous phase was then transferred to a fresh tube and mixed with 2.5 volumes of ice cold 100% ethanol and 0.1 volumes of 3 M sodium acetate. The contents of the tube were mixed by inverting. The DNA was then pelleted by centrifugation at 20,000

x *g* for 15 min at 4°C. The DNA pellet was washed with 70% ethanol, air dried and resuspended in nuclease free water.

2.1.12 BioCAP

For each BioCAP experiment, 50 µl of Neutravidin bead slurry (Invitrogen) was aliquoted into a 1.5 ml microcentrifuge tube. The bead slurry was centrifuged, and the beads were washed in 1 ml of BC150 (20 mM HEPES pH 7.9 ± 0.1, 150 mM KCl, 1 mM beta-mercaptoethanol, 0.1 mM $ZnCl_2$, 10% (v/v) glycerol). The beads were then incubated with 25 µg of biotinylated CXXC protein (expressed and purified from Rosetta 2 (DE3) pLysS *E.coli* cells in-house using procedure described in (Blackledge et al., 2012)) diluted in 475 µl of BC150 for 1 hour at 4°C at 200 rpm on a rotator. After the incubation, the beads were centrifuged at 750 x *g*, and the flowthrough was discarded. The beads were washed 3 times in 1 ml CAP 1000 (1 M NaCl, 20 mM HEPES pH 7.9 ± 0.1, 12.5% (v/v) glycerol, 0.1% Triton X-100) and 1 time in 1 ml CAP100 (same composition as CAP1000 but with 100 mM NaCl). The input genomic DNA was sonicated to a fragment size range of 150-350 bp at 4°C in TE buffer (10 mM Tris-HCl pH 7.9 ± 0.1 ,1 mM EDTA) using Diagenode Bioruptor Pico (20 cycles of 30 seconds on and 30 seconds off or until the required fragment size was achieved). The size of sonicated DNA was checked by agarose gel electrophoresis. 8 µg of sonicated genomic DNA was diluted in 950 µl of CAP100 and incubated with neutravidin conjugated CXXC protein for 1 hour at 4°C. The tubes were centrifuged at 750 x *g*, the flowthrough was collected in a new eppendorf and stored at -20°C. DNA bound to CXXC protein-neutravidin resin was then eluted using a step gradient of NaCl concentration (300 mM to 1000 mM). For every elution, 50 µl of the buffer was added, incubated with the resin for 10 min at room temperature at 500 rpm in a thermomixer and then collected after centrifugation. This process was carried out twice for CAP300, 500, 700 and 1000 buffers (same composition as CAP100 with 300, 500, 700 or 1000 mM NaCl respectively). The elutions from the same buffer were pooled and cleaned using Zymo Research ChIP clean and concentrator-5 kit, along with 50 µl of unbound flowthrough (collected after incubation of DNA with resin) and input DNA.

2.1.13 Methyl binding domain affinity purification (MBD)

MBD-affinity purification was performed using the Epimark Methylated DNA Enrichment kit manufactured by NEB (Cat no. E2600S). The protocol was carried out as per the manufacturer's guidelines. Genomic DNA was sonicated to a fragment size range of 100-300 bp at 4°C in TE buffer using Diagenode Bioruptor Pico (as above). 5 µg of sonicated genomic DNA was used as input and

mixed with 150 ng of DNA from HeLa cells (3% of input DNA from mouse cells) which served as an internal calibration for downstream quantitative bioinformatic analysis. Genomic DNA from Hela cells was isolated and sonicated in bulk and the same stock was used as control for all MBD experiments. Enrichment of methylated DNA was then validated using qPCR analysis (described below) using primers targeted against CpG islands of imprinted loci *Daz1* and *Ezhip* which I found methylated in our model cell lines (the primer sequences are mentioned in Table 2.5).

2.1.14 qPCR validation of MBD and BioCAP

The pulldown efficiency for MBD and BioCAP experiments were validated using qPCR. The qPCR was performed using a 20 µl reaction containing 2X SensiFAST SYBR No-ROX master mix (Bioline), 10 µM of primers and diluted DNA (1:50 dilution in nuclease free water) from input and immunoprecipitated (IP) fraction using Rotor-Gene Q machines (Qiagen). The qPCR program details are in Table 2.4 and the primers used for qPCR validation are given in the Table 2.5.

Cycles	Temperature (°C)	Time	Step
1	95	2-3 min	Polymerase activation
40	95 60 72	10 s 10 s 20 s	Denaturation Annealing Extension
Hold	50	10 min	-

Table 2.4: Program used for qPCR validation/estimation of IP efficiency of MBD and BioCAP.

Primer	Sequence	Purpose
Bmp7_Gene Body_F_KY31	TGGGATTCATGAGCACAGAG	Primer for BioCAP qPCR validation
Bmp7_Gene Body_R_KY32	GTTGACGTCTGGGAGGAAGA	Primer for BioCAP qPCR validation
Bmp7_CGI_F_KY33	CGAGCCAGCAAGACTTATCC	Primer for BioCAP qPCR validation
Bmp7_CGI_R_KY34	CTGTGGTCCACCTTTCCATT	Primer for BioCAP qPCR validation
Suv420h1_Gene Body_F_KY35	TGAGGCTCTCAGCAAGACTG	Primer for BioCAP qPCR validation
Suv420h1_Gene Body_R_KY36	ATCTTCCAGGAGAACGAGCA	Primer for BioCAP qPCR validation
Suv420h1_CGI_F_KY37	GACGTGGTTTCTTTGGTGGT	Primer for BioCAP qPCR validation
Suv420h1_CGI_R_KY38	CGGGAAGAGGCTGAGAGAC	Primer for BioCAP qPCR validation
Fabp7_Gene Body_F_KY39	CGAAATCCCTTTCAAAGCAG	Primer for BioCAP qPCR validation
Fabp7_Gene Body_R_KY40	GGATCCAAGAAAAACGCATC	Primer for BioCAP qPCR validation
Fabp7_CGI_F_KY41	TGAGCAAATCACAAGGAGGA	Primer for BioCAP qPCR validation

Primer	Sequence	Purpose
Fabp7_CGI_R_KY42	TGGAGGAACTCGGGTCTTAC	Primer for BioCAP qPCR validation
Adamts1_exon8_F_KY92	TCTTGCATGTGGAACCGTTT	Primer for MBD qPCR validation
Adamts1_exon8_R_KY93	GCTGTGATCGCATCATAGACTC	Primer for MBD qPCR validation
Pisd_exon7_F_KY94	GGATACGGATAGAGCCCACA	Primer for MBD qPCR validation
Pisd_exon7_R_KY95	CCTGGCATGGCCCGATG	Primer for MBD qPCR validation
Chli_intron3_R_KY96	AATTCCCGTTTCCAACGAAT	Primer for MBD qPCR validation
Chli_intron3_F_KY97	ACAAGAGAAAGTATAGAAACGGG	Primer for MBD qPCR validation
Daz1_CGI_F_KY98	GCGTTATCCCGTCAGCAAAG	Primer for MBD qPCR validation
Daz1_CGI_R_KY99	CAAGGCCTGCTCTCTAGTCGT	Primer for MBD qPCR validation
Ctse_intron5_F_KY100	ACAGGCGCAAGCCAAAC	Primer for MBD qPCR validation
Ctse_intron5_R_KY101	AGGTGTGCGCCACCAC	Primer for MBD qPCR validation
Itga9_intron21_F_KY102	TAGCATGGCTCAGTAGACATGAG	Primer for MBD qPCR validation
Itga9_intron21_R_KY103	CAGAGCTGTGAACATAGGCAGT	Primer for MBD qPCR validation
Itga9_intron22_F_KY104	CATTTCCAACCAGTCAACATTAC	Primer for MBD qPCR validation
Itga9_intron22_R_KY105	GAAAACGGACCAGAACAGTTCC	Primer for MBD qPCR validation
Ezhip_CGI_F_KY106	AGCGGAGGGCCTCGG	Primer for MBD qPCR validation
Ezhip_CGI_R_KY107	CCTGACCAAAGCCACCAT	Primer for MBD qPCR validation
Ezhip_CGI_F_KY108	TGCTCCTACGACCGAAAGG	Primer for MBD qPCR validation
Ezhip_CGI_R_KY109	CCCGATGATGATGATGACG	Primer for MBD qPCR validation

Table 2.5: Primers used for qPCR validation/estimation of IP efficiency of MBD and BioCAP.

2.1.15 Native Chromatin immunoprecipitation (nChIP)

For preparation of native chromatin for H3K9me3, DNMT3A and DNMT3B ChIP experiments, I used 5 x 10^7 (ES cells or well-suspended EBs), (counted using LUNA-II automated cell counter (Logos Biosystems)). Cells were resuspended and lysed using RSB (10 mM Tris pH 8, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% NP-40) with gentle inversion. The resuspended cells were centrifuged for 5 minutes at 1500 x *g* at 4°C. The nuclei were then resuspended in 1 ml of RSB with 0.25 M sucrose and 3 mM $CaCl_2$ and treated with 200 U MNase (Fermentas) at 37°C for 5 minutes with occasional inversions. MNase activity was then inhibited using 4 µl of 1 M EDTA. The released chromatin was collected by centrifugation at 2000 x *g* for 5 minutes and 1 ml of supernatant (called S1) was moved to a fresh tube. The remaining pellet was resuspended in 300 µl of nucleosome release buffer (10 mM Tris pH 7.5, 10 mM NaCl, 0.2 mM EDTA) and incubated at 4°C for 1 hour, rotating at 20 rpm. After 1 hour, the resuspended pellet was drawn 5 times through a 27G needle and centrifuged at 2000 x *g* for 5 minutes. The supernatant containing chromatin (S2) was mixed with S1. As a validation for the proper digestion of chromatin predominantly into mononucleosomes, DNA was extracted from 25 µl of chromatin

using ChIP clean and concentrator-5 kit (Zymo Research) and visualized using agarose gel electrophoresis. For each ChIP experiment, 100 µl of chromatin was diluted in native ChIP dilution buffer (10 mM Tris pH 7.5, 70 mM NaCl, 2 mM $MgCl_2$, 2 mM EDTA, 0.1% Triton X-100) to make up the volume to 1 ml and incubated with antibody overnight. For performing the IP, 40 µl of Sepharose beads (GE Healthcare) or protein A Dynabeads (Thermofisher) per sample, were blocked with 1 mg/ml BSA and 1 mg/ml yeast tRNA for 1 hour at 4^{o}C, washed with native ChIP incubation buffer and added to the samples. The IP was allowed to continue for 1 hour at 4^{o}C. After the incubation, flowthrough was collected (for western blots), and the beads were washed 5 times with cold native ChIP wash buffer (20 mM Tris pH 7.5, 2 mM EDTA, 125 mM NaCl, 0.1% Triton X-100) and once with TE (pH 8). Elution buffer (1% SDS, 100 mM $NaHCO_3$) was then added to the washed beads and the elution was performed at 25^{o}C at 1100 rpm for 30 minutes in a thermoshaker. For sequencing, the eluted DNA was cleaned using ChIP clean and concentrator-5 kit and before sequencing the IP efficiency was tested using qPCR. For western blot, Laemmli buffer (31.5 mM Tris-HCl pH 6.8, 10% glycerol, 1% SDS, and 0.005% Bromophenol Blue) was added to the beads, which were then boiled at 95^{o}C for 10 minutes, and loaded on SDS gel (5% stacking, 15% resolving gels for histone modifications). For T7-DNMT3A and T7-DNMT3B ChIP, I pooled 5 x 10^6 of Hela cells expressing T7-tagged SMC1 (T7-SMC1) (a gift from Rob Klose's Lab) with the cell pellet, to serve as a calibration for downstream quantitative analysis and followed the protocol described above.

2.1.16 NGS library preparation, validation, and quantification

The fragment size for the pulled down DNA after MBD, BioCAP and ChIP experiments was checked using Bioanalyzer 2100 (Agilent) with High sensitivity DNA Chip protocol to ensure that the fragments were within the range of 150-800 bp. If larger, the DNA was sonicated using Diagenode Bioruptor Pico (detailed above). DNA was quantified using Qubit double stranded DNA high sensitivity kit (Thermofisher) for batch library preparation. Quantity of starting/input DNA from the samples was diluted in TE buffer to be within the same range (± 10 ng). NGS DNA library was prepared using NEBNext Ultra II DNA library preparation kit from Illumina as per manufacturer's recommendations. After library preparation, the libraries were quantified using Qubit and diluted to 2 ng/µl in TE buffer. Size of the libraries was checked to be between 300 and 1000 bp, by running them on Bioanalyser. Libraries were quantified using qPCR with KAPA library quantification DNA standards (Roche) before pooling (using a 20 µl reaction containing 2X SensiFAST SYBR No-ROX master mix (Bioline), 20 µM of Illumina primers and 1:2000 dilution of 2 ng/µl library on Rotor-Gene Q machines (Qiagen), program is described in Table 2.6 and primer sequences are in Table 2.7). After pooling, the library pool was

again quantified with qPCR alongside the library pools from sequenced NGS runs to serve as a reference. The libraries were sequenced using 150 cycle, paired-end NextSeq 500/550 v2.5 Kits from Illumina at University of Oxford, Department of Zoology Nextseq sequencing facility on Illumina NextSeq500 (FC-404-2002) managed by Amanda Williams.

Cycles	Temperature (°C)	Time	Step
1	95	3 min	Polymerase activation
40	95 60 75	10 s 10 s 15 s	Denaturation Annealing Extension
Hold	50	10 min	-

Table 2.6: Program used for quantification of NGS libraries using qPCR

Primer	Sequence
Illumina_qPCR_F_TNK300	CAAGCAGAAGACGGCATACGA
Illumina_qPCR_R _TNK301	AATGATACGGCGACCACCGA

Table 2.7: Primers used for quantification of NGS libraries using qPCR

2.1.17 Salt extractions to estimate the chromatin binding strength of DNMT3B and its mutants

12 x 10^6 cells were collected, centrifuged and divided equally into 6 pellets of 2 x 10^6 cells each (which would later be treated with a different salt concentrations). The pellet was resuspended with 500 µl of RLB (10 mM KCl ,1.5 mM $MgCl_2$, 10 mM Tris (pH 7.5), 0.1% NP-40, 0.1% Tween20, supplemented with protease inhibitors tablet (complete EDTA-free, Roche). 500 µl of RLB + 24% sucrose was carefully added to the bottom of the tube to form a sucrose cushion. The pellets were then centrifuged (700 x *g* for 5 minutes at 4°C), this step removes the cytoplasmic fraction from the nuclei, which pellet at the bottom. The nuclei in each pellet were then resuspended in 100 µl of nuclear extraction buffer (10 mM HEPES pH 7.5, 5% glycerol, 1.5 mM $MgCl_2$, 0.2 mM EDTA, NaCl and protease inhibitors tablet) containing different salt concentrations (I used 75, 150, 250, 300 and 400 mM NaCl). The nuclear

fraction was extracted for 30 minutes at 4°C. After extraction, the resuspended nuclei were centrifuged (1500 *x g* for 3 minutes). The supernatant (nuclear soluble fraction) was saved in a fresh tube. The pellet (chromatin bound fraction) was resuspended in 100 µl of nuclear extraction buffer (containing 150 mM NaCl) supplemented with 200 U of Benzonase (Merck) to release the chromatin bound proteins. As a control, one pellet was used to prepare a whole nuclear extraction. For whole nuclear extraction, nuclei were resuspended in 200 µl of nuclear extraction buffer (with 150 mM NaCl and 200 U benzonase) and placed at 4°C alongside other pellets for extraction for 30 minutes. The nuclear soluble, chromatin-bound fractions and whole nuclear extraction were quantified using Bradford test (Bradford reagent BioRad), and 25 µg of protein was loaded on SDS gel after boiling with Laemmli buffer for 10 minutes for western blot. As a confirmation of the salt extraction experiment, a commercial subcellular protein cell fractionation kit my Thermofisher (Cat no. 78840) was also used as per manufacturer's recommendations.

2.1.18 Sucrose gradient fractionation of chromatin-bound proteins

A sucrose freeze gradient (5-16%) was formed by carefully layering 500 µl of sucrose buffer solutions (starting from 16% – 5% with increments of 1.1%, containing 10 mM Tris pH 7.5, 0.25 mM EDTA, 300 mM NaCl, protease inhibitor tablet) in 5 ml tube for Beckman SW 40Ti rotor i.e., 16% sucrose solution was placed in tube, allowed to freeze using dry ice, and then 14.9% sucrose buffer was pipetted on top of frozen 16% sucrose buffer and so on. 1 ml of native chromatin was prepared as described earlier and carefully placed on top of the sucrose gradient set in the rotor tube. The sucrose gradient fractionation of the chromatin was carried out for 16 hours at 4°C at 40,000 x *g* in a Beckman ultracentrifuge using a SW 40Ti rotor. After the centrifugation, fractions of 250 µl were carefully collected from the top of the tube with a pipette. For each fraction, 100 µl was used for DNA extraction and the rest was used for protein precipitation and western blot. DNA was extracted by adding 1 volume of isopropanol to the collected fraction, performing gentle inversions and centrifuging at 10,000 rpm for 3 minutes. The DNA pellet was then washed with 70% ethanol, air-dried and resuspended in nuclease free water. This DNA was used to test the fractionation of chromatin by size, and its reproducibility among the samples, using agarose gel electrophoresis. To precipitate and concentrate proteins in collected fractions, 0.4 volume of trichloroacetic acid (TCA, Sigma-Aldrich) was added to the collected fraction and vortexed to allow proper mixing. The samples were then placed on ice for 30 minutes. The precipitated proteins were then separated by centrifugation 13,000 rpm at 4°C. The precipitated proteins formed a layer along the wall of the tube which was scraped off with a pipette tip, 1 volume of ice-cold acetone was added to the precipitate, vortexed and air dried. The

dried pellet was then resuspended in 1 M Tris (pH 8) and an equal volume from each fraction boiled with Laemmli buffer and used for western blot analysis.

2.1.19 Chromatin extraction under high salt

This protocol was adapted from the publication by (Henikoff et al., 2009) with slight modification for mouse ES cells. 1 x 10^8 ES cells were collected and divided into 6 equal parts (each to be treated with different salt concentration). Native chromatin preparation was carried out as usual up to collection of S1 fraction (described in section 2.1.15). Each pellet was then resuspended in 200 µl of nucleosome release buffer supplemented with 10, 150, 400, 600, 800 or 1000 mM NaCl and incubated at 4^{o}C for 1 hour on a rotor at 20 rpm. The suspension was then passed through 27G needle five times and centrifuged at 13,000 rpm for 2 minutes to collect chromatin in supernatant. To test if any chromatin remained in the pellet after the chromatin extraction, the pellet was resuspended in 100 µl of TNE buffer (10 mM Tris pH 7.5, 200 mM NaCl and 1 mM EDTA). For each salt condition/ treatment, 25 µl of the extracted fraction was used for DNA extraction and visualization (using Zymo Research ChIP and concentrator-5 kit). 10 µl of the collected fraction was boiled with Laemmli buffer and used for western blot.

2.1.20 Immunoprecipitation of T7-DNMT3A and T7-DNMT3B from nuclear lysates

For each immunoprecipitation ~1 ml of packed cell volume (pellet volume) was used which was collected prior to the experiment and frozen at -80^{o}C. Cell pellets were thawed on ice and resuspended in 10 volumes for buffer A (10 mM HEPES pH 7.9, 1.5 mM $MgCl_2$, 10 mM KCl, freshly supplemented with 0.5 mM phenylmethylsulfonyl fluoride (PMSF), 0.5 mM Dithiothreitol (DTT), and 1X complete EDTA-free proteinase inhibitor (Roche), before use). The resuspended pellets were incubated on ice for 10 minutes, after which cells were recovered by centrifugation at 1500 x *g* for 5 minutes at 4^{o}C. Pellets were then resuspended 3 volumes of buffer A with 0.1% NP-40, and again incubated on ice for 15 minutes with gentle inversion. Nuclei were recovered with centrifugation (1500 x *g* for 5 minutes at 4^{o}C) and the pellet was washed with 5 volumes of ice-cold PBS (Gibco). Nuclei were then resuspended in lysis buffer (20 mM Tris pH 7.5, 200 mM NaCl, 1% Triton X-100, 0.1% sodium deoxycholate, with 1 mM PMSF, 1 mM DTT and and 1X complete EDTA-free proteinase inhibitor (Roche) added before use). 2 µl of benzonase (Merck, 250 U/µl) was added to the lysates to digest DNA. The nuclear lysates were passed 5 times through 21G and then 23G needles to homogenize the

lysate. Lysates were incubated on ice for 1 hour 20 minutes to allow digestion of DNA. Nuclei were then pelleted at 21000 x *g* for 45 minutes at 4°C and supernatant (clear nuclear lysate) was collected in a fresh tube. Nuclear lysate was quantified using Bradford assay. For performing the IP, nuclear lysates were diluted in lysis buffer (without salt) to bring the final salt concentration to 150 mM NaCl. The lysates were again centrifuged at 21000 x *g* for 5 minutes to pellet down any insoluble precipitates. 100 µl of the nuclear lysate was pipetted in a new tube to serve as input. The remaining nuclear lysate was used for immunoprecipitation; 20 µl of T7 antibody (Cell signalling technology) was added and lysate (1 ml in volume) was incubated for 3 hours at 4°C with 12 rpm rotation. 50 µl Protein A Dynabeads (Thermofisher) per sample were washed 3 times with lysis buffer (with 150 mM NaCl) and blocked with 1.5 mg/ml bovine serum albumin (BSA, NEB) at 4°C for 2 hours. After blocking, the beads were again washed with lysis buffer (with 150 mM NaCl) and resuspended in lysis buffer (with 150 mM NaCl) to return beads to their original volume. 50 µl of blocked Protein A Dynabeads were added to the sample and the immunoprecipitation was incubated for 2 hours at 4°C with 12 rpm rotation. After incubation, flowthrough (FT) was collected for western blots. The beads were washed 5 times with wash buffer (20 mM Tris pH 7.5, 150 mM NaCl, 0.1% sodium deoxycholate, with 1 mM PMSF, 1 mM DTT). One final wash was done with 1 ml PBS (Gibco) and then the beads were suspended in 1 ml PBS. 300µl of suspended beads were separated for western blot and silver staining. From the other 700 µl of bead suspension, all the liquid was removed, beads were snap frozen on dry ice and stored at -80°C (to be send off to Oxford South Parks mass spectrometry facility). Laemmli buffer was added to input, FT and IP (300 µl fraction), boiled and used for downstream western blot and silver staining. For silver staining the samples were run on NuPAGE 4-12% BisTris (ThermoFisher) gels and silver staining was performed using Pierce High sensitivity Silver Stain kit (Thermofisher) as per manufacturer's guidelines.

2.1.21 Chromatin RNA extraction and library preparation for sequencing (chrRNA-seq)

Cell pellets ($1 \times 10^5 – 3 \times 10^6$) were washed with PBS (Gibco), resuspended in RLB (10 mM Tris pH 7.5, 10 mM KCl, 1.5 mM $MgCl_2$, and 0.1% NP-40) and incubated on ice for 5 minutes. RLB with 24% sucrose was then added to the suspended pellets and cytoplasmic fraction was separated from the nuclei by centrifugation at 2800 x *g* for 10 minutes at 4°C. The nuclei were resuspended in NUN1 (20 mM Tris pH 7.5, 75 mM NaCl, 0.5 mM EDTA, 50% glycerol, 0.1 mM DTT). Nuclear lysis was conducted by addition of equal volume of NUN2 (20 mM HEPES pH 7.9, 300 mM, 7.5 mM $MgCl_2$, 0.2 mM EDTA, 1 M Urea, 0.1 mM DTT) and incubation for 15 minutes on ice with occasional mixing. The chromatin

fraction was isolated as an insoluble pellet by centrifugation at 2800 x *g* for 10 minutes at 4°C. The RNA was extracted by suspending and homogenizing chromatin pellet in TRIzol (Invitrogen). RNA was then extracted using chloroform and precipitated with isopropanol. RNA pellet was washed with 70% ethanol and resuspended in nuclease free water. RNA was then treated with Turbo DNAse (Thermofisher scientific) and quantified using Qubit RNA HS Assay Kit. This RNA was used for NGS library preparation using TruSeq stranded total RNA kit form Illumina as per manufacturer's recommended protocol.

2.1.22 Nuclear extract preparation

10-20 x 10^6 cells were pelleted and washed with PBS. The cell pellet was resuspended in 10 volumes of cold buffer A (10 mM HEPES pH 7.9, 1.5 mM $MgCl_2$, 10 mM KCl, supplemented with 0.5 mM PMSF, 0.5 mM DTT, and 1X complete EDTA-free proteinase inhibitor (Roche), before use). The resuspended pellets were incubated on ice for 10 minutes and centrifuged at 1500 x *g* for 5 minutes at 4°C. Pellets were again resuspended in 3 volumes of buffer A with 0.1% NP-40 and again incubated on ice for 10 minutes with occasional inversion. Nuclei were recovered with centrifugation (400 x *g* for 5 minutes at 4°C). The pelleted nuclei were then resuspended in 1 volume of buffer C (250 mM NaCl, 5 mM HEPES pH 7.9, 26% glycerol, 1.5 mM $MgCl_2$, 0.2 mM EDTA pH 8, with freshly added 0.5 mM DTT and complete EDTA-free PIC). 5 M NaCl was added to the lid of the 1.5 ml tube so that the final concentration of NaCl in the extract (resuspended nuclei and buffer C) becomes 400 mM and mixed by quickly inverting to avoid any precipitation. The extraction was allowed to proceed for 1 hour at 4°C. The nuclear extracts (liquid fraction) were recovered using centrifugation (16000 x *g* for 20 minutes at 4°C). The nuclear extract was quantified using Bradford assay (BioRad) and 25 µg of protein was used for western blot analysis.

2.1.23 RNA extraction and cDNA preparation for quantitative PCR analysis

For qPCRs, 2-3 x 10^6 cells were pelleted and RNA was extracted using RNeasy kit (Qiagen). The extracted RNA was quantified using nanodrop and treated with Turbo DNAse (Thermofisher scientific) as per the manufacturer's instructions. cDNA was then synthesized using the Superscript III first strand synthesis system (Thermofisher) using random hexamers as per manufacturers guidelines and protocol. A RT minus control (without superscript III reverse transcriptase) was included to detect DNA contamination in the RNA preparation. This protocol was used for the estimation of expression levels of pluripotency factors (as a proxy for efficient differentiation) or *Dnmt3b* where chrRNA-seq analysis

in complementation experiments could not be performed. The primers used are listed below in Table 2.8

Primer	Sequence	Gene
Dnmt3b_qPCR_F1_KY69	GAGGGGAGGATGAGGAGAGC	*Dnmt3b*
Dnmt3b_qPCR_R1_KY70	TTCTTTCCACAGGACAAACAGCG	*Dnmt3b*
Dnmt3b_qPCR_F2_KY71	TGCTGCTGTGCAGTAACACA	*Dnmt3b*
Dnmt3b_qPCR_R2_KY72	TGCAGCTTGGCATCCTCAG	*Dnmt3b*
GAPDH_qPCR_F_KY73	TGTAGTGAGCCCCAGGCTAT	*GAPDH*
GAPDH_qPCR_R_KY74	ATAGCTGATGGCTGCAGGTT	*GAPDH*
Actin_qPCR_F_KY75	CTAAGGCCAACCGTGAAAAG	*Actin*
Actin_qPCR_R_KY76	AGCCTGGATGGCTACGTACA	*Actin*
Oct4_qPCR_F_KY77	AGCCGACAACAATGAGAACC	*Oct4*
Oct4_qPCR_R_KY78	TGGTCTCCAGACTCCACCTC	*Oct4*
Nanog_qPCR_F_KY79	ATGCCTGCAGTTTTTCATCC	*Nanog*
Nanog_qPCR_R_KY80	GAGGCAGGTCTTCAGAGGAA	*Nanog*
Klf4_qPCR_F_KY81	AGAACAGCCACCCACACTTG	*Klf4*
Klf4_qPCR_R_KY82	GTGGTAAGGTTTCTCGCCTGT	*Klf4*
Klf4_qPCR_F_KY83	GGCTCAGGTACCCCTCTCTC	*Klf4*
Klf4_qPCR_R_KY84	GGGAAGTCGCTTCATGTGAG	*Klf4*

Table 2.8: Primers used for estimating gene expression levels.

2.1.24 XX karyotyping for cell lines using immunofluorescence (IF) and allele specific PCR

Female mouse ES cells can frequently lose an X chromosome and attain XO karyotype. Therefore, after generating cell lines and during differentiation timecourse experiments, I regularly tested the cells for XX karyotype. I used two methods in parallel (1) Allele-specific PCR and (2) H3K27me3 and H2AK119ub1 IF to visualise the Xi and score the percentage of the cell population containing an Xi (Figure 2.1 C). The allele-specific PCR was optimized by Dr. Tatyana Nesterova and uses two forward primers that either anneal to Xist sequence (amplifying wt allele) or the Tet-On promoter (amplifying the allele containing inducible Xist) and a common reverse primer that anneals to 3' homology arm

(Figure 2.1 A and B). The sequence of the primers used for the allele-specific PCR are given in the Table 2.9.

Primer	Sequence	Purpose
AA102	CGCGGATCAGTTAAAGGCGT	Forward (wt allele)
TNK63	TGACCTCCATAGAAGACACCG	Forward (Tet-On allele)
AA103	AACCACGGAAGAACCGCAC	Reverse

Table 2.9: Primers used for XX karyotyping of the cell lines.

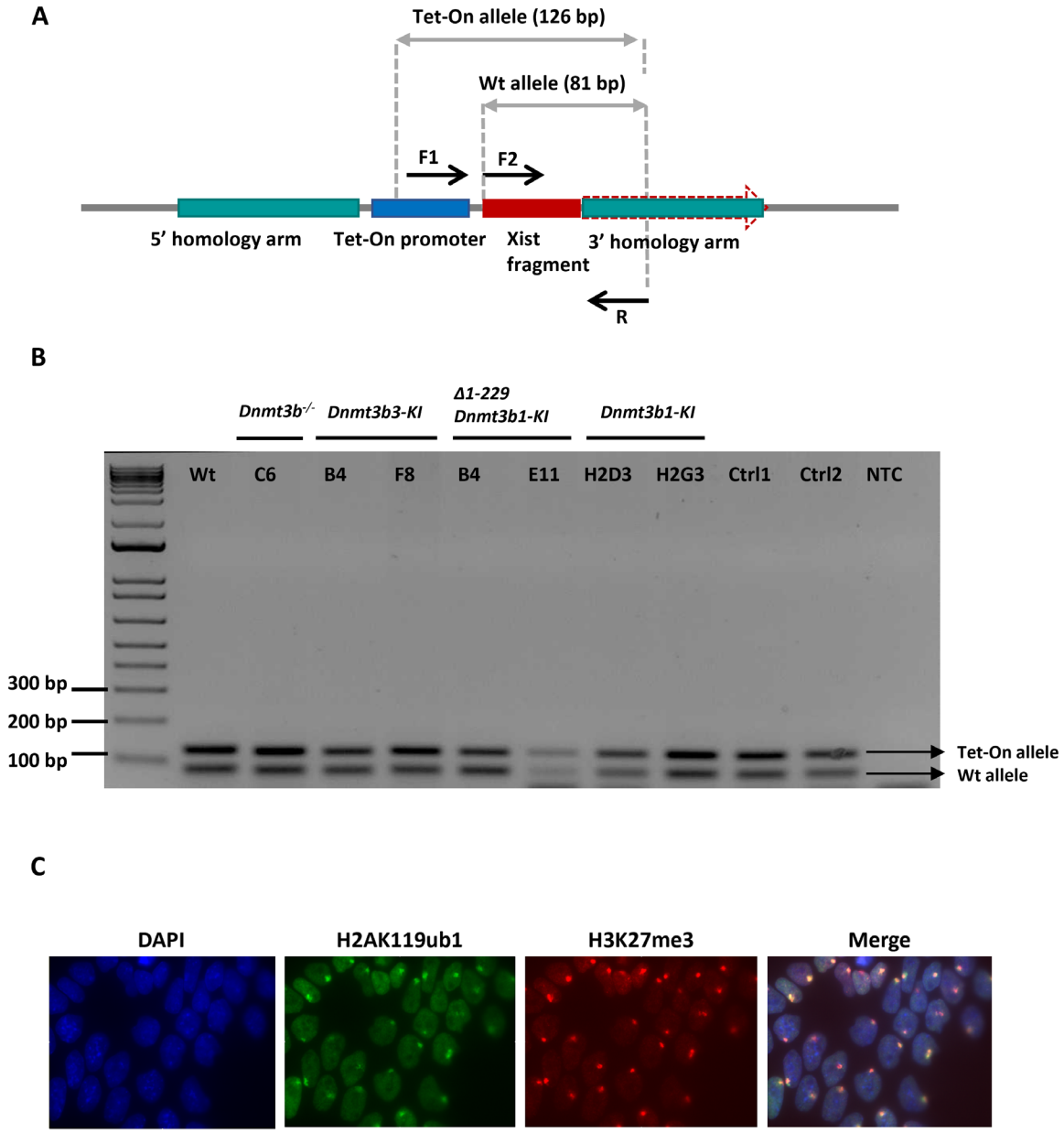


Figure 2.1: XX karyotyping of cell lines using allele specific PCR and IF (shown on the previous page).

A) Schematic illustrating allele specific PCR strategy: Two forward primers (F1 and F2) anneal to Tet-On promoter or wt Xist sequence respectively. A common reverse primer (R) anneals to the 3' homology arm.
B) Agarose gel electrophoresis of XX karyotyping PCR. The XX karyotype status of cell lines was monitored at all stages of derivation. The upper band evident on the 1% agarose gel comes from the amplification of Tet-On promoter elements inserted upstream Xist. The bottom band comes from the amplification of wt Xist allele. The bands appear in roughly equal ratio on the gel in a cell line that has retained XX karyotype.
C) A representative image of H2AK119ub1 and H3K27me3 IF routinely used for estimating the percentage of cell population carrying Xi. This IF was used in combination with allele-specific PCR to ensure that cells retained their XX karyotype in culture and throughout the differentiation timecourse the cell line

Before any NGS experiments, these two methods were used to ensure that the cell lines possessed an XX karyotype. After sequencing, the cell lines were again tested for XX karyotype computationally by analysing the number of reads originating from either X chromosomes in the untreated/input samples (see section 2.2.1 and Appendix 1).

2.2 Data analysis softwares and packages

2.2.1 Mapping of fastq files for BioCAP-, ChIP- and MBD-seq and assigning reads to specific alleles

Nextseq fastq files were mounted on the Oxford Biochemistry genomics server using a tool called Basemount from Illumina. Basemount is synchronized with the sequencing run management platform from Illumina called BaseSpace, which saves the data during the sequencing run. The mapping criteria slightly varied for the three methods. For BioCAP-seq, the fastq files were mapped to N-masked mm10 genome reference (called GTF file, provided by Dr. Guifeng Wei in the lab) using bowtie2 V2.3.5.1 (Langmead and Salzberg, 2012) with "--sensitive", "--no-mixed" and "--no-discordant" parameters. The "--maxins" parameter sets the threshold for the distance (in bp) between the paired-end reads and it is dependent on the DNA fragment size used in the method. For ChIP- and BioCAP-seq "--maxins" was set to 2000 while for MBD-seq, "--maxins" was set to 800. For calibrated MBD-seq, the fastq files were mapped to N-masked mm10 genome concatenated with human hg38 genome using

bowtie2. PCR duplicates introduced during the library preparation were removed from alignment files using picard-tool "MarkDuplicates" script. The aligned reads were then split among parental genomes, *M.musculus Castaneous* (Cast) or *Mus.musculus Domesticus* (129S) with SNPsplit (Krueger and Andrews, 2016) using "paired" parameter and SNP file containing the information regarding the 23,005,850 SNPs between Cast and 129S genomes (UCSC , provided by Dr. Guifeng Wei in the lab). For ChIP-seq the low quality and multiple mapped reads were also discarded to obtain uniquely mapped reads before performing SNPsplit, this was carried out using samtools view "-q 7" and "-v "XS:i"" criteria (Li et al., 2009). The binary aligned files (called bam files) were then used for downstream analysis and generation of bigwig files for visualization of data as "tracks" on UCSC genome browser. Generation of normalized bigwig files was done using deeptools "bamCoverage" with the parameter "scaleFactor (1 /read count)" where readcount is the number of reads successfully aligned to the reference genome for each sample, counted using samtools.

2.2.2 Counting reads in regions of interest

After mapping the paired-end reads, the bam files were used to count the number of reads overlapping regions of interest, for instance CGIs. The tool featureCounts V1.5.2 (Liao et al., 2014) takes an annotation file (defining the genomic coordinates for the regions of interest) and bam file as input, and generates a table containing the number of reads in a sample/library assigned to each individual region in the annotation file (in this example, every individual CGI) . This table (or count matrix) was used for downstream analysis such as finding enrichment levels in ChIP-seq (reads aligned to given region in pulldown divided those in input, IP/ Input), the gain of signal in differentiated cells as compared to undifferentiated cells (reads in EBs/reads in corresponding ES sample) and calculating allelic ratio.

For allelic ratio calculation of BioCAP- and MBD-seq, the reads overlapping the CGIs were counted by running featureCounts on allelically split bamfiles. After obtaining the count matrix from featureCounts describing the number of reads aligning CGIs in both the parental genomes, the CGIs covered in low sequencing depth (defined as reads from Xi + Xa < 8) were removed which could contribute noise to the dataset. The allelic ratio of filtered CGIs was then calculated (Xi/ (Xi +Xa)) and used as a measure of change in the signal of the Xi relative to the Xa, and subsequently to plot box and whisker plots. In BioCAP- and ChIP-seq data, the two X chromosomes had equal reads (giving an allelic ratio of 0.5) in ES cells and then the signal became allelically biased. For MBD-seq, unmethylated CGIs on ES cells gave no reads as the assay only captures methylated CpGs (i.e., Xi + Xa = 0). Therefore,

for MBD-seq data the filter of (Xi+ Xa ≥ 7) was applied at the last three time points of differentiation timecourse. All of this analysis downstream of featureCounts was performed in R.

2.2.3 Normalization of MBD- and ChIP-seq data using spiked-in human DNA

For MBD- and ChIP-seq data acquired from experiments calibrated with human DNA (or cells in the case of ChIP-seq), fastq files were aligned to genomic sequence containing concatenated mouse and human genomes using bowtie (as described in section 2.2.1). After the removal of PCR duplicates, the number of reads mapping to human and mouse genome were counted using samtools. A "normalization factor" was calculated using the formula derived and described in detail by (Hu et al., 2015). For clarity, the equation used is given below:

$$\text{Normalization factor} = \frac{\text{IP mouse}}{\text{IP human}} \text{ X } \frac{\text{Input human}}{\text{Input mouse}}$$

Where:

- IP mouse = number of reads in MBD-/ChIP-seq sample mapped to mouse genome
- IP human = number of reads in MBD-/ChIP-seq sample mapped to human genome
- Input mouse = number of reads in corresponding input sample mapped to mouse genome
- Input human = number of reads in corresponding input sample mapped to human genome

For every comparison, datasets were further normalized to a control. For example, when comparing the relative gain in CGI methylation during EB timecourse, ES cells at day 0 were considered as control with normalization factor 1, and other timepoints (day 4, day 8, day 10 and day 12) were normalized to day 0. In case of T7-DNMT3B ChIP, the wt T7-DNMT3B1 was used as normalization reference.

2.2.4 Peak-calling using MBD-seq data

Peaks are regions representing enrichment of signal (BioCAP-, MBD- or ChIP-seq) in an experiment in comparison to input. To call peaks, I used a software called MACS2 V2.2.7.1 (Zhang et al.) with parameters for paired-end bam file and q value of 0.01 ("-f BAMPE" and "-q 0.01") using the bam file for experiment as test "-t" and that for the corresponding input as control "-c". MACS2 provides a file containing a list of genomic coordinates (chromosome number, start and end of peak) as an output, termed as "bed file". The overlap among peaks or other regions of interest (such as CGIs or NMIs) was tested using a script/tool called bedintersect (Quinlan and Hall, 2010).

2.2.5 Finding differentially methylated regions between two conditions

Differentially methylated regions in this work were the differential MBD-seq peaks in the two tested conditions (i.e., ES versus EBs or wt vs knockout). This analysis was performed using a bioconductor package called DiffBind (Stark and Brown, 2011) which gives a pipeline starting from bamfiles to log2 fold changes of differential peaks and their corresponding significance values . A sample sheet was first prepared in which comma separated values (a .csv file), detailed the paths to IP bamfiles, input bamfiles, spike-in control bamfiles, MACS2 called peaks for each sample, and information like replicate and condition. This sample sheet was loaded in R and downstream analysis was performed as recommended by the authors of DiffBind (Stark and Brown). Briefly, upon loading the sample sheet, "dba.overlap" script was used to create a list of "consensus peaks" for analysis which included peaks replicated in at least 2 samples per condition. These peaks were then centred, their summits determined, and regions flanking 200 bp of the summit was used to make the list of consensus peaks. The samples were then normalized using "dba.normalize", which normalizes samples with calibration/spike-in control if provided, or by number of reads if the samples do not have a calibration control. For MBD-seq since all the samples did not include human spike-in, the samples were normalized by number of reads. The reads overlapping these consensus peaks were counted for all the samples using "dba.count" command. For performing comparative analysis, contrast was defined (such as wt versus mutant) on the basis of conditions specified in the sample sheet. Differential analysis was conducted using DESeq2 script inbuilt in DiffBind. The differential regions (MBD-seq peaks in this case) were defined as statistically significant if the false discovery rate (FDR) was less than 0.01. The log2 fold changes and FDR values were used to plot volcano and MA plots.

2.2.6 Finding genes associated with CGIs

CGIs as defined by (Illingworth et al., 2010), were allotted to the closest gene using a HOMER software (Heinz et al., 2010) command "annotatePeaks.pl". The output consists of a table describing the distance to the closest gene (as defined by RefSeq , (Pruitt et al., 2005)), the names of the gene and the genomic feature (exon, intron, promoter, intergenic etc) for every individual CGI. The CGIs found within 2 kb of a gene transcription start site (TSS) were defined as "TSS or promoter associated CGIs". This script was also applied to calculate distance from the nearest gene for "CGI-like elements" and "rising peaks" introduced in chapter 3.

2.2.7 Mapping chrRNA-seq fastq files

The procedure for mapping chrRNA-seq fastq files has been detailed in (Nesterova et al., 2019). Briefly, reads corresponding to rRNA contaminants were removed by aligning the paired-end reads to mm10 rRNA sequence using bowtie2 (Langmead and Salzberg, 2012) and discarding the aligned reads. The remaining reads were then mapped to N-masked mm10 genome using STAR (Dobin et al., 2013) with "--outFilterMultimapNmax 1", "--outFilterMismatchNoverReadLmax 0.06" , " --alignEndsType EndToEnd". As for the MBD-seq data, the aligned reads were then assigned to *M.musculus Castaneous* (Cast) or *Mus.musculus Domesticus* (129S) genomes using SNPsplit (Krueger and Andrews, 2016). Reads overlapping genes, both before and after splitting i.e., unsplit and split reads, were counted and assigned to the given genes using featureCounts (Liao et al., 2014) using "-t transcript" , " -g gene_id" and "-s 2" parameters. The gene annotation used for this analysis was obtained from Refseq (Pruitt et al., 2005).

2.2.8 Differential expression analysis

Bioconductor DESeq2 package was used to assess the statistically significant changes in the gene expression between various cell lines (for example wildtype and *Dnmt3b*$^{-/-}$ lines) (Love et al., 2014). The parameter p-adj < 0.05 was used to define the threshold for a gene to be called significantly differentially expressed. DESeq2 calculates a normalized mean (based on the average of normalized count divided by the size of library, calculated across all samples), termed "BaseMean", log2 fold changes from the BaseMean for a given sample and a corresponding p-value for each gene. The results for this analysis were visualized using MA and volcano plots generated using R and ggplot2.

2.2.9 Generation of heatmaps and profile plots

Profile plots (or metaplots) for the visualization of average MBD- or BioCAP-seq signal at CGIs was generated using deeptools V3.5.1 (Ramírez et al., 2014). First, the "computeMatrix scale-regions" command with parameter "binSize 50" on size normalized bigWig file was used. The output from computeMatrix was then used to run deeptools commands "plotProfile" or "plotHeatmap" using "region file" containing the genomic coordinates for the regions of interest (e.g., CGIs defined by Illingworth et al., 2010).

2.2.10 Defining statistically significant DNMT3A- or DNMT3B-associated proteins from mass spectrometry data after immunoprecipitation

The data obtained from the mass spectrometry after IP was analysed by two methods. (1) The mass spectrometry facility provided results using the software 'Perseus', which provides a comparative analysis (say between control and DNMT3B IP) by performing t-test and gives output as a volcano plot. (2) I then analysed the raw (intensity) data using a Bioconductor package called DEP which provides a standardized method to perform reproducible differential proteomics (Zhang et al., 2018). The pipeline first excluded proteins that are not identified/captured in a minimum of 2 out of 3 replicates for at least one condition i.e., setting threshold "thr=1" for "filter_missval" (Figure 2.2 B and C shows the proteins before and after applying this filter). The data was then normalized using two further pipelines that differed in their normalization and imputation methods. The first pipeline normalized the data with a widely used method called variance stabilizing transformation i.e., "normalize_vsn". The missing values were then imputed. Since my control cells do not express T7 tagged DNMT3A or DNMT3B, I expect the interactors of DNMT3A and DNMT3B to be absent in control. However, in order to perform differential analysis, these zero values need to be allotted a value. In the first pipeline, I imputed missing values with minimum probability, using DEP::impute with parameters "fun = "MinProb" and "q = 0.01". This method picks up a value from a Gaussian distribution centred on the minimum value. Afterwards, a differential analysis is performed between test and control (say DNMT3A and "blank/negative control" iXist-ChrX control) which gives log 2-fold changes and the corresponding p- and p-adjusted values (otherwise known as q-values/FDR) for all the protein hits.

In the second pipeline I used a package called MSnbase, widely used for proteomics analysis in combination with Bioconductor DEP. Data normalization using this pipeline is carried out using the median (i.e., MSnSet with "method=diff.median"). For the entire dataset a combined median is calculated and all the samples are normalized such that the median for all individual samples is equal to this combined median (Figure 2.2 D). The missing values are imputed using the minimum value for that sample (DEP::impute with fun="MinDet") (Figure 2.2 E and F). The differential analysis from this step onwards is similar to the first pipeline. Both the pipelines I used showed a good overlap in terms of the protein hits identified as potential interactors, between each other, as well as between the analysis performed by the mass spectrometery facility (Figure 2.2 G). Figure 2.2, 5.9 and 5.10, present the data that was obtained using the second pipeline (MSnSet, normalization method "diff.median", and imputation using "MinDet").

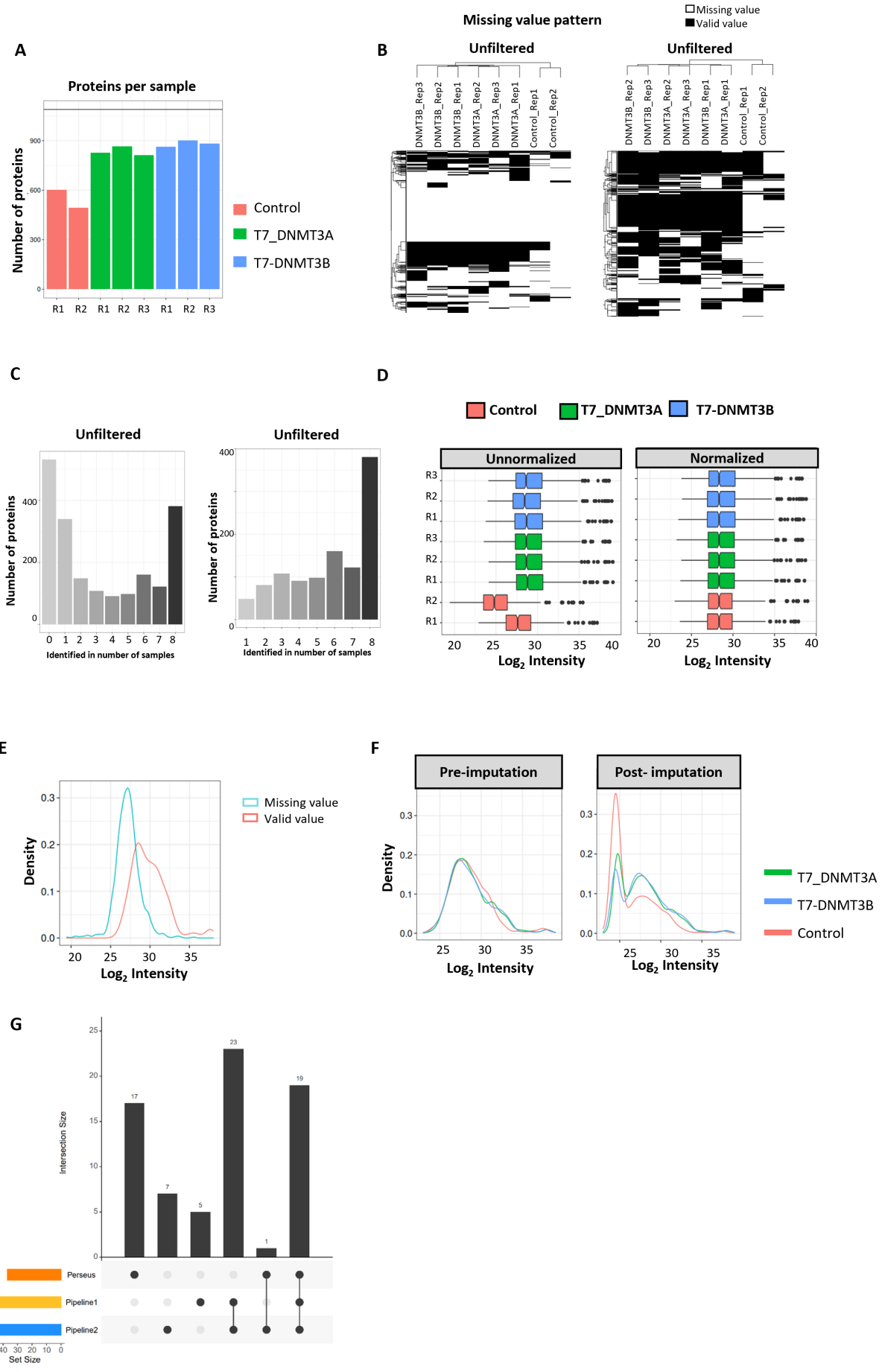

A
Proteins per sample
Number of proteins
Control
T7_DNMT3A
T7-DNMT3B
R1 R2 R1 R2 R3 R1 R2 R3
B
Missing value pattern
Missing value
Valid value
Unfiltered
Unfiltered
DNMT3B_Rep3
DNMT3B_Rep2
DNMT3B_Rep1
DNMT3A_Rep2
DNMT3A_Rep3
DNMT3A_Rep1
Control_Rep1
Control_Rep2
C
Unfiltered
Unfiltered
Number of proteins
Identified in number of samples
D
Control
T7_DNMT3A
T7-DNMT3B
Unnormalized
Normalized
Log2 Intensity
E
Missing value
Valid value
Density
Log2 Intensity
F
Pre-imputation
Post- imputation
Density
Log2 Intensity
T7_DNMT3A
T7-DNMT3B
Control
G
Intersection Size
Perseus
Pipeline1
Pipeline2
Set Size

Figure 2.2: Steps of mass spectrometry data analysis to identify statistically significant proteins associated with DNMT3A or DNMT3B after IP from nuclear lysates (shown on the page).

A) Abundance of proteins in all IP samples as detected by mass spectrometry.
B) Abundance of proteins in samples before and after filtering for low abundance proteins (see text for details).
C) Plot showing the proteins common among samples before and after filtering for the proteins not found in 2 out of 3 replicates for a given IP.
D) Data before and after normalization by median. A combined median is calculated for all samples in the analysis and all the samples are normalized such that their median is equal to the common median.
E) Plot showing the skew of proteins with missing values towards low intensity indicating non-random trend of the proteins missing which is expected when an IP is compared with negative control not expressing the IP-ed "bait" protein.
F) Density plot showing the distribution of intensity values for samples before and after imputation.
G) Overlap among the significantly enriched proteins in a representative IP from the data analysed by three pipelines. Pipelines from DEP analysis yielded larger set of interactors (see set size of pipeline 1 and 2 on the horizontal scale). The hits in the two pipelines showed 23 common interactors. 19 interactors were common between all the analysis pipelines.

2.2.11 Softwares and tools used in routine for data analysis

Software	Reference
MACS2 (v 2.2.7.1)	(Zhang et al., 2008)
Deeptools (v 3.5.1)	(Ramírez et al., 2014)
Samtools (v1.12)	(Li et al., 2009)
BEDtools (v 2.27.1)	(Quinlan, 2014)
HOMER	(Heinz et al., 2010)
DESeq2	(Love et al., 2014)
DiffBind	(Stark and Brown, 2011)
DEP	(Zhang et al., 2018)
Bowtie2	(Langmead and Salzberg, 2012)
SNPsplit	(Krueger and Andrews, 2016)
UCSC genome browser	(Kent et al., 2002)
R Bioconductor	(Huber et al., 2015)
Software	**Reference**

Star	(Dobin et al., 2013)
R ggplot2	(Wickham, 2016)
FeatureCounts (v 1.5.2)	(Liao et al., 2014)

Table 2.10: Softwares and packages used in routine NGS data analysis

2.2.12 Publicly available datasets

In this study I have made use of several datasets. The references for these are provided in table below.

Reference	GEO accession	Data
(Gdula et al., 2019) GSE115984	GSE122094	Table with methylation levels (as determined by WGBS) on X-linked CGIs in wt and SMCHD1 mutant MEFs provided by Dr. Michal Gdula upon request.
(Illingworth et al., 2010) Datasets S1	N/A	List/coordinates of CGIs provided in Dataset S1.
(Long et al., 2013) GSE43512	GSM1064680	List/coordinates for NMIs.
(Baubec et al., 2015) GSE57413	GSM1382257 GSM1382258 GSM1382259 GSM1382260 GSM1382265 GSM1382266	DNMT3A ChIP-seq ES cells (replicate 1) DNMT3A ChIP-seq ES cells (replicate 2) DNMT3B ChIP-seq ES cells (replicate 1) DNMT3B ChIP-seq ES cells (replicate 2) DNMT3B ChIP-seq NPC (replicate 1) DNMT3B ChIP-seq NPC (replicate 2)
(Bowness et al., 2022) GSE185871	GSE185871	List of fast, medium, and slow silencing genes during XCI in iXist-ChrX line determined by chrRNA-seq, provided upon request by Dr. Joseph Bowness. List of gene expressed at high, intermediate, and low levels in iXist-ChrX ES cells as determined by chrRNA-seq, also provided upon request by Dr. Joseph Bowness.

Table 2.11: Published datasets used in this work.

2.2.13 Datasets from other lab members analysed

In this study I have made use of several datasets generated by other members of the lab that are unpublished. Description of these datasets is as follows.

Dataset	Dataset generated by
ATAC-seq (wt iXist-ChrX ES cells)	Dr. Joseph Bowness
H3K4me1 ChIP-seq (wt iXist-ChrX ES cells)	Dr. Joseph Bowness
H3K9me3 ChIP-seq (wt iXist-ChrX ES and day 10 EBs)	Dr. Tatyana Nesterova
chrRNA-seq (wt and *Dnmt3b*$^{-/-}$ iXist-ChrX ES and day 15 NPCs)	Dr. Tatyana Nesterova

Table 2.12: Unpublished datasets generated by other lab members used in this work.

Chapter 3

Establishment of CGI methylation during differentiation



3.1 Introduction

As discussed in 1.3.3.3, CGI methylation occurs late in XCI and ensures the maintenance of transcription silencing in a mitotically stable manner (Blewitt et al., 2008). I have also described that an earlier study from our lab has shown that CGIs on Xi can be categorized into two classes, a small class of CGIs that acquire DNA methylation with fast dynamics, and a larger group of slow methylating CGIs (Gendrel et al., 2012). These dynamic classes were first defined based on a few candidate CGIs on Xi in female mouse ES cells differentiated to embryoid bodies (EBs). Fast and slow methylating CGIs gained significant levels of methylation either at day 7 or day 10 of EB differentiation, respectively. In order to define these two classes, somatic XX and XY lines (which represented the complete and unmethylated states of CGIs respectively) were used as a reference to compare the gain of methylation at these CGIs. These findings were also confirmed in mouse embryos where the fast CGIs gained methylation earlier than slow CGIs, at E9.5 and E15.5 respectively (Gendrel et al., 2012). Another key finding of that study was the discovery of the specificity of DNMT3B for CGI methylation on Xi as evident from the hypomethylation of CGIs in *Dnmt3b* null E9.5 embryos.

This prior study however has a few limitations. (1) The experimental system made use of an XX ES cell line undergoing random X inactivation facilitated/triggered by EB differentiation. This system has the limitation of not allowing allelic analysis to study the status of Xa and Xi independently. Additionally, the system gives rise to a heterogenous population of cells that initiate XCI at different times during the timecourse, which in turn may affect the accuracy with which the gain of methylation at CGIs is measured. (2) Throughout the study, different model systems have been used (such as EBs, MEFs, and mouse embryos), for exploring various aspects of CGI methylation on Xi. For instance, the specificity

of DNMT3B was analysed in mouse embryos while the contribution of SMCHD1 in methylation of X-linked CGIs was studied in MEFs. These experimental systems vary in their differentiation state and biology, and so are not directly comparable with each other. Thus, there is a need of a unified experimental system in which the findings from (Gendrel et al., 2012) can be validated. (3) Finally, the study makes vast use of an obsolete technique, Sequenom EpiTyper analysis (described in Table 1.1, a method that employs bisulphite treatment of genomic DNA followed by mass spectrometry for the analysis of methylation status of regions of interest), for profiling CGI methylation. Sequenom EpiTyper, due to the need for special primer design for every region of interest, allowed the study of only a limited number of CGIs. For instance, the specificity of DNMT3B was analysed only at 7 selected X-linked CGIs using Sequenom EpiTyper analysis. Towards the end of their study (Gendrel et al., 2012) make use of high throughput sequencing of MBD affinity purification (MBD-seq), but only for late stage EBs (day 7 and day 10). The study, therefore, did not take into account the methylation status of the CGIs at the beginning of the timecourse and so did not accurately gauge the gain in methylation achieved during XCI. Thus, an improved model system must be developed that overcomes the shortcoming of the previous study. Importantly, the advancement in the techniques for CGI methylation profiling and better affordability of NGS technology, allow a more in-depth analysis of CGI methylation than in prior work. For instance, with the advent of BioCAP and improved MBD affinity purification, CGIs can be studied using less DNA and affinity module (recombinant CXXC or MBD) in contrast to tedious column-chromatography-based methods used in past.

To improve the analysis of CGI methylation in XCI, I set out to develop a novel system to obtain high-quality/resolution datasets capable of capturing the dynamics of methylation of individual CGIs on Xi. This system also provides allelic analysis of CGI methylation gained on Xi (whereby Xa serves as an internal control), and genome-wide during differentiation in mouse ES cells. An in-depth analysis of the behaviour of CGI methylation would then act as a standard against which I can compare the CGI methylation dynamics in cell lines engineered with deletions and mutations of the DNA methyltransferases, the key players for determining CGI methylation on Xi. This chapter characterizes the basics of CGI methylation during XCI, the impact of deletion of *Dnmt3b* on CGI methylation on Xi and describes the rationale for the computational analysis methods which set the groundwork for the experiments in chapter 4.

3.2 Tet-On Xist mouse ES model cell line (iXist-ChrX mouse ES cells)

To overcome the limitations in previous studies I made use of a female mouse ES cell line engineered in the lab (Nesterova et al, 2019). In this cell line, the native Xist promoter on one X chromosome has been replaced with a Tet-inducible promoter using CRISPR/Cas9-assisted homologous recombination (thus termed iXist-ChrX ES cells). The gene for Tet trans-activator (rtTA) protein under the control of pCAG promoter has been inserted into the TIGRE locus on chromosome 9 to allow constitutive expression of rtTA (Zeng et al., 2008) (Figure 3.1 A). This system allows cells to synchronously express Xist RNA upon treatment with doxycycline, added as a medium supplement, and overcomes the heterogeneity in cell models where Xist is expressed from its native promoter. Importantly, the parental line for these cells has been derived from the F1 generation of a cross between two diverged mouse strains *Mus musculus Domesticus* (129/SvlmJ) and *Mus musculus Castaneus* (CAST/EiJ). This allows the two alleles to be distinguished based on single nucleotide polymorphism (SNPs) when NGS data is analysed. Notably, the allelic analysis enables us to discriminate NGS reads mapping to the Xi, *Castaneus* (CAST/ EiJ) allele carrying the Tet-inducible Xist, upon doxycycline induction, relative to the Xa which serves as an internal control for any variability introduced in experiment and sample preparation.

Since DNA methylation is known to be a differentiation-coupled process, I have differentiated these iXist ES cells using either a heterogenous EB (Figure 3.1 B), or a more homogenous neuronal progenitor cell (NPC) differentiation method. EB differentiation makes use of the natural tendency of ES cells to spontaneously differentiate in culture in absence of LIF (leukaemia inhibitory factor), an ES media component responsible for maintaining the pluripotent state of ES cells. In NPC differentiation, however, the cells are treated with growth factors and supplements that direct them into the neuronal lineage. Both of these methods were tested and compared as part of the optimization of this experimental system to study CGI methylation subsequently in engineered cell lines.

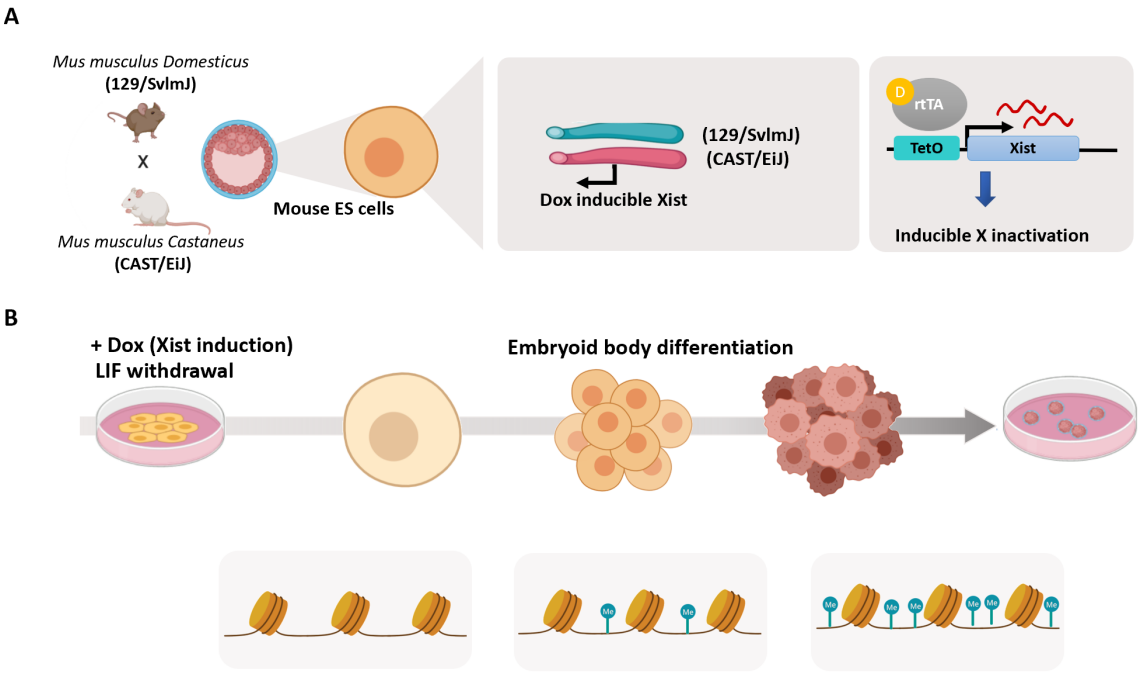


Figure 3.1: The iXist-ChrX model cell line.

A) Schematic depicting the key features of the iXist-ChrX model cell line. This study makes use of interspecific female mouse ES cell line engineered to contain a Tet-inducible Xist promoter on the *Castaneus* originating X chromosome.

B) Experimental setup for embryoid body differentiation (EB) of iXist-ChrX cells. Cells can be collected at various stages of differentiation to study the dynamics of CGI methylation during differentiation.

3.3 BioCAP-seq - a method to study CGI methylation dynamics on Xi

For the development of a robust experimental system to study CGI methylation, I experimented with two techniques, BioCAP- and MBD-seq. In section 1.2.7, I have compared the molecular techniques used to study CpG methylation. WGBS is considered a gold standard for studying cytosine methylation due to its genomewide coverage and ability to provide a quantitative measure of the methylation levels at a given locus. WGBS however requires large sequencing depth which makes it expensive especially if applied to a large number of samples. On the other hand, IP-based techniques like BioCAP- and MBD-seq enrich for CpG-rich genomic DNA, which makes them an ideal tool for the analysis of CGIs. Owing to the small proportion of CpG-rich DNA, this technique recovers a small fraction of the

genome, thus allowing an effective study of the methylation status of individual CGIs using NGS at much lower sequencing depth than required by techniques like WGBS. This in turn makes MBD- and BioCAP-seq highly affordable for my study which requires numerous CGI methylation profiling experiments. Thus, although MBD- and BioCAP-seq do not offer a precise quantitative measure of the methylation levels on CGIs, their high affordability makes them suitable for the purposes of this study.

The BioCAP-seq method was developed in 2012 to study the methylation status of CGIs (Blackledge et al., 2012). The workflow is illustrated in Figure 3.2 A. The principle of this technique is the affinity purification of unmethylated CpG-rich DNA using zinc finger CXXC domain protein. The original BioCAP protocol makes use of the CXXC domain derived from human H3K36 demethylase KDM2B, which *in vivo* recruits this protein to unmethylated CGIs. The CXXC-bound unmethylated DNA can then be eluted using a high salt step gradient. Biotin conjugated CXXC domain protein used in BioCAP, was expressed using a plasmid vector coding for a fragment of KDM2B CXXC domain (600-750 amino acids) with a C-terminus Avi-tag, and a gene for BirA. BirA is an enzyme that conjugates biotin to lysine residue in Avi-tag sequence. This allows the CXXC domain protein to get biotinylated while being expressed in the *E.coli* expression strain BL21. The biotinylation of the CXXC domain enables the use of neutravidin beads to facilitate affinity pulldown and subsequent elution of unmethylated CGIs.

Since the BioCAP reagents are not commercially available, I purified CXXC-Avi using the *E.coli* expression strain Rosetta 2 (DE3) pLysS, following the protocol described in the paper (Blackledge et al., 2012), with the help of a colleague in the lab, Dr. Antonio Biasutto. The cloned plasmid for the expression of CXXC-Avi protein was kindly provided by Prof. Rob Klose, Oxford Biochemistry. As a test, I performed BioCAP-seq on the iXist-ChrX line differentiated to EBs for 10 days. For the test experiment, the eluted fractions were collected with 500-, 700- and 1000-mM salt (represented as elution 2, 3, and 4, E2-E4) and analysed via NGS sequencing. In BioCAP-seq data, the enriched hypomethylated CpG-rich DNA is visible as “peaks” on the genome browser tracks (Figure 3.2 C). As shown in Figure 3.2 B and C, the CGIs were captured at lower salt (E2), however, the peaks are slightly broader. E3 and E4 efficiently capture the CGIs; this agrees with previous observations (Blackledge et al., 2012). In a subsequent study from the Klose lab (Long et al., 2013), E3 and E4 were combined as immune-purified “IP” fraction and used for the downstream analysis. Since in this test experiment, CGIs are mostly captured in E3 and E4 as well (Figure 3.2 B), I have followed that same strategy for the subsequent experiments.

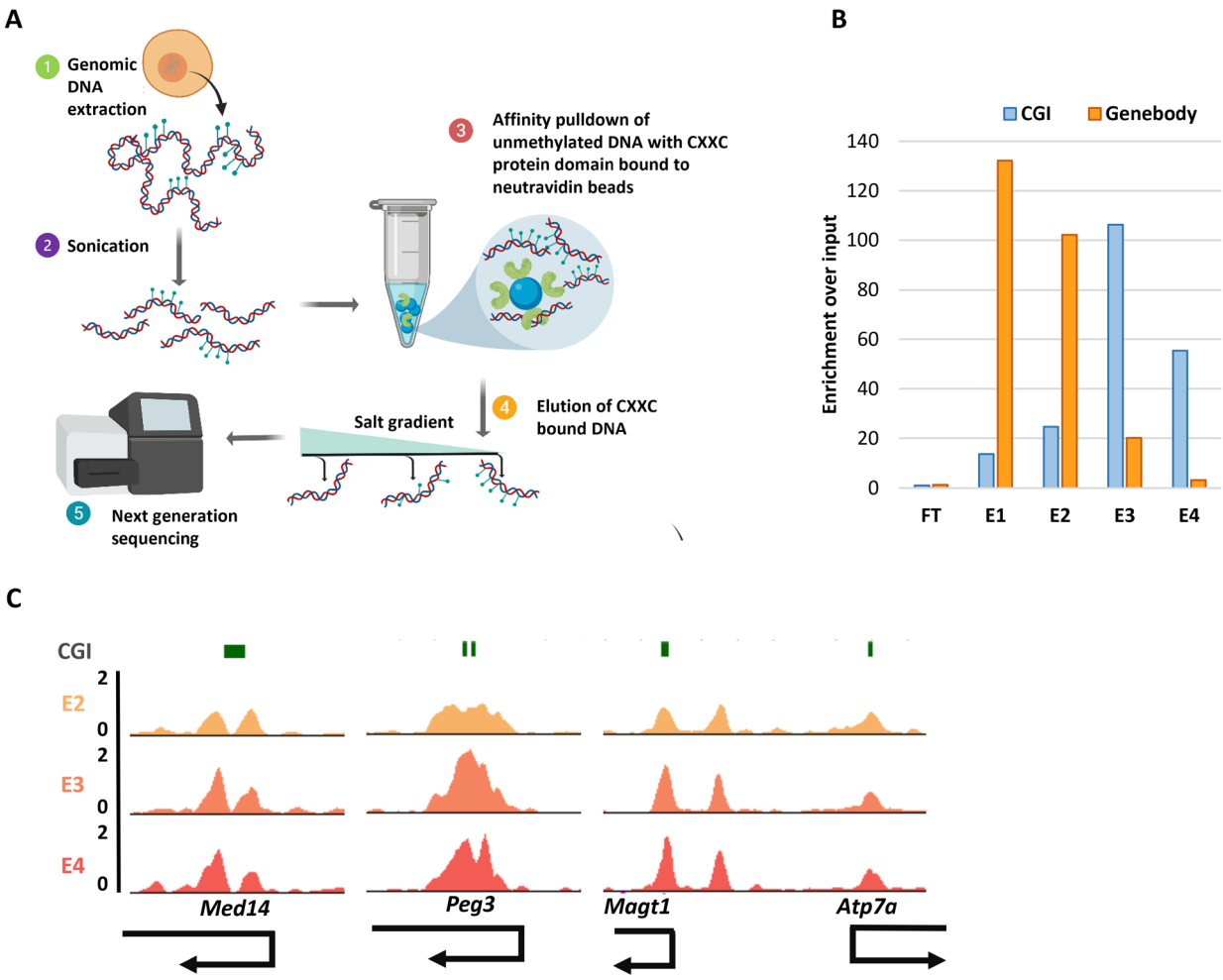


Figure 3.2: BioCAP-seq for studying methylation of CGIs.

A) Schematic illustrating the principle and working of BioCAP-seq. Sonicated genomic DNA is used as input for affinity purification of hypomethylated CpG-rich regions by neutravidin bead-conjugated CXXC domain. The CXXC-bound DNA is eluted with salt step gradient and used for NGS.

B) qPCR validation of BioCAP-seq. Prior to sequencing, qPCR was used to estimate the signal-to-noise ratio of BioCAP. Hypomethylated CGIs are expected to bind firmly to CXXC protein and therefore elute at higher salt concentrations compared to methylated gene body regions. This plot shows the qPCR amplifying CGI and CpG-rich gene body region of *Kmt5b* (also called *Suv420h1*). (FT: unbound flowthrough, E1-E4: Elution 1-4. Salt concentration ranges from 100mM in FT to 300 mM, 500 mM, 700 mM, and 1 M in E1-E4).

C) Genome browser tracks showing BioCAP-seq signal at few representative CGIs in eluted fractions E2-E4. From the qPCR and genome browser tracks, it can be seen that E3 and E4 show higher enrichment and specificity for unmethylated CGIs. The position and span of CGIs are depicted by green bars. Panels (B) and (C) show representative data from two well corelated replicates (see Appendix 2).

To test the pulldown efficiency of my CXXC-Avi batch (Test-CXXC) relative to the previously published result (Long et al., 2013), I performed BioCAP-seq with CXXC protein stock (generously gifted to me by the Rob Klose lab) in parallel with mine to serve as a reference (Ref-CXXC). Figure 3.3 A (left) shows the overlap between the peaks called in BioCAP-seq experiment performed using Ref-CXXC and those described in Long et al., 2013 publication (termed by authors as non-methylated islands, NMIs). Here peaks can be defined as the genomic loci showing enrichment of mapped NGS reads in IP, using input as baseline control. This analysis was conducted using a software called MACS2, routinely used for the analysis of ChIP-seq datasets. The clear overlap between the Ref-CXXC peaks and NMIs shows the BioCAP-seq experiment worked well (Figure 3.3 A, left). The small variability observed can be explained by differences in cell lines and variations in pulldown efficiency which is expected in IP-based experiments.

The Test-CXXC experiment demonstrated a comparable efficiency to Ref-CXXC as indicated by the overlap between the respective peaks (Figure 3.3 A middle panel compares BioCAP-seq peaks between Test-CXXC and NMIs (Long et al., 2013), while Figure 3.3 A right shows Venn diagram comparing BioCAP-seq peaks in Test-CXXC and Ref-CXXC performed in iXist-ChrX cells). Figure 3.3 B shows a genome browser screenshot for selected CGIs. It can be appreciated that the signal originating from the Test-CXXC is smaller than the Ref-CXXC, which could suggest that the Test-CXXC batch was not performing pulldowns as efficiently. Although these genome browser tracks have been size-normalized, it is hard to draw a strong conclusion regarding the relatively poor efficiency of Test-CXXC protein without having a quantifiable measure such as the inclusion of exogenous DNA that could serve as a (spike-in) control for measuring variability introduced during IP, or library preparation steps.

The assay maintained a good signal-to-noise ratio upon splitting the data and assigning it to parent-of-origin using SNPs, validating the suitability of the technique for studying CGIs on Xi (Figure 3.3 C and D). Figure 3.3 D includes a track for WGBS data (Gdula et al., 2019), showing representative autosomal CGIs. WGBS data agrees well with BioCAP-seq, whereby the hypomethylated CGIs in WGBS (visible as “dips/depressions”) align with peaks in BioCAP-seq track.

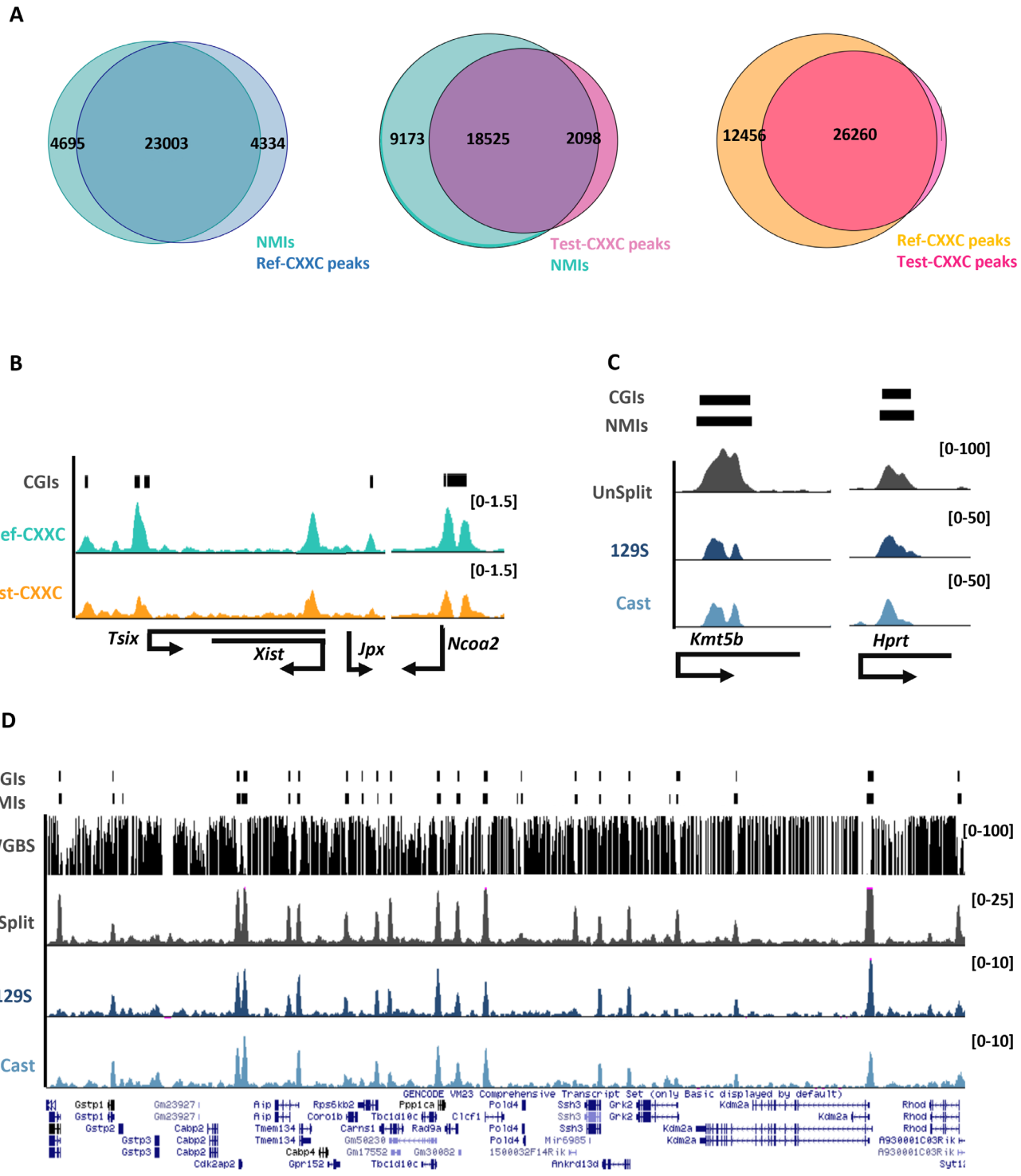

A
4695
23003
4334
NMIs
Ref-CXXC peaks
9173
18525
2098
Test-CXXC peaks
NMIs
12456
26260
Ref-CXXC peaks
Test-CXXC peaks
B
CGIs
Ref-CXXC
[0-1.5]
Test-CXXC
[0-1.5]
Tsix
Xist
Jpx
Ncoa2
C
CGIs
NMIs
[0-100]
UnSplit
[0-50]
129S
[0-50]
Cast
Kmt5b
Hprt
D
CGIs
NMIs
WGBS
[0-100]
UnSplit
[0-25]
129S
[0-10]
Cast
[0-10]
GENCODE VM23 Comprehensive Transcript Set (only Basic displayed by default)

Figure 3.3: Optimization of BioCAP-seq for studying CGIs on Xi (shown on the previous page).

A) Venn Diagrams showing the overlap of peaks called in BioCAP-seq performed using Ref-CXXC with NMIs reported in Long et al 2013 (left). (Middle and right) Venn diagrams compare the peaks called in BioCAP-seq performed using Test-CXXC (expressed for this work) with NMIs and Ref-CXXC in the iXist-ChrX line respectively. BioCAP-seq (using Test-CXXC or Ref-CXXC) was performed in the iXist-ChrX line differentiated to EBs.
B) Genome browser tracks comparing BioCAP-seq signal performed using Ref-CXXC and Test-CXXC (top and bottom track respectively). Both experiments show comparable enrichment profiles.
C) Genome browser tracks showing BioCAP-seq data from Test-CXXC. Unsplit data (top track) can be further split to assign reads to parent-of-origin on the basis of SNPs. The bars on the top indicate the position of CGIs (defined in the publication by Illingworth et al., 2010), and NMIs (defined in Long et al., 2013).
D) BioCAP-seq peaks align well with unmethylated regions in WGBS. Comparison with a sensitive technique like WGBS further validates the efficiency of BioCAP-seq. CGIs and NMIs mostly overlap but NMIs are an accurate definition of regions expected to enrich in BioCAP-seq. Region shown is Chr19: 4,005,985-4,553,949. The figure shows representative data form two well correlated replicates for Test-CXXC (Appendix 2). BioCAP-seq for Ref-CXXC was performed only once.

3.4 BioCAP-seq on iXist-ChrX line to study CGI methylation dynamics during EB differentiation timecourse

To study the dynamics of CGI methylation during XCI, I collected cells at 5 timepoints during EB differentiation, starting from ES cells (day 0) till day 12, and performed BioCAP-seq (Figure 3.4 A). To measure the relative gain in CGI methylation on Xi, I use an allelic ratio metric. As mentioned earlier, the iXist-ChrX experimental system allows us to compare Xi with Xa due to allele-specific SNPs. Allelic ratio is the ratio of sequencing reads from Xi to the total reads from X chromosome (i.e., allelic ratio = Xi / (Xi +Xa)). Xa provides an internal control that is fairly robust to variability introduced by pulldown experiments and library preparation and is, therefore, a good metric to make comparisons between timepoints. On day 0, the signal from both X chromosomes (Xa and Xi) is equal, which gives an allelic ratio of 0.5. As Xi is expected to gain CGI methylation during XCI, the BioCAP-seq signal, a measure of the hypomethylated status of CGIs, is expected to drop. Hence, with the progress of EB differentiation, the allelic ratio is expected to decrease from 0.5. For this analysis, I used the list of NMIs defined by (Long et al., 2013) as a proxy for CGIs. As shown in Figure 3.3 A, NMIs were efficiently captured in my BioCAP-seq. NMIs also show a significant overlap with CGIs defined by (Illingworth et al., 2010) from CAP-seq (Figure 3.3 C and D). As expected, the allelic ratio dropped during differentiation, albeit only

to a nominal degree, i.e., from 0.5 in ES cells to 0.4 on day 12 EB (Figure 3.4 B left). This small drop in allelic ratio could indicate that a majority of NMIs on Xi stay unmethylated during EB differentiation. The unmethylated NMIs, if predominant, would therefore mask those undergoing methylation. Another factor to consider is the NMIs that are not covered in sufficient sequencing depth; these NMIs are amenable to large changes in the number of reads between the timepoints and skewing allelic ratio, which adds noise to the data. Thus, the NMIs used in the analysis must be filtered. Out of 606 annotated NMIs, 386 NMIs were filtered such that the total reads from X chromosomes per NMI were at least 8 (i.e. $Xi + Xa \geq 8$); a relaxed criterium to accommodate for the loss of large proportion of reads upon their assignment to parent of origin based on SNPs, while avoiding large skews in allelic ratio (Figure 3.4 B, middle panel). In this group of 386 NMIs, a large majority of NMIs displayed ambiguous profiles where the allelic ratio did not follow any proper trend during the timecourse; these NMIs were excluded from the analysis as well (Figure 3.4 D). Even with this thorough filtering, the median allelic ratio at day 12 of EB differentiation was only 0.3 (Figure 3.4 B, right, and 3.4 C).

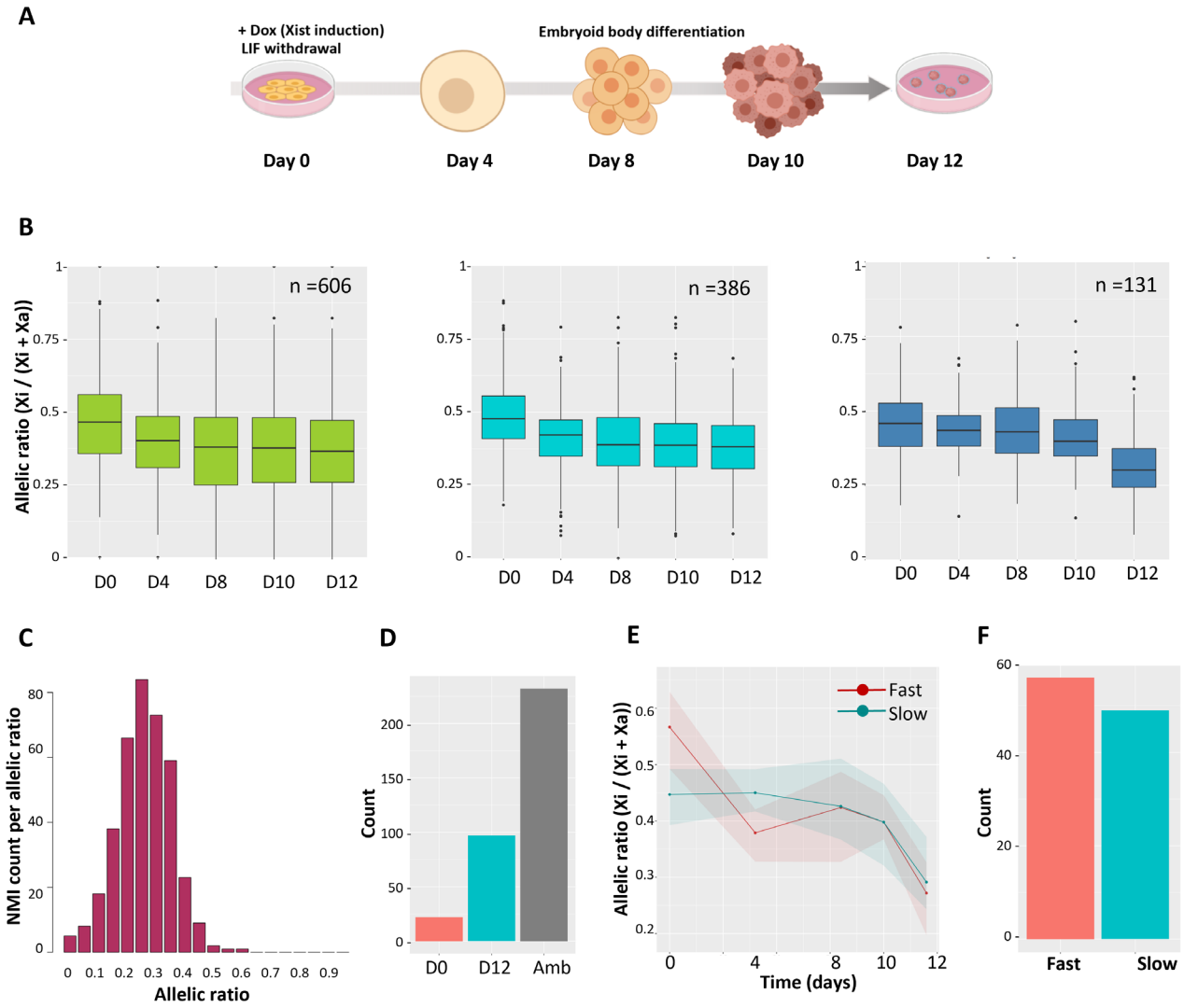


Figure 3.4: BioCAP-seq on EB timecourse in the iXist-ChrX cell line (shown on the previous page).

A) Schematic showing experimental design and the timepoints collected for BioCAP-seq experiment during EB differentiation.
B) Boxplots showing the allelic ratio change at the 5 timepoints collected during EB differentiation of iXist-ChrX line. (left) shows all NMIs which were then filtered to keep the ones with suitable depth (middle). (Right) shows the NMIs that depict a constant trend of decreasing allelic ratio, which could be used to study methylation dynamics. The number of NMIs in each group is detailed at the top-right corner of the plot. D0-D12 indicate day 0-12 of EB differentiation.
C) Histogram showing the distribution of NMIs for the given value of allelic ratio. Out of the filtered 386 NMIs (as shown in B (middle)), half of the NMIs have an allelic ratio below 0.3.
D) Counts of NMIs plotted against the day at which they show the minimum value of allelic ratio, i.e., highest level of DNA methylation. D0 (Day 0) represents the group of NMIs that did not gain any methylation and stayed unmethylated, while D12 (Day 12) is the group of NMIs that kept gaining methylation till the end of the differentiation. A majority of NMIs did not show any proper trend; these are annotated here as ambiguous (Amb).
E) Line graph showing the allelic ratio changes during the timecourse for "fast" and "slow" classes of NMIs. The solid line represents the median allelic ratio, and the ribbon shows the interquartile range. X-axis of the plot shows time in days of EB differentiation timecourse.
F) Counts for the NMIs classified in "fast" or "slow" methylation dynamic classes. This figure shows data from one of the two well correlated replicates (see Appendix 2).

I then looked at the dynamics at which NMIs were getting methylated. For each NMI, I selected the earliest timepoint at which the allelic ratio dropped by 0.1. Two main trends emerged, about half (n = 56 out of 131 used in this analysis) showed a "Fast" dynamic and a drop in allelic ratio at day 4, while the "slow" NMIs (n = 43) showed a progressive decline in allelic ratio up to EB day 12 (Figure 3.4 E and F and Figure 3.5 A). In Figure 3.5 B, the heatmaps show individual NMIs in each category (fast, slow and ambiguous). Figure 3.5 C shows genome browser tracks for examples of "fast" and "slow" dynamic NMIs, *Ndufb11* and *Med14* respectively.

I compared the NMIs assigned to both the dynamic categories with those described previously (Gendrel et al., 2012). *Mtm1*, reported to exhibit fast dynamics in the study of Gendrel et al., was indeed found to be methylated before day 8. *Slc25a14*, another "fast" NMI, however, showed slow dynamics in this experiment. Some well characterized "slow" NMIs such as *Hprt, Atrx* and *Mecp2* were filtered into the "ambiguous" category and showed no proper trend in this experiment (Figure 3.5 D). Some disagreement between the methylation timing for individual NMIs in my experiment and those reported previously (Gendrel et al., 2012), can be expected as the model systems used for the two

studies are different and EB differentiation is a heterogenous system which can vary from experiment to experiment. Nevertheless, the overall small drop in allelic ratio observed in BioCAP-seq, combined with variability introduced in EB differentiation, indicates that the method may not be robust enough to be used routinely for studying the dynamics of individual CGI methylation on Xi.

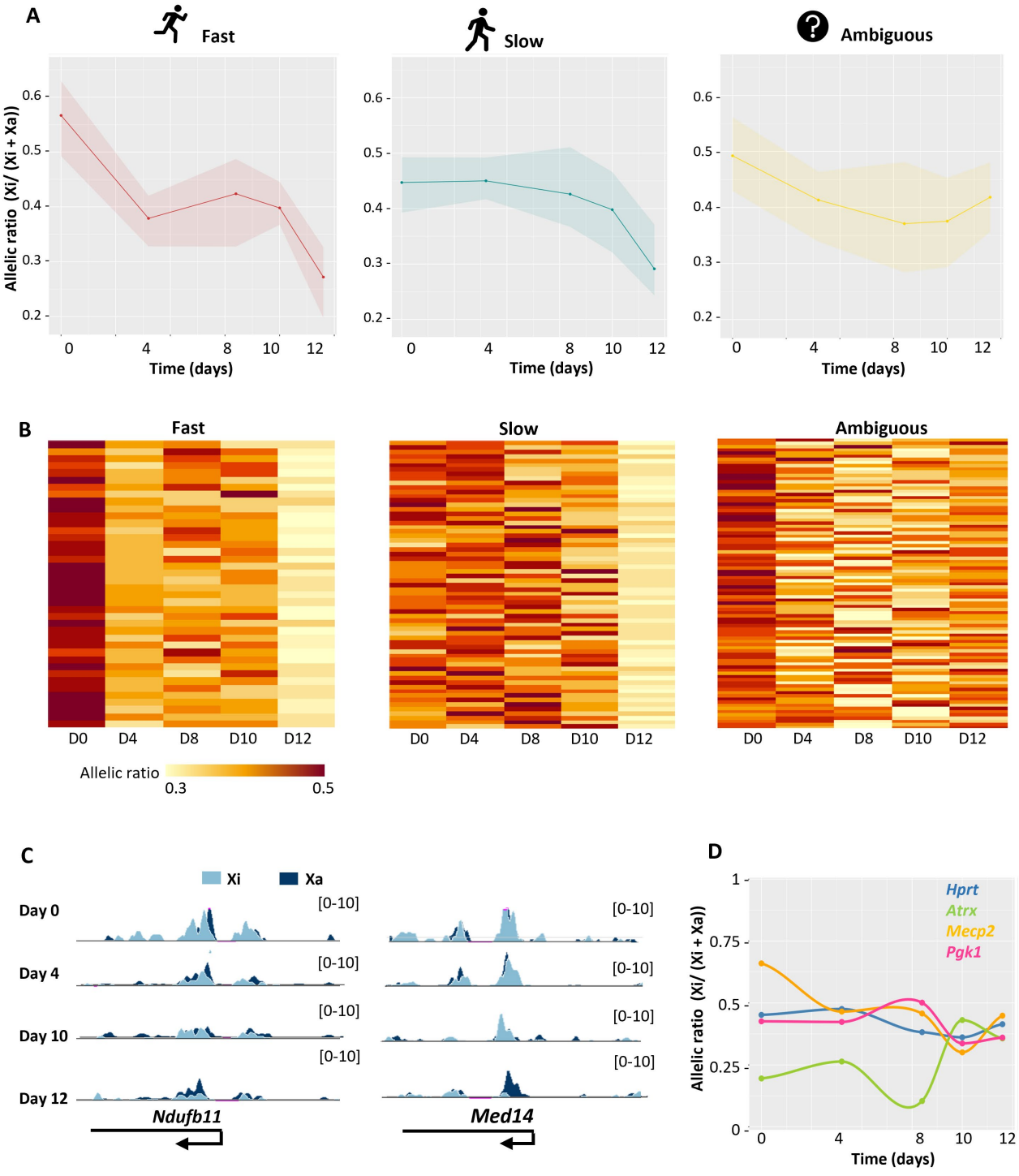


Figure 3.5: Methylation dynamics of NMIs during EB differentiation (shown on the previous page).

A) Line plots showing the changes in allelic ratio for NMI dynamic classes during EB differentiation. (Left) "Fast" methylating NMIs show a sharp drop in allelic ratio at day 4 and continue to be methylated until day 12. (Centre) "Slow" NMIs gain methylation at day 12. (Right) NMIs in "ambiguous" class show a wide variation around day 8 and do not show a clear trend towards lower allelic ratio during the timecourse. In these plots, the solid line shows the median allelic ratio, and the shaded ribbon shows the interquartile range for the NMIs in the group. X-axis is time in days for EB differentiation timecourse.
B) Heatmaps showing the allelic ratio (used to define the methylation dynamics) for individual NMIs in the three categories shown in (A). Each line or tile represents a single NMI. The "fast" NMIs show a drop in allelic ratio at day 4 (visible as a change to light colour of the tiles), in contrast to "slow" NMIs, which do so at day 10. Allelic ratios of "ambiguous" class of NMIs show no proper trend during the timecourse. Being the largest group, ambiguous NMIs are responsible for adding noise to the data.
C) Genome browser tracks showing a representative "fast" and "slow" NMIs. Each track shows signal from Xa and Xi. In *Ndufb11* the Xa (dark blue track) has a larger signal than Xi (light blue) at day 4 which lowers the allelic ratio form 0.5. in *Med14*, Xa and Xi signal are comparable until day 12.
D) Line plots showing allelic ratio for a few exemplary NMIs filtered as "ambiguous" in this analysis. This figure shows data from one of the two well correlated replicates for BioCAP-seq performed on iXist-ChrX lines during EB differentiation (see Appendix 2).

3.5 Testing the sensitivity of BioCAP-seq using MEFs

To further test the sensitivity of BioCAP-seq for analysing CGI methylation in XCI and to exclude the variability introduced by differentiation, I performed BioCAP-seq in an XX MEF line. This MEF line was derived from E9.5 embryos, where XCI has already been completed and a significant number of CGIs show high levels of methylation. The MEF line was also interspecific (Fvb/Cast) and was subcloned from primary embryo fibroblasts to select a clone with, in this case, *Castaneus* X chromosome as Xi, allowing allelic analysis of sequencing data similar to the model iXist-ChrX ES cell line described above. The rationale behind using MEFs to test the sensitivity of BioCAP-seq is that (1) they represent a fully differentiated cell type where XCI is complete, making them a model system of choice to study late-stage processes of XCI. (2) Most of the CGIs in MEFs have obtained a high level of methylation and therefore provide a good "end-point" reference for CGI methylation on Xi. (3) The availability of WGBS data for this line (Gdula et al., 2019) meant that I had high-resolution, quantitative information

regarding the methylation status of individual CGIs to provide as a reference. In the WGBS data set 418 CGIs were informative (out of a total of 514 CGIs on the X chromosome), and 366 CGIs could be seen to be methylated by 50% or more (Figure 3.6 A). I also calculated the relative methylation on Xi (i.e., Xi/Xa) after filtering out the CGIs that had no reads on Xa in order to avoid undefined/infinite values for the ratio. 248 CGIs showed at least 2-fold more methylation on Xi relative to Xa (Figure 3.6 B, majority of CGIs show Xi/Xa ≥ 2).

Since IP-based methods are more limited in resolution, depth, and coverage in comparison to WGBS, I used an alternative IP-based assay, MBD-seq to serve as a comparison for BioCAP-seq. As described in section 1.2.7, MBD-seq gets its name from the MBD proteins (see chapter 1, section 1.2.6) which are used for affinity purification of methylated CpG-dense DNA (the assay is thus called MBD affinity purification or as MBD-seq in short). For MBD assay, I used a commercially available kit from NEB. The kit employs a fragment of human MBD2 protein which is fused to the Fc fragment of a human antibody IgG1 (the fusion protein called MBD-Fc). The IP is performed using magnetic protein A beads which bind to the Fc domain of the MBD-Fc protein. The input DNA for both BioCAP- and MBD-seq came from the same culture/population of Fvb/Cast MEF line, the same high-purity DNA preparation, and sonication to avoid any confounding variability. Before sequencing, I performed qPCR to validate that the IP has worked well for both BioCAP and MBD (Figure 3.6 D).

As before, for the allelic data analysis, I used the allelic ratio metric (Xi / (Xi + Xa)) introduced in section 3.4. Since BioCAP enriches for unmethylated CpGs, more signal originates from Xa in comparison to Xi upon CGI methylation, and the allelic ratio decreases from 0.5 as the CGI methylation is gained. MBD, on the other hand, pulls down methylated DNA, thus for methylated CGIs, Xi would have more signal than Xa, increasing the allelic ratio. In order to represent the gain in CGI methylation from these distinct methods, I used [1-allelic ratio] for analysis of BioCAP-seq data in this experiment; thus, the gain in methylation would cause an increase from 0.5 in both cases. CGIs were defined using the complied list of regions published by (Illingworth et al., 2010) as this definition was used previously for the analysis of WGBS data (Gdula et al., 2019). The CGIs defined by Illingworth et al., are a subset of NMIs that I used in the last section (section 3.4).

MBD-seq data had a much-improved signal-to-noise ratio in comparison to BioCAP-seq. Figure 3.6 C compares the proportion of reads assigned to annotated CGIs, which is 10-fold higher in MBD-seq than BioCAP-seq (~30% MBD-seq and ~3% in BioCAP-seq). This suggests that more reads in BioCAP-seq contribute towards the background. This higher signal-to-noise ratio of MBD-seq was also apparent from the genome browser view. Figure 3.6 E shows BioCAP- and MBD-seq signal on two X-linked CGIs, a hypomethylated CGI of *Kmt5b* and hypermethylated CGI of *Ezhip*. Even after allelically

splitting the data, MBD-seq maintained a much better signal-to-noise ratio as compared to BioCAP-seq. After allelic split, BioCAP-seq signal diminishes, almost levelling with the background (example shown in Figure 3.6 F and combined average profile for X-linked CGIs is shown in 3.6 H). 30% of reads assigned to CGIs in MBD-seq is an underestimate of the specificity of the assay i.e., not all the remaining 70% reads comprised background. As I describe later, MBD-seq reproducibly showed enrichment/peaks at regions not annotated as CGIs. I will discuss the biology of this "non-CGI MBD-seq signal" in detail in later sections.

In BioCAP-seq a total of 251 X-linked CGIs were covered in sufficient depth to be used for allelic analysis. The median value for [1-allelic ratio] on these 251 CGIs was 0.7 (Figure 3.6 G, right panel). On the contrary, in MBD-seq a larger number of X-linked CGIs were captured (351 CGIs) for which the data could be analysed allelically. In MBD-seq the median allelic ratio was 1 (Figure 3.6 G, left panel). Thus, for the same input DNA, MBD-seq offered a much more sensitive read-out for methylation. This drastic change in allelic ratio in MBD-seq can be understood from Figure 3.6 F. For methylated CGIs (such as *Hprt*), almost the entire MBD-seq signal comes from Xi, causing the allelic ratio which is Xi/(Xi +Xa) to become 1 (since Xa = 0, allelic ratio becomes Xi/Xi+0, and hence Xi/Xi, which is 1). BioCAP-seq, although presenting essentially the same information (i.e., Xa is more hypomethylated than Xi), shows signal from both alleles which diminishes the [1-allelic ratio]. The BioCAP-seq signal on Xi allele of methylated CGIs could represent the signal contributed by cells in the population where this CGI is not fully methylated. The sensitivity of BioCAP for hypomethylated DNA could also explain the high background observed in mouse ES cells and MEFs both of which have been reported to have hypomethylated genomes upon prolonged culture.

Since WGBS provides a quantitative measure for DNA methylation, I wanted to see what degree of methylation was detected using BioCAP- and MBD-seq. For this, I looked at Xi/Xa ratio in WGBS for the CGIs captured in BioCAP- and MBD-seq. Xi/Xa for CGIs captured in MBD-seq was lower (median 11.03978) as compared to BioCAP-seq (median 17.72635) (Figure 3.6 I). This indicates that indeed MBD-seq was able to capture CGIs with lower methylation levels in comparison to BioCAP-seq. Furthermore, for the CGIs captured, MBD-seq showed a much larger change in allelic ratio than BioCAP-seq. Taking into consideration this comparison, I have used MBD-seq for all subsequent experiments.

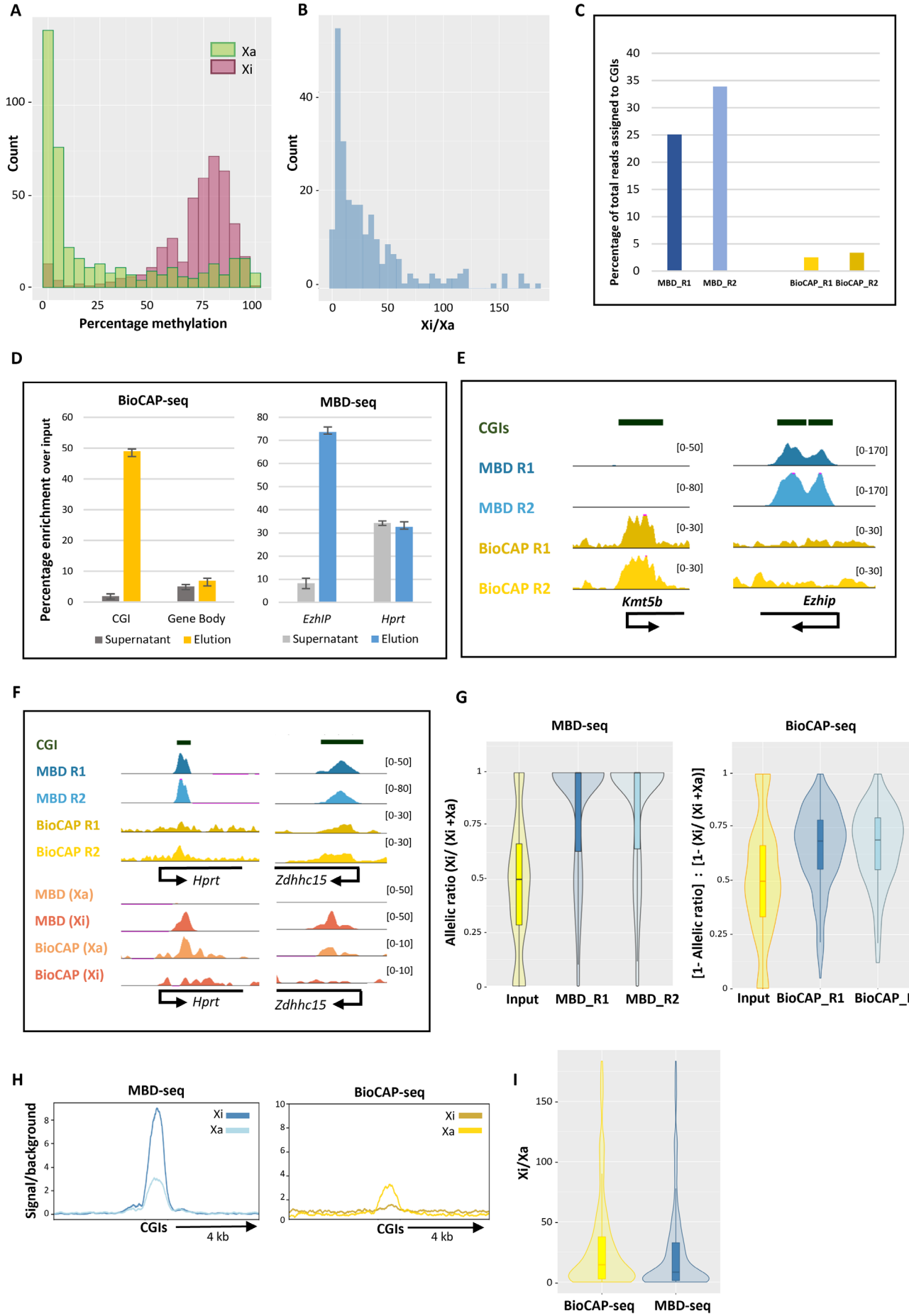

A
Xa
Xi
Count
Percentage methylation
B
Count
Xi/Xa
C
Percentage of total reads assigned to CGIs
MBD_R1
MBD_R2
BioCAP_R1
BioCAP_R2
D
BioCAP-seq
MBD-seq
Percentage enrichment over input
CGI
Gene Body
EzhIP
Hprt
Supernatant
Elution
E
CGIs
MBD R1
MBD R2
BioCAP R1
BioCAP R2
Kmt5b
Ezhip
F
CGI
MBD R1
MBD R2
BioCAP R1
BioCAP R2
MBD (Xa)
MBD (Xi)
BioCAP (Xa)
BioCAP (Xi)
Hprt
Zdhhc15
G
MBD-seq
BioCAP-seq
Allelic ratio (Xi/ (Xi +Xa)
[1- Allelic ratio] : [1- (Xi/ (Xi +Xa)]
Input
MBD_R1
MBD_R2
BioCAP_R1
BioCAP_R2
H
MBD-seq
BioCAP-seq
Signal/background
Xi
Xa
CGIs
4 kb
I
Xi/Xa
BioCAP-seq
MBD-seq

Figure 3.6: Comparison between BioCAP-seq and MBD-seq for studying methylated CGIs (shown on the previous page).

A) Plot showing the distribution of the percentage methylation of CGIs on Xi and Xa in Fvb/Cast MEF line in WGBS data from Gdula et al., 2019. The majority of the CGIs on Xi are more than 50% methylated.
B) Distribution of relative methylation of CGIs on the Xi, represented by Xi/Xa. The CGIs showing at least 2-fold methylation on Xi relative to Xa were subsequently used for the analysis of MBD- and BioCAP-seq data. WGBS data used for this plot is from Gdula et al., 2019.
C) Plot shows the percentage of total NGS reads aligned to regions of interest (i.e., CGIs) as a measure of signal-to-noise ratio for both assays. In MBD-seq, up to 30% of the total reads constitute CGIs which is ~10 fold greater than BioCAP-seq. The next sections discuss the significance of non-CGI signal in MBD-seq. (R1 and R2 is replicate1 and replicate 2)
D) qPCR validation of MBD- and BioCAP-seq. For BioCAP-seq hypomethylated CGI and hypermethylated gene body of autosomal gene *Kmt5b* (also known as *Suv420h1)* was amplified which acted as a positive and negative control, respectively. For MBD-seq two X-linked CGIs, highly methylated *Ezhip* and moderately methylated *Hprt*, were amplified. Supernatant shows DNA in unbound fraction and elution shows DNA in bound fraction.
E) Genome browser tracks showing the regions used for qPCR in (D). At the hypomethylated *Kdm5b* CGI, the BioCAP-seq signal has a larger signal than MBD-seq. At hypermethylated *EzhIP* CGI, MBD-seq has a higher signal. MBD-seq shows a higher signal over background in comparison to BioCAP-seq, as evident from the scale on the right of the track. (R = replicate).
F) Genome browser tracks for representative X-linked CGIs before (top) after allelic split (bottom). As one allele is methylated, both assays are enriching for one allele (Xi in the case of MBD-seq and Xa for BioCAP-seq). MBD-seq shows much higher enrichment in comparison to BioCAP-seq (discernible from the scale on right). The signal-to-noise ratio is well maintained in MBD-seq upon allelic split. (R1 and R2 are replicate 1 and 2).
G) Box-violin plots showing the allelic ratio for MBD-seq and 1-allelic ratio in the case of BioCAP-seq, for the group of CGIs selected in (B). (R1 and R2 are replicate 1 and 2).
H) Metaplots showing the average profile of Xi and Xa signal on CGIs, used to calculate the allelic ratio in (G).
I) Violin plot showing Xi/Xa (calculated from WGBS data, Gdula et al, 2019) for CGIs captured in BioCAP-seq and MBD-seq. Xi/Xa ratio was plotted for CGIs captured in sufficient depth in MBD-seq and BioCAP-seq (defined as reads in Xi + Xa ≥ 8). The lower median of the MBD-seq plot shows that MBD-seq captured CGIs that showed lower methylation as compared to BioCAP-seq and therefore represents a more sensitive assay. Panels (H) and (I) show data from one of two reproducible replicates.

3.6 Analysis of CGI methylation dynamics on Xi during EB differentiation.

MBD-seq data on the MEF line described in the last section showed that MBD-seq is an effective method for studying CGI methylation with a much better signal-to-noise ratio in comparison to

BioCAP-seq. To study the dynamics of CGI methylation, I performed MBD-seq on EB differentiation timecourse for the iXist-ChrX line using the same input DNA as used for BioCAP-seq in section 3.4, enabling a direct comparison between these two datasets. MBD-seq data from this experiment was reproducible with other replicates collected later in this study, and it is therefore a representative dataset (Figure 3.7 A). In this work, I have analysed the MBD-seq data with two methods. (1) using allelically split MBD-seq data which would indicate the relative gain of CGI methylation on Xi as compared to Xa. This method is however unable to inform about the absolute gain in CGI methylation during the timecourse. (2) This issue is addressed in the second method which makes use of unsplit MBD-seq data to gauge the absolute gain of CGI methylation during EB differentiation (described in section 3.7).

For the allelic analysis, CGIs were filtered such that the sum of reads from both the X chromosome alleles for a given CGI (Xi + Xa, the denominator for the allelic ratio) was at least 7 at day 12, giving 285 CGIs that I could process in this analysis. This filter discards the CGIs covered in low depth which could introduce large skews in the allelic ratio. The reason this filter ($Xi + Xa \geq 7$) was applied on day 12 and not earlier, was that at day 0 all the CGIs are expected to be unmethylated and therefore completely lack MBD-seq signal ($Xi + Xa < 7$). Even at day 4, only 104 CGIs (out of 514 X-linked CGIs) qualified this criterion, indicating that the majority of CGIs are methylated at later timepoints.

Figure 3.7 B shows the allelic ratio changes during the EB differentiation timecourse. The allelic ratio undergoes a profound shift from the median of 0 in ES cells to 1 at EB day 4, showing that the CGIs on Xi overall begin to acquire DNA methylation rapidly. The shape of violin plots which indicates the spread of the data becomes concentrated at allelic ratio 1 at day 8 and stays consistent afterwards which suggests that the majority of CGIs gain methylation between day 4-day 8 and plateau afterwards (Figure 3.7 B). As discussed in section 3.5, the large shift in the allelic ratio is due to the absence of an MBD-seq signal on Xa, which makes the allelic ratio a sensitive measure to study the dynamics of CGI methylation during XCI, where any small gain in CGI methylation would be visible as a large shift of allelic ratio above 0.5.

For studying the methylation dynamics, I categorized the CGIs based on the day at which they reached the maximum allelic ratio value. A CGI was defined as methylated if its allelic ratio was greater than 0.5. Judging from the violin plots, the majority of the CGIs reached an allelic ratio of 1. A small group of 56 CGIs reached an allelic ratio of 1 (AR = 1) at day 4 and plateaued till day 12 (Figure 3.7 C); I have termed CGIs in this group as “fast” CGIs. Most CGIs could be grouped as “slow” CGIs, showing an allelic ratio of 1 at day 8 (Figure 3.7 C and D). This data agrees with previous reports where CGIs of *Hprt* (Figure 3.7 E) and *Mecp2* (not shown) displayed slower dynamics while *Mtm1, Nup62cl,* and *Gabre* for

instance, were classified in the “fast” dynamic category (Gendrel et al., 2012) (Figure 3.7 F). *Mtmr1* however, was categorized as “slow” CGI in this work and was reported to show fast dynamic previously (Gendrel et al., 2012). A small group of 20 CGIs showed signal on both the Xa and Xi, and therefore gain XCI-independent methylation, i.e., the number of reads from Xi + Xa increased during the timecourse, but allelic ratio stayed 0.5 (Figure 3.7 G). Notable CGIs in this group, such as *Dusp9, Rhox13, Taf7l, AU022751,* are associated with germline-specific genes and have been reported previously to get methylated after mouse embryo implantation (Auclair et al., 2014).

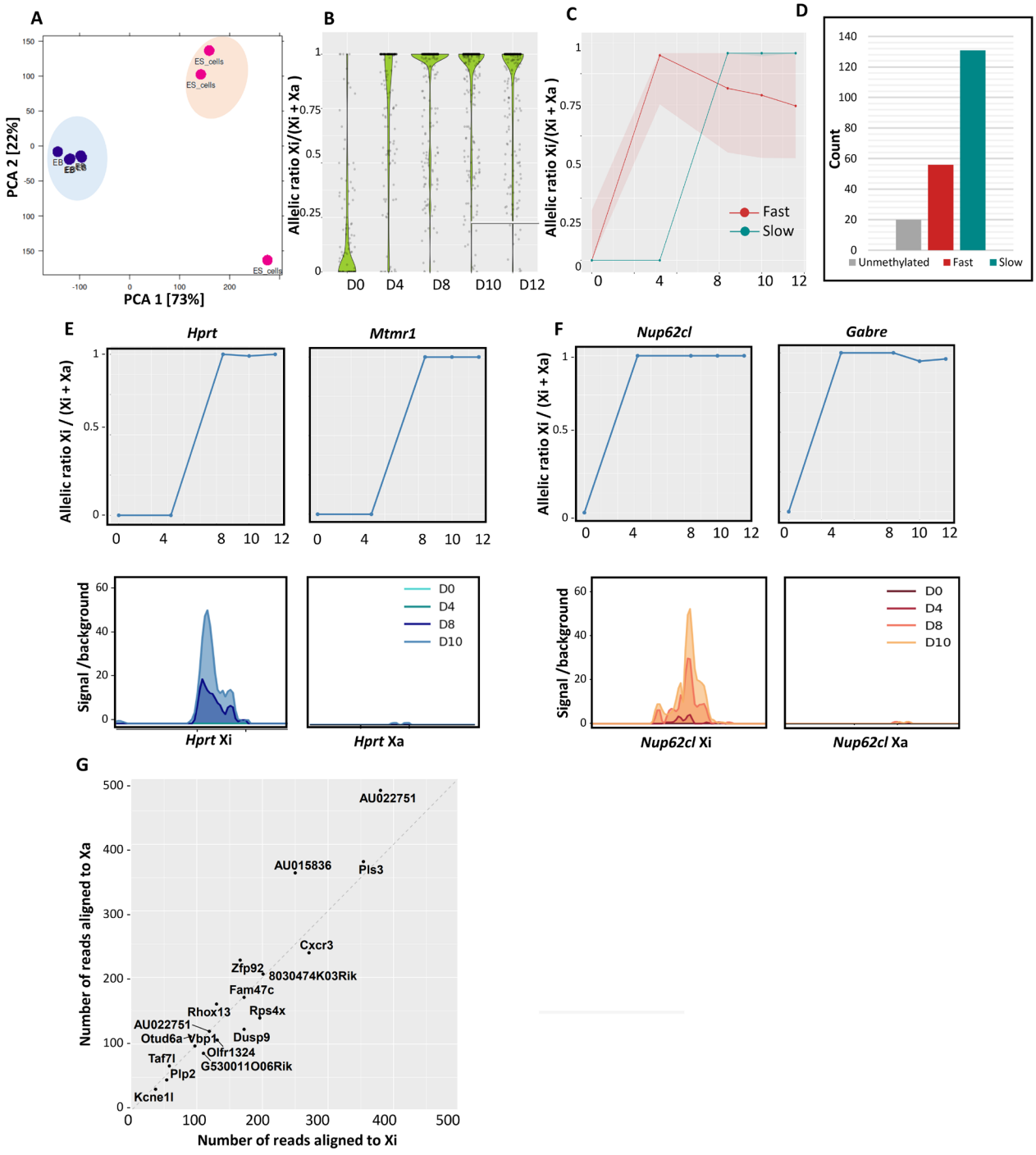


Figure 3.7: Methylation dynamics of CGIs during EB timecourse of the iXist-ChrX line (shown on the previous page).

A) PCA plot to illustrate the reproducibility between replicates for MBD-seq data on iXist-ChrX EB timecourse. ES cells and day 12 EBs form well-separated clusters.
B) Violin plot depicting allelic ratio on 285 informative CGIs at mentioned days during the EB timecourse (D0 – D12 represent day 0 – day12). The allelic ratio shows a shift from 0 to 1 at day 4. At day 8, the majority of the CGIs are methylated. Individual CGIs are represented by dots on the violin plot.
C) CGIs can be classified as "fast" (red) and "slow" (green), based on the timing at which they attain an allelic ratio of 1 during the timecourse. X-axis shows time in days during EB differentiation.
D) Counts for the CGIs in "fast" and "slow" dynamic categories. This graph shows the numbers for CGIs that could be used in the allelic analysis and confidently categorized as "fast" or "slow" CGIs. Several CGIs that did not have sufficient coverage for calculating allelic ratio at day 4, for instance, were excluded. The CGIs showing an allelic ratio of 0.5 were categorized as "unmethylated", since using the allelic ratio it is not possible to distinguish if these are gaining methylation.
E) Line plots showing examples of "slow" methylating CGIs. Below is one example for each, showing the gain of signal during the timecourse on Xi (left) and Xa (right). The Xa has a very small signal, causing the profound shift in the allelic ratio (from 0 to1) as observed in (B).
F) Line plots as in (E), but for "slow" methylating CGIs.
G) Dot plot showing CGIs that have MBD-seq signal on both Xi and Xa. Germline-specific genes such as *Rhox13, Taf7l, AU022751* fall into this category. Panels (B-G) show data from one of two well-correlated replicates for day 4- day 10 (see Appendix 2) and 5 reproducible replicates for day 12 timepoint as shown in (A).

EB differentiation progresses heterogeneously compared to directed differentiation protocols. Therefore, in order to validate my observations regarding CGI methylation dynamics, I experimented with the more homogenous NPC differentiation system (Conti et al., 2005; Splinter et al., 2011) (Figure 3.8 A). For this I collected NPCs from day 12 and day 15 since at these timepoints the NPCs look uniform and gain neural lineage morphology. Figure 3.8 B shows allelic ratio for CGIs during NPC timecourse. Day 12 represents an intermediate stage where some CGIs achieve an allelic ratio of 1 while others are unmethylated (allelic ratio < 0.5). By comparing Figure 3.7 B, it can be seen that the onset of CGI methylation in NPC differentiation is much later compared to EBs. Here too, I divided CGIs into dynamic categories of "fast" and "slow". The CGIs that gained methylation (allelic ratio > 0.5) at day 12 were categorized as "fast" while those showing methylation at day 15 were defined as "slow" (counts of CGIs in these two classes are shown in Figure 3.8 C). A large number (114 out of 285 CGIs amenable to allelic analysis) of CGIs stayed unmethylated even at day 15 (Figure 3.8 C). Figure 3.8 D shows the rate of methylation for "fast" and "slow" CGIs in NPC differentiation timecourse.

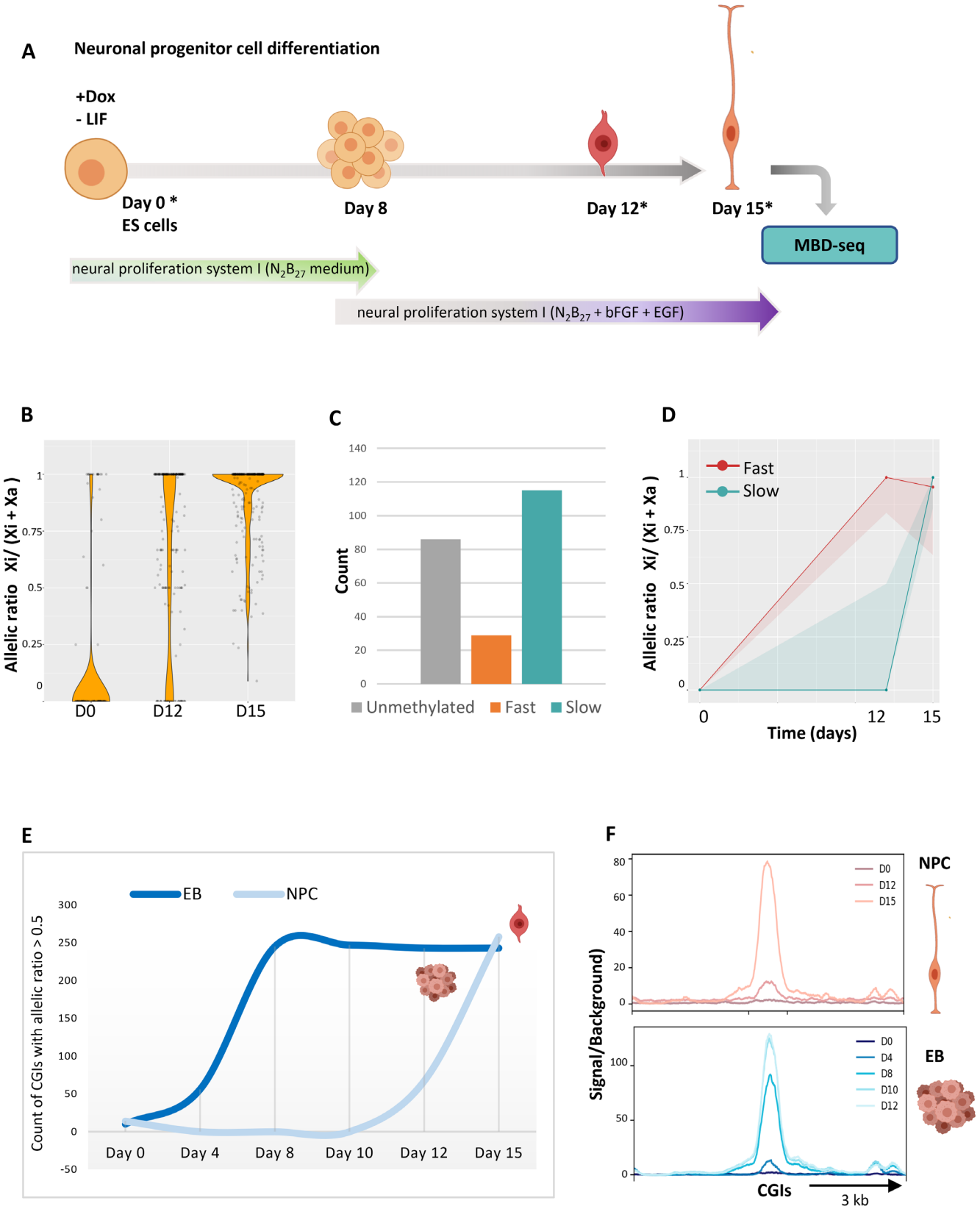

A
Neuronal progenitor cell differentiation
+Dox
- LIF
Day 0 *
ES cells
Day 8
Day 12*
Day 15*
MBD-seq
neural proliferation system I (N2B27 medium)
neural proliferation system I (N2B27 + bFGF + EGF)
B
Allelic ratio Xi/ (Xi + Xa)
D0
D12
D15
C
Count
Unmethylated
Fast
Slow
D
Fast
Slow
Allelic ratio Xi/ (Xi + Xa)
Time (days)
E
EB
NPC
Count of CGIs with allelic ratio > 0.5
Day 0
Day 4
Day 8
Day 10
Day 12
Day 15
F
NPC
D0
D12
D15
EB
D0
D4
D8
D10
D12
Signal/Background)
CGIs
3 kb

Figure 3.8: CGI methylation dynamics during NPC differentiation of the iXist-ChrX cell line (shown on the previous page).

A) Schematic showing NPC differentiation. NPC differentiation uses media supplemented with growth factors to direct cells to differentiate towards the neuronal lineage. Asterisks at timepoints represent the days at which cells were collected for this experiment.
B) Violin plot showing allelic ratio at day 0, 12 and 15 (D0-D15). At day 12, around half of the CGIs have achieved the allelic ratio of 1.
C) Counts of CGIs assigned to "fast" and "slow" methylating categories.
D) Line plots showing the allelic ratio for "fast" (red) and "slow" (green) CGIs during NPC differentiation. The solid line shows the median while the shaded ribbon shows the interquartile range for allelic ratio. X-axis shows time in days during differentiation.
E) The number of CGIs methylated (defined as allelic ratio > 0.5) at various stages/days during EB or NPC differentiation. In EB differentiation, the number of CGIs that are methylated plateau after day 8, and further differentiation does not result in the methylation of more CGIs. However, the number of CGIs gaining methylation in the NPC differentiation continues to rise until the end of the timecourse.
F) Metaplots showing the average profile of CGIs during NPC and EB differentiation. D0-D15 represent day 0 -day 15 of differentiation. Due to better suitability of EB differentiation for downstream experiments, MBD-seq on NPC differentiation was not repeated (see text for details).

Owing to the presence of an internal control in the form of Xa, the allelic analysis, is robust to experimental variability and allows effective comparison between timepoints or between independent experiments. However, allelic splitting comes at a cost of significant data loss because the NGS reads mapping to regions without SNPs are discarded. This in turn limits our analysis to regions that contain informative SNPs, so that the reads can be assigned to either parental allele. I set out to test if the NPC differentiation timecourse offered the advantage of a greater number of CGIs that can be studied by allelic analysis. I counted the number of CGIs that were amenable to allelic analysis (i.e., the sum of reads from Xi +Xa ≥ 8) in both EB and NPC timecourse and plotted the CGIs gaining methylation (defined as allelic ratio > 0.5) for all the timepoints of EB and NPC timecourse (Figure 3.8 E). The total number of CGIs amenable to allelic analysis was comparable in both differentiation systems (265 in EB and 285 in NPC). In EB differentiation, the number of CGIs that get methylated plateaus at day 8, while in NPC differentiation, it did not achieve plateau even at day 15 (Figure 3.8 D). This indicates that the total number of CGIs that can be analysed using allelic analysis in NPC differentiation is greater if this differentiation is allowed to progress further and therefore NPC differentiation can allow us to study the methylation of a greater number of CGIs (i.e., greater number of datapoints) on the X chromosome as compared to EBs. EB and NPC differentiation however show similar dynamics of CGI methylation. Importantly, despite the differences in timescale, "slow" CGIs represent a larger group in both EB and NPC differentiation experiments (i.e., "slow" CGIs achieve methylation between day 4-8 in EB differentiation and between day 12-15 in NPC differentiation)

(Figure 3.8 F). Furthermore, EB differentiation has the advantage of showing quicker CGI methylation and presents a simple and cost-effective method. Therefore, I have adhered to this procedure for the remainder of this work.

3.7 Spike-in MBD-seq gives quantitative information on the methylation status of X-linked CGIs

Allelic analysis of MBD-seq data is informative for studying the methylation of CGIs on Xi relative to Xa, however, it does not indicate the absolute level of the methylation gained. As shown in section 3.6, the allelic ratio for the majority of the CGIs in EBs plateaus at day 8, but the CGIs on average continue to gain methylation until day 12 (as visible from the MBD-seq signal getting higher along the timecourse in Figure 3.8 F). Therefore, although the allelic analysis is a powerful tool to study the relative gain of methylation at CGIs on Xi, it does not present a full picture of the levels of CGI methylation. Furthermore, due to the precise capture of CpG-rich regions by MBD-seq, the total size of the dataset (as measured by the number of reads mapped) is small. The size of the data is further reduced during the allelic split, which in turn decreases the number of informative CGIs that can be analysed (as described above). Hence analysis of unsplit data provides more computational power by providing a greater number of reads aligned per CGI, and detection of lowly methylated CGIs that are excluded in the allelic analysis due to poor coverage.

To gain insight into methylation levels on individual CGIs, the inclusion of exogenous DNA that serves as a "spike-in" control is essential. For performing quantitative comparisons of the methylation levels gained at CGIs during the differentiation timecourse, I used human DNA (extracted from Hela cells). This human DNA serves as internal calibration to correct any variability in IP efficacy or library preparation of the samples. Similar internal calibration controls are frequently used for ChIP-seq and there is now a well-accepted convention for the normalization of such "spiked-in/ calibrated" datasets, described by (Hu et al., 2015) and mentioned in section 2.2.3.

Using unsplit MBD-seq data, I was able to analyse 448 CGIs on the X chromosome which were covered in good/sufficient sequencing depth. Out of these 448 CGIs, 382 gained methylation in comparison to day 0 (as indicated by the gain in signal during the timecourse). CGIs on X chromosomes vary in methylation levels by magnitudes which makes it difficult to present them on the same scale (Figure 3.9 A). Therefore, for looking at dynamics, I used a metric called max ratio (defined as methylation level on a given CGI at given timepoint/maximum methylation for that CGI, denoted as IP/Max_IP plotted in Figure 3.9 B). Max ratio provides values from a minimum of 0 to a maximum of 1. As shown

in Figure 3.9 B, most of the MBD-seq-captured methylation signal is gained between day 4 and day 8 which agrees with allelic analysis. The maximum MBD-seq signal for most of these CGIs (331/382) was detected at day 12 (Figure 3.9 B and D) and the methylation was beginning to plateau at this point (as can be seen from the ribbon plot in Figure 3.9 B and the examples *Suv39h1* and *Mtmr1* shown in Figure 3.9 E). However, EB timecourse experiments were stopped at day 12 to avoid overgrowth of a proportion of undifferentiated ES cells in the culture. Figure 3.9 C shows a heatmap displaying the changes in max ratio during the EB timecourse for individual CGIs.

I then looked at the methylation status of CGIs before the onset of XCI and tested if any CGIs were already methylated in the ES cells i.e., "premethylated". CGIs were defined as "premethylated" if they exhibited more than half of the maximum levels of methylation accumulated during the differentiation timecourse at day 0 (i.e., IP/Max_IP ≥ 0.5 at day 0). A group of 66 CGIs were "premethylated" in ES cells. These premethylated CGIs can be seen to be more methylated in day 0 ES cells as compared to day 4 (CGIs on the bottom of the heatmap shown in Figure 3.9 C). After day 4, these premethylated CGIs gained methylation as differentiation progressed. CGIs associated with germline-specific genes such as *Dusp9, AU022751, G530011O06Rik* (described in section 3.6 to show biallelic methylation) and *Bcor* are included in this category (MBD-seq signal on *Bcor* during the EB differentiation is shown in Figure 3.9 E).

Since overall, the CGI methylation levels keep rising until day 12, it is possible that CGIs that start to gain methylation earlier (i.e., "fast" methylating CGIs defined in allelic analysis) will accumulate more methylation than those that start methylating late (i.e., "slow" CGIs). This indeed was the case for *Hprt* ("slow" CGI) and *Nup62cl* ("fast" CGI). Figure 3.9 F shows that *Hprt* gained less methylation in comparison to *Nup62cl*. On average, the "fast" CGIs showed higher levels of methylation at day 12 in comparison to "slow" CGIs (Figure 3.9 G). I then looked at additional features that determine the fast and slow methylation dynamics of CGIs. Dr. Joseph Bowness in the lab has previously shown that the genes expressed at low levels in ES cells silence earlier during XCI. Given the antagonism between CGI methylation and transcription (described in chapter 1), I hypothesized that the CGIs associated with lowly expressed genes in ES cells would attain CGI methylation earlier during differentiation. For his analysis, Dr. Joseph Bowness categorized genes into three categories (low, intermediate, and high) based on their expression levels in iXist-ChrX ES cells; I used his categorization for this analysis. Figure 3.9 I shows that "fast methylating" CGIs indeed are frequently associated with lowly expressed genes. There is also a weak correlation of "slow methylating" CGIs with genes with high or intermediate expression levels.

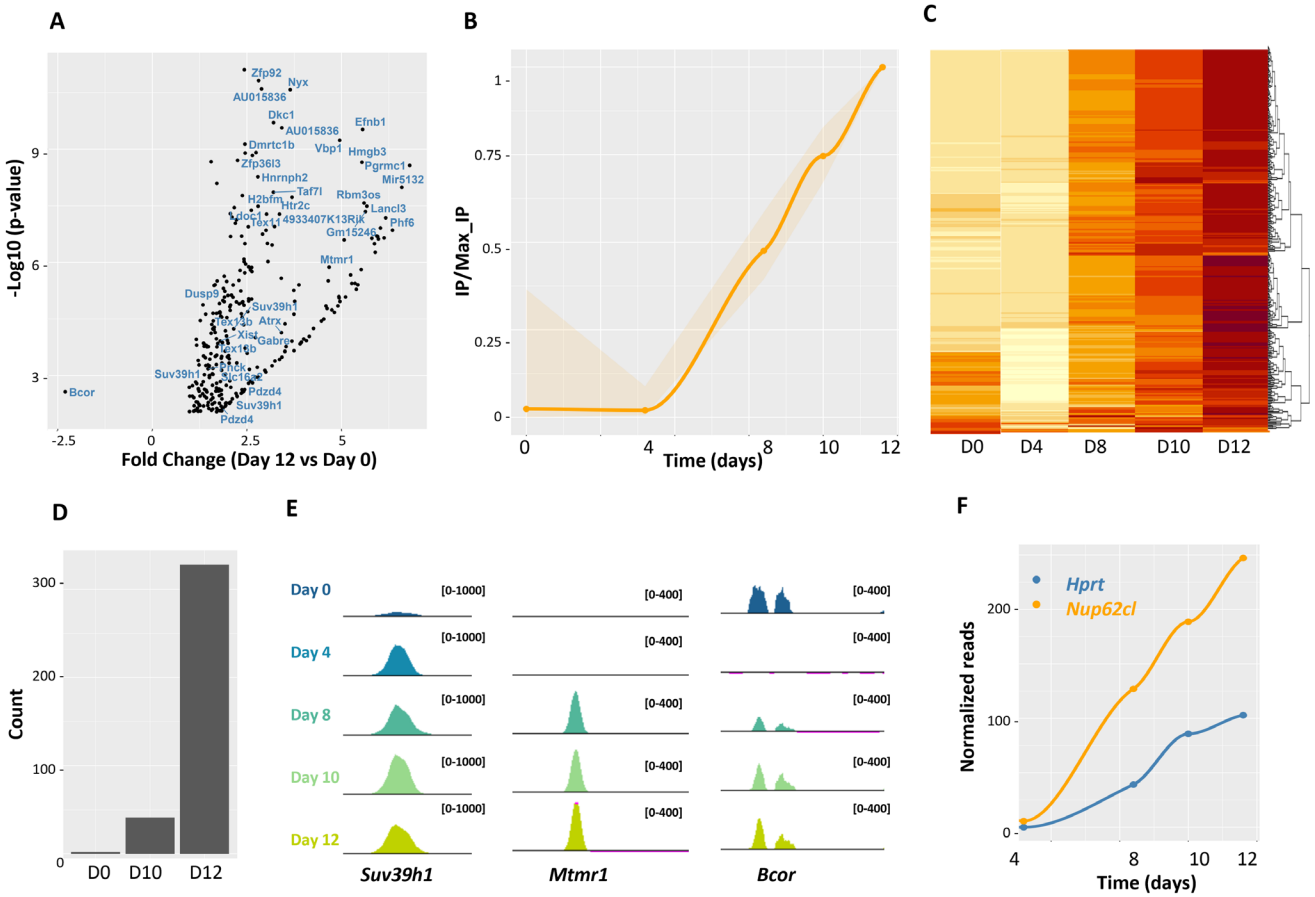

A
-Log10 (p-value)
Fold Change (Day 12 vs Day 0)
Zfp92
Nyx
AU015836
Dkc1
Efnb1
AU015836
Dmrtc1b
Vbp1
Hmgb3
Zfp36l3
Pgrmc1
Hnrnph2
Mir5132
H2bfm
Taf7l
Rbm3os
Htr2c
Lancl3
Tex14
4933407K13Rik
Phf6
Gm15246
Mtmr1
Dusp9
Suv39h1
Atrx
Xist
Gabrg3
Tex14b
Suv39h1
Bcor
Pdzd4
Suv39h1
Pdzd4
B
IP/Max_IP
Time (days)
C
D0
D4
D8
D10
D12
D
Count
D0
D10
D12
E
Day 0
Day 4
Day 8
Day 10
Day 12
[0-1000]
[0-400]
Suv39h1
Mtmr1
Bcor
F
Hprt
Nup62cl
Normalized reads
Time (days)
G
p = 5.1 x 10-8
***
Normalized reads
Fast CGIs
Slow CGIs
H
p = 0.298
ns
Distance to nearest TSS
Fast CGIs
Slow CGIs
I
Expression level of associated gene
Low
Intermediate
High
Fast CGIs
Slow CGIs
J
Speed of silencing of associated gene during XCI
Fast
Medium
Slow
Escapee
Fast CGIs
Slow CGIs

Figure 3.9: Calibrated MBD-seq reveals the level of DNA methylation gained by CGIs on the Xi (shown on the previous page).

A) Volcano plot showing the fold changes in methylation levels between ES (Day 0) and EBs (Day 12). Highlighted are examples of CGIs with low or high fold increase in methylation levels during the differentiation timecourse. The fold changes are representative of 5 MBD-seq datasets acquired from independent differentiation timecourse experiments. The CGIs indicated here were found methylated in all replicates with a false discovery rate < 0.05 (FDR < 0.05).
B) Plot showing the methylation /MBD-seq signal gained on CGIs on the X chromosome during EB differentiation timecourse. To put CGIs on one scale, the MBD-seq signal is normalized to maximum methylation signal gained by a given CGI during the EB timecourse (IP/Max_IP). Here, the solid line depicts the median while the shaded ribbon shows the interquartile range. Some CGIs are "premethylated" and can be found to have MBD-seq signal at day 0, which is lost at day 4 and then again gained further into differentiation.
C) Heatmap for MBD-seq signal (expressed as IP/Max_IP) at individual CGIs shown in (B) during EB differentiation. Each line/tile represents a single CGI. The bottom cluster of CGIs represents premethylated CGIs. (D0-D12 represent day 0 – day12).
D) Count of CGIs showing maximum MBD-signal at the mentioned day. Most of the CGIs, continue to gain signal till day 12 (D12). Panels (B)-(D) summarize data from two replicates for day 4 - day 10 (see Appendix 2) and 5 replicates for day 12 timepoints respectively (see Figure 3.7 A).
E) Genome browser tracks showing CGIs with low (*Suv39h1*), high (*Mtmr1*) and negative (*Bcor*) fold change in methylation during EB differentiation.
F) Absolute methylation levels on one example of "slow" CGI (*Hprt*) and "fast" CGI (*Nup62cl*) during EB timecourse.
G) Boxplot showing absolute methylation levels at day 12 on "fast" and "slow" CGIs. "Fast" CGIs accumulate more methylation in comparison to "slow" CGIs. Asterisks depict that the difference between methylation levels of "fast" and "slow" CGIs is statistically significant, p-value = 5.1×10^{-8}.
H) Boxplot comparing the distance of "fast" and "slow" CGIs from the nearest TSS. Distance from the nearest TSS is comparable for both "fast" and "slow" CGIs. (ns= non-significant).
I) Pie charts showing the proportion of genes with high, intermediate, and low expression levels associated with "fast methylating" and "slow methylating" CGIs. "Fast methylating" CGIs are frequently associated with genes expressed at low levels in ES cells.
J) Pie charts showing the proportion of genes with fast, medium, and slow silencing dynamics during XCI associated with "fast methylating" and "slow methylating" CGIs. "Slow methylating" CGIs are mostly associated with slow silencing genes. Panels (A), (E)-(J) show data from one of the 5 reproducible replicates for MBD-seq on wt iXist-ChrX cells at day 12 EBs (Figure 3.7 A).

I also observed that "slow methylating" CGIs are associated with genes that are silenced slower/later during XCI (shown in Figure 3.9 J). The categorization of genes based on the silencing speed in XCI was performed by Dr. Joseph Bowness based on RNA-seq data attained from iXist-ChrX lines before and after NPC differentiation. I also tested if the distance of a CGI from the TSS impacted its methylation

dynamic. Figure 3.9 H shows that both “fast” and “slow” CGIs did not differ much in their distance to nearest TSS.

Hence calibrated MBD-seq provided insights into additional aspects of CGI methylation on the X chromosome. CGIs on Xi show a large variability in the absolute levels of methylation gained during XCI and differentiation which can be partly explained by the expression levels of the associated genes in ES cells. Although calibrated MBD-seq does not provide accurate quantitative measure of the methylation levels on a given CGI, as for techniques like WGBS, it allows robust comparison of methylation levels between timepoints and gives useful insights into the gain of methylation during XCI.

3.8 CGI methylation on Autosomes

In section 1.2.12, I have described a few exceptional instances where CGIs gain methylation. These CGIs are generally associated with genes which function in development and morphogenesis (Auclair et al., 2014). Similarly tissue specific CGI methylation has also been reported (Illingworth et al., 2008). I tested if the EB differentiation system could recapitulate CGI methylation observed in mouse embryos. If these reported CGIs indeed gain methylation in EB differentiation, it would be reassuring that this experimental system, despite being cell-culture based, is near physiological. Characterization of autosomal CGIs that gain methylation during differentiation is also relevant for analysis of downstream experiments which involve ablation/mutations of important determinants of CGI methylation (described below and in chapters 4 and 5).

Taking advantage of the spike-in MBD-seq, I filtered out the autosomal CGIs that gained methylation during EB differentiation. Differentially methylated CGIs between day 0 and day 12 EBs were searched using a R Bioconductor package called DiffBind (Stark and Brown, 2011). DiffBind uses DESeq2, a well-recognized algorithm for differential expression analysis (Love et al., 2014), to find the statistically significant differential regions between the two tested conditions (in this case CGI methylation between ES and EBs). A detailed analysis of CGI methylation during mouse embryogenesis was published previously (Auclair et al., 2014). In agreement with Auclair et al 2014., the majority of CGIs remain unmethylated. I was able to filter 5592 CGIs (out of a total 23078 CGIs) that gained methylation during EB differentiation. Figure 3.10 A shows a heatmap for these 5592 CGIs that gain methylation and shows reproducible trends among the replicates (FDR < 0.05). As in the case of the Xi chromosome, autosomal CGIs continued to gain methylation until day 12 at which point, they started to plateau (Figure 3.10 B). There was good overlap of CGIs methylated in EBs with those reported for

E8.5 mouse embryos (Auclair et al., 2014). Some loci filtered in this analysis include imprinted loci (*Igf2r, Peg3, Meg3*) and genes encoding pluripotency factors (*Klf3, Dppa3, Nanog*), some genes from the *H*ox locus, genes of Wnt/βcatenin pathway and protocadherins (Figure 3.10 C shows the gain of methylation as a function of log fold change, and D shows genome browser tracks for representative CGIs). CGIs associated with genes coding for TFs such as *Bcl6b, Foxo6*, and some zinc finger proteins (for instance *Zfp560, Zfp41, Zfp647*) were also found methylated. Surprisingly, a number of CGIs showed the highest signal at day 0 and lose methylation as differentiation proceeds (Figure 3.10 A, bottom cluster). To conclude CGIs associated with developmental genes that have been reported to gain methylation in mouse embryos also show methylation during EB differentiation.

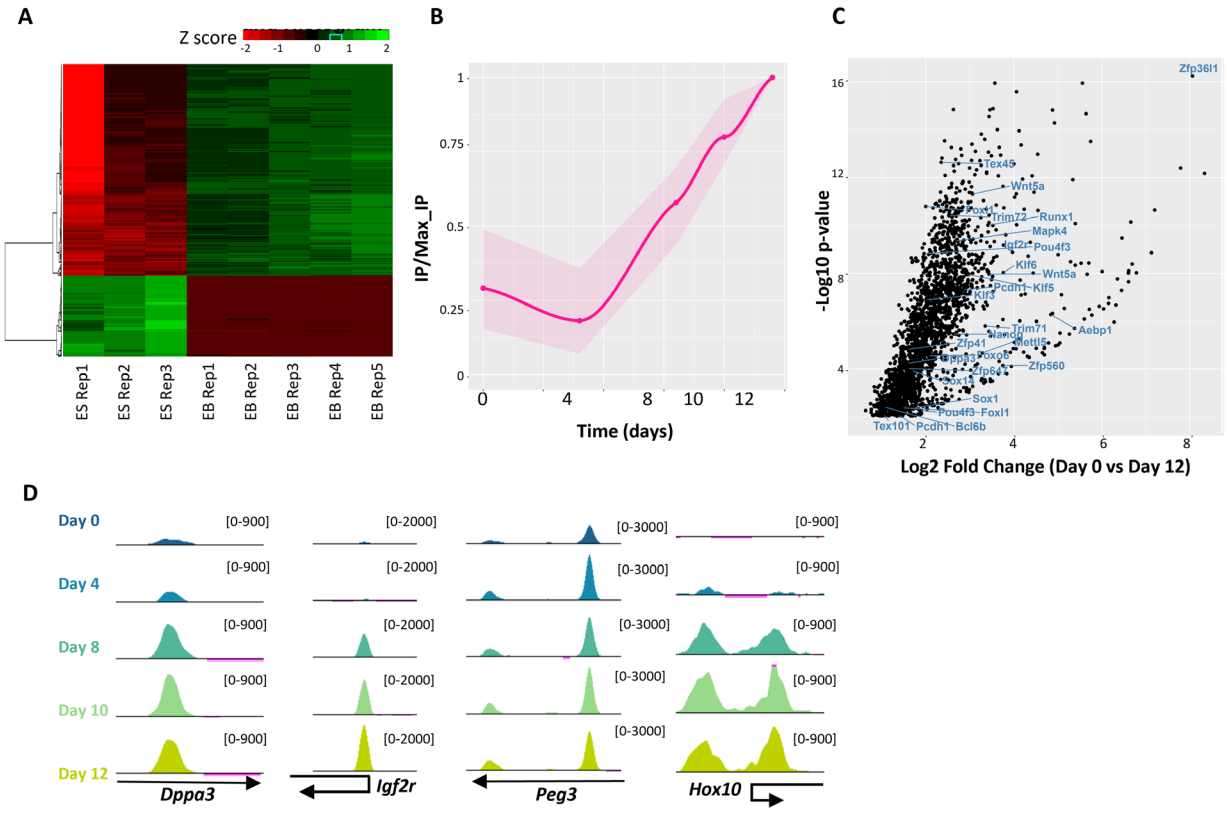


Figure 3.10: Differentiation-coupled methylation at autosomal CGIs (continued to the next page).

A) Heatmap for differentially methylated autosomal CGIs between ES cells and 12-day differentiated EBs obtained from DiffBind analysis. Although he majority of CGIs gain methylation during EB differentiation, a group of CGIs is methylated at the ES cell stage which loses methylation during the timecourse (bottom cluster).

B) Line plot showing the gain of methylation at autosomal CGIs during EB timecourse. The timing of methylation at autosomes resembles that of the Xi. To display the methylation dynamics, the absolute levels of methylation at CGIs have been normalized to maximum methylation gained during the timecourse as described in section 3.7.

C) Volcano plot showing autosomal CGIs that are methylated during EB differentiation (expressed as log2 fold change in methylation between ES cells and EBs). A few examples of CGIs associated with genes involved in developmental regulation and chromatin modification have been highlighted.
D) Genome browser tracks showing some examples of CGIs that gain methylation during EB differentiation. Panels (B) and (D) summarize data from two replicates for day 4- day 10 and 5 replicates for day 12 timepoints respectively.

3.9 Gain of methylation at non-CGI CpG-rich elements on the Xi

Visualization of the MBD-seq signal on the genome browser shows that a significant proportion of the signal originates from regions that are not annotated CGIs; I have termed this as "non-CGI signal". Since this signal was observed consistently between MBD-seq samples, I investigated its origin and how it changes during differentiation.

To perform this analysis, I compiled a list of reproducible MBD-seq peaks from day 0 ES cells and day 12 EBs so that all the signal, irrespective of the differentiation state, could be studied (Figure 3.11 A and B show the total peaks reproducibly found in MBD-seq data). As expected, some of the peaks represented annotated CGIs that I have used for my analysis thus far (9363 of the total peaks) (Illingworth et al., 2010). Among the non-CGI signal, 1922 peaks flanked TSS (residing within 2 kb of TSS) while 4890 overlapped with regions with high chromatin accessibility (detected in ATAC-seq performed by Dr. Joseph Bowness in the lab on iXist-ChrX cell line), indicating these represent potential gene expression regulatory regions. Most peaks (67054 of a total of 73866) however, originated from regions far from genes (Figure 3.01 B). The MBD-seq signal from regions overlapping TSS and regions of high chromatin accessibility had GC content comparable to CGIs, were hypomethylated in ES cells and gained methylation during differentiation which suggests that these most likely represent a group of CGIs that are unannotated (Figure 3.11 C-E). It is possible that due to different cell line (in comparison to those used in studies by Long et al., or Illingworth et al., for annotation of NMIs and CGIs respectively), or because of the low signal and different thresholding criteria in the analysis, these regions were not included in the annotations in the earlier studies (shown in Figure 3.11 A, the count of non-CGI MBD-seq signal plotted against the distance from the nearest TSS. Peaks within 2 kb of TSS are highlighted in the inset with dotted lines).

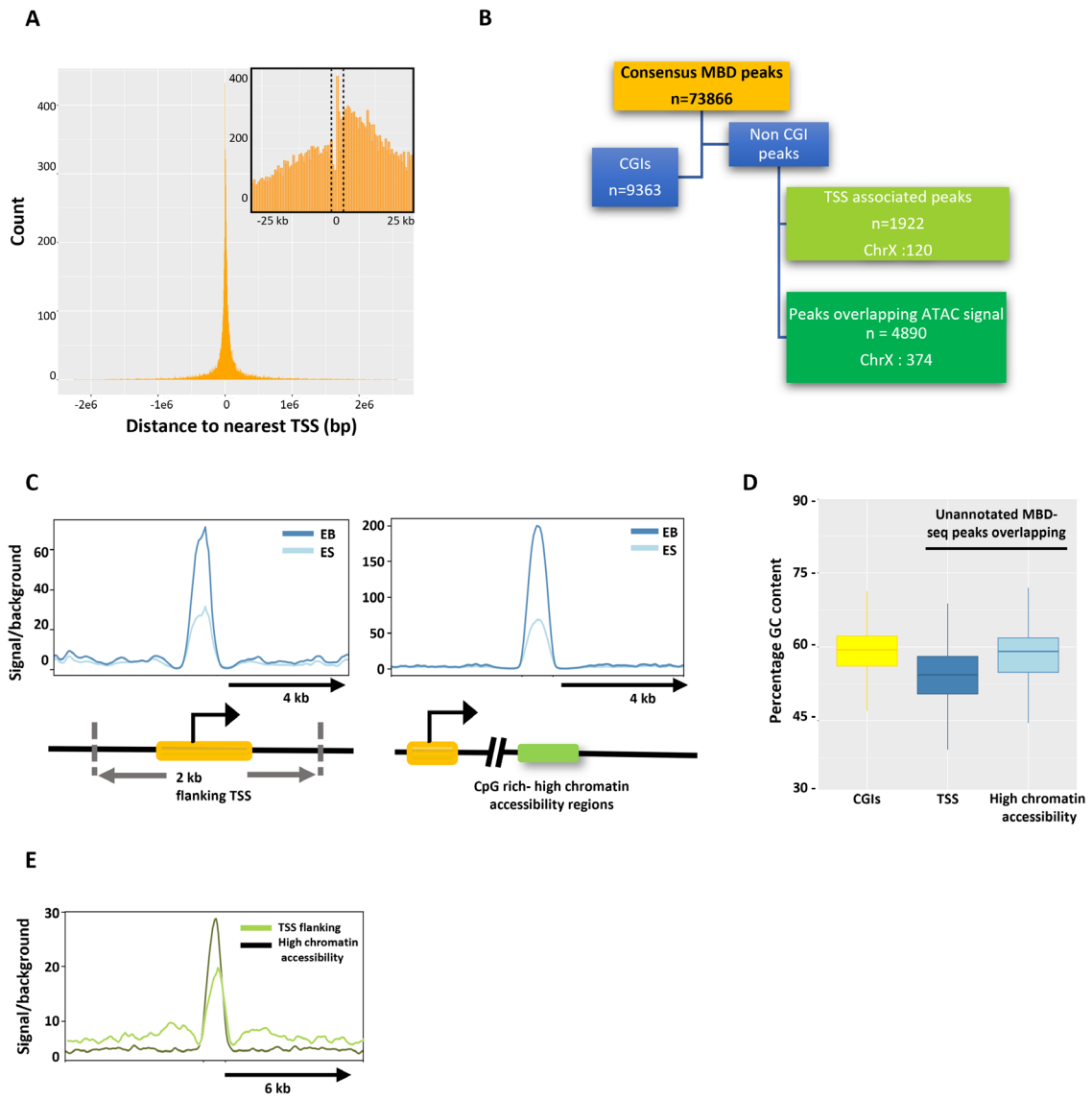


Figure 3.11: Characterization of the "Non-CGI signal" in MBD-seq (continued to the next page).

A) MBD-seq peaks plotted against distance from the nearest TSS. Inset shows the peaks within 25 kb from TSS. The dotted lines indicate a 2 kb distance spanning TSS. The MBD-seq peaks within 2 kb of TSS are likely to be unannotated CGIs.

B) Breakdown of MBD-seq peaks called in undifferentiated and 12-day differentiated iXist-ChrX cells. The peaks in proximity (2 kb) of the TSS sites and those overlapping regions of high chromatin accessibility can be characterized as potential CGIs.

C) MBD-seq signal at the "non-CGI" regions overlapping TSS (left), and regions of high chromatin accessibility (right). Both the classes of these regions gain methylation during differentiation.

D) Plot shows CpG content of the two classes of non-CGI regions (regions flanking TSS or regions of high chromatin accessibility) in comparison to annotated CGIs. This plot shows that both the classes represent CG-rich regions, further validating the capture by MBD-seq. Panels (A)-(D) summarize data from 5 highly reproducible replicates (see Figure 3.7 A)

E) BioCAP-seq signal in ES cells on the "non-CGI" regions flanking TSS or regions of high chromatin accessibility. The peaks in ES cells indicate that these regions are hypomethylated in ES cells. The metaplots show data from one of two well correlated replicates (see Appendix 2)

To study the relevance of this "non-CGI" MBD-seq signal in XCI, I tested for allelic bias (i.e., different methylation levels on Xi as compared to Xa). For this analysis, I called peaks from allelically split MBD-seq data from day 0 ES cells and day 12 EBs to account for the change in the methylation status of these regions pre- and post-XCI. I was able to see three trends among MBD-seq signal on the X chromosome (Figure 3.12 A shows a schematic for the types of MBD-seq signal). The MBD-seq peaks that can be seen in ES cells and persist in day 12 EBs; were termed "default peaks". Another class of peaks that occurs in ES cells but gains more methylation during differentiation, was termed "rising peaks". The third group of peaks emerged specifically at day 12 EBs on the X chromosome (these were termed "*de novo* peaks"). "*De novo* peaks" were then further classified as "differentiation-coupled peaks" or "XCI-specific peaks" depending on if they were found equally methylated on both the alleles or methylated specifically on Xi (Figure 3.12 B). As expected, the majority of the "XCI-specific peaks" were annotated CGIs that resided close to TSS/promoters (located within 2 kb of a TSS, 182/281 were annotated CGIs). The remaining "XCI-specific peaks" were far from TSS (non-CGI XCI-specific peaks) (Figure 3.12 C). Figure 3.12 D shows a representative non-CGI XCI-specific peak which is reproducibly found in various experimental systems and shows a specific signal over the background. Annotation for these regions mostly revealed that they originated from intergenic regions and transposons (Figure 3.12 E).

Interestingly, these non-CGI XCI-specific peaks were CG-dense regions (Figure 3.12 G) and showed preferential methylation on Xi as compared to Xa during differentiation (Figure 3.12 F). I then tested if these non-CGI XCI-specific peaks overlapped with regions of high chromatin accessibility or chromatin modifications such as H3K4me1/3. For this analysis, I used the ATAC-seq and H3K4me1 ChIP-seq datasets generated in the lab by Dr. Joseph Bowness in the iXist-ChrX ES cells. Interestingly, non-CGI XCI-specific peaks overlapped with ATAC-seq and H3K4me1 ChIP-seq signal, indicating that these might be potential enhancers or cis regulatory elements (CREs) (Figure 3.12 H). Illingworth et al., and Long et al., have previously proposed that the overlap of unannotated CAP-/BioCAP-seq peaks with these chromatin features qualifies them as potential CREs and categorized the regions they found as CGIs (or NMIs in Long et al., 2013). Since these non-CGI XCI-specific peaks are hypomethylated in ES cells and share CRE-like features, I have used the term "CGI-like elements" for these regions. Further investigation will inform about the functions "CGI-like elements" perform. "Rising peaks",

defined above as a subgroup of peaks that existed in ES cells and gained methylation during differentiation, like the “CGI-like elements” showed a bias towards Xi (as apparent from allelic ratio > 0.5) and also emerged from repeat elements (Figure 3.12 I).

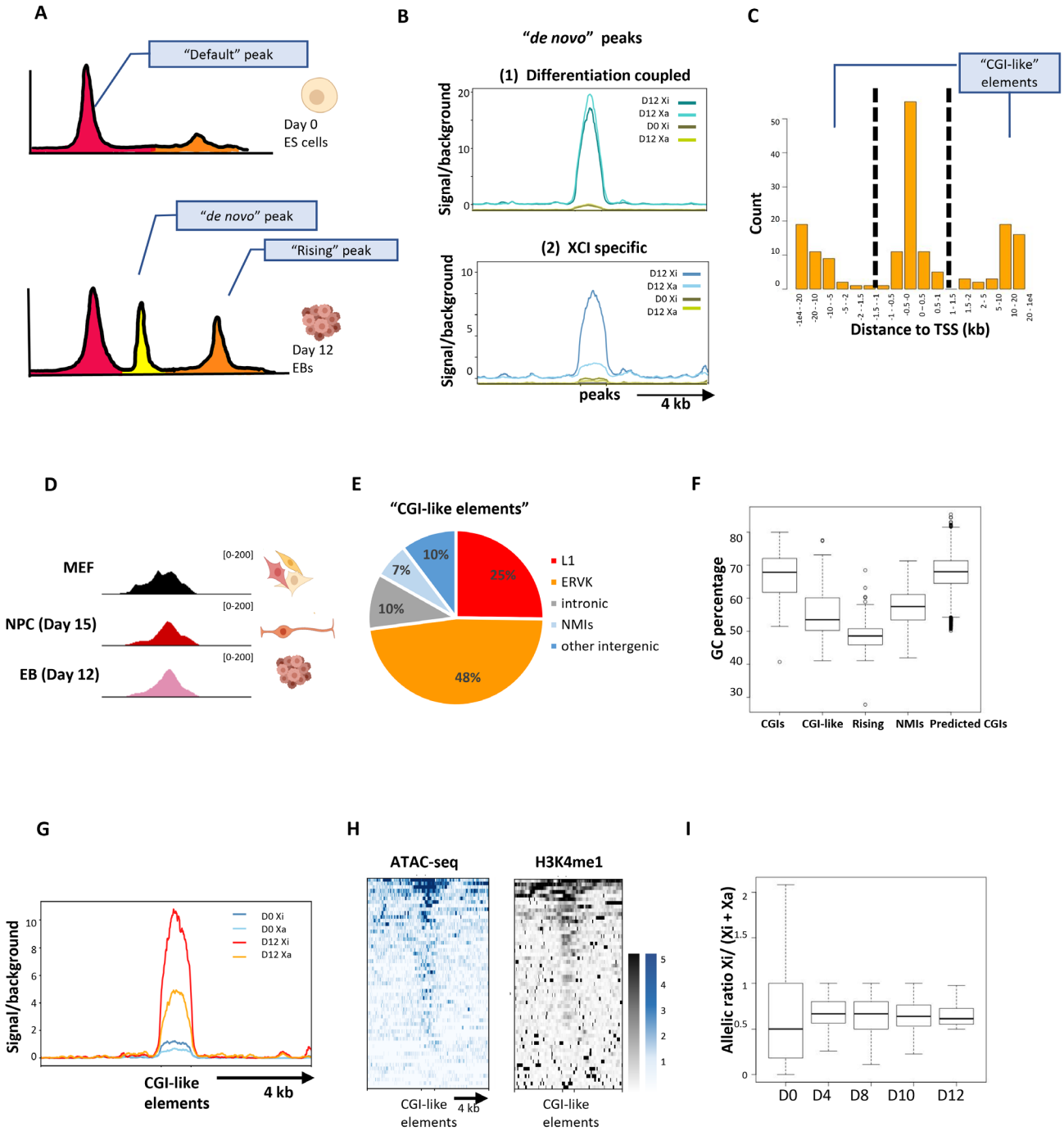


Figure 3.12: Characterization of the novel "CGI-like elements" on the X chromosome (shown on previous page).

A) Schematic showing the pipeline for categorizing MBD-seq signal on X chromosome. The peaks that were appearing on day 12 (*de novo*) were further classified as shown in (B).
B) MBD-seq peaks on X chromosome observed specifically at day-12 EBs (*de novo* peaks) were classified as "differentiation coupled" or "XCI-specific" depending on if the signal originated from both alleles (former) or only Xi (latter). (D0 = Day 0 ES cells, D12= Day 12 EBs).
C) XCI specific peaks plotted against the distance from nearest TSS. The peaks within 2 kb of TSS overlapped with CGIs. The non-CGI peaks found further away were termed as "CGI-like elements".
D) Genome browser track showing the reproducibility of "CGI-like elements" among different data sets and differentiation systems.
E) Pie chart showing the distribution of "CGI-like elements" among genomic features using a software 'HOMER'. "CGI-like elements" were mostly associated with transposons and intergenic regions.
F) GC percentage of MBD-seq peak classes (see text for details).
G) Metaplot showing the gain of MBD-seq signal on "CGI-like elements" during EB differentiation.
H) Heatmaps showing ATAC-seq and H3K4me1 signal at "CGI-like elements". "CGI-like elements" mostly overlap with H3K4me1 modified chromatin and regions of high chromatin accessibility.
F) Boxplot for allelic ratio at peaks categorized as "rising peaks". "Rising peaks" were defined as the MBD-seq signal in ES cells that is gained on Xi during EB differentiation. This figure summarizes data from 5 highly reproducible replicates (see Figure 3.7 A)

3.10 Impact of *Dnmt3b* deletion on XCI-linked CGI methylation in iXist-ChrX cells

To further validate my ES cell-based model and to dissect the role of DNMT3B in CGI methylation in XCI, I went on to generate a *Dnmt3b* knockout in the iXist-ChrX background (iXist-ChrX *Dnmt3b*$^{-/-}$, denoted as *Dnmt3b*$^{-/-}$ for simplicity). Several studies in the past have analysed *Dnmt3a* and *Dnmt3b* knockout mouse ES cells. The absence of *Dnmt* proteins has no discernible impact on the physiology of the knockout ES cells. I therefore used CRISPR/Cas9 mediated double stranded breaks to generate a constitutive knockout of *Dnmt3b* in the iXist-ChrX parental line. Figure 3.13 A shows a schematic for the strategy used, indicating the targeting sites for the small guide RNAs (sgRNAs). Western blot

analysis demonstrates complete loss of DNMT3B protein in selected *Dnmt3b* knockout lines (Figure 3.13 A bottom panel).

I went on to test if the specificity of DNMT3B for CGIs methylation on Xi, observed previously in mouse embryos (Auclair et al., 2014; Gendrel et al., 2012) can be recapitulated in differentiated iXist-ChrX *Dnmt3b*$^{-/-}$ line. I differentiated *Dnmt3b*$^{-/-}$ cells for 12 days using EB differentiation protocol and performed MBD-seq. *Dnmt3b*$^{-/-}$ cells behaved similarly to the wildtype cells in culture, and I observed no changes in morphology or growth rate throughout the timecourse. The MBD-seq data acquired from different replicates, all of which constitute independent differentiation timecourse experiments, was reproducible. The reproducibility among MBD-seq datasets can be appreciated from the PCA plot shown in Figure 3.13 B; the MBD-seq replicates for *Dnmt3b*$^{-/-}$ and wildtype iXist-ChrX in day 12 EBs (shown as KO_EB and wt_EB respectively) cluster together. The PCA plot in Figure 3.13 B also separates wt from for *Dnmt3b*$^{-/-}$ iXist-ChrX (KO_EB), indicating the change in CpG methylation upon loss of DNMT3B.

I performed allelic ratio analysis for the X-linked CGIs that normally get methylated, in the *Dnmt3b*$^{-/-}$ cells during the EB timecourse (as performed in wild type cells in the section 3.6). This revealed that CGIs on Xi are overall hypomethylated in the absence of DNMT3B (Figure 3.13 C). Since *Dnmt3b*$^{-/-}$ cells have normal levels of DNMT3A (Figure 3.13 A, lower panel), this experiment validates that DNMT3A is not redundant with DNMT3B for methylation of the majority of CGIs on Xi (Figure 3.13 D). A small group of CGIs (15 out of 265 CGIs that are amenable to allelic analysis, characterized before in section 3.6) however did get methylated in the absence of DNMT3B. CGIs such as *Dusp9, Kcne1l, G530011O0Rik* and *Zfp92* which are methylated in *Dnmt3b*$^{-/-}$ line have been described in section 3.6 to not show allelic bias and so are methylated independent of XCI (Figure 3.13 E and F lower panel). Only 5 out of these 15 DNMT3B-independent CGIs reached an allelic ratio of 1, indicating strong allelic bias towards Xi (Figure 3.13 E, shows the skew of allelic ratio to 1 in wt and highlights the DNMT3B-independent CGIs on X chromosome). Thus, DNMT3B is the most predominant (but not the sole) *de novo* methyltransferase involved in CGI methylation on Xi (Figure 3.13 F and H). Genome browser tracks for two examples each of DNMT3B-dependent and DNMT3B-independent CGIs on X chromosome are shown in (Figure 3.13 G).

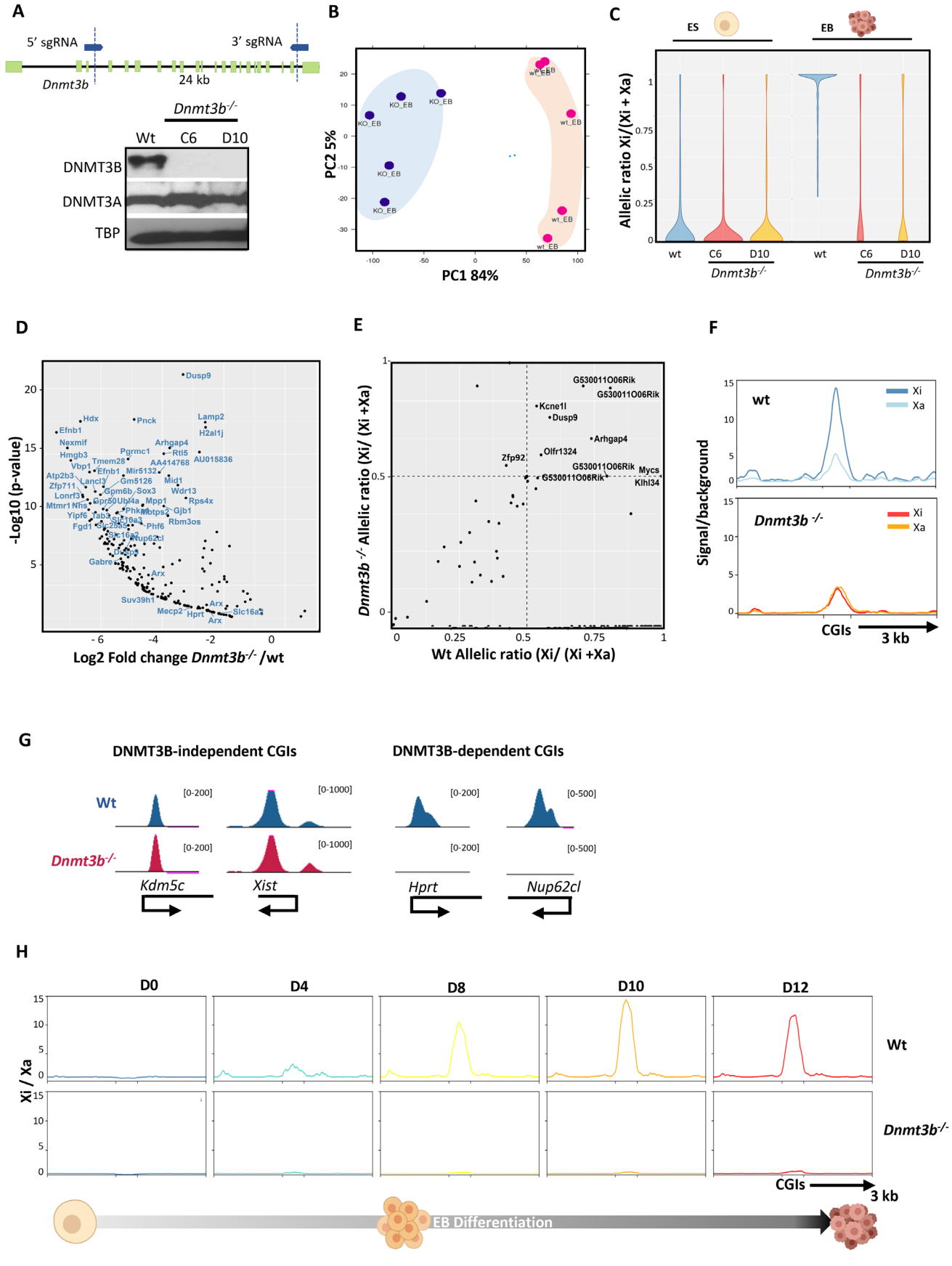

A
5' sgRNA
3' sgRNA
Dnmt3b
24 kb
Dnmt3b-/-
Wt
C6
D10
DNMT3B
DNMT3A
TBP
B
PC2 5%
PC1 84%
KO_EB
wt_EB
C
ES
EB
Allelic ratio Xi/(Xi + Xa)
wt
C6
D10
Dnmt3b-/-
D
-Log10 (p-value)
Log2 Fold change Dnmt3b-/- /wt
Dusp9
Hdx
Pnck
Lamp2
H2al1j
Efnb1
Nexmif
Arhgap4
Pgrmc1
Rtl5
Hmgb3
Tmem28
AA414768
AU015836
Vbp1
Efnb1
Mir5132
Atp2b3
Lancl3
Gm5126
Mid1
Zfp711
Gpm6b
Sox3
Wdr13
Lonrf3
Mtmr1
Nhs
Gpr50
Ubl4a
Mpp1
Rps4x
Yipf6
Phkg1
Mbtps2
Gjb1
Fgd1
Slc10a3
Rbm3os
Slc16a2
Phf6
Nup62cl
Dusp9
Gabre
Arx
Suv39h1
Mecp2
Hprt
Slc16a2
E
Dnmt3b -/- Allelic ratio (Xi/ (Xi +Xa)
Wt Allelic ratio (Xi/ (Xi +Xa)
G530011O06Rik
Kcne1l
Dusp9
Arhgap4
Olfr1324
Zfp92
Mycs
Klhl34
F
Signal/background
wt
Dnmt3b -/-
Xi
Xa
CGIs
3 kb
G
DNMT3B-independent CGIs
DNMT3B-dependent CGIs
Wt
Dnmt3b-/-
[0-200]
[0-1000]
[0-500]
Kdm5c
Xist
Hprt
Nup62cl
H
D0
D4
D8
D10
D12
Xi / Xa
Wt
Dnmt3b-/-
CGIs
3 kb
EB Differentiation

Figure 3.13: CGI methylation in the *Dnmt3b*$^{-/-}$ cell line (shown on the previous page).

A) Schematic showing the sgRNAs used for deletion of the *Dnmt3b* gene. The sgRNA are targeted to delete the region between exon 3 and 24. Western blot shows the successful knockout of *Dnmt3b* in iXist-ChrX line in clones C6 and D10.
B) PCA plot shows replicates for wt and *Dnmt3b*$^{-/-}$ cell lines differentiated to EBs for MBD-seq.
C) Violin plots show the allelic ratio at CGIs for wt and *Dnmt3b*$^{-/-}$ cells, before and after 12-day EB differentiation (denoted as ES and EB respectively). Most of CGIs stay unmethylated in *Dnmt3b*$^{-/-}$ cells.
D) Volcano plot shows the log2 fold change in MBD-seq signal between *Dnmt3b*$^{-/-}$ vs wt i.e., negative fold change indicates higher signal in wt in comparison to *Dnmt3b*$^{-/-}$.
E) Allelic ratio of X-linked CGIs in wt plotted against *Dnmt3b*$^{-/-}$ line in day 12 EBs, each dot equates to one CGI. Most CGIs attain an allelic ratio of 1 in wt EBs and do not gain methylation in *Dnmt3b*$^{-/-}$ line (dots clustered around 1 on the horizontal axis). Some CGIs on the X chromosome gain methylation independent of DNMT3B (examples labelled). Fewer of these, gain methylation specifically on Xi, as indicated by allelic ratio > 0.5 (upper right quadrant).
F) Metaplots showing average profile for CGIs characterized in section 3.6. The CGIs that get methylated in *Dnmt3b*$^{-/-}$ line on average do not show an allelic bias.
G) Genome browser tracks for some DNMT3B-independent (left) and dependent (right) CGIs on the Xi.
H) Relative gain of CGI methylation on the Xi during the timepoints studied in this experiment and a comparison of wt with *Dnmt3b*$^{-/-}$ line. CGIs gain methylation during EB differentiation predominantly between day 4-8, as described in the section 3.6. In *Dnmt3b*$^{-/-}$ cells, the CGIs that get methylated on average do not show an allelic bias (resulting in Xi-Xa ~0). The 5 CGIs that show allelic bias in *Dnmt3b*$^{-/-}$ line contribute to the small signal in EBs. D0-D12 represent day 0 – day 12 of EB differentiation. Panels (E)-(H) show one representative dataset out of 5 highly reproducible replicates (see Figure 3.13 B).

Interestingly "CGI-like elements" that I defined in section 3.9 were not affected by the absence of DNMT3B (Figure 3.14 A and B). Since these regions gained methylation on Xi during EB differentiation (Figure 3.12 G), it shows that these regions are also DNMT3B-independent for their methylation. A similar behaviour was observed for "rising peaks" that originated from repeat elements (section 3.9) (heatmap shown in Figure 3.14 C compares methylation of CGIs, "CGI-like elements" and "rising peaks" between wt and *Dnmt3b*$^{-/-}$). A genome browser track for "CGI-like elements" is shown in Figure 3.14 D.

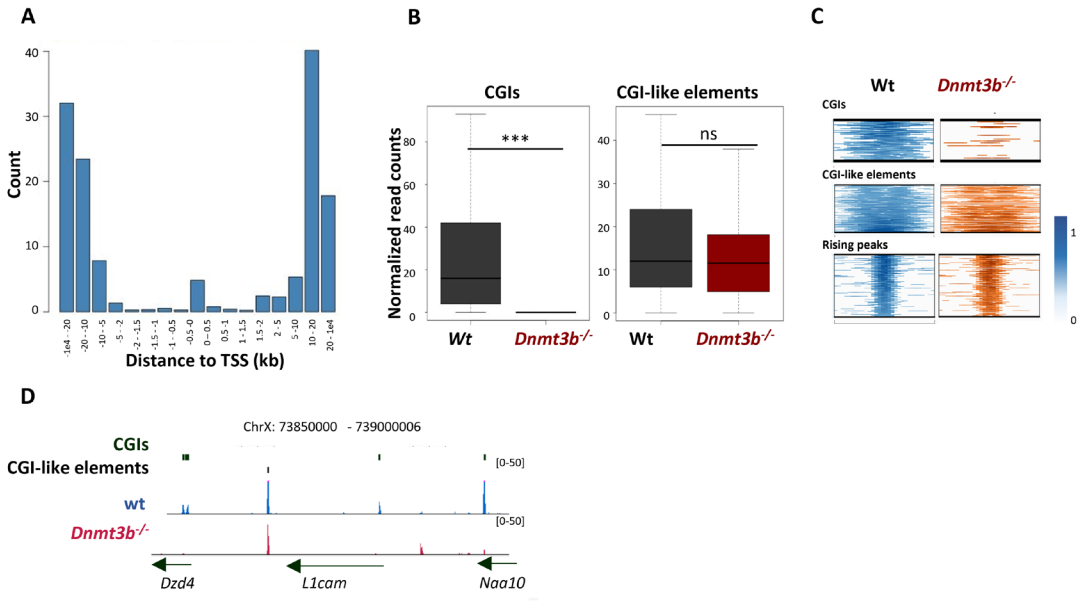


Figure 3.14: "CGI-like elements" gain methylation independent of DNMT3B.

A) Peaks called in *Dnmt3b*$^{-/-}$ cells after 12 days of EB differentiation on the X chromosome plotted against the distance from the nearest TSS. CGIs (found in proximity of TSS) are absent. "CGI-like elements", defined as the MBD-seq peaks further then 2 kb from TSS are largely unaffected in *Dnmt3b* knockout cells.

B) Boxplots comparing the MBD-seq signal on CGIs and "CGI-like elements" between wt and *Dnmt3b*$^{-/-}$ line. Methylation of "CGI-like elements" is unaffected in *Dnmt3b*$^{-/-}$ cells. (ns = not significant).

C) Heatmaps showing all classes of MBD-seq peaks defined in section 3.9. "Rising peaks" and "CGI-like elements" gain methylation in the absence of DNMT3B. CGIs are the most affected group in *Dnmt3b*$^{-/-}$ cells.

D) Genome browser tracks showing an example of "CGI-like elements". The region displayed in the screenshot is indicated on top of the figure. This figure shows one representative dataset of 5 highly reproducible replicates for MBD-seq on 12 days differentiated wt and *Dnmt3b*$^{-/-}$ cells (see Figures 3.7 A and 3.13 B).

Absence of DNMT3B in mice leads to embryonic lethality at E11.5 indicating that DNMT3B plays an important role in development (Okano et al., 1999). DiffBind analysis for differentially methylated regions between differentiated wt and *Dnmt3b*$^{-/-}$ identified 2428 differentially methylated regions genome-wide (219 on X chromosome), which stayed unmethylated in the *Dnmt3b*$^{-/-}$ line (Figure 3.15 A and B). These included genes involved in development such as *Dppa3, Wnt, Notch, Hoxb1* and *Igf2r* (Figure 3.15 C and D). *Igf2r* locus presents an interesting case of two closely located CGIs, one of *Igf2r* and another for the antisense transcript *Airn.* The former gains methylation in wt cells while *Airn* CGI gets methylated in *Dnmt3b*$^{-/-}$ line (Figure 3.15 D). Using this analysis, I was able to compile a list of CGIs

that are dependent upon DNMT3B for methylation. Intergenic CpG-rich regions which gained methylation in DNMT3B-dependent manner during differentiation were also included in this list. Figure 3.15 F shows a simplified schematic for defining DNMT3B-dependent intergenic MBD-seq signal and Figure 3.15 G shows the average signal on these loci in wt and *Dnmt3b*$^{-/-}$ cell line. As mentioned in section 3.9 these could indicate unannotated regulatory elements. Interestingly, DNMT3B-dependent CGI methylation showed similar dynamics on both autosomes and X chromosome (Figure 3.15 H) where most of the methylation takes places between day 4 and 8. The DNMT3B-independent regions gain methylation earlier in the differentiation timecourse (Figure 3.15 H).

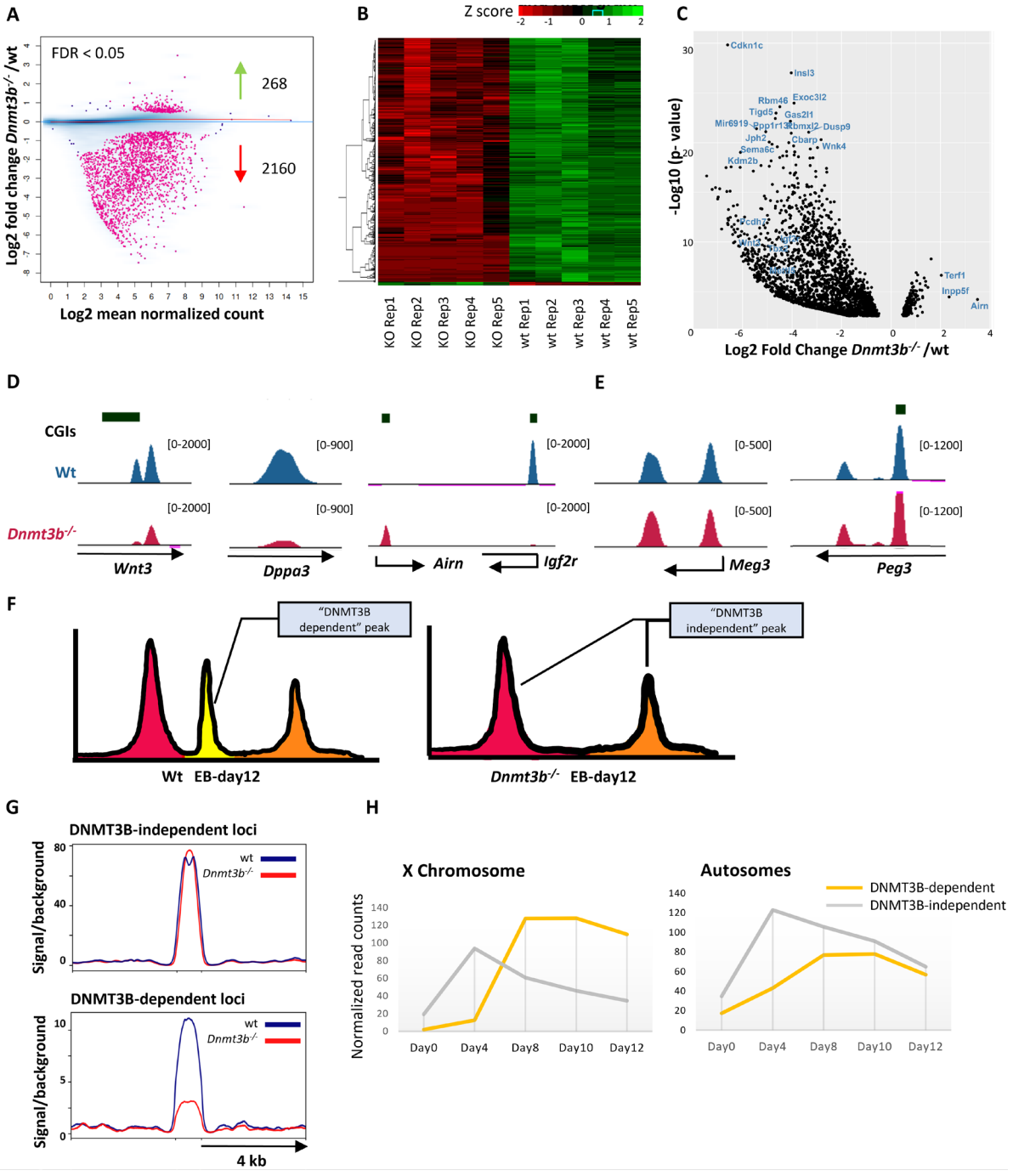


Figure 3.15: CGI methylation at autosomal CGIs in *Dnmt3b*$^{-/-}$ cells (shown on the previous page).

A) MA plot showing differentially methylated regions for 12-day differentiated wt and *Dnmt3b*$^{-/-}$ cells.
B) Heatmap showing the differentially methylated regions as illustrated in (A). The majority of the regions lose methylation upon loss of *Dnmt3b* (represented by KO). Replicates of wildtype and *Dnmt3b*$^{-/-}$ have been shown as wt Rep1-5 or KO Rep1-5 respectively.
C) Differentially methylated regions include developmental genes as well as imprinted regions. Some examples have been labelled in the volcano plot.
D) Genome browser tracks showing examples of DNMT3B-dependent CGIs (defined as regions showing compromised methylation upon knockout of *Dnmt3b*).
E) Genome browser tracks showing examples of DNMT3B-independent CGIs.
F) Schematic showing strategy of defining DNMT3B-dependent and -independent MBD-seq peaks. The MBD-seq peaks that were detected in day 12 EBs in wt but absent in *Dnmt3b*$^{-/-}$ line were termed as DNMT3B-dependent while the regions that were methylated in *Dnmt3b*$^{-/-}$ line in EBs were termed DNMT3B-independent. As described in section 3.9, these regions included annotated CGIs as well as intergenic and intragenic CpG-rich regions that could be associated with repeats or CREs.
G) Average MBD-seq signal on DNMT3B-independent (top) and -dependent (bottom) in wt and *Dnmt3b*$^{-/-}$ cells at day 12. The plot shows 4 kb regions flanking the regions of interest.
H) Plots show the dynamics of DNMT3B-dependent and -independent regions on the X chromosome and autosomes. DNMT3B-independent regions show faster dynamics. Panels (D), (G) and (H) show one representative dataset of 5 highly reproducible replicates for MBD-seq on 12 days differentiated wt and *Dnmt3b*$^{-/-}$ cells (see panel B).

Thus, my data builds on the earlier work from Gendrel et al 2012. By using high throughput sequencing data, I have provided a thorough analysis of the methylation gained on individual CGIs and the dynamics of CGI methylation on X chromosome and genome-wide. Overall, this chromosome-wide data agrees with the previous conclusion, obtained on a set of selective CGIs (Gendrel et al 2012), unequivocally proving that DNMT3B is crucial for methylation of most CGIs on Xi. In addition, I have identified exceptional CGIs and a group of unannotated novel "CGI-like elements" that are independent of DNMT3B for their methylation.

3.11 Loss of *Dnmt3b* does not affect XCI or gene expression in general

In section 1.2.11, I have provided background for the inverse correlation between transcriptionally active gene promoters and CGI methylation. The enrichment of H3K4me3 on actively transcribed promoters and the ADD domain-mediated autoinhibition of DNMT3B indicate the role of transcription

in antagonizing the establishment of DNA methylation. It is therefore important to investigate if the observed loss in CGI methylation in *Dnmt3b*$^{-/-}$ cells is a secondary effect of (1) a defect in X chromosome silencing and (2) inefficient differentiation of ES cells to EBs. The latter is of importance as Xi CGI methylation is known to be coupled with differentiation due to dependence on SMCHD1, which is recruited to Xi only in differentiated cells (section 1.3.3.1 and (Bowness et al., 2022)) (discussed in section 3.12).

In order to test the efficiency of silencing and differentiation, I performed chromatin RNA-seq (chrRNA-seq) on wt and *Dnmt3b*$^{-/-}$ ES cells and 12-day differentiated EBs. ChrRNA-seq captures nascent RNA transcripts which offers two main advantages: (1) detection of low abundance transcripts and (2) inclusion of unspliced transcripts bearing intronic regions, therefore increasing the number of SNPs for allelic analysis. Both factors add strength to the allelic analysis used to study the relative silencing on Xi using the allelic ratio. Deletion of *Dnmt3b* results in no noticeable changes in the global gene expression as compared to wt in EBs (Figure 3.16 A), implying that *Dnmt3b*$^{-/-}$ cells differentiate as efficiently as wt line. Furthermore, the result of allelic transcription analysis of *Dnmt3b*$^{-/-}$ vs wt cells (shown in Figure 3.16 B) indicate that both the wt and *Dnmt3b*$^{-/-}$ cells show efficient gene silencing on the Xi and *Dnmt3b*$^{-/-}$ cells show no noticeable loss of X chromosome silencing.

With regards to differentiation efficiency, in Figure 3.16 C, I have shown the expression levels of well-known pluripotency genes. As expected, these pluripotency genes are downregulated upon differentiation in wt and *Dnmt3b*$^{-/-}$ cells. The differences in *Oct4* and *Nanog* levels wt and *Dnmt3b*$^{-/-}$ cells at day 12 in Figure 3.16 C, were not statistically significant when taking replicates into account (Figure 3.16 A).

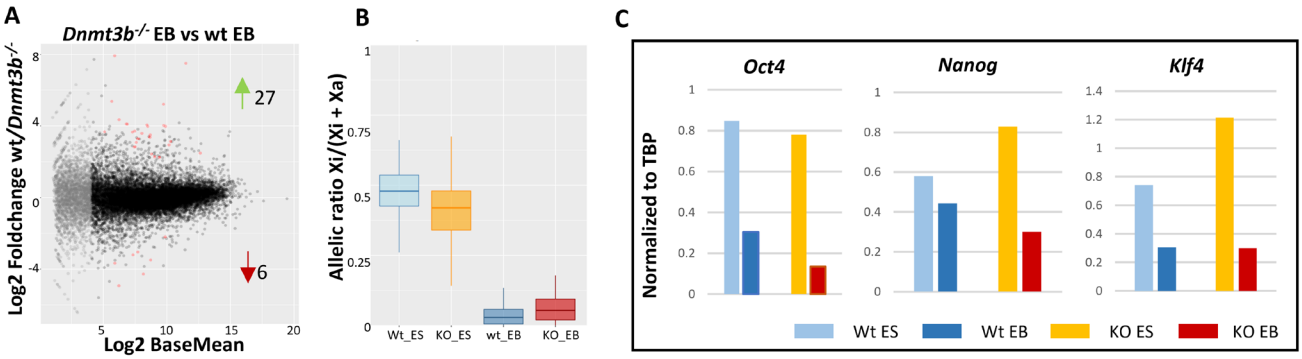


Figure 3.16: Interplay between CGI methylation and transcriptional silencing on the Xi (shown on the previous page).

A) MA plot comparing gene expression between *Dnmt3b*$^{-/-}$ and wt EBs. The significantly up- and down-regulated genes are shown as red dots and the counts are shown beside the arrows. Overall, the absence of DNMT3B has a negligible impact on gene expression. This differential expression analysis was carried out on two technical replicates for *Dnmt3b*$^{-/-}$ and wt cells using DESeq2, a Bioconductor package used extensively for this purpose.

B) Boxplots showing allelic ratio in wt and *Dnmt3b*$^{-/-}$ cells before and after differentiation. Wt_ES and EB denote allelic ratio for chrRNA-seq in wildtype ES and day 12 EBs respectively. Similarly, KO_ES and EB represent chrRNA-seq in *Dnmt3b*$^{-/-}$ line at undifferentiated ES and EB stage. Loss of *Dnmt3b* does not compromise transcriptional silencing during XCI.

C) Plots showing the expression levels of pluripotency genes *Oct4, Nanog,* and *Klf4,* all of which are downregulated during differentiation in both wt and *Dnmt3b*$^{-/-}$ (represented as KO for knockout) cells. The expression levels of these genes have been normalized to the housekeeping gene TATA-box binding protein *(TBP).* Panels (B) and (C) summarize data form two technical replicates for *Dnmt3b*$^{-/-}$ and wt cells.

3.12 Role of SMCHD1 in CGI methylation during XCI

The role of SMCHD1 in CGI methylation on Xi is well established (Blewitt et al., 2008; Blewitt et al., 2005; Gdula et al., 2019; Gendrel et al., 2012). A recent study from our lab presented high quality WGBS data for *Smchd1* wt and knockout MEF lines, demonstrating that the loss of *Smchd1* results in strongly reduced methylation at a significant number of CGIs on Xi (Gdula et al., 2019). I therefore looked at the overlap between the CGIs defined as SMCHD1-dependent in this study and my data. Indeed, all SMCHD1-dependent CGIs defined by Gdula and colleagues overlapped with DNMT3B-dependent CGIs on Xi (Figure 3.17 A). Gendrel and colleagues found that enrichment of SMCHD1 on Xi during EB differentiation occurred after the methylation of fast dynamic CGIs (Gendrel et al., 2012). To support this, they examined the methylation status of "fast" CGIs in *Smchd1* knockout MEFs and found that the methylation at these regions was indeed unaffected. I tested if the classification of dynamic classes of CGIs that I performed in section 3.6, also follows this trend. I indeed found that SMCHD1-dependent CGIs mostly lie in the "slow" dynamic category and that they get methylated between day 4-8 (shown in Figure 3.17 B and C). However, as the datasets being compared originate from different experimental systems (E9.5 MEF and iXist-ChrX cells, the latter of which is expected to have faster dynamics due to induction of high levels of Xist), this analysis must be repeated using definition of SMCHD1-dependent CGIs obtained from an experiment using the iXist-ChrX model system.

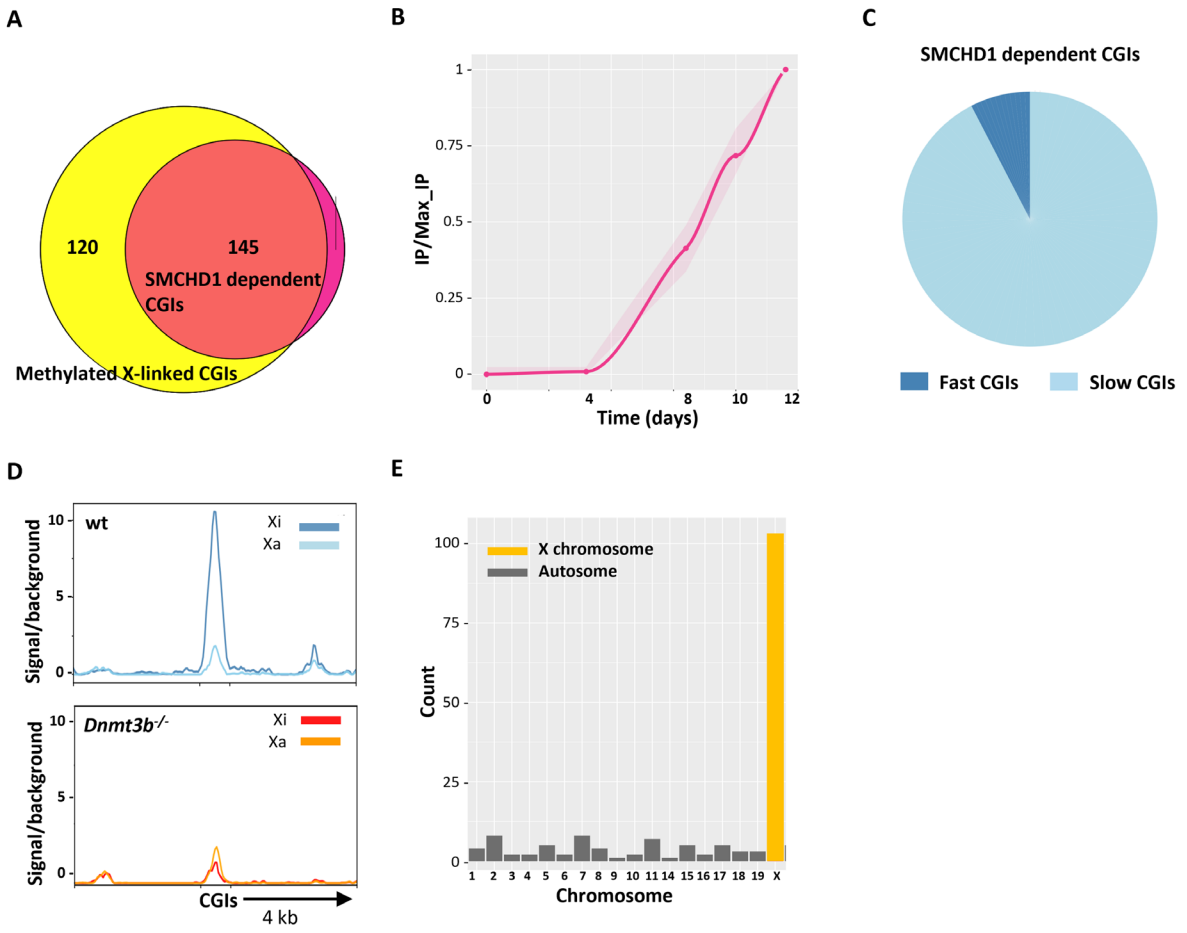


Figure 3.17: Role of SMCHD1 in CGI methylation.

A) Venn diagram showing the overlap between CGIs methylated during differentiation and SMCHD1-dependent CGIs. The latter were defined by differential analysis of WGBS data collected from wt and *Smchd1* mutant MEF by Gdula et al, 2019.

B) Line plot showing the gain of methylation at SMCHD1-dependent CGIs. The absolute levels of methylation (MBD-seq reads at various timepoints during differentiation) have been normalized to the maximum value of methylation, to obtain a ratio with values between 0 and 1 (IP/MaxIP, introduced in section 3.7). The solid line represents the median of the IP/MaxIP ratio, and the ribbon shows the interquartile range. The plot includes 2 replicates for day 0 - day 10 and 5 replicates for day 12 timepoints respectively.

C) Pie chart shows the proportion of "fast" and "slow" CGIs among SMCHD1-dependent CGIs on the X chromosome. The majority of the SMCHD1-dependent CGIs are "slow" CGIs.

D) Metaplots showing the MBD-seq signal on Xi and Xa in day 12 EBs in wt and *Dnmt3b*$^{-/-}$ cells (representative of 5 reproducible datasets , see Figure 3.15 B).

E) Bar plot showing the genomic distribution of CGIs that are both DNMT3B- and SMCHD1-dependent for their methylation. Here, SMCHD1-dependent CGIs were defined as CGIs found hypomethylated in *Smchd1* mutant cells differentiated to NPCs for 15 days in comparison to wt in MBD-seq. Most of these CGIs reside on the X chromosome.

The requirement of SMCHD1 for CGI methylation on Xi is intriguing as SMCHD1 has no known methyltransferase activity and is not known to interact with DNA methyltransferases. Given the clear dependence of the methylation of X-linked CGI on DNMT3B, I tested if this link (i.e., co-dependence of CGI methylation on DNMT3B and SMCHD1) occurs at other sites in the genome. At the time of this study, Dr. Tatyana Nesterova performed MBD-seq on an *Smchd1* mutant iXist-ChrX cell line that was differentiated to NPCs for 15 days. I defined SMCHD1-dependent CGIs as the CGIs that gained methylation during differentiation in wt and stayed unmethylated in the *Smchd1* mutant cell line. I then found the SMCHD1-dependent peaks that overlapped with DNMT3B-dependent peaks. The majority of these (103/ 172) were on the X chromosome (Figure 3.17 E), but there are nevertheless some autosomal loci, for example *L3mbtl4, Notch3, Pchd4* and *Adamts20* for which DNA methylation is both SMCHD1- and DNMT3B-dependent. However, it can be concluded that in this experimental system, the dependence of DNMT3B-mediated CGI methylation on SMCHD1 is predominantly an X chromosome-associated phenomenon.

3.13 Summary and conclusions

I have developed a new cell culture-based model system to investigate the mechanism of DNMT3B action in CGI methylation during XCI. The optimised model system uses MBD-seq to characterise the dynamics of CGI methylation on Xi using the interspecific iXist-ChrX mouse ES cell line and EB differentiation. Inducible Xist expression, increased sensitivity and genome-wide coverage of MBD-seq, and the ability to perform allele-specific analysis provide advantages over previously used model systems.

MBD-seq shows high specificity for methylated CpG-rich DNA. This feature means that for CGIs that get methylated during XCI, there is a nominal signal (if at all) originating from Xa. This sensitivity offered by MBD-seq is useful to effectively capture allelic CGI methylation even at relatively low levels. I have compiled a group of 265 CGIs that were reproducibly found to gain methylation during XCI. Consistent with the previous work, I have shown that CGIs on Xi can be divided into two groups, (1) those exhibiting fast methylation dynamics (reaching an allelic ratio of 1 at day 4) and (2) “slow” CGIs (allelic ratio of 1 at day 8 of EB differentiation). The latter represents a larger class. My observations validate and extend observations by Gendrel et al., which were conducted on selected CGIs.

Reassuringly, the MBD-seq datasets in this study show high reproducibility between independent experiments and differentiation timecourses.

My experimental system is further improved by the use of internal calibration with human DNA (thus "calibrated MBD-seq"). Human DNA serves as a reference to normalize for variability introduced due to IP efficiency and library preparation steps and therefore allows comparison of methylation levels between timepoints during differentiation. This improvement revealed that absolute levels of methylation keep rising on CGIs on Xi until day 12, where they begin to plateau. I have also briefly examined those autosomal CGIs that gain methylation during differentiation, a group composed of pluripotency genes, developmental genes, and imprinted CGIs. Reassuringly, the autosomal CGIs found methylated in EBs were reported as methylated in E8.5 mouse embryos which shows that despite being an *in vitro* cell culture-based model system, iXist-ChrX system is "near physiological" in terms of CGI methylation profile (Auclair et al., 2014).

In addition to the annotated CGIs, my MBD-seq reproducibly captures some regions that reside near gene promoters or regions with high chromatin accessibility and H3K4me1 enrichment, indicating that these regions could be methylated unannotated CGIs and/or CREs (Figure 3.11 A). These regions are hypomethylated in ES cells and gain methylation during differentiation. In their paper, Illingworth and colleagues comment that the number of CGIs detected highly varied with the experimental conditions for the CAP-seq used for compiling a list of CGIs. Afterwards, Long et al., using an improved BioCAP-seq detected additional regions that they denoted as NMIs; so it is possible that these additional regions represent CGIs that escaped detection in previous studies which were based on CAP-seq approaches and were detected in EBs upon their methylation by a much more sensitive MBD-seq assay. Methylation of enhancers is reported to occur during mouse and human embryogenesis, especially for those associated with cellular differentiation and pluripotency genes such as *Sox2* and is thought to facilitate temporal gene regulation during development (Li et al., 2018).

Interestingly, some MBD-seq peaks found in intergenic regions (termed "CGI-like elements" in section 3.9) are specifically methylated on Xi. These "CGI-like elements" are also reproducibly observed between different MBD-seq experiments and overlapped with ATAC-seq and H3K4me1 signals and therefore show CRE-like chromatin features. I have referred to these as "CGI-like elements" since these are located far from genes. Further investigation is required to unveil the function performed by this methylation. I speculate that a proportion of "CGI-like elements" represent CREs, where methylation could occur during XCI as a step towards silencing of the associated genes. A large proportion of "CGI-like elements" originate from transposons (Figure 3.12 E). These are likely to be young transposable elements that possess weak transcription potential as chrRNA-seq shows that

"CGI-like elements" are associated with some low-level transcription (Figure 3.18), which is silenced on Xi in EBs. The preferential methylation of these transposon-associated "CGI-like elements" on Xi could be explained by the silencing of these transcripts which could in turn be due to chromosome-wide gene silencing, rendering them permissive to methylation on Xi. Finally, the function of the methylation of "CGI-like elements" could be to antagonize CTCF binding and help in the loss of TADs observed on Xi during the late stages of XCI; this can be experimentally tested by performing CTCF ChIP on ES and 12 day differentiated iXist-ChrX cells.

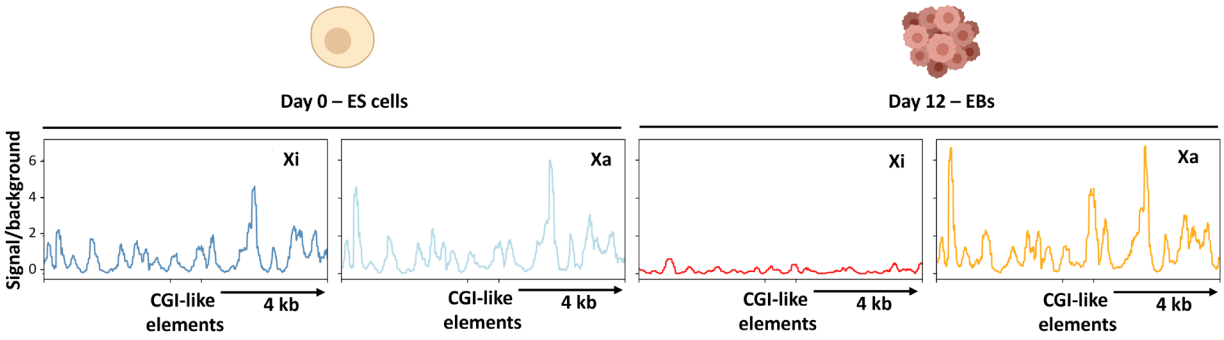


Figure 3.18: "CGI-like elements" are associated with low-level transcripts.

ChrRNA-seq signal in an 8 kb window flanking the "CGI-like elements" in ES cells and EBs. The low-level transcripts associated with "CGI-like elements" are silenced/downregulated on the Xi as opposed to the Xa upon differentiation. Shown here is a representative profile of two replicates.

I went on to test the role of DNMT3B in CGI methylation on Xi in the new model system. Using calibrated MBD-seq, I was able to identify individual CGIs on the X chromosome and autosomes that are dependent on DNMT3B for their methylation. DNMT3B is the main *de novo* methyltransferase bringing about CGI methylation on Xi. I have discovered that a minority of CGIs on the Xi become methylated independent of DNMT3B. Intriguingly, I found that methylation of "CGI-like elements" are unaffected in the absence of DNMT3B. On autosomes, the CGIs found dependent on DNMT3B were in coherence with those previously reported in mouse embryos (Auclair et al., 2014). The thorough genome-wide analysis of CGI methylation on Xi in *Dnmt3b*$^{-/-}$ cells and the identification of CGIs dependent on DNMT3B for their methylation during EB differentiation is novel to this work and forms the foundation for the experiments in the next chapter.

The absence of DNMT3B has a negligible effect on global gene expression patterns which overcomes the need to dissect the primary impact of DNMT3B loss from secondary effects on CGI methylation

and simplifies the analysis of complementation experiments described later. My analysis demonstrates that the methylation dynamics of CGIs exhibit some correlation with the silencing dynamics of genes during XCI. For instance, I have shown that "slow methylating" CGIs are mostly associated with genes that show slow silencing dynamics during XCI (Figure 3.9 J). This observation makes sense given the antagonistic role of transcription in CGI methylation. However, this correlation was not very clear in the opposite case, i.e., "fast methylating" CGIs were not strictly associated only with fast silencing genes. It is likely that other processes occurring between transcriptional silencing (initiating a few hours after Xist induction) and DNA methylation (shown in this study to start around 4 days into differentiation) play a more direct role in determining the methylation levels gained at individual CGIs.

One such process is the recruitment of SMCHD1, which has been shown in several studies to be indispensable for CGI methylation on Xi (Gdula et al., 2019; Gendrel et al., 2012). I have described that most of the CGIs that are methylated on Xi, overlap with CGIs shown to be dependent on SMCHD1 (Gdula et al 2019). As further validation of my experimental system, I found that the majority of the SMCHD1-dependent CGIs show slow methylation dynamics; this is in agreement with the results reported earlier in MEFs (Gendrel et al 2012). Unexpectedly, a small proportion of SMCHD1-dependent CGIs were "fast methylating" CGIs in my experiments. This difference could originate from the induction of Xist from the Tet-On promoter which results in higher than physiological expression of Xist and potentially alters the dynamics of processes in XCI when compared to MEFs where XCI occurs under physiological conditions.

Currently, it is not known how SMCHD1 influences CGI methylation on Xi. One hypothesis is that SMCHD1 facilitates DNA methylation by creating nucleosome-free regions. This hypothesis is supported by the structural and biochemical studies showing that DNMT3B catalysis occurs on linker DNA and not on nucleosome-occupied DNA (Xu et al., 2020) (also described in section 1.2.3). Previously, a chromatin remodeller LSH (also called HELLS) has also been shown to be crucial for DNA methylation for the same reason (Lyons and Zilberman, 2017). SMCHD1 could also be affecting chromatin recruitment or the residence of DNMT3B on chromatin (further investigated in chapter 5). DNMT3B ChIP in *Smchd1* mutant cells can show whether or not DNMT3B is delocalized from Xi in these cell lines.

In this chapter, I have shown that iXist-ChrX largely recapitulates CGI methylation that occurs under physiological conditions on both the X chromosome and autosomes. While the iXist-ChrX system allows a synchronous expression of Xist and overcomes the heterogeneity in cell cultures undergoing random X chromosome inactivation, a drawback of this system is the likelihood that induction of Xist

from Tet-On promoter results in higher than physiological expression of Xist. This in turn could affect the timing of gene silencing and earlier onset of CGI methylation. However, the iXist-ChrX line successfully overcomes the limitations of the earlier study (Gendrel et al 2012) and faithfully captures/reconfirms the observations made in MEFs and mouse embryos, both of which represent experimental systems where XCI occurs *in vivo*/under physiological conditions. The thorough characterization of CGI methylation genome-wide and the validation of the specificity of DNMT3B for CGI methylation on Xi, sets the ground for the downstream experiments aimed at understanding the underlying mechanisms for DNMT3B specificity.

Chapter 4

Investigating DNMT3B specificity mechanisms

4.1 Introduction

As detailed in chapter 1, DNMT3A and DNMT3B are evolutionarily well-conserved homologs and act redundantly at the majority of genomic loci, but with some important exceptions. The data in the previous chapter confirmed that CGI methylation in XCI is an exception with a clear dependence on DNMT3B in the majority of instances. Furthermore, embryonic lethality of *Dnmt3b*$^{-/-}$ (E11.5) in comparison to *Dnmt3a*$^{-/-}$ embryos which survive until birth, also points to a unique role of DNMT3B during development. Some of the DNMT3B-dependent genes identified in my experiments represent key developmental regulators, for example *Hox* genes, *Wnt, Igf2r, Dppa3* (section 3.8) and this could provide an explanation for the relatively early lethality. What mechanisms govern DNMT3B specificity is, therefore, a key question.

Some recent studies have investigated if the specificity of DNMT3 proteins for their target sites could be governed by distinct enzymatic properties or expression patterns, as discussed in chapter 1, section 1.4.1. The high processivity of DNMT3B, for instance, has been suggested to be a factor for the specificity of DNMT3B for its target sites. DNMT3B has been shown to possess a less flexible target recognition domain (TRD) which contributes to more stable DNA binding and, in turn, more processivity in comparison to DNMT3A (Lin et al., 2020). This property of DNMT3B has been suggested to be the reason behind the indispensable role of DNMT3B in the methylation of pericentric repeats which are composed of large regions carrying sequences with high GC content (Lin et al., 2020).

Although insightful, prior studies do not fully explain the specificity of *de novo* DNMT3 proteins for developmental loci. Furthermore, both DNMT3A and DNMT3B have several isoforms generated by alternative splicing. However, DNMT3B, unlike DNMT3A, codes for several inactive isoforms that are distinct from the catalytically active counterparts, in both structure and function. These inactive isoforms have been reported to regulate the methylation activity of the catalytically active DNMT3A and DNMT3B isoforms (Duymich et al., 2016). One catalytically inactive isoform of DNMT3B, DNMT3B3, is well known to act as a coenzyme and enhance the catalytic activity of DNMT3A and

DNMT3B (Gordon et al., 2013; Zeng et al., 2020). Accordingly, the reason that knockout of the *Dnmt3b* gene impacts developmental gene expression could be the loss of DNMT3B3 protein impacting both DNMT3A- and DNMT3B-mediated methylation. Therefore, in this chapter I explore two themes; dissect the contribution of two predominantly expressed isoforms of DNMT3B in CGI methylation on Xi as a means to decipher if DNMT3B specificity is due to a specific isoform, and investigate the regions of DNMT3B that could govern its specificity for CGI methylation on Xi.

4.1.1 Isoform model of DNMT3B specificity for CGI methylation on Xi

The *Dnmt3b* locus undergoes extensive alternative splicing. In humans up to 30 isoforms of DNMT3B have been reported, several of which are exclusively expressed in cancers (Ostler et al., 2007). The functions of only a few of these inactive isoforms have been well researched, implicating specific isoforms either in activation or inhibition of catalytically active DNMT3A and DNMT3B (Chen et al., 2002; Gordon et al., 2013; Weisenberger et al., 2004). The relatively abundant isoforms of DNMT3B in mouse are DNMT3B1-8 (Figure 4.1 A). Two of these, DNMT3B1-2, are catalytically active. In ES cells DNMT3B1, the longest isoform coded by *Dnmt3b* gene, is predominant. DNMT3B2-5 have 20 residues missing between PWWP and ADD domains (Figure 4.1 A). Among the inactive isoforms, DNMT3B3 has been widely researched. Interestingly, during development DNMT3B3 is upregulated and its expression is maintained in somatic cells (Figure 4.1 B). DNMT3B3 has been suggested to replace DNMT3L in somatic cells (DNMT3L is downregulated in early development), serving as an accessory factor to enhance the catalytic activity of DNMT3A and DNMT3B (Chedin et al., 2002; Gordon et al., 2013) (Figure 4.1 C). A recent structural study suggests that stimulation of the catalytic activity of DNMT3A and DNMT3B by DNMT3B3 can be explained by its ability to bind the acidic patch of the nucleosome. This in turn places the TRD of catalytically active DNMT3A or DNMT3B, which is heterodimerized with DNMT3B3, on the linker DNA where the former then carries out methylation (Xu et al., 2020) (Figure 4.1 D). Thus, both the catalytically active and inactive isoforms are important for DNA methylation.

Based on this background, two scenarios can be assumed via which DNMT3B contributes towards CGI methylation on Xi: (1) catalytically active DNMT3B1 establishes CGI methylation or (2) catalytically inactive DNMT3B3 is required as a cofactor to promote Xi CGI methylation by catalytically active DNMT3 (A or B) (schematic shown in Figure 4.2 E). It is also possible that both of these mechanisms are at play. Indeed, a recent study has suggested that DNMT3B3 preferentially enhances the function of DNMT3B rather than DNMT3A (Zeng et al., 2020). Notably, the earlier study from our lab that

established the specificity of DNMT3B for CGI methylation on Xi, made use of *Dnmt3b* null embryos that do not express any of the DNMT3B isoforms (Gendrel et al., 2012). The contribution of individual isoforms of DNMT3B in XCI has thus not been explored.

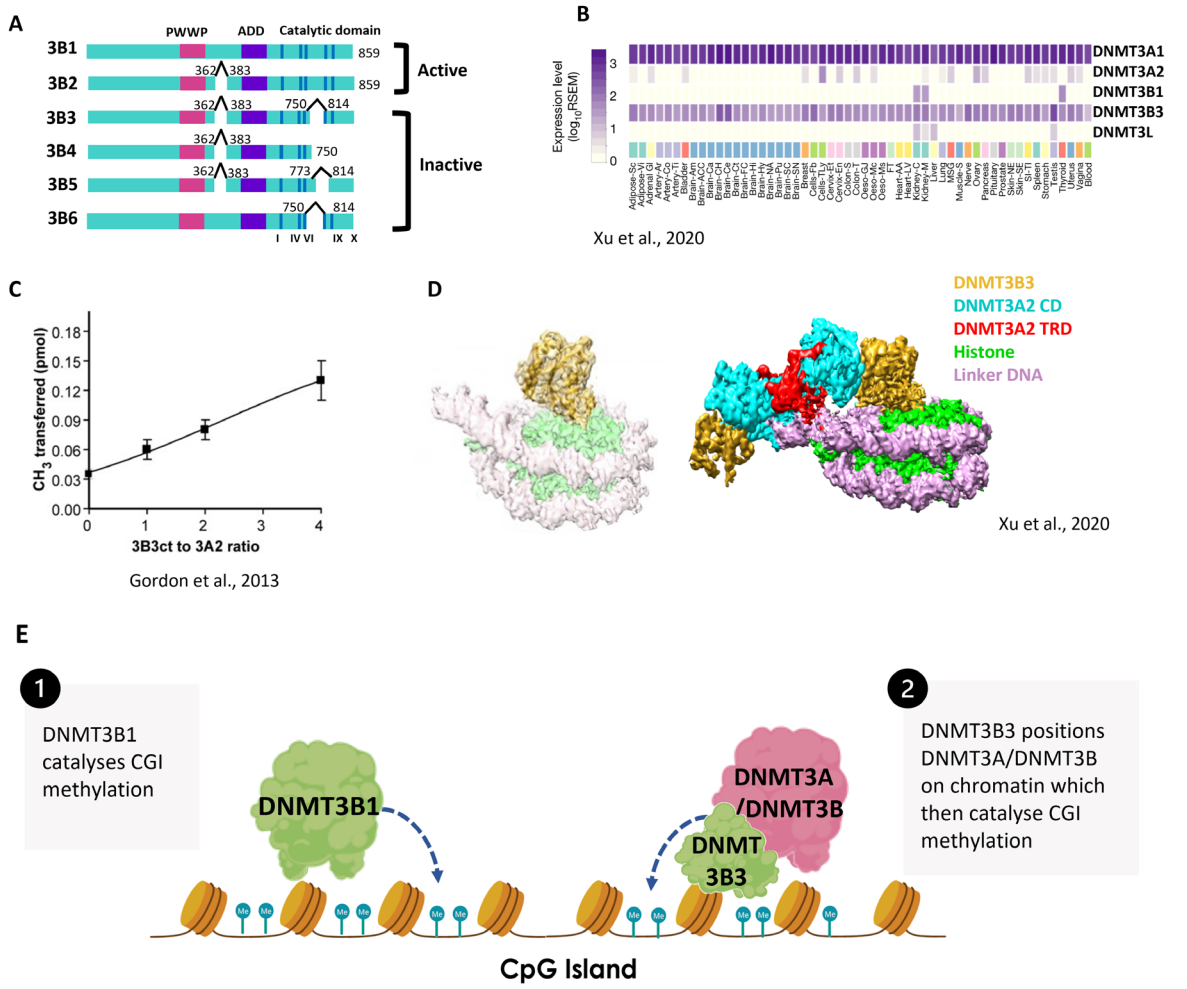


Figure 4.1: The *Dnmt3b* gene codes for multiple isoforms (continued to the next page).

A) Schematic showing isoforms of DNMT3B. DNMT3B1 and 3B2 are both catalytically active. DNMT3B2-3B5 lack amino acids 362-383. The catalytically inactive isoforms of DNMT3B lack residues from the conserved motifs (denoted by the roman numerals on the bottom of the schematic) in the catalytic domain as shown.

B) Adapted from Xu et al., 2020. The heatmap shows the expression of human DNMT3A1, 3A2 and DNMT3B1 and 3B3 in various human somatic tissues. DNMT3A1 and 3B3 are widely expressed in somatic cells, while DNMT3A2 and DNMT3B1 are downregulated.

C) Graph adapted from Gordon et al., 2013 shows that upon the addition of DNMT3B3, the catalytic activity of DNMT3A2, measured by the CH3 transfer from radio-labelled methyl donor, S-adenosyl methionine (SAM) to the substrate DNA, is enhanced.

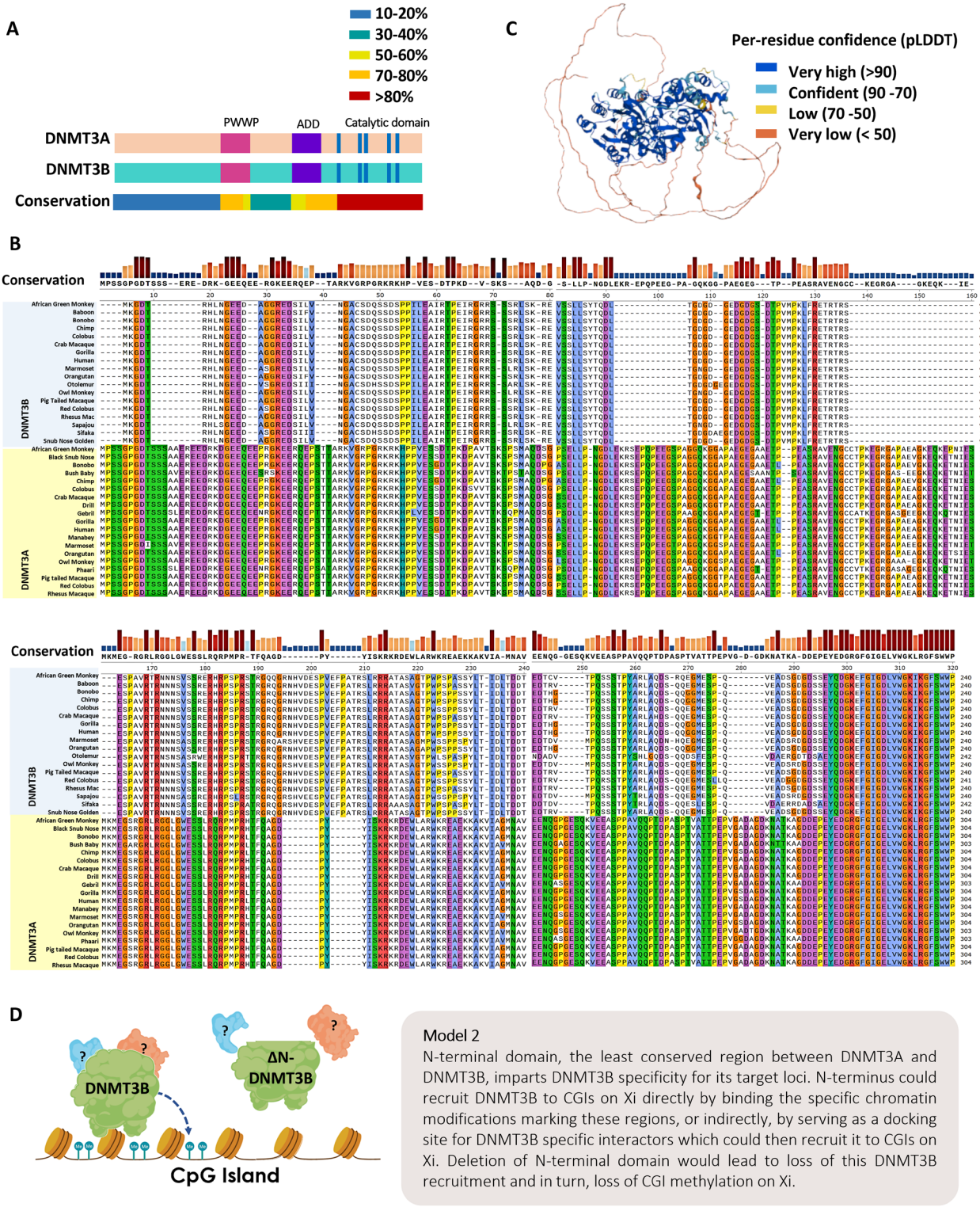
D) Figure from Xu et al., 2020 elucidating the structure of catalytically inactive DNMT3B3 and active DNMT3A2 in complex with the nucleosome. The left panel shows DNMT3B3 (yellow) bound to the acidic patch of nucleosome. The right panel shows the nucleosome-DNMT3B3-DNMT3A2 complex. The nucleosome-bound DNMT3B3 is heterodimerized with DNMT3A2 (blue). Binding of DNMT3B3 to nucleosome acidic patch leads to positioning of DNMT3A2 TRD (red) on linker DNA.

E) First model for DNMT3B-mediated CGI methylation on Xi. Scenario 1 assumes that DNMT3B1 catalyses the methylation on CGIs while scenario 2 assumes that DNMT3B specificity is conferred by DNMT3B3, which is pivotal to position/localize active DNMT3A or DNMT3B (which could be redundant) to then carry out CGI methylation.

4.1.2 Specificity-imparting feature model of DNMT3B

Structural studies on various domains of DNMT3A and DNMT3B have revealed subtle differences between the two highly conserved homologues. For instance, the PWWP domain of DNMT3B has been shown to have a higher affinity of H3K36me3 as compared to its counterpart in DNMT3A (Baubec et al., 2015). Similarly the ADD domain of DNMT3B forms a weaker interaction with the catalytic domain and the different conformation of the autoinhibitory loop has been proposed to form an interface for additional protein-protein interactions in DNMT3B (Boyko et al., 2022). However, the N-terminal domain represents the least conserved domain between the homologues, making it a candidate region that could be responsible for DNMT3B specificity (Figure 4.2 A). A recent publication showed that the N-terminal domain of DNMT3A and DNMT3B is under positive divergence pressure (Molaro et al., 2020). Intriguingly, despite the lack of conservation in the N-terminus among the paralogues (DNMT3A and DNMT3B), the N-terminal domains of both DNMT3A and DNMT3B are well conserved among the orthologues in primates (Molaro et al., 2020) (Figure 4.2 B). This conservation of the N-terminal domains of DNMT3A and DNMT3B with their orthologues is perplexing since the N-terminal domain is predicted to form a flexible unstructured region (Figure 4.2 C). Until recently, this domain was poorly studied and there was much less known regarding the functions of this part of DNMT3 proteins.

Jeong and colleagues suggested that the N-terminus of DNMT3B has a role in stabilization and the general localization/anchorage of DNMT3B on chromatin (Jeong et al., 2009). They showed that in sucrose gradients, the N-terminal deletion of DNMT3B caused it to dissociate from chromatin. Another group showed later that DNMT3A2, the isoform of DNMT3A with the shorter N-terminus, has lower DNA binding activity compared to DNMT3A1, the isoform with a long N-terminus (Suetake et al., 2011). They also showed the N-terminal domain alone could bind DNA and pointed out that basic residues in the N-terminus of DNMT3A could be responsible for this DNA binding. Thus, both studies suggested the role of the N-terminal domain of DNMT3A and DNMT3B in general chromatin / DNA binding.



Model 2

N-terminal domain, the least conserved region between DNMT3A and DNMT3B, imparts DNMT3B specificity for its target loci. N-terminus could recruit DNMT3B to CGIs on Xi directly by binding the specific chromatin modifications marking these regions, or indirectly, by serving as a docking site for DNMT3B specific interactors which could then recruit it to CGIs on Xi. Deletion of N-terminal domain would lead to loss of this DNMT3B recruitment and in turn, loss of CGI methylation on Xi.

Figure 4.2: N-terminal domain of DNMT3A and DNMT3B represents the most diverged regions between the two homologues (show on the previous page).

A) Conservation between mouse DNMT3A and DNMT3B measured by sequence alignment. DNMT3A and DNMT3B share a high degree of conservation as evident from the same linear order of composite domains. In contrast, the N-terminus of DNMT3A and DNMT3B is not conserved between the two homologues. The color-coded bar underneath the DNMT3A and DNMT3B schematic defines the regions that on average show similar levels of conservation. The linker regions between PWWP and ADD domain (138 amino acids) and the N-terminus (229 amino acids for DNMT3B) are the least conserved regions between the two homologues.
B) Protein sequence alignment and conservation between DNMT3A and DNMT3B from various mammals. The N-terminal domain despite being less conserved between DNMT3A and DNMT3B, is fairly conserved within the group. The sequence alignment was performed using Clustal Omega.
C) Structure of mouse DNMT3B as generated by alpha fold. Alpha fold calculates a per-residue confidence score which shows how accurate the structure prediction is. The entire N-terminal domain has very low confidence score and is shown as an unstructured loop.
D) Schematic explaining the model and rationale for the complementation experiment with N-terminal deletion of DNMT3B.

Some recent studies have highlighted the role of the N-terminal domain of DNMT3A in determining its recruitment to target genomic loci. Manzo and colleagues showed that the two isoforms of DNMT3A which vary only in the length of their N-terminal domains exhibit distinct binding patterns in the genome (Manzo et al., 2017). The group of David Allis has recently reported that the N-terminal domain of DNMT3A determines its recruitment to CGIs modified with H2AK119ub1; this mode of recruitment competes with PWWP domain-mediated recruitment to H3K36me3-modified chromatin found on gene bodies of actively transcribing genes (Weinberg et al., 2021). The N-terminal domain-mediated recruitment of DNMT3A to H2AK119ub1-enriched regions was also observed by Goodel's group (Gu et al., 2022). However, Goodel and colleagues did not observe the competition between N-terminus and PWWP domains. They showed that DNMT3A localizes on H2AK119ub1-modified regions even in the presence of PWWP domain and this localization is abolished upon the deletion of the N-terminus (Gu et al., 2022). Hence, recent reports point towards the role of the N-terminal domain in recruitment of DNMT3A to specific target location i.e., at Polycomb-enriched genomic loci. Given the recent evidence of the importance of the N-terminal domain in determining the genomic localization of DNMT3A, the N-terminal domain of DNMT3B could be performing an analogous function of determining its specific localization and role in CGI methylation on Xi. The lack of conservation between N-terminal domains of DNMT3A and DNMT3B could explain the distinct target loci of the two enzymes.

4.2 A complementation strategy to study the role of DNMT3B isoforms in Xi CGI methylation

Since DNMT3B isoforms share most of the coding sequence of the gene, engineering mutations of the *Dnmt3b* locus such that only one isoform is deleted is not possible (Figure 4.1 A). Any manipulation in the gene would affect several coded transcripts, creating potential unpredictable secondary effects. Even if the manipulation of a single isoform were possible, the developmental control of the alternative splicing of Dnmt3b transcript and our incomplete understanding of the functions of DNMT3B isoforms would complicate the study. To overcome this complexity, I have used a complementation strategy where I introduce one DNMT3B isoform at a time in *Dnmt3b*$^{-/-}$ cells generated in iXist-ChrX parent line.

For the generation of complementation lines, cDNA encoding DNMT3B isoforms were introduced at a homing site, the TIGRE locus on chromosome 9 (Zeng et al., 2008), under a constitutive pCAG promoter (schematic shown in Figure 4.3 A). As described in chapter 3, one allele of the TIGRE locus carries the sequence coding tet-transactivator (rtTA) which allows doxycycline-inducible expression of Xist in the iXist-ChrX line. The complementary lines, therefore, carry a single copy of the respective rescue protein being tested. To test the ability of DNMT3B complementation constructs to rescue CGI methylation on Xi, I differentiated complementary cell lines into EBs for 12 days where I observe maximal CGI methylation (section 3.7) and performed MBD-seq on day 0 ES cells and day 12 EBs. The CGI methylation pre- and post-differentiation in these lines can be compared with wildtype and *Dnmt3b*$^{-/-}$ cells (Figure 4.3 B). This strategy was used to perform all of the complementation assays described in this chapter.

The analysis of MBD-seq data from these complementation lines builds on the work described in chapter 3. Specifically, I characterized the CGIs that are hypomethylated in *Dmnt3b*$^{-/-}$ cells. CGIs on Xi are predominantly hypomethylated in this cell line (250 out of 265 CGIs that are amenable to allelic analysis are DNMT3B-dependent and found methylated during EB differentiation in wt cell line). The *Dnmt3b*$^{-/-}$ cell line, therefore, forms a 'baseline' upon which any changes/rescue of CGI methylation, resulting from the introduction of isoforms/mutants of DNMT3B, can be studied with the MBD-seq assay (Figure 3.12 H). CGIs that were unaffected in *Dnmt3b*$^{-/-}$ cells, on the other hand, serve as a good control to test the efficiency of MBD-seq.

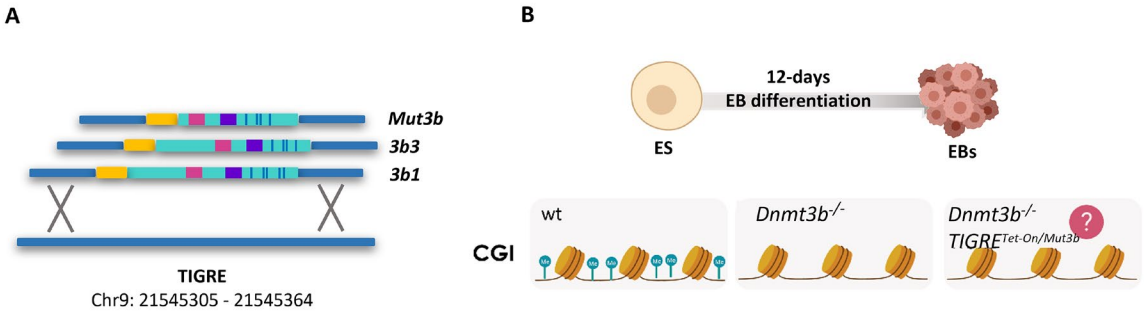


Figure 4.3: Complementation assay to study the contribution of DNMT3B isoforms and /or the protein features that confer specificity for Xi CGI methylation.

A) Schematic illustrating the targeting of DNMT3B isoforms or mutants into the TIGRE locus of iXist-ChrX line for generation of complementary lines. The coordinates of the targeted locus are shown at the bottom.
B) Since the establishment of DNA methylation requires differentiation, iXist-ChrX *Dnmt3b*$^{-/-}$ complementation lines are differentiated for 12 days to EBs. The methylation state of CGI in these lines can then be studied via MBD-seq and compared with wt and *Dnmt3b*$^{-/-}$ control.

4.3 The catalytically active isoform DNMT3B1 fully rescues CGI methylation on Xi

Using CRISPR/Cas9-assisted homology recombination, I generated *Dnmt3b1* complementation cell line in iXist-ChrX *Dnmt3b*$^{-/-}$ clone C6 (iXist-ChrX *Dnmt3b*$^{-/-}$ *TIGRE*$^{Tet\text{-}On/Dnmt3b1}$ which I have referred to as *Dnmt3b1* knock-in or *Dnmt3b1-KI* for simplicity). I confirmed that three clones were expressing DNMT3B1 (Figure 4.4 A). Loss of X chromosomes among female ES cells when grown in culture, resulting in XO karyotype is a well-reported issue in the XCI field. Only 2 of the three clones (A6 and H2) were stable and maintained both X chromosomes. Since *Dnmt3b1-KI* cell line showed a high propensity to become XO in culture, I subcloned the two positive clones to ensure that I had a stable XX line for this experiment. Described here are the results for two subclones (H2D3 and H2G3) that maintained XX karyotype throughout the differentiation timecourse (also see Appendix 1).

Since the establishment of CGI methylation depends on differentiation, it is crucial to ensure proper differentiation of the complementation lines to derive reliable conclusions from the MBD-seq results. I, therefore, performed chrRNA-seq on these lines alongside MBD-seq. This allowed me to compare how well the *Dnmt3b1-KI* line differentiates in comparison to the parental *Dnmt3b*$^{-/-}$ lines. ChrRNA-

seq also allowed me to study any changes in gene expression caused by *Dnmt3b1* complementation. Figure 4.4 C shows the comparison of the gene expression profile of wt and *Dnmt3b1-KI* lines at day 12 EB. Only a negligible proportion of genes are differentially expressed in the *Dnmt3b1-KI* complementation line, and these do not show any bias (i.e., favourably up or downregulated) in their expression pattern (Figure 4.4 C). Figure 4.4 B shows the expression levels of three pluripotency genes *Oct4, Nanog,* and *Klf4*. As expected, these are downregulated equally in the rescued clones compared to wt day 12 EBs. Taken together chrRNA-seq data showed that the *Dnmt3b1-KI* line differentiates as efficiently as the wt parental line.

ChrRNA-seq also validated that the expression level of the *Dnmt3b1* complementation construct was comparable to that in wt during differentiation (Figure 4.4 D, left panel). As per previous reports, both *Dnmt3b* and *Dnmt3l* are downregulated during differentiation (Figure 4.4 D, left and right panel) (Bourc'his et al., 2001; Chen et al., 2002). *Dnmt3a* on the contrary is known to be expressed in somatic cells (Chen et al., 2002). During differentiation, DNMT3A2 is replaced by the longer isoform, DNMT3A1, which is expressed even in terminally differentiated cells (Chen et al., 2002). It is reported that both DNMT3A1 and DNMT3A2 are expressed in EB differentiation between day 4-day 12 which could explain the upregulation of *Dnmt3a* observed in EBs (Chen et al., 2002), (Figure 4.4 D, middle panel).

I then looked at the X-linked CGIs that are methylated during XCI in EBs. The methylation of these CGIs was rescued in the *Dnmt3b1* complementation lines (Figure 4.4 E) validating that the DNMT3B1 isoform is sufficient to fully complement the methylation of X-linked CGIs. The DNMT3B1 isoform also rescued autosomal CGI methylation as characterized in section 3.10 (Figure 4.4 F). I also performed an in-depth DiffBind analysis to search for regions differentially methylated between wt and *Dnmt3b1-KI* EBs captured in MBD-seq data. A small number of regions (711 out of 56295 MBD-seq peaks captured, 1.26%) were found to be differentially methylated between differentiated wt and *Dnmt3b1-KI* line (Figure 4.4 G, left). Only 215 of these 711 regions were CGIs and these were only ~1 fold hypermethylated in the *Dnmt3b1-KI* line (Figure 4.4 G right, X-linked CGIs are shown in red). The regions hypomethylated in *Dnmt3b1-KI* line as compared to wt were mostly intergenic, as evident from greater distance to TSS (Figure 4.4 H). I conclude that this low level of differential methylation among the samples is a background effect possibly related to the heterogenous nature of EB differentiation and variability in pulldown efficiency of MBD-seq experiment. I did not find any CGI that is methylated in wt cells and that failed to get methylated in the *Dnmt3b1* complementation line in this analysis. Overall, DNMT3B1 faithfully complements the CGI methylation lost due to *Dnmt3b* knockout.

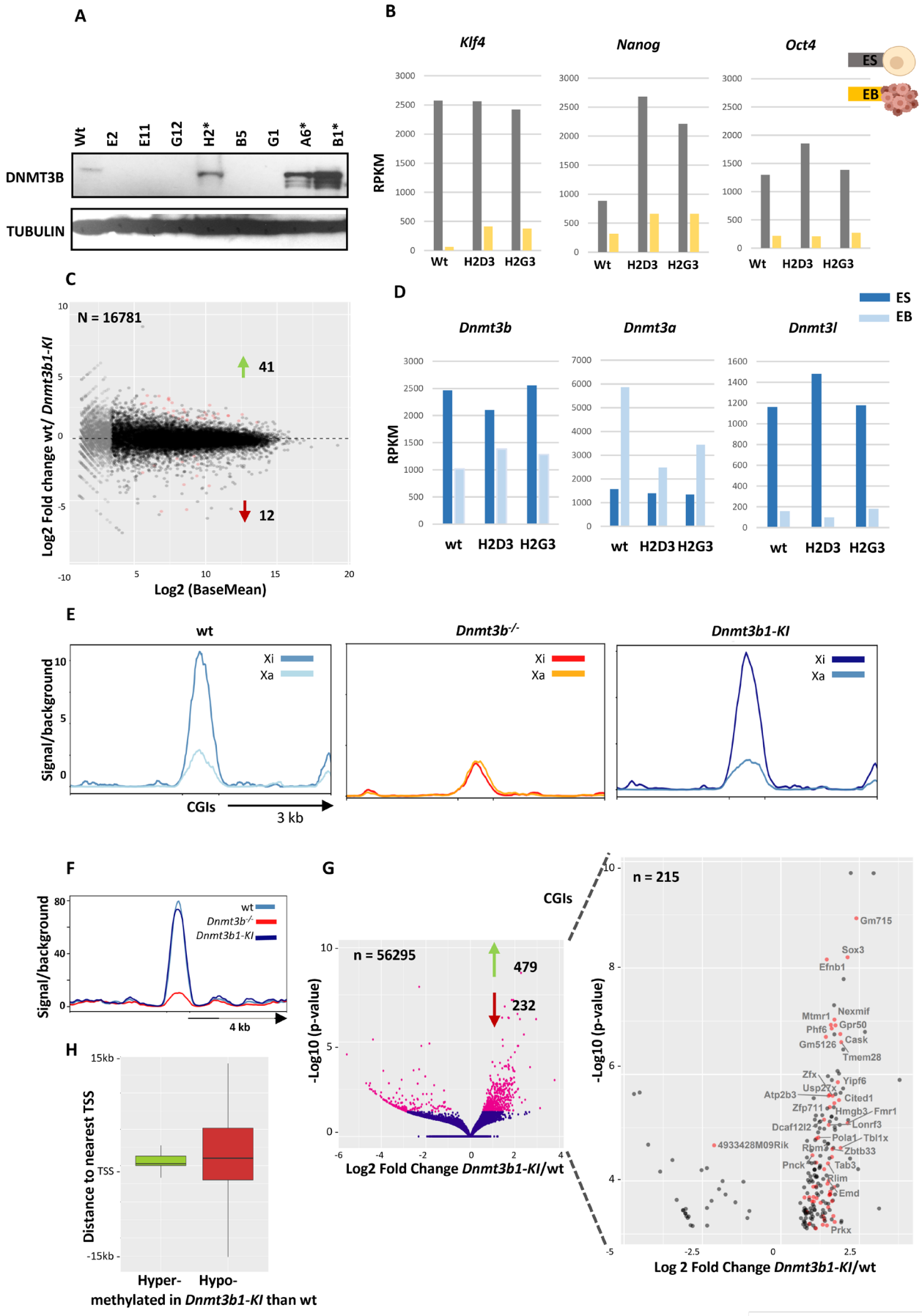

A
Wt
E2
E11
G12
H2*
B5
G1
A6*
B1*
DNMT3B
TUBULIN
B
Klf4
Nanog
Oct4
ES
EB
RPKM
Wt
H2D3
H2G3
C
N = 16781
41
12
Log2 Fold change wt/ Dnmt3b1-KI
Log2 (BaseMean)
D
ES
EB
Dnmt3b
Dnmt3a
Dnmt3l
RPKM
wt
H2D3
H2G3
E
wt
Dnmt3b-/-
Dnmt3b1-KI
Xi
Xa
Signal/background
CGIs
3 kb
F
wt
Dnmt3b-/-
Dnmt3b1-KI
Signal/background
4 kb
G
CGIs
n = 56295
479
232
-Log10 (p-value)
Log2 Fold Change Dnmt3b1-KI/wt
n = 215
Gm715
Sox3
Efnb1
Nexmif
Mtmr1
Gpr50
Phf6
Cask
Gm5126
Tmem28
Zfx
Yipf6
Usp27x
Atp2b3
Cited1
Zfp711
Hmgb3
Fmr1
Dcaf12l2
Lonrf3
Pola1
Tbl1x
4933428M09Rik
Rbm3
Zbtb33
Pnck
Tab3
Rlim
Emd
Prkx
Log 2 Fold Change Dnmt3b1-KI/wt
H
15kb
TSS
-15kb
Distance to nearest TSS
Hyper-
Hypo-
methylated in Dnmt3b1-KI than wt

Figure 4.4: Complementation of the *Dnmt3b*$^{-/-}$ line with *Dnmt3b1* (shown on the previous page).

A) Western blot validation of *Dnmt3b1-KI* complementation line. H2, A6 and B1 show successfully targeted clones. A6 and H2 were further subcloned to obtain stable XX lines. The lower protein level of DNMT3B in wt could be due to technical issues in the western blot (such as poor transfer) as the protein levels in wt and *Dnmt3b-KI* lines were comparable (see panel D).

B) Expression levels of pluripotency genes (*Klf4, Nanog* and *Oct4*) in ES and EBs in wt and *Dnmt3b1* complementation lines as determined by chrRNA-seq. Since each differentiation experiment is an independent event which is prone to variability, the RNA levels of pluripotency genes cannot be expected to be identical between experiments but maintain a clear trend. One chrRNA experiment was performed for each *Dnmt3b1-KI* clone, giving two replicates for the *Dnmt3b1-KI* cell line.

C) MA plot comparing genome-wide expression between wt and *Dnmt3b1* complementation line at day 12 EB. The plot shows results for differential gene expression analysis (using DESeq2) of 2 replicates for each of wt and *Dnmt3b1-KI* cell lines.

D) Expression levels of *Dnmt3a, Dnmt3b* and *Dnmt3l* in ES and EB in wt and *Dnmt3b1* complementation lines determined by chrRNA-seq.

E) Average profile for CGIs on the X chromosome that gain methylation during EB differentiation (characterized in chapter3, section 3.6). Left panel shows the signal for Xa and Xi in wt cell line at day 12. Middle panel shows *Dnmt3b*$^{-/-}$ line and the right panel shows *Dnmt3b1-KI* complementation line. The MBD-seq profile is representative of three replicates *for Dnmt3b1-KI* line (H2D3 and two technical replicates for the clone H2G3). MBD-seq data from H2D3 and H2G3 show high correlation (see Appendix 2).

F) Average profile at DNMT3B-dependent loci (characterized in section 3.10) at day 12 EBs in wt, *Dnmt3b*$^{-/-}$ and *Dnmt3b1-KI* lines.

G) Volcano plot showing differentially methylated regions in *Dnmt3b1-KI* versus wt. The number of differentially methylated regions showing higher or lower methylation in *Dnmt3b1-KI* line as compared to wt are indicated next to the arrows. Significantly differentially methylated regions (defined as FDR < 0.01) are shown in pink. Right panel shows CGIs among the differentially methylated regions identified in the left panel. Significantly differentially methylated CGIs on the X chromosome are shown in red, and a few examples are labelled. Most of the significantly differentially methylated regions show hypermethylation in *Dnmt3b1-KI* line (indicated by log 2 fold change > 0). Data from three replicates was included in this analysis.

H) Plot showing distance of differentially methylated regions between wt and *Dnmt3b1-KI* lines to the nearest TSS. Hypomethylated regions in *Dnmt3b1-KI* line are farther away from TSS as compared to those found hypermethylated. Panels (F) and (H) show representative data from one of three highly reproducible replicates (two replicates for H2G3 and one for H2D3, see Appendix 2).

4.4 The catalytically inactive isoform DNMT3B3 is not required for CGI methylation on Xi

The previously described experiment does not rule out a role for catalytically inactive DNMT3B3 functioning together with DNMT3A in Xi and autosomal CGI methylation. To test the contribution of DNMT3B3 I generated complementation lines expressing *Dnmt3b3* from the TIGRE locus in iXist-ChrX *Dnmt3b*$^{-/-}$ cells. Figure 4.5 A shows western blot characterisation of the clones with successful insertion of *Dnmt3b3* (iXist-ChrX *Dnmt3b*$^{-/-}$ *TIGRE*$^{Tet\text{-}On/Dnmt3b3}$ line which is referred to as *Dmnt3b3-KI* line). For this experiment, I used B4 and F8 clones which showed either high or moderate levels of *Dnmt3b3* expression, respectively (Figure 4.5 A and D, left panel). Overall, these lines did not show any noticeable difference in morphology or growth rate compared to the wt iXist-ChrX lines when grown in culture. A parallel chrRNA-seq experiment was performed to see if the *Dnmt3b3-KI* cells differentiated well. ChrRNA-seq showed comparable expression of pluripotency genes which were downregulated upon differentiation as expected (Figure 4.5 B). Figure 4.5 D shows the expression levels of *Dnmt3a, Dnm3l* and *Dnmt3b (Dnmt3b3* transgene in case of *Dnmt3b3-KI* line) in differentiated wt and *Dnmt3b3-KI* lines. As expected *Dnmt3l* is downregulated upon differentiation in wt and *Dnmt3b3-KI* lines (Figure 4.5 D right panel). *Dnmt3a* had lower expression in the *Dnmt3b3-KI* line in day 12 EBs as compared to wt (Figure 4.5 D middle panel). Overall, only a small proportion of genes were differentially expressed in *Dnmt3b3-KI* EBs when compared with wt iXist-ChrX EBs (3% of total genes were differentially expressed, 263 up- and 316 down-regulated) (Figure 4.5 C). Interestingly, many among these were genes associated with DNMT3B-dependent autosomal CGIs that were described in section 3.10 (Figure 4.5 C, genes associated with DNMT3B-dependent autosomal CGIs are labelled). These however did not display a uniform trend in terms of expression change and could be seen to be both up and down-regulated in *Dnmt3b3-KI* complementation lines. Notably, these cell lines did not show any silencing defect on Xi, shown as the allelic ratio (Figure 4.5 E).

Analysis of MBD-seq data showed that CGI methylation on Xi in *Dnmt3b3-KI* lines was not rescued (Figure 4.5 G shows average profiles for CGIs on the X chromosome in wt, *Dnmt3b*$^{-/-}$, *Dnmt3b1-KI* and *Dnmt3b3-KI* lines). *Dnmt3b3-KI* lines have a profile very similar to *Dnmt3b*$^{-/-}$. Figure 4.5 F shows some representative DNMT3B-independent CGIs which are equally methylated in wt and *Dnmt3b3-KI* line indicating that the assay worked equally well among the tested cells lines and that the lack of CGI rescue on Xi is not due to less efficient MBD-seq. DiffBind analysis of the MBD-seq data from *Dnmt3b3-KI* line did not show any significant differentially methylated regions from the parental *Dnmt3b*$^{-/-}$ line (data not shown) despite the lower expression of *Dnmt3a* in *Dnmt3b3-KI* EBs (Figure 4.5 D, middle

panel). This analysis indicated that DNMT3B3 is completely dispensable for the establishment of CGI methylation on Xi and for other DNMT3B-dependent DNA methylation on autosomal loci.

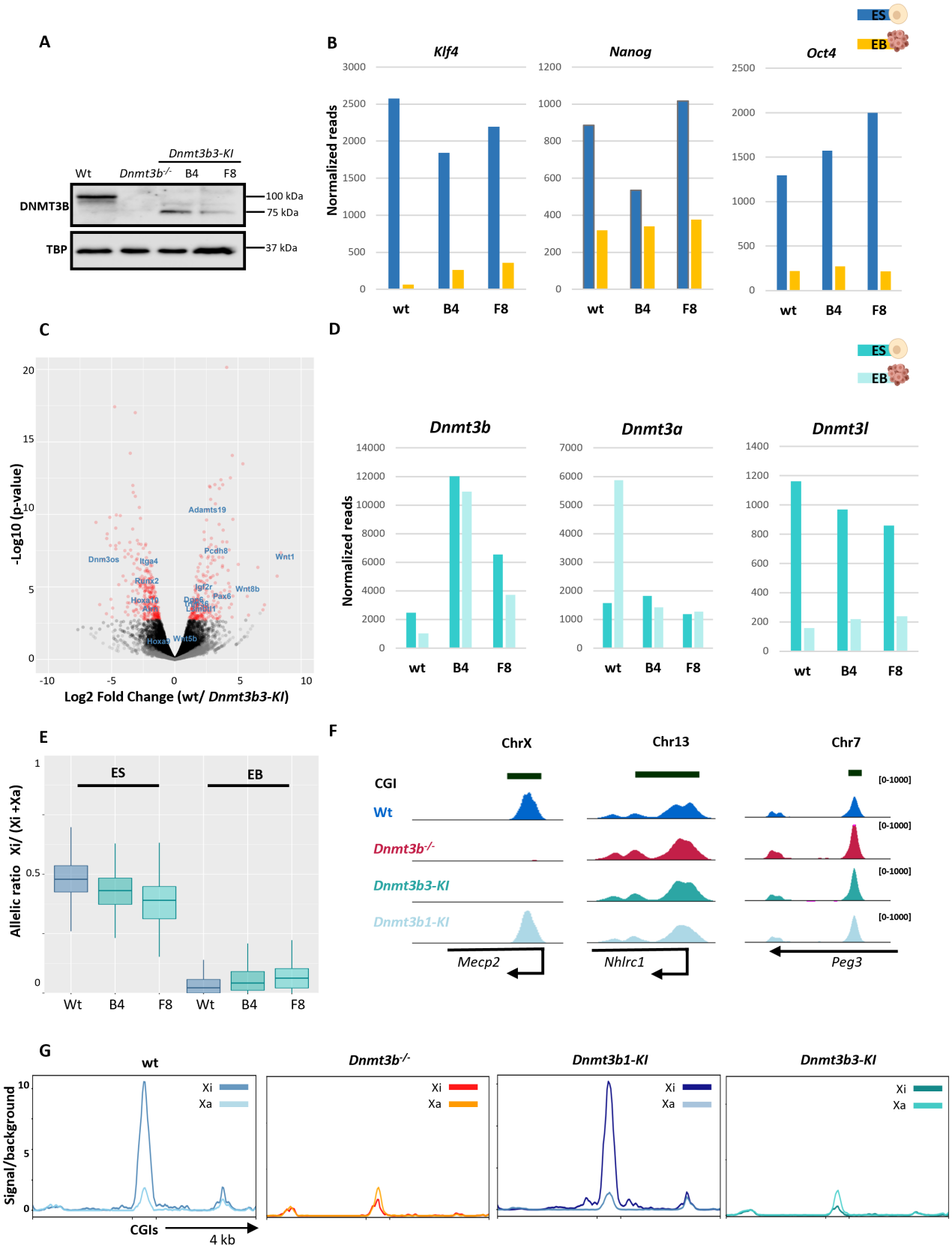


Figure 4.5: Complementation of *Dnmt3b*$^{-/-}$ line with *Dnmt3b3* (*Dnmt3b3-KI*) (shown on the previous page).

A) Western blot validation for successful targeting of *Dnmt3b3* in TIGRE locus. Clones B4 and F8 were chosen for the differentiation experiment. DNMT3B3 is 87.6 kDa.
B) Expression levels of *Klf4, Nanog* and *Oct4* in ES cells and day 12 EBs for wt and *Dnmt3b3-KI* line as determined by chrRNA-seq. One chrRNA experiment was performed for each *Dnmt3b3-KI* clone giving two replicates for the *Dnmt3b3-KI* cell line.
C) Differential expression analysis (using DESeq2, FDR < 0.05) between *Dnmt3b3-KI* and wt in day 12 EBs (2 replicates each). The significantly up- or down-regulated genes are shown in red.
D) Expression levels of *Dnmt3b* (or *Dnmt3b3* transgene in *Dnmt3b3-KI* line), *Dnmt3a,* and *Dnmt3l*, at day 0 and day 12 EBs in wt and *Dnmt3b3-KI* lines as determined by chrRNA-seq.
E) Boxplot showing allelic ratio at day 0 and day 12 in wt and *Dnmt3b3-KI* line. *Dnmt3b3* shows no defect in X chromosome silencing.
F) Genome browser tracks show an example of DNMT3B-dependant CGI and two DNMT3B-independent CGIs. All the cell lines show comparable enrichment on DNMT3B-independent regions suggesting MBD-seq pulldown and sequencing have worked equally well for the cell lines being compared.
G) Average profiles for Xi and Xa signal on CGIs on the X chromosome. *Dnmt3b3-KI* line shows no rescue of CGI methylation on Xi. Panels (F) and (G) show representative data from one of three highly reproducible replicates (two replicates for B4 and one for F8, see Appendix 2).

A recent publication has suggested accessory functions of DNMT3B (Nowialis et al., 2019). This study reported that mouse embryos expressing DMNT3B with two-point mutations in the catalytic domain (DNMT3B$^{P656V/C657D}$, referred to as catalytically inactive DNMT3B or DNMT3B^{CI}) largely restores methylation of CpG sites hypomethylated in *Dnmt3b* knockout embryos. I tested if I could see complementation of CpG methylation seen in this study in the *Dnmt3b3-KI* line. The rescued CpGs shown by Nowialis et al., were characterized based on RRBS which has a larger genomic coverage of CpGs in comparison to MBD-seq. Of 20703 CpGs shown in their data, most of which were hypomethylated in *Dnmt3b*$^{-/-}$ embryos, I was only able to capture 4508 CpGs in MBD-seq (Figure 4.6 A). Contrary to Nowialis et al., most of these regions were methylated in my *Dnmt3b*$^{-/-}$ line. In agreement with the genome-wide differential methylation analysis (which revealed no differentially methylated regions between *Dnmt3b*$^{-/-}$ and *Dnmt3b3-KI* line), I was unable to see any rescue of DNA methylation at these CpG-rich sites in *Dnmt3b3-KI* line (Figure 4.6 A and B bottom panel). It is important to mention that although both DNMT3B3 and DNMT3B^{CI} are catalytically inactive, these are different proteins and have distinct structures. Furthermore, the evidence for the lack of catalytic activity of DNMT3B^{CI} mutant presented in the paper is limited and the authors note that they could not rule out some residual catalytic activity in their mutants.

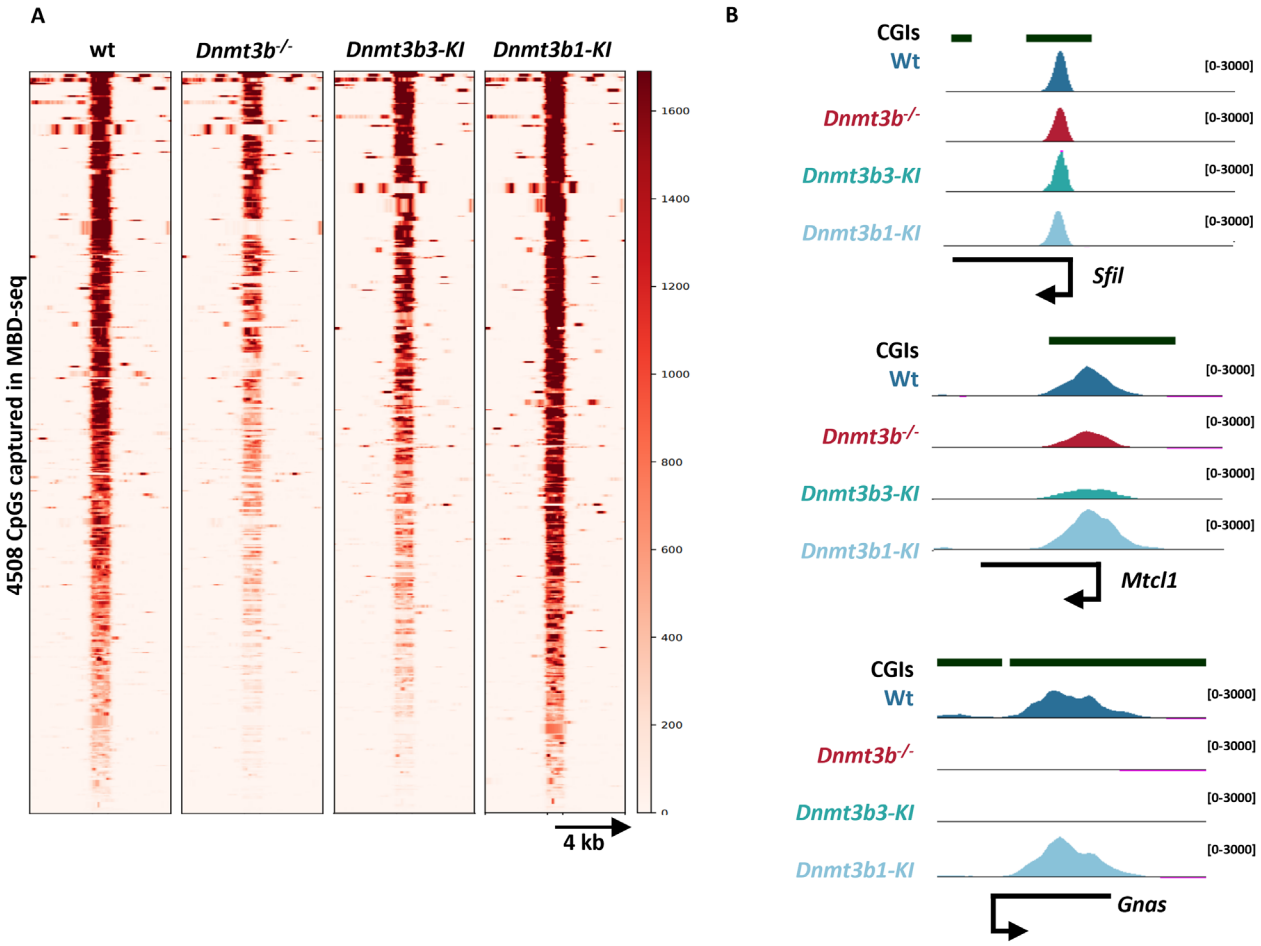


Figure 4.6: Testing "accessory" functions of DNMT3B in *Dnmt3b3-KI* complementation line.

A) Heatmap plotted on regions claimed to have been rescued/methylated in mouse embryos expressing DNMT3B^{Cl} (Nowialis et al., 2019) that were captured using MBD-seq. More than half of these CpGs were methylated in my *Dnmt3b*$^{-/-}$ and complementation lines in day 12 EBs. The regions found towards the bottom of the heatmap were only methylated in wt and *Dnmt3b1-KI* complementation line. I did not see any rescue of these regions upon complementation with *Dnmt3b3*.

B) Genome browser tracks showing one CGI from the top, middle, and bottom third of the heatmap shown in (A) in the wt, *Dnmt3b*$^{-/-}$, *Dnmt3b1-KI* and *Dnmt3b3-KI* complementation lines. As indicated in (A), the regions that were hypomethylated in *Dnmt3b*$^{-/-}$ line were not rescued in *Dnmt3b3-KI* line suggesting that DNMT3B3 does not have any accessory function that can rescue DNA methylation on these regions. Panels (A) and (B) show representative data from one of five highly reproducible replicates for wt and *Dnmt3b*$^{-/-}$ (Figure 3.13 B) and three replicates for *Dnmt3b1-KI* and *Dnmt3b3-KI* cells (see Appendix 2).

4.5 DNMT3B1 N-terminus is required to complement Xi CGI methylation in *Dnmt3b*$^{-/-}$ cells

The results discussed in section 4.4 indicate that the non-catalytic isoform of DNMT3B is not relevant in the context of Xi CGI methylation, therefore providing support for model 2, i.e., that there are unique features specific to the catalytic DNMT3B1 enzyme that confer specificity for this function. As discussed above, the linker regions between PWWP and ADD domain (138 amino acids) and the N-terminus (229 amino acids) are the least conserved regions between the two homologs (Figure 4.2 A). In this section, I investigate the role of the DNMT3B1 N-terminal region.

The N-terminus of DNMT3A and DNMT3B show no sequence homology (Figure 4.2 A and C). The PWWP domain, on the contrary, is highly conserved between DNMT3A and DNMT3B, with the sequence of the first 18 amino acids being identical between the two. The 10 amino acids preceding the PWWP domain also show high conservation (86.27%), and since the N-terminus of the PWWP domain forms a beta strand, these preceding residues may form a stabilizing loop (Figure 4.2 C residues 291-301). With these considerations in mind, I generated two N-terminus deletions in *Dnmt3b1*, Δ1-229, and Δ1-218 *Dnmt3b1*. As for the *Dnmt3b1-KI* complementation line, I targeted these constructs to the TIGRE locus to generate iXist-ChrX *Dnmt3b*$^{-/-}$ *TIGRE* $^{Tet\text{-}On/\Delta 1\text{-}229\ Dnmt3b1}$ and *TIGRE* $^{Tet\text{-}On/\Delta 1\text{-}218\ Dnmt3b1}$ complementation lines and differentiated them for 12 days to study the role of N-terminus domain of DNMT3B1 in CGI methylation on Xi. Similar to the complementation lines described before, these lines are denoted as *Δ1-229 Dnmt3b1-KI* and *Δ1-218 Dnmt3b1-KI* lines for simplicity.

DNMT3B1 is stably expressed upon the truncation of its N-terminal domain as evident from the successful targeting of *Δ1-229 Dnmt3b1* in iXist-ChrX line (Figure 4.7 A). For *Δ1-229 Dnmt3b1-KI*, I chose 2 clones with different expression levels of Δ1-229 DNMT3B1 to see if expression levels contribute to any difference in CGI methylation (Figure 4.7 A, chosen clones are indicated with asterisks, the two clones with one technical replicate for B4 in total make 3 replicates for this experiment). As with the experiments mentioned in sections 4.3 and 4.4, I performed chrRNA-seq to gauge any difference in gene expression between wt and *Δ1-229 Dnmt3b1-KI* line and to estimate the

proper differentiation of these lines. Figure 4.7 B shows the expression of pluripotency genes and Figure 4.7 D shows the expression of *Dnmt3a, Dnmt3b* (or the complementation construct *Δ1-229 Dnmt3b1*) and *Dnmt3l*. One clone, E11, behaved differently, with higher expression of *Nanog* and *Dnmt3l* at day 12 suggesting it did not differentiate as efficiently as B4. E11 also showed much higher expression of the Δ1-229 DNMT3B1 construct and incomplete X chromosome silencing in comparison to B4 (as evident from the higher allelic ratio for expression of X-linked genes shown in Figure 4.7 E). Broadly, however, *Δ1-229 Dnmt3b1-KI* lines had negligible differential gene expression when compared to wt iXist-ChrX line (Figure 4.7 C).

Intriguingly, the *Δ1-229 Dnmt3b1-KI* line showed a complete lack of CGI methylation rescue observed in full-length *Dnmt3b1* complementation lines (Figure 4.7 H). The *Δ1-229 Dnmt3b1-KI* line was also unable to rescue methylation at DNMT3B-dependent autosomal CGIs defined in section 3.10 (Figure 4.7 I). A Diffbind comparison between *Δ1-229 Dnmt3b1-KI* and *Dnmt3b*$^{-/-}$ revealed only a small number of differentially methylated regions genome-wide (118/30129 = 3.9% to total MBD-seq peaks); implying that *Δ1-229 Dnmt3b1-KI* largely behaved like *Dnmt3b*$^{-/-}$ line in EBs. The small number of differentially methylated regions between *Dnmt3b*$^{-/-}$ and *Δ1-229 Dnmt3b1-KI* line is expected among independent differentiation experiments (Figure 4.6 F and G).

The lack of CGI methylation rescue in *Δ1-229 Dnmt3b1-KI* could indicate partial/complete loss of catalytic activity due to large-scale conformational changes in DNMT3B1 protein upon the deletion of the N-terminal domain. I, therefore, tested a smaller deletion of amino acids 1 to 218. This deletion does not truncate the conserved residues before the start of PWWP domain which could be of importance for the proper functioning/folding of the protein. Furthermore, this deletion has been shown in previous publications to retain catalytic activity as measured *in vitro* (Qiu et al., 2002; Zeng et al., 2020) and to respond to the presence of DNMT3B3 by the enhancement of the catalysis (Zeng et al., 2020).

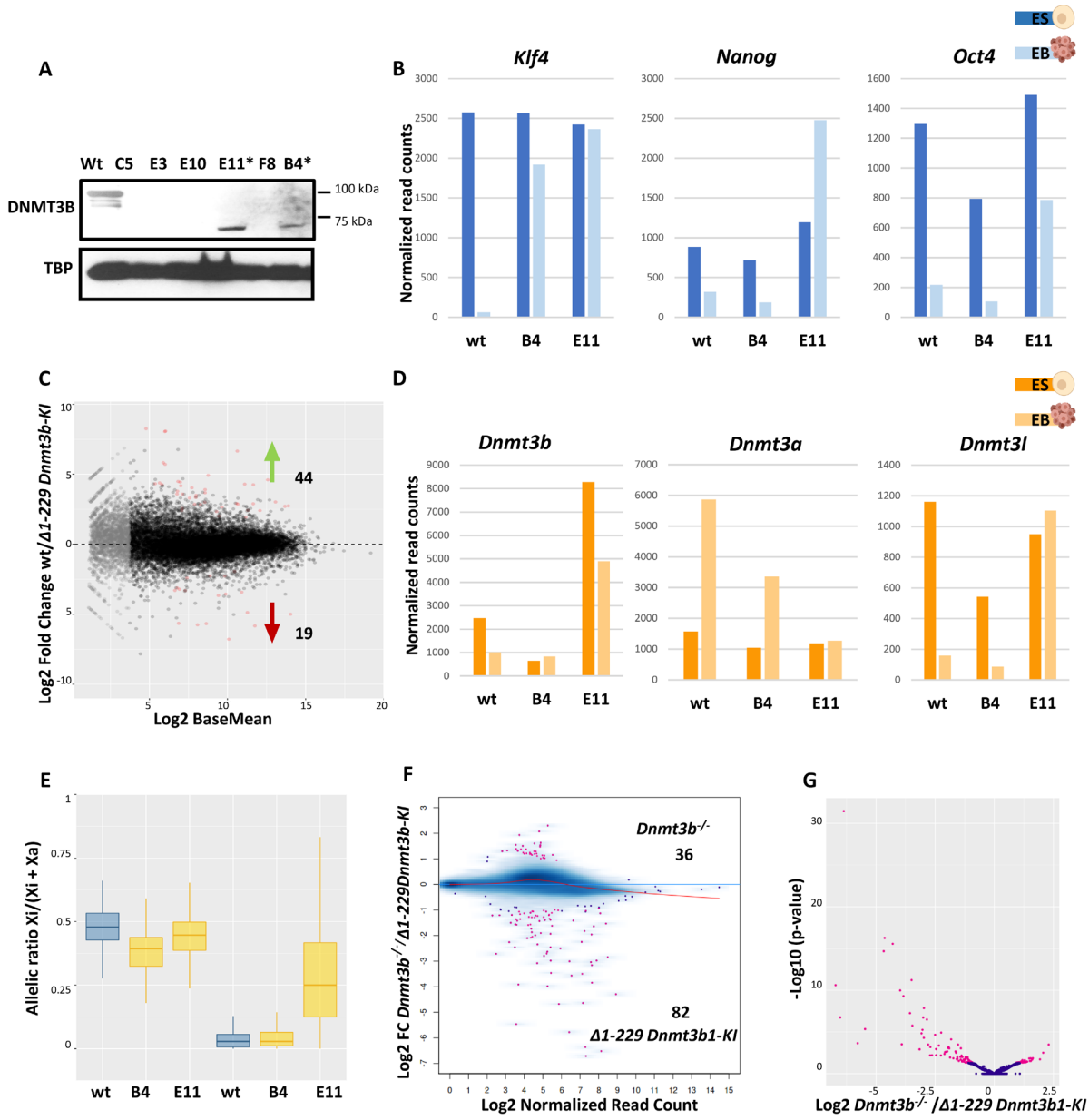

A
Wt C5 E3 E10 E11* F8 B4*
DNMT3B
TBP
100 kDa
75 kDa
B
Klf4
Nanog
Oct4
ES
EB
Normalized read counts
wt B4 E11
C
Log2 Fold Change wt/Δ1-229 Dnmt3b-KI
Log2 BaseMean
44
19
D
Dnmt3b
Dnmt3a
Dnmt3l
E
Allelic ratio Xi/(Xi + Xa)
wt B4 E11 wt B4 E11
F
Dnmt3b-/-
36
82
Δ1-229 Dnmt3b1-KI
Log2 FC Dnmt3b-/-/Δ1-229Dnmt3b-KI
Log2 Normalized Read Count
G
-Log10 (p-value)
Log2 Dnmt3b-/- /Δ1-229 Dnmt3b1-KI
H
wt
Dnmt3b-/-
Dnmt3b1-KI
Δ1-229 Dnmt3b1-KI
Xi
Xa
Signal/background
CGIs
4 kb
I
DNMT3B-dependent peaks
DNMT3B-independent peaks
Signal/background
4 kb
wt
Dnmt3b-/-
Dnmt3b3-KI
Δ1-229 Dnmt3b1-KI
Dnmt3b1-KI

Figure 4.7: Complementation of *Dnmt3b*$^{-/-}$ line with *Δ1-229 Dnmt3b1* (shown on the previous page).

A) Western blot indicating the clones with successful targeting of *Δ1-229 Dnmt3b1*. B4 and E11 were chosen for further experiments. The expected molecular weight of Δ1-229 DNMT3B1 is 71 kDa.
B) Expression levels of pluripotency genes at day 0 ES and day 12 EBs in wt and *Δ1-229 Dnmt3b1-KI* lines (determined by chrRNA-seq).
C) MA plot comparing chrRNA-seq data of *Δ1-229 Dnmt3b1-KI* and wt line (2 replicates each). Significantly up- and downregulated genes (FDR < 0.05) are shown in red, and their numbers are indicated. This analysis took into account data from both E11 and B4 clones. Overall, very few genes showed differential expression.
D) Expression levels of *Dnmt3a, Dnmt3b* and *Dnmt3l*. E11 shows a 4-fold higher expression of *Δ1-229 Dnmt3b1* rescue construct as compared to *Dnmt3b* levels in wt. E11 can be seen to show a lower expression of *Dnmt3a* and a higher expression of *Dnmt3l* in comparison to B4.
E) Boxplots showing allelic ratio for X-linked gene expression. The allelic ratio of B4 shows complete silencing with a value comparable to wt. E11 shows incomplete silencing of X-linked genes consistent with inefficient differentiation.
F) MA plot showing the comparison of MBD-seq signal between *Dnmt3b*$^{-/-}$ and *Δ1-229 Dnmt3b1-KI* lines. The significantly differential regions (FDR < 0.01) are shown in pink and only make 3.9% of total regions in the analysis. The numbers on top and bottom represent the regions hypo or hypermethylated in *Δ1-229 Dnmt3b1-KI* line as compared to the *Dnmt3b*$^{-/-}$ line. This analysis was performed using 5 MBD-seq replicates for *Dnmt3b*$^{-/-}$ and 3 replicates for *Δ1-229 Dnmt3b1-KI* lines.
G) Volcano plot for the MBD-seq signal for *Dnmt3b*$^{-/-}$ and *Δ1-229 Dnmt3b1-KI* lines. Volcano plot shows that among the differentially methylated regions (shown in F), an even smaller fraction is statistically significant (upper left quadrant of the graph, points shown in pink represent significantly differential regions i.e., FDR < 0.01). Thus *Δ1-229 Dnmt3b1-KI* lines largely resemble *Dnmt3b*$^{-/-}$ in terms of CpG methylation profile.
H) Average profiles for methylation of CGIs on the X chromosome. Wt and *Dnmt3b1-KI* are also shown for comparison. Deletion of the N-terminal domain leads to complete loss of rescue observed upon complementation with full-length *Dnmt3b1*.
I) Average profiles for methylation on DNMT3B-dependent and -independent regions for the complementation lines tested thus far. *Δ1-229 Dnmt3b1-KI* line did not rescue methylation of DNMT3B-dependent autosomal loci. Panels (H) and (I) show representative data from one of three highly reproducible replicates (two replicates for B4 and one for E11, see Appendix 2).

I generated complementation lines expressing *Δ1-218 Dnmt3b1-KI* in *Dnmt3b*$^{-/-}$ cell. I used two clones (D9 and G6, with two technical replicates for D9) successfully targeted with *Δ1-218 Dnmt3b1* (Figure 4.8 A). Similar to *Δ1-229 Dnmt3b1-KI*, *Δ1-218 Dnmt3b1-KI* did not rescue CGI methylation on Xi to any degree (Figure 4.8 B shows the average profile for MBD signal on CGIs on the X chromosome). Since the deletion of 1-218 amino acids of DNMT3B have been previously shown to be dispensable for the catalytic activity of DNMT3B *in vitro*, this data indicates that the N-terminal domain must be crucial for its recruitment to target loci. It is however also important to consider that the conventional *in vitro* catalytic assays frequently use naked DNA as substrate instead of chromatin, the physiological

substrate of *de novo* DNMT enzymes, and therefore do not fully predict the catalytic ability of DNMTs *in vivo*.

Delta variants of DNMT3B are known to be expressed in non-small lung cancer. These delta variants of DNMT3B lack 200 amino acids from the N-terminus due to the initiation of transcription from an alternative promoter between intron 4 and exon 5. The delta variants of DNMT3B have been reported to cause aberrant DNA methylation potentially due to unspecific targeting and interference with the catalysis of active DNMT3B and DNMT3A (Wang et al., 2007; Wang et al., 2006a; Wang et al., 2006b). I, therefore, tested if deletion of the N-terminal domain led to ectopic- or hypermethylation elsewhere in the genome. I performed a differential methylation analysis between the *Dnmt3b*$^{-/-}$ and *Δ1-218 Dnmt3b1-KI* line to find regions methylated by the complementation construct. As before, this analysis was performed using the DiffBind package which takes into account all the MBD-seq signal (CGIs and other CpG-rich regions captured) and finds the regions that significantly differ in their methylation status between the two tested cell lines.

In EBs, the DNA methylation profile of *Δ1-218 Dnmt3b1-KI* largely resembled that of *Dnmt3b*$^{-/-}$ with only 0.85% of total MBD-seq peaks (665/78189) differentially methylated (Figure 4.8 D). Intriguingly, in ES cells, 2393 MBD-seq peaks were found to be significantly differentially methylated between *Dnmt3b*$^{-/-}$ and *Δ1-218 Dnmt3b1*-KI line (Figure 4.8 C, the statistically significant differential MBD-seq peaks i.e., FDR < 0.01, are shown in pink, total peaks included in the analysis are 35985). These 2393 differential peaks showed a clear trend where 2378 were hypermethylated in *Δ1-218 Dnmt3b1-KI* line in comparison to *Dnmt3b*$^{-/-}$ (Figure 4.8 C). Most of these peaks resided within genes or intergenic regions (Figure 4.8 E) and had lower GC content in comparison to CGIs (Figure 4.8 H). The heatmap in Figure 4.8 F compares the MBD-seq signal at regions found to be hypermethylated in *Δ1-218 Dnmt3b1-KI* in ES cells and EBs. Upon differentiation, the majority of these differential peaks gain methylation in the *Dnmt3b*$^{-/-}$ line such that in EBs only 505 (of 2378 hypermethylated regions in ES cells) are hypermethylated in *Δ1-218 Dnmt3b1-KI* as compared to *Dnmt3b*$^{-/-}$ line (Figure 4.8 D and F), of which only 6 were on the X chromosome. Figure 4.8 G shows examples of CGIs among the regions hypermethylated in *Δ1-218 Dnmt3b1-KI* ES cells. These regions are also methylated in wt and *Dnmt3b1-KI* line. Thus, in ES cells *Δ1-218 Dnmt3b1* behaves like a hypermorphic DNMT3B affecting exons and intergenic regions. The absence of methylation at DNMT3B target CGIs in *Δ1-218 Dnmt3b1-KI* EBs is therefore intriguing and implies a failure of Δ1-218 DNMT3B1 to accurately localize in differentiated cells (investigated further later). The significance of CpG methylation at the intra and intergenic regions found hypermethylated in *Δ1-218 Dnmt3b1-KI* line is not clear since the absence of DNMT3B does not lead to gene expression changes. ChrRNA-seq was not performed in *Δ1-218 Dnmt3b1-KI* cells in ES cells or EBs but complementation with *Dnmt3b1* or *Δ1-229 Dnmt3b1-KI*, for

instance, did not show any significant changes in gene expression either. To conclude, both N-terminal deletions (henceforth denoted together as a combined term, *ΔN-Dnmt3b1* complementation lines) fail to rescue methylation of DNMT3B-dependent regions on the X chromosome as well as autosomes. Given the evidence for intact catalytic activity of Δ1-218 DNMT3B1 protein, the findings from this experiment indicate the failure of Δ1-218 DNMT3B1 to localize to the DNMT3B target loci.

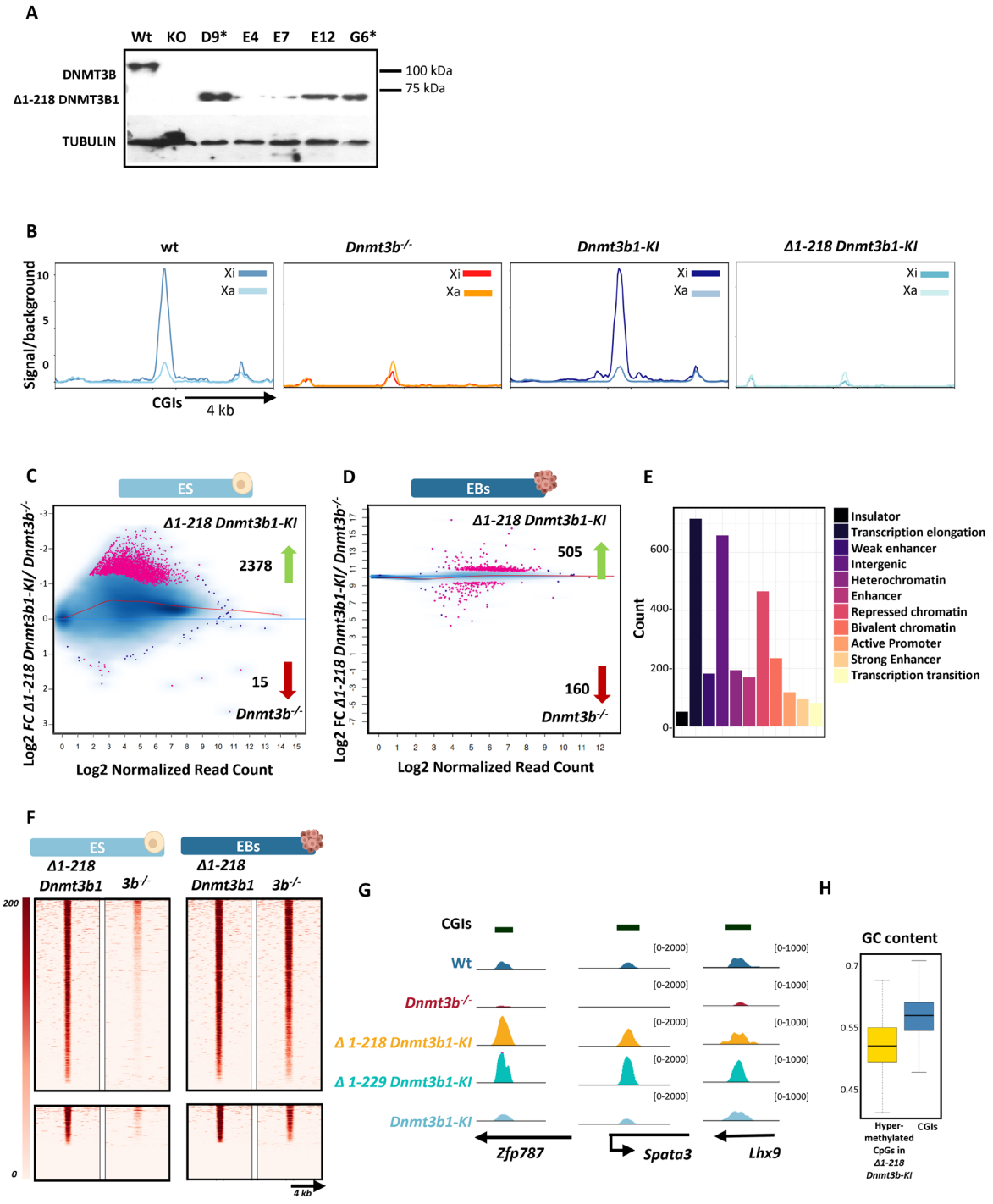


Figure 4.8: Complementation of *Dnmt3b*$^{-/-}$ line with *Δ1-218 Dnmt3b1* (*Δ1-218 Dnmt3b1-KI*), (shown on the previous page).

A) Western blot indicating the clones targeted with *Δ1-218 Dnmt3b1*. D9 and G6 were chosen for further experiments. Δ1-218 DNMT3B1 is 73 kDa in size.
B) Average profiles showing methylation levels of CGIs on the X chromosome. Wt and *Dnmt3b1-KI* are also shown for comparison. Deletion of the N-terminal domain (*Δ1-218 Dnmt3b1-KI*) leads to complete loss of rescue observed upon complementation with full length *Dnmt3b1*. The plot shows representative data from one of three well correlated replicates (two replicates for D9 and one for G6, see Appendix 2).
C) MA plot comparing MBD-seq signal of *Δ1-218 Dnmt3b1-KI* and *Dnmt3b*$^{-/-}$ lines in ES cells. A much larger number of differentially methylated regions were observed in ES cells in comparison to EBs (6.5% of total evaluated regions in comparison to 0.85% in EBs). Most of the differentially methylated regions were hypermethylated in *Δ1-218 Dnmt3b1-KI*. This analysis was performed using the data from three MBD-seq replicates. Significantly up and downregulated genes (FR < 0.01) are shown in pink.
D) MA plot comparing MBD-seq signal of *Δ1-218 Dnmt3b1-KI* and *Dnmt3b*$^{-/-}$ lines in day 12 EBs. Significantly up and downregulated genes (FR < 0.01) are shown in pink. 0.85% of the regions show differential methylation (see text for details). Panels (C) and (D) consider data from 5 replicates for *Dnmt3b*$^{-/-}$ lines and 3 replicates for *Δ1-218 Dnmt3b1-KI* lines.
E) Distribution of hypermethylated regions found in *Δ1-218 Dnmt3b1-KI* among various genomic features, as determined by chromHMM. Most of these regions reside within gene bodies, intergenic regions, or repressed chromatin.
F) Heatmap showing the regions found hypermethylated in *Δ1-218 Dnmt3b1-KI* ES cells (top) and EBs (bottom) in DiffBind analysis in comparison to *Dnmt3b*$^{-/-}$ line. The MBD-seq signal at these differential regions is plotted in ES (left) and EBs (right) for *Δ1-218 Dnmt3b1-KI* and *Dnmt3b*$^{-/-}$ lines. Most of these regions gain methylation in *Dnmt3b*$^{-/-}$ upon differentiation (as evident from the smaller number of differentially methylated regions in EBs in comparison to ES cells)
G) Genome browser tracks showing examples of hypermethylated regions in *Δ1-218 Dnmt3b1-KI* ES cells. In the examples, MBD-seq peaks associated with *Zfp787* and *Lhx9* are intragenic while that of *Spata3* is on the gene promoter. These are methylated on wt cells in contrast to *Dnmt3b*$^{-/-}$, indicating that Δ1-218 DNMT3B1 behaves like full-length DNMT3B1 in ES cells.
H) *Δ1-218 Dnmt3b1-KI* hypermethylated regions on average have low GC content that CGIs as indicated in the boxplot. *Δ1-218 Dnmt3b1-KI* line did not rescue methylation of DNMT3B-dependent autosomal loci. Panels (F)-(H) show representative data from one of three well correlated replicates (Appendix 2).

4.6 Further analysis of the N-terminus of DNMT3B1 to determine the regions required for CGI methylation rescue

In section 4.1.2 I discussed that protein structure predictions indicate that the N-terminus of DNMT3B forms an unstructured flexible loop (Figure 4.2 C). However, it is possible that this region becomes structured by PTMs, in the vicinity of appropriate interacting partners or chromatin modifications

(Figure 4.2 D). Based on this assumption, if the N-terminus of DNMT3B1 does indeed determine recruitment of DNMT3B1 to the X chromosome, then some regions of the N-terminus that, for example, form binding interfaces with protein interactors or chromatin, would be essential while those away from these binding interfaces would have a minimal effect on the recruitment of DNMT3B1. It is also possible that the N-terminus of DNMT3B1 has a functional subdomain that ensures proper localization of DNMT3B1 to its target loci as proposed in the case of DNMT3A (Gu et al., 2022; Weinberg et al., 2021). Therefore, to gain further insight into the function of the N-terminal domain of DNMT3B1, and search for the critical regions/functional subdomains within it, I complemented the *Dnmt3b*$^{-/-}$ line with *Dnmt3b1* bearing smaller deletions in the N-terminus. To determine the regions of the N-terminus to delete, I divided the region into 5 roughly equal parts while avoiding cutting any predicted alpha helices or beta strands (shown by letters A to E in Figure 4.9 A). Since the loss of a large fragment of the N-terminal domain (1-218 aa) has no impact on the catalytic activity of DNMT3B1, it can be speculated that these smaller deletions would also not compromise catalysis.

Figure 4.9 B shows the clones targeted with *Dnmt3b1* with smaller N-terminal deletions. Two clones for each deletion were selected for further experiments. I also tested if these complementation constructs of *Dnmt3b1* were stably expressed throughout the differentiation timecourse. Figure 4.9 C shows the expression of these constructs in 12-day differentiated EBs. Four of the N-terminal deletions did not have any impact on the stability of DNMT3B1 and were expressed throughout differentiation (Figure 4.9 C). ΔC however rendered DNMT3B1 unstable as it can be seen to be expressed (qPCR shown in Figure 4.9 D) but with very low corresponding protein levels in EBs (as shown in Figure 4.9 C). I also ensured efficient differentiation of these cell lines by testing the expression levels of a pluripotency gene *Oct4* by qPCR in day 12 EBs (Figure 4.9 E). All the cell lines showed an equal or lower level of *Oct4* expression in comparison to wt in EBs, indicating proper differentiation.

Interestingly, all the smaller deletions of the N-terminus showed the same outcome as that of full N-terminus deletion. Figure 4.9 F shows genome browser tracks for DNMT3B-dependent (*Hprt, Dppa3, Igf2r*) and independent loci (*Peg3*). Even though all these lines had most of the N-terminal domain intact, the small deletions failed to rescue CGI methylation on Xi or DNMT3B-dependent regions on autosomes (heatmaps of 4.9 G). These results suggest that the N-terminal domain is not modular and that all regions are required for its faithful recruitment to DNMT3B1 target loci.

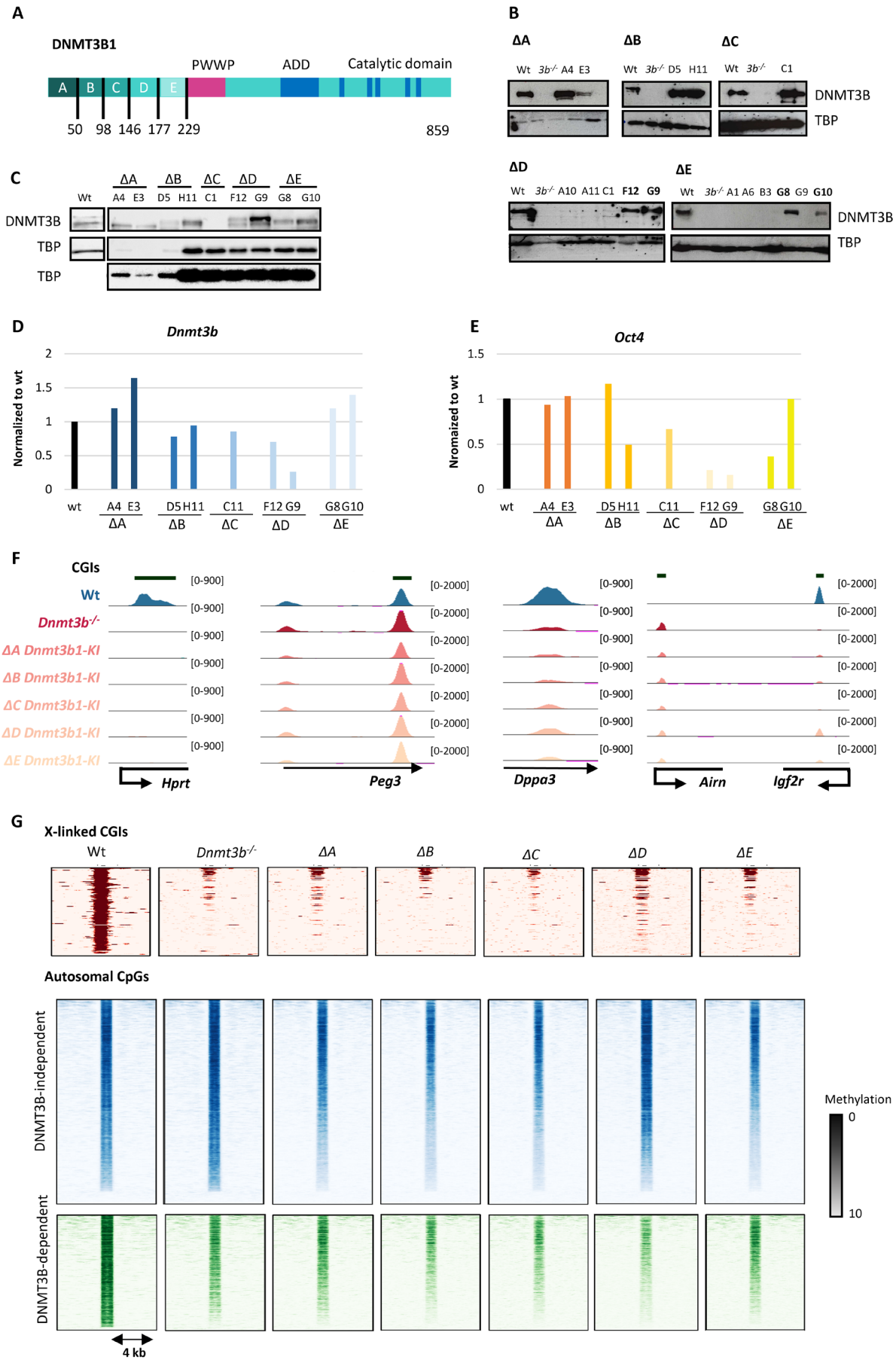

A
DNMT3B1
PWWP
ADD
Catalytic domain
A B C D E
50 98 146 177 229
859
B
ΔA
ΔB
ΔC
Wt 3b-/- A4 E3
Wt 3b-/- D5 H11
Wt 3b-/- C1
DNMT3B
TBP
ΔD
ΔE
Wt 3b-/- A10 A11 C1 F12 G9
Wt 3b-/- A1 A6 B3 G8 G9 G10
C
ΔA ΔB ΔC ΔD ΔE
Wt A4 E3 D5 H11 C1 F12 G9 G8 G10
DNMT3B
TBP
TBP
D
Dnmt3b
Normalized to wt
2
1.5
1
0.5
0
wt A4 E3 D5 H11 C11 F12 G9 G8 G10
ΔA ΔB ΔC ΔD ΔE
E
Oct4
Nromaized to wt
1.5
1
0.5
0
wt A4 E3 D5 H11 C11 F12 G9 G8 G10
ΔA ΔB ΔC ΔD ΔE
F
CGIs
Wt
Dnmt3b-/-
ΔA Dnmt3b1-KI
ΔB Dnmt3b1-KI
ΔC Dnmt3b1-KI
ΔD Dnmt3b1-KI
ΔE Dnmt3b1-KI
[0-900]
[0-2000]
Hprt
Peg3
Dppa3
Airn
Igf2r
G
X-linked CGIs
Wt
Dnmt3b-/-
ΔA
ΔB
ΔC
ΔD
ΔE
Autosomal CpGs
DNMT3B-independent
DNMT3B-dependent
4 kb
Methylation
0
10

Figure 4.9: Complementation of *Dnmt3b*$^{-/-}$ line with *Dnmt3b1* bearing small deletions in the N-terminus (shown on the previous page).

A) Schematic of DNMT3B1, highlighting the fragments of the N-terminal domain deleted in *Dnmt3b1* rescue constructs (labelled A-E) for the generation of complementation lines.
B) Western blot showing the targeting of *Dnmt3b1* rescue constructs with smaller deletions in the N-terminal domain as indicated in (A).
C) The expression of DNMT3B1 rescue constructs in differentiated EBs. The lower TBP lane of the western blot illustrates the unequal/lower loading in the first three lanes. The expression levels of these lines are expected to be greater than that shown in the first lane.
D) Expression levels of *Dnmt3b* in the complementation cell lines as compared to wt in day 12 EBs as determined by qPCR.
E) Expression levels of *Oct4* in wt and complementation lines determined by qPCR in 12 day differentiated EBs to estimate proper differentiation of the complementation lines in comparison to wt.
F) Genome browser tracks showing examples of DNMT3B-dependent and DNMT3B-independent CGIs. The complementation lines ΔA-E do not rescue DNMT3B-dependent CGIs.
G) Heat map of CGIs on the X chromosome and DNMT3B-dependent and independent regions on autosomes in the complementation lines show that the MBD-seq signal is similar to *Dnmt3b*$^{-/-}$. Panels (F) and (G) show representative data from one of three well correlated replicates (Appendix 2).

4.7 CGI methylation in complementation lines expressing DNMT3A with N-terminal domain of DNMT3B1

The failure of DNMT3B1 with N-terminal domain truncation to rescue CGI methylation indicates the importance of the N-terminal domain of DNMT3B in determining its recruitment to the target loci. This observation agrees with the earlier reports showing the crucial role of the N-terminus of DNMT3A in determining its localization on chromatin (Gu et al., 2022; Weinberg et al., 2021). Small tiling deletions along the N-terminal domain of DNMT3B1 revealed that DNMT3B1 requires its N-terminal domain to be intact in order to rescue CGI methylation. Working on the assumption that DNMT3A and DNMT3B are closely related enzymes with a variable N-terminal domain that governs their specific recruitment, switching the N-terminal domains could lead to DNMT3 proteins being recruited to one another's target loci. Thus, replacing the N-terminal domain of DNMT3A with that of DNMT3B could allow DNMT3A to be recruited to CGIs on the X chromosome, and thereby rescue CGI methylation (Figure 4.10 A1). Weinberg and colleagues performed a similar experiment and showed that replacing the N-terminal domain of DNMT3B with that of DNMT3A, did indeed cause DNMT3B to be recruited to DNMT3A target loci (Weinberg et al., 2021). I therefore complemented *Dnmt3b*$^{-/-}$ cells with a chimeric DNMT3A with N-terminal domain of DNMT3B (*Dnmt3b(N)3a-KI*). For this experiment I

exchanged the amino acids 1-66 of DNMT3A2 (N-terminal to PWWP domain), - the predominant isoform of DNMT3A expressed in ES cells -, with amino acids 1-229 of DNMT3B (Figure 4.10 B shows the lines targeted with *Dnmt3b(N)3a-KI*). Since the first few residues of PWWP domains of DNMT3A and DNMT3B are the same, this "graft" of DNMT3B is expected to merge with DNMT3A with minimal structural interference.

I also generated a complementation line expressing the reciprocal *Dnmt3a(N)3b* construct, i.e., DNMT3B in which the N-terminal domain had been exchanged with that of DNMT3A. The rationale behind this experiment was that, unlike DNMT3A, prior studies have not examined the role of DNMT3B N-terminal domain in context-specific recruitment. A popularly cited paper from Jeong et al., (mentioned in section 4.1.2) reported that without its N-terminal domain, DNMT3B forms a loose association with chromatin, and dissociates more readily than full-length DNMT3B in a sucrose gradient (Jeong et al., 2009). This loss of chromatin association would also explain why ΔN-DNMT3B1 rescue construct failed to rescue CGI methylation (on X chromosome and autosomes), despite having an intact catalytic activity tested *in vitro* using naked DNA as a template. Thus, another model to consider would be that the N-terminal domains of DNMT3A and DNMT3B form a general chromatin binding 'anchor' or 'hook' and are hence indispensable for functionality *in vivo*. The N-terminal domains of DNMT3A and DNMT3B could thus be interchangeable between the two DNMT3 proteins (Figure 4.10 A, right). An earlier study by Suetake et al., supports the hypothesis that the N-terminal domain of DNMT3A functions as a general DNA/chromatin binding domain by showing that the basic residues in DNMT3A N-terminal domain bind DNA (Suetake et al., 2011). Considering this rationale, I created *Dnmt3a(N)3b-KI* line. Based on the *Dnmt3b(N)3a* construct, the N-terminus of DNMT3B1 until just before the PWWP domain (1-229 aa), was replaced by the N-terminus of DNMT3A2 (1-66 aa). Fusion of the N-terminal domain of DNMT3A2 should theoretically rectify the loss of chromatin binding in ΔN-DNMT3B1 protein. Figure 4.10 C shows the *Dnmt3b*$^{-/-}$ lines targeted with *Dnmt3a(N)3b* (denoted as *Dnmt3a(N)3b-KI* line).

The MBD-seq data from day 12 EBs in both the complementation lines showed no CGI methylation on the X chromosome (average profiles of the X linked-CGIs is shown in Figure 4.10 E). Only the DNMT3B-dependent regions were found hypomethylated in these cell lines (Figure 4.10 D shows DNMT3B-dependent X-linked CGI *Hprt* and DNMT3B-independent autosomal CGIs, *Nhlrc1* and *Peg3*). Thus, the lack of CGI methylation on Xi was not due to lower MBD-seq pulldown efficiency.

The lack of CGI methylation rescue in the *Dnmt3b(N)3a-KI* complementation line suggests that the N-terminal domain of DNMT3B1 is essential for its faithful localization, but not sufficient to recruit DNMT3A to DNTM3B target sites. On the other hand, the fusion of the N-terminal domain of

DNMT3A2 with DNMT3B1 (*Dnmt3a(N)3b* construct), which should theoretically rectify the loss of chromatin association observed in *ΔN-Dnmt3b1-KI* lines also failed to rescue CGI methylation on Xi. This indicates that loss of CGI methylation observed in *ΔN-Dnmt3b1-KI* is not due to dissociation of DNMT3B1 from chromatin. Importantly, although both the chimeric DNMT3 constructs were stably expressed and could be found bound to chromatin (Figure 4.10 B and C right panel), the effect of the fusion of N-terminal domain on the overall structure of DNMT3A and DNMT3B is unknown. I cannot exclude the possibility that the fusion of N-terminus disturbed the precise arrangement of N-terminal domain relative to the remaining/other DNMT3B or DNMT3A domains, rendering the resulting construct inactive. It is therefore important to test if these proteins are catalytically active. For this, I tried two catalytic assays previously used in work with DNMT3A and DNMT3B. However, these experiments proved inconclusive, and I am not confident to claim the intact catalytic activity of these constructs based on these experiments (see Appendix 3).

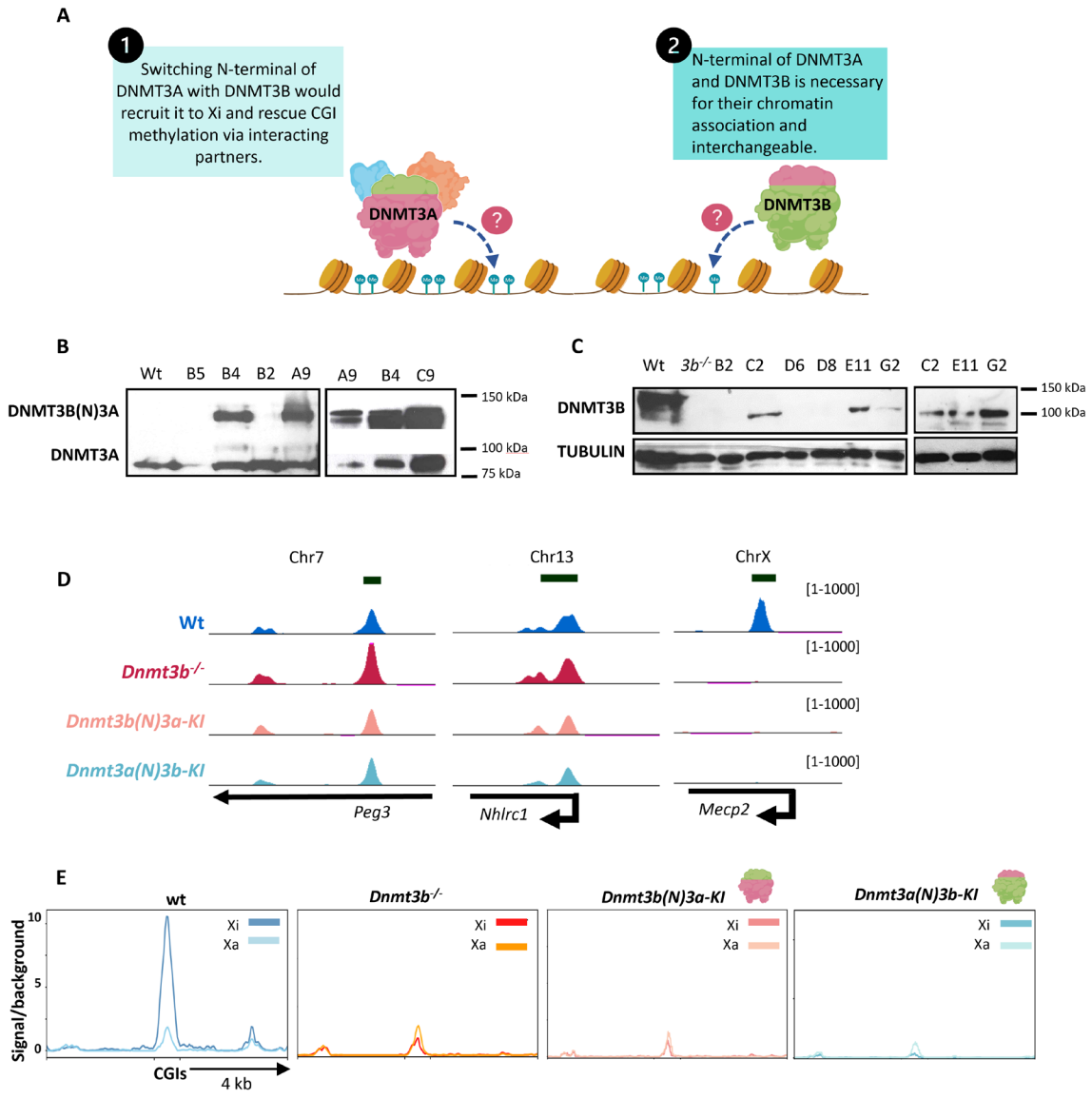


Figure 4.10: Complementation of *Dnmt3b*$^{-/-}$ cells with chimeric DNMT3 proteins (shown on previous page)

A) Schematic explaining the rationale behind the complementation experiment. See text for details.
B) Western blot showing the successful targeting of *Dnmt3b(N)3a* construct (expected protein size is 79.5 kDa). Right panel shows the proteins levels of DNMT3B(N)3A in chromatin-bound nuclear fraction in day 12 EBs.
C) Same as (B) for *Dnmt3a(N)3b-KI* line. Expected size of DNMT3A(N)3B protein is 96.8 kDa.
D) Genome browser tracks showing examples of DNMT3B-independent and DNMT3B-dependent CGIs. Both the complementation lines expressing chimeric DNMT3 constructs failed to rescue CGI methylation on DNMT3B-dependent CGIs.
E) Average profile of the MBD-seq signal on CGIs on the X chromosome. *Dnmt3b(N)3a* and *Dnmt3a(N)3b* complementation lines fail to rescue CGI methylation on Xi. Profiles for wt and *Dnmt3b*$^{-/-}$ cell lines are shown as a reference. Panels (D) and (E) show representative data from one of three well correlated replicates (Appendix 2).

4.8 Role of the N-terminal domain in chromatin binding of DNMT3B1

In light of the evidence that the N-terminal domain of DNMT3A and DNMT3B is required for chromatin binding, I wanted to determine if the N-terminal domain deletions that I generated had a detectable effect on the chromatin binding of DNMT3B. Initially, I investigated this using salt extraction, a technique widely used to test the chromatin binding strength of proteins. For this experiment I separated nuclear soluble and chromatin-bound fractions at a range of salt concentrations (75, 150, 250, 300, and 400 mM) for wildtype iXist-ChrX and *Δ1-218 Dnmt3b1-KI* complementation line. Figure 4.11 A shows a schematic for this protocol. If the N-terminal domain deletion indeed compromises the chromatin binding of DNMT3B1, it would be evident as Δ1-218 DNMT3B1 dissociating from chromatin at a lower salt concentration. In Figure 4.11 B full-length DNMT3B1 can be seen to be extracted into the soluble fraction at high salt concentrations (250 mM and higher) but only to a small degree. Even at the stringent salt concentration of 400 mM, most of the protein is stably bound to the chromatin. I did not observe any difference between Δ1-218 DNMT3B1 and full-length DNMT3B1 in this experiment, suggesting that at 400 mM salt, Δ1-218 DNMT3B1 is stably bound to chromatin (Figure 4.11 B). Like the Δ1-218 DNMT3B1, the DNMT3B1 constructs with smaller N-terminal deletions showed high chromatin binding strength and most of the protein stayed bound to chromatin at 400 mM salt condition (Figure 4.11 C). Thus, the inability of DNMT3B1 with N-terminal domain deletions to rescue CGI methylation on Xi cannot be explained due to their weak association with chromatin.

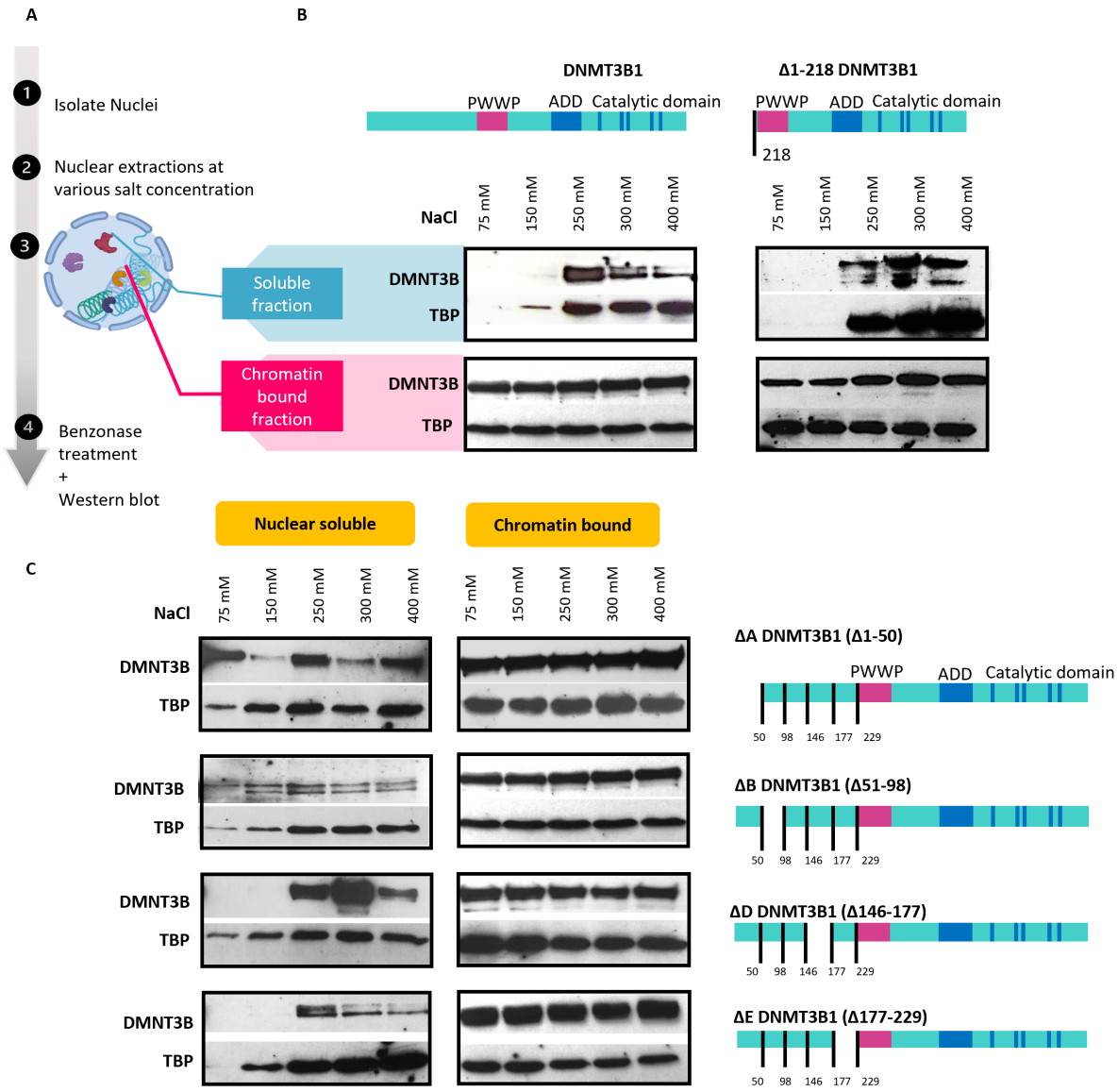


Figure 4.11: Salt extractions to test the chromatin binding strength of DNMT3B1 with N-terminal domain deletion (continued to next page).

A) A schematic showing the working principle of salt extractions. The protocol works by isolation of the nuclei and then extraction with different salt concentrations. This step releases the soluble contents of the nucleus. The remaining pellet is then treated with benzonase, an endonuclease, to digest DNA, releasing all DNA-bound proteins. As the salt concentration is increased at the stage of nuclear extraction, the proteins which are loosely bound dissociate from the chromatin. Only the tightly bound proteins stay associated with chromatin at higher salt concentrations.

B) The top panel shows the schematic for the full-length and Δ1-218 DNMT3B1. The bottom panel shows the nuclear soluble (upper) and chromatin-bound (lower) fractions for the two tested constructs. As the salt concentration is increased, the amount of DNMT3B1 extracted in the soluble fraction becomes higher but a larger fraction of DNMT3B1 is associated with chromatin even at 400 mM NaCl. Δ1-218 DNMT3B1 behaves similar to the full-length DNMT3B1.
C) Nuclear soluble (left) and chromatin-bound (right) fractions after salt extraction from complementation lines expressing DNMT3B1 with smaller N-terminal domain deletions. All complementation constructs can be seen to be bound to chromatin at 400 mM NaCl. Δ1-50 and Δ51-98 DNMT3B1 (ΔA and ΔB DNMT3B1) appear in nuclear soluble fraction at low salt concentration however, a greater fraction of the protein resides within the chromatin-bound fraction.

Theoretically, upon further increasing salt concentration, there would come a point at which the full-length DNMT3B1 dissociates from chromatin; and I could then compare if the Δ1-218 DNMT3B1 shows comparable chromatin binding strength as full-length DNMT3B1. However, salt concentration higher than 400 mM has been reported to cause nuclear lysis in ES cells which would make it impossible to discriminate between the nuclear soluble and chromatin-bound fractions. I, therefore, used a native chromatin isolation protocol reported previously (Henikoff et al., 2009), and described in the methods chapter in section 2.1.19. Briefly, the nuclei are isolated from the cells and treated with MNase. After this step, the supernatant containing mononucleosomes is collected. The remaining pellet of MNase-treated nuclei is then treated with buffers containing variable concentrations of salt, which releases the bulk of the chromatin in solution. Importantly, the buffer with high salt creates stringent chromatin extraction conditions which can be used to test the chromatin binding strength of DNMT3B1 (schematic shown in 4.12 A). This method allowed me to exceed the maximal 400 mM NaCl concentration used in the salt extraction protocol.

For performing this experiment, I divided the cell pellet into 6 equal parts, extracted nuclei from each and then performed salt extraction using defined salt concentrations (indicated in Figure 4.12 A) after MNase treatment. Figure 4.12 B shows full-length DNMT3B1 associated with chromatin at given salt concentrations (10, 150, 400, 600, 800 and 1000 mM). Δ1-218 DNMT3B1 showed similar chromatin binding strength as full-length DNMT3B1 (Figure 4.12 C). Using this method, I was able to see almost all of the chromatin extracted at 600 mM NaCl, which agrees with the original publication (Henikoff et al., 2009) (agarose gel in Figure 4.12 D and E). To show the chromatin content of selective fractions in this experiment (marked as red in Figure 4.13 B and C), I have shown the agarose gel of DNA extracted from the given fractions (Figure 4.12 D and E). I also tested for DNMT3B1 and chromatin

remaining in the insoluble fraction left after high salt washes of the MNase-treated nuclei. Figure 4.12 B-E show that the high salt washes successfully extracted the bulk of chromatin and associated DNMT3B1 (the lanes ending with (P) indicate the pellet left after the wash of MNase-treated nuclei with the given salt concentration).

As mentioned earlier, Jeong et al., first reported that the N-terminal domain of DNMT3B is responsible for its stable association with chromatin. They showed that deletion of the N-terminus of DNMT3B caused it to dissociate from chromatin upon sucrose fractionation of mononucleosomes at high salt (5-25% sucrose gradient at 300 mM NaCl). Without the N-terminus, DNMT3B appeared in lighter sucrose fractions which were devoid of chromatin as opposed to full-length DNMT3B which could be seen in fractions containing chromatin (as indicated by the presence of H3 in the fractions containing DNMT3B). To be confident of my findings and to exclude the possibility that the methods I used were not sensitive enough to capture small changes in chromatin binding strength between the full-length and Δ1-218 DNMT3B1, I set out to reproduce the results by Jeong et al. In order to clearly resolve the sucrose fractions containing dissociated Δ1-218 DNMT3B1 as observed by Jeong et al., I used a 5-16% gradient (schematic shown in Figure 4.12 F, the photograph shows the tubes for the sucrose fractionation, high sucrose solution was coloured red to clearly visualise the gradient). Chromatin extracted from the wt iXist-ChrX and *Δ1-218 Dnmt3b1-KI* cells was fractionated through this gradient using ultracentrifugation. Afterwards, fractions were collected from the top (i.e., light fractions were collected first). Figure 4.12 G shows the results from the sucrose fractionation experiment. The western blot shows every second fraction collected, and the numbers indicate the fraction number (from light to heavy in ascending order). I was unable to see Δ1-218 DNMT3B1 separated out in lighter sucrose fractions further reinforcing the results obtained thus far (Figure 4.12 G).

To exclude the possibility that reduced chromatin binding of Δ1-218 DNMT3B1 is only evident in differentiated cells, I used a cell fractionation assay provided by a commercial kit from Thermofisher, to analyse chromatin in differentiated EBs. This kit also allows the separation of nuclear soluble, and chromatin-bound fractions. In agreement with all the results presented thus far, I did not find a lower chromatin association of Δ1-218 DNMT3B1 even in differentiated cells. The general distribution of Δ1-218 DNMT3B1 was comparable to full-length DNMT3B1 (shown in section 4.11, Figure 4.16 D, lower panel). Thus, I find no evidence that deletion of the DNMT3B1 N-terminus results in overall lower chromatin binding strength, at least in the cell models used in this study. The methods mentioned in this section however cannot reveal subtle differences in chromatin localization between the full-length and Δ1-218 DNMT3B1 which would become apparent in sensitive methods like ChIP-seq.

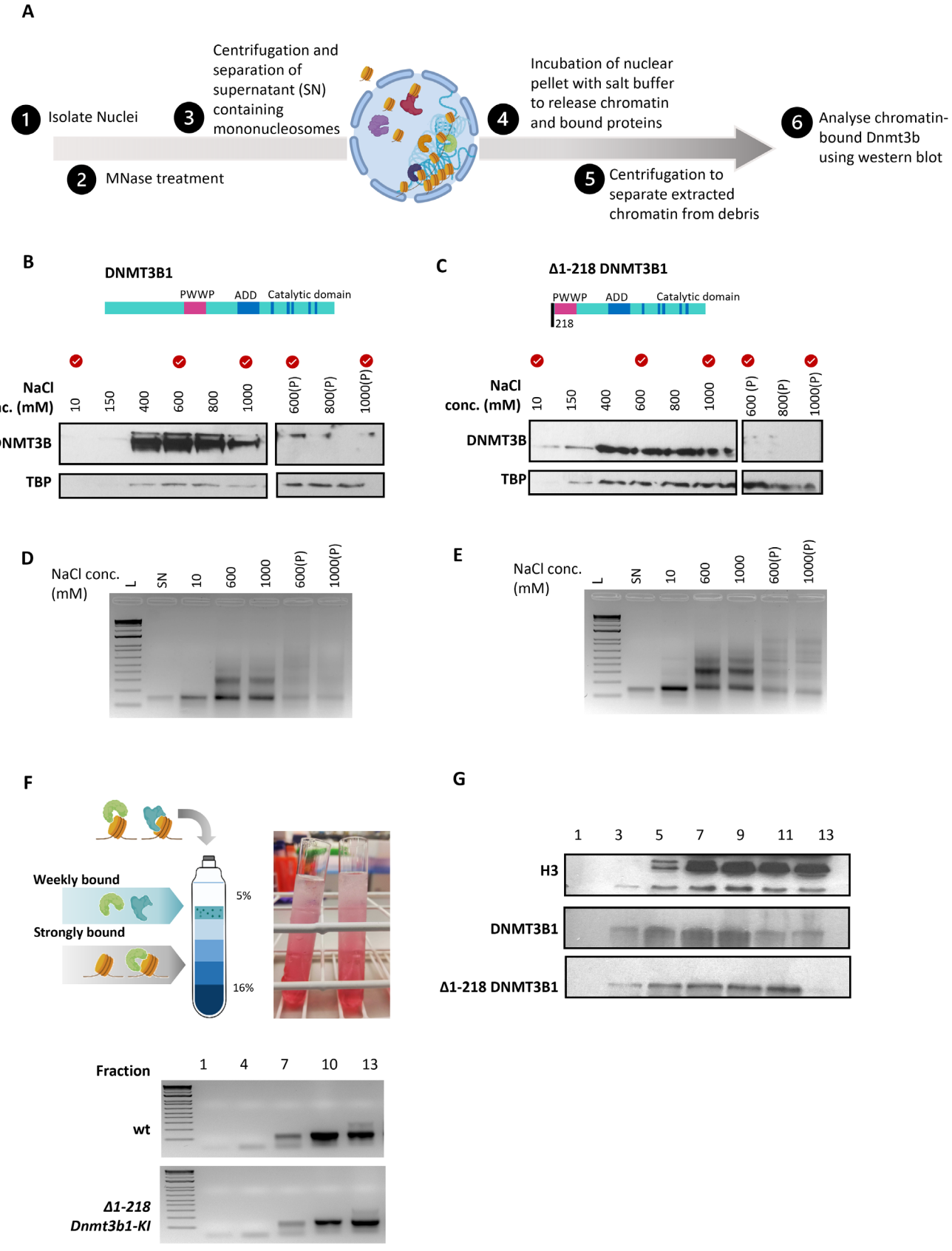

A
1 Isolate Nuclei
2 MNase treatment
3 Centrifugation and separation of supernatant (SN) containing mononucleosomes
4 Incubation of nuclear pellet with salt buffer to release chromatin and bound proteins
5 Centrifugation to separate extracted chromatin from debris
6 Analyse chromatin-bound Dnmt3b using western blot
B
DNMT3B1
PWWP
ADD
Catalytic domain
NaCl conc. (mM)
10
150
400
600
800
1000
600(P)
800(P)
1000(P)
DNMT3B
TBP
C
Δ1-218 DNMT3B1
PWWP
ADD
Catalytic domain
218
NaCl conc. (mM)
10
150
400
600
800
1000
600 (P)
800(P)
1000(P)
DNMT3B
TBP
D
NaCl conc. (mM)
L
SN
10
600
1000
600(P)
1000(P)
E
NaCl conc. (mM)
L
SN
10
600
1000
600(P)
1000(P)
F
Weekly bound
Strongly bound
5%
16%
Fraction
1
4
7
10
13
wt
Δ1-218 Dnmt3b1-KI
G
1
3
5
7
9
11
13
H3
DNMT3B1
Δ1-218 DNMT3B1

Figure 4.12: High salt chromatin extraction and sucrose gradient fractionation of DNMT3B1 to test DNMT3B-chromatin-binding strength (shown on the previous page)

A) Schematic showing the high salt chromatin extraction protocol; (1) isolation of the nuclei followed by (2) MNase treatment (3) centrifugation and separation of the mononucleosomal fraction (4) incubation of MNase treated nuclei with high salt to release the bulk of chromatin (salt concentration used for this step is indicated in (B) and (C)). Since this high salt extraction creates stringent conditions, it allows us to test the strength of association of DNMT3B1 with chromatin. (Method published by Henikoff et al 2009).

B) The result of the high salt chromatin extraction from wt iXist-ChrX line. A single cell pellet was equally divided into 6 parts. For each of these pellets, nuclei were isolated and treated with MNase. The MNase-treated nuclei were then washed with different salt concentrations as indicated (10, 150, 400 ,600, 800 and 1000 mM). The western blot shows DNMT3B1 is found in the high salt washes of MNase-treated nuclei. For selected fractions (marked by red ticks), DNA was extracted from the chromatin and analysed by gel electrophoresis (shown in (D)).

C) The same high salt chromatin extraction experiment as in (B) shown for the Δ1-218 DNMT3B1 expressing complementation line.

D) Agarose gel showing DNA content in selected fractions (marked as red in western blot shown B). Most of the chromatin was extracted at 600 and 1000 mM. SN is the supernatant collected after MNase treatment, which largely contains mononucleosomes released from the nuclei before high salt chromatin extractions.

E) Agarose gel DNA content in selected fractions (marked as red in western blot shown C). As for (D), most of the chromatin was extracted at 600 and 1000 mM.

F) (Upper panel) Schematic explaining the rationale of sucrose gradient fractionation. When chromatin is loaded onto a sucrose gradient and subjected to ultracentrifugation, the proteins which are weakly bound separate from the chromatin-containing fractions. The photograph shows the tubes for the sucrose fractionation, the sucrose solution was coloured red to give a visual estimate that the gradient was uniform. (Lower panel) Agarose gel shows the separation of chromatin among fractions after ultracentrifugation (from lighter to heavier). Resolution of chromatin among the fractions for both wt and *Δ1-218 Dnmt3b-KI* lines is uniform showing that the sucrose gradient fractionation worked equally well.

G) Western blot showing results from the sucrose gradient fractionation performed on chromatin extracted from wt iXist-ChrX and *Δ1-218Dnmt3b1-KI* line. Both full-length and Δ1-218 DNMT3B1 stay associated with the bulk of chromatin (as indicated by H3 western blot).

4.9 ChIP-seq analysis of DNMT3B1 after deletion of the N-terminal domain

The above analyses indicate that full-length DNMT3B1 and Δ1-218 DNMT3B1 show broadly similar chromatin localization. However, from the MBD-seq data, it is clear that Δ1-218 DNMT3B1 fails to carry out methylation at DNMT3B-dependent/ target sites, suggesting that Δ1-218 DNMT3B1 fails to localize to these regions. To investigate this point, I performed a ChIP-seq experiment to compare the localization of full-length and Δ1-218 DNMT3B1. Due to the lack of commercial ChIP-grade antibodies against DNMT3B, I made use of an N-terminal epitope tag (3X-T7) included in *Dnmt3b1* and *Δ1-218 Dnmt3b1* constructs integrated into the TIGRE locus in *Dnmt3b*$^{-/-}$ cells to generate *T7-Dnmt3b1-KI* and *T7-Δ1-218 Dnmt3b1-KI* lines. To perform quantitative comparisons between full-length DNMT3B1 and Δ1-218 DNMT3B1, I calibrated the ChIP using human Hela cells expressing 3X-T7-SMC1, a component of the cohesin complex, generously provided to me by lab of Professor Rob Klose in Oxford Biochemistry. DNMT3B1 ChIP shows a blanket signal broadly distributed across the length of chromosome (Figure 4.13 A and C) with large domains of enrichment overlapping gene-rich regions; this agrees with literature regarding the localization of DNMT3B on gene bodies due to co-transcriptional deposition of H3K36me3 (Baubec et al., 2015). Within the gene bodies, DNMT3B1 ChIP signal showed a gradual rise towards the transcription termination site (TTS) mimicking the accumulation of H3K36me3 (Figure 4.13 B, top panel). Overall the DNMT3B1 ChIP showed a good overlap with published data ((Baubec et al., 2015; Weinberg et al., 2019) and see Figure 4.13 A). Upon differentiation, the distribution of DNMT3B1 between the X chromosome alleles changes, with higher signal originating from the active (Xa) as compared to Xi (apparent as low allelic ratio) (Figure 4.13 D and E). This observation is consistent with the loss of H3K36me3 on gene bodies as a result of gene silencing on Xi (Figure 4.13 D and E). Consistent with the results presented in section 4.8, in ES cells, both the full-length and Δ1-218 DNMT3B1 show comparable enrichment on the X chromosome and autosomes (Figure 4.13 F, left). Interestingly, upon differentiation, Δ1-218 DNMT3B1 shows a lower signal on the X chromosome in comparison to full-length DNMT3B1 (Figure 4.13 F, right). This observation suggests a specific effect linked to the chromatin state of the inactive X chromosome that is only present in the differentiated cells.

The reduced occupancy of Δ1-218 DNMT3B1 on the Xi chromosome in EBs was confirmed on allelically split data which provides an internal control (Xa) and a more robust quantitative comparison (Figure 4.13 G). In comparison to full-length, Δ1-218 DNMT3B1 showed a lower allelic ratio suggesting that the deletion results in an even lower distribution of Δ1-218 DNMT3B1 on Xi. I then looked at CGIs to

see if Δ1-218 DNMT3B1 shows poor localization compared to full-length DNMT3B1. Given the gain of CGI methylation on Xi, I expected that the DNMT3B1 would show an enrichment at CGIs in EBs. On the contrary, the DNMT3B1 ChIP-seq signal does not change much between ES cells and EBs on the X chromosome (Figure 4.13 G). The lack of expected enrichment of DNMT3B1 on CGIs in EBs could be due to the fast reaction dynamics of DNMT3B1 which would make it difficult to be captured by techniques like ChIP that work better for slow and stably bound proteins. Interestingly, Δ1-218 DNMT3B1 shows lower localization on CGIs in both ES cells and EBs, suggesting that the N-terminal domain of DNMT3B1 is important for its residence on CGIs (Figure 4.13 G).

Within the gene bodies, the DNMT3B1 ChIP signal strongly correlates with the H3K36me3 signal and gene expression levels. For instance, the genes with the highest expression showed the highest enrichment of DNMT3B1 and vice versa (Figure 4.13 B, top). Δ1-218 DNMT3B1 ChIP does not show this correlation with H3K36me3 (Figure 4.13 B, bottom). The N-terminus deletion might in some way disrupt the function of the PWWP domain required for reading H3K36me3 in gene bodies. An *in vitro* study has shown that the peptide composed of the PWWP domain DNMT3B is not sufficient to bind H3K36me3 modified chromatin alone and that this specificity indeed requires the N-terminal domain (Baubec et al., 2015). However, this dependence of PWWP domain on the N-terminus has not been studied in context of full-length DNMT3B and the mechanism underlying this observation has not been investigated.

Intriguingly, my ChIP-seq data suggests a compromised binding of Δ1-218 DNMT3B1 on CGIs and gene bodies which make up a large portion of DNMT3B bound regions, yet the overall chromatin-bound Δ1-218 DNMT3B1 is similar to full-length DNMT3B1 (as indicated by salt extractions, chromatin extraction and sucrose gradient). It is possible that the lower binding of Δ1-218 DNMT3B1 on CGIs and gene bodies is balanced by higher binding on other regions. Due to the high background of the DNMT3B1 ChIP, I could not find any significantly differentially enriched regions between DNMT3B1 and Δ1-218 DNMT3B1 ChIP. However, it is possible that repetitive elements (which show poor mappability in genomic data) bind Δ1-218 DNMT3B1 more efficiently. I explore this point in the following sections.

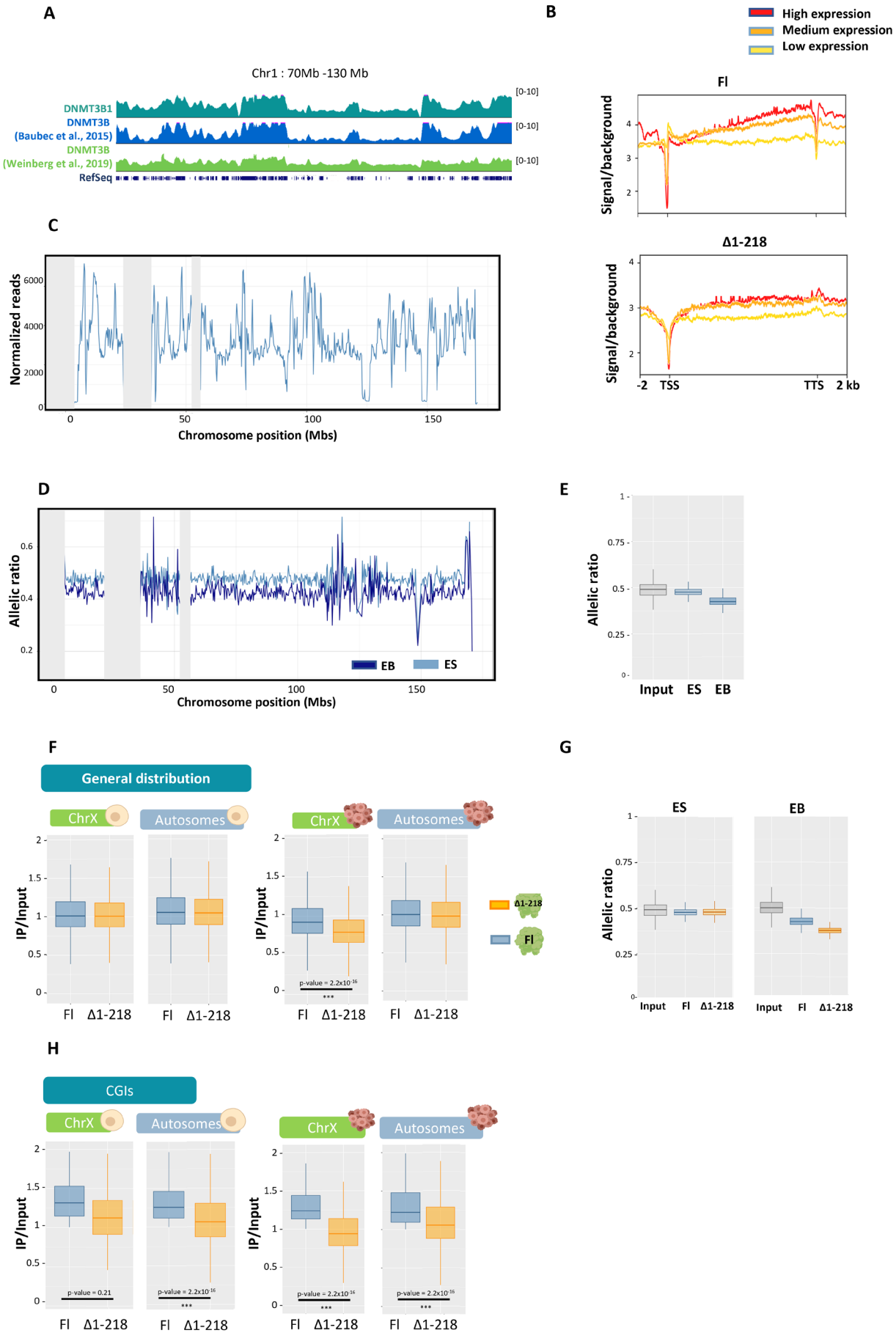

A
Chr1 : 70Mb -130 Mb
DNMT3B1
DNMT3B
(Baubec et al., 2015)
DNMT3B
(Weinberg et al., 2019)
RefSeq
[0-10]
[0-10]
[0-10]
B
High expression
Medium expression
Low expression
Fl
Signal/background
Δ1-218
-2 TSS
TTS 2 kb
C
Normalized reads
Chromosome position (Mbs)
D
Allelic ratio
EB
ES
Chromosome position (Mbs)
E
Allelic ratio
Input ES EB
F
General distribution
ChrX
Autosomes
IP/Input
Fl Δ1-218
p-value = 2.2x10-16
***
G
ES
EB
Allelic ratio
Input Fl Δ1-218
H
CGIs
p-value = 0.21
p-value = 2.2x10-16

Figure 4.13: DNMT3B1 ChIP reveals differential localization of Δ1-218 DNMT3B1 (shown on the previous page)

A) Tracks showing the DNMT3B1 ChIP signal performed in this work (top) and published by other studies (indicated). The ChIP performed for this work was comparable to the studies published earlier and shows large domains of enrichment on gene-dense regions. Genes are shown in the bottom track labelled as RefSeq.
B) DNMT3B1 ChIP (or Δ1-218 DNMT3B1) on gene bodies sorted by expression levels in ES cells. DNMT3B1 mimics the H3K36me3 pattern and enriches towards the transcription termination site (TTS). Δ1-218 DNMT3B1 does not follow this pattern. Fl denotes full-length DNMT3B1 ChIP.
C) DNMT3B1 ChIP signal on the X chromosome. The grey blocks mask the chromosomal regions with poor mappability.
D) Line plot showing allelic ratio for DNMT3B1 ChIP calculated over 250 kb bins along the length of the X chromosome.
E) Boxplot showing the same data as in (D).
F) DNMT3B1 ChIP signal (shown as a ratio of ChIP over input, IP/Input) for full length (Fl) and Δ1-218 DNMT3B1 calculated on 5 kb bins on the X chromosome or all autosomes in ES cells (left) or EBs (right) (indicated with icons). Enrichment of Δ1-218 DNMT3B1 is comparable to full-length DNMT3B1 in ES cells but gets lower on the X chromosome in EBs.
G) Boxplots showing allelic ratio for DNMT3B1 (or Δ1-218 DNMT3B1) enrichment on the X chromosome in ES cells or day 12 EBs.
H) DNMT3B1 (or Δ1-218 DNMT3B1) enrichment on CGIs in EBs (left) or ES cells (right) indicated by icons on top. p-values for comparison between DNMT3B1 and Δ1-218 DNMT3B1 ChIP were calculated using Wilcox rank sum test. This figure shows data from two well correlated ChIP-seq replicates for both *T7-Dnmt3b1-KI* and *T7-Δ1-218 Dnmt3b1-KI* lines (see Appendix 2).

4.10 A point mutation in the DNMT3B1 PWWP domain found in ICF results in complete loss of CGI methylation on the X chromosome

The CGIs on Xi have been reported to be hypomethylated in female patients with the congenital disorder ICF. As introduced in chapter 1, the most prevalent subtype of ICF is caused by mutations in human gene *DNMT3B*. It has been shown that CGIs for three X-lined genes (*G6PD, SYBL1* and *FMR1*) were 10-50 fold less methylated in ICF patients in comparison with the healthy control group, which correlated with the escape of *G6PD* from silencing in fibroblasts (Hansen et al., 2000). Therefore, to further investigate the features of DNMT3B protein that have crucial role in CGI methylation on Xi, I explored the mutations in *DNMT3B* that result in ICF. Most of the mutations in ICF fall in two classes:

(1) nonsense mutations occurring early in the transcript and resulting in complete loss of DNMT3B, or (2) missense and splice site mutations in the catalytic domain of DNMT3B (reviewed in (Hamidi et al., 2015)). Since the complete loss of DNMT3B is lethal (Okano et al., 1999), and ICF is considered to be a hypomorphic phenotype. Thus, no patient has been reported to carry nonsense or frameshift mutations on both alleles. In 2002, three ICF patients (two male and one female) with a missense mutation in the PWWP domain (S282P) of *DNMT3B* were reported ((Shirohzu et al., 2002). The patients with S282P *DNMT3B* depict all hallmarks of ICF (i.e., hypomethylation of satellite repeats causing multiradical karyotype of chromosome 1, 9 and 16, vulnerability to infections and facial abnormalities such as flat nasal bridge and low set ears). Since the PWWP domain and this mutation, do not affect the catalytic domain of DNMT3B *in vitro*, S282P (equivalent mutation in mouse is S277P) presents a curious case (Shirohzu et al., 2002). CGI methylation on Xi was not assessed in these patients. I, therefore, set out to explore if CGI methylation on Xi is compromised due to this mutation in *Dnmt3b* gene.

To perform this study, I complemented *Dnmt3b*$^{-/-}$ iXist-ChrX cells with *S277P Dnmt3b1*. In order to make this cell line amenable to ChIP and imaging, the *S277P Dnmt3b1* construct included a 3X T7 tag on the N-terminus. Figure 4.14 A shows the western blot confirming the targeting of *T7-S277P Dnmt3b1*. Three clones (A8, D5, and F3) were chosen for further experiments. I differentiated these lines into EBs (12 days) and performed MBD-seq. Figure 4.14 B shows the downregulation of *Oct4* and *Klf4* in EBs, indicating efficient differentiation.

The expression of S277P DNMT3B1 failed to rescue CGI methylation on Xi (Figure 4.14 C) and on DNMT3B-dependent autosomal CpGs (Figure 4.14 E). Comparison of total MBD-seq signal between *S277P Dnmt3b1-KI* and *Dnmt3b*$^{-/-}$ showed negligible differentially methylated regions (Figure 4.14 D). Some examples of DNMT3B-dependent and independent CGIs are shown in Figure 4.14 F (*Mtmr1* and *Hprt* are X-linked DNMT3B-dependent loci, while *Peg3* and *Meg3* are DNMT3B-independent loci). The complete loss of rescue of CGI methylation is a very profound finding since the complete deletion of DNMT3B PWWP domain has been shown to be dispensable for the catalytic activity of DNMT3B *in vitro* (Qiu et al., 2002).

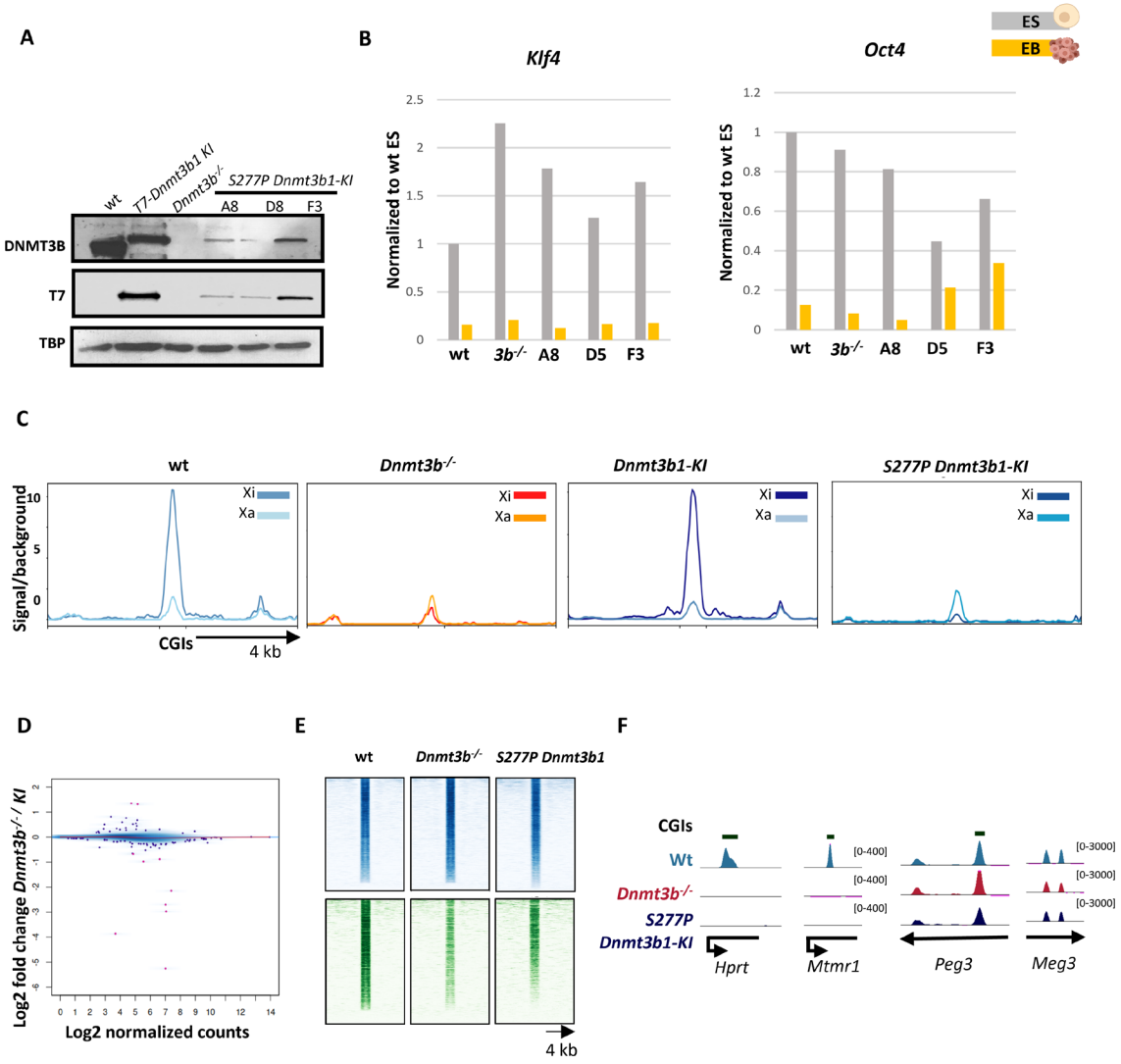


Figure 4.14: MBD-seq on PWWP point mutant *S277P Dnmt3b1* complementation line (continued to next page).

A) Western blot validation of targeting of *T7-S277P Dnmt3b1* in *Dnmt3b*$^{-/-}$ line to generate *S277P Dnmt3b1-KI* line. A8, D5, and F3 were used for further analysis.

B) Expression levels of *Klf4* and *Oct4* in ES and EB in *S277P Dnmt3b1-KI* complementation lines as determined by qPCR. Wt and *Dnmt3b*$^{-/-}$ lines are also shown for reference.

C) Average MBD-seq signal on CGIs on the X chromosome in wt, *Dnmt3b*$^{-/-}$, *Dnmt3b1-KI and S277P Dnmt3b1-KI*. *S277P Dnmt3b1-KI* does not rescue CGI methylation on Xi. Shown is the representative profile of three highly reproducible replicates (see Appendix 2).

D) MA plot comparing MBD-seq signal between *Dnmt3b*$^{-/-}$ and *S277P Dnmt3b1-KI* line (denoted as KI for knock-in). The significant (FDR < 0.01) differential regions which are very few are shown in pink.

E) Heatmaps showing DNMT3B-independent (top, blue) and -dependent (bottom, green) autosomal CGIs. *S277P Dnmt3b1-KI* (denoted as *S277P Dnmt3b1* in short) shows a profile similar to *Dnmt3b*$^{-/-}$.

F) Genome browser tracks for MBD signal at DNMT3B-dependent (*Hprt, Mtrm1*) and DNMT3B-independent (*Meg3, Peg3*) CGIs in wt, *Dnmt3b*$^{-/-}$ and *S277P Dnmt3b1-KI* lines. Panels (D)-(F) show data from one of three well correlated replicates (Appendix 2).

4.11 Differences in pericentric localization of DNMT3B1 for Δ1-218 DNMT3B1 and S277P DNMT3B1 mutations

DNMT3B has been shown to localize to pericentric chromatin as evident from DNMT3B signal overlapping DAPI dense regions in immunofluorescence (IF) analysis (Chen et al., 2004). It has been reported that S277P mutation in DNMT3B leads to the loss of DNMT3B localization to pericentric chromatin (Chen et al., 2004). This phenotype is linked to the resulting hypomethylation of major satellites on chromosomes 1, 9, and 16, a hallmark of ICF (Shirohzu et al., 2002). I therefore performed IF to see if DNMT3B1 localisation was altered in differentiated complementation lines expressing DNMT3B1, Δ1-218 DNMT3B1, or S277P DNMT3B1. Full-length/wt DNMT3B1 showed a punctate signal overlapping the DAPI dense regions (Figure 4.15 A, top panel). As reported, S277P DNMT3B1 showed a diffuse signal in the nucleus (Figure 4.15 A bottom panel). Intriguingly Δ1-218 DNMT3B1 also showed punctate signal on pericentric chromatin as indicated by DAPI dense regions, and moreover, it appeared that the pericentric signal was higher than that for wt DNMT3B1 (Figure 4.15 A and B). This observation indicates that Δ1-218 DNMT3B1 and S277P DNMT3B1 mutations operate through distinct mechanisms. While the N-terminal domain of DNMT3B1 is crucial for its localization to CGIs, it is dispensable for its recruitment to pericentric chromatin.

The overall nuclear IF signal seen for S277P DNMT3B1 was noticeably weaker in comparison to wt DNMT3B1 and Δ1-218 DNMT3B1 which prompted me to test if this mutation results in protein instability. Indeed, the *S277P Dnmt3b1* complementation lines expressed either comparable, or, in the case of clone F3, significantly more Dnmt3b mRNA as compared to wt iXist-ChrX and *T7-Dnmt3b1-KI* line, while the S277P DNMT3B1 protein level was much lower (Figure 4.15 C and D top panel). Cell fractionation showed that a relatively larger proportion of S277P DNMT3B1 was chromatin-bound as for wt DNMT3B1, but the protein levels were much lower in comparison (Figure 4.14 D, bottom panel). Similarly, as for the wt DNMT3B1, S277P DNMT3B1 was downregulated with differentiation, resulting in further reduced levels of the protein (Figure 4.15 C). The low levels of S277P DNMT3B1 in EBs could explain why the punctate signal of S277P DNMT3B1 on pericentric chromatin was undetectable and why CGI methylation on Xi is not rescued upon *S277P Dnmt3b1* complementation.

I also performed ChIP-seq for comparing the genomic localization of wt with S277P DNMT3B1 (data not shown). S277P DNMT3B1 showed similar localization to wt, albeit with a lower signal, which is expected. Due to low protein levels of S277P DNMT3B1, the ChIP signal was close to the background which results in low confidence for these datasets.

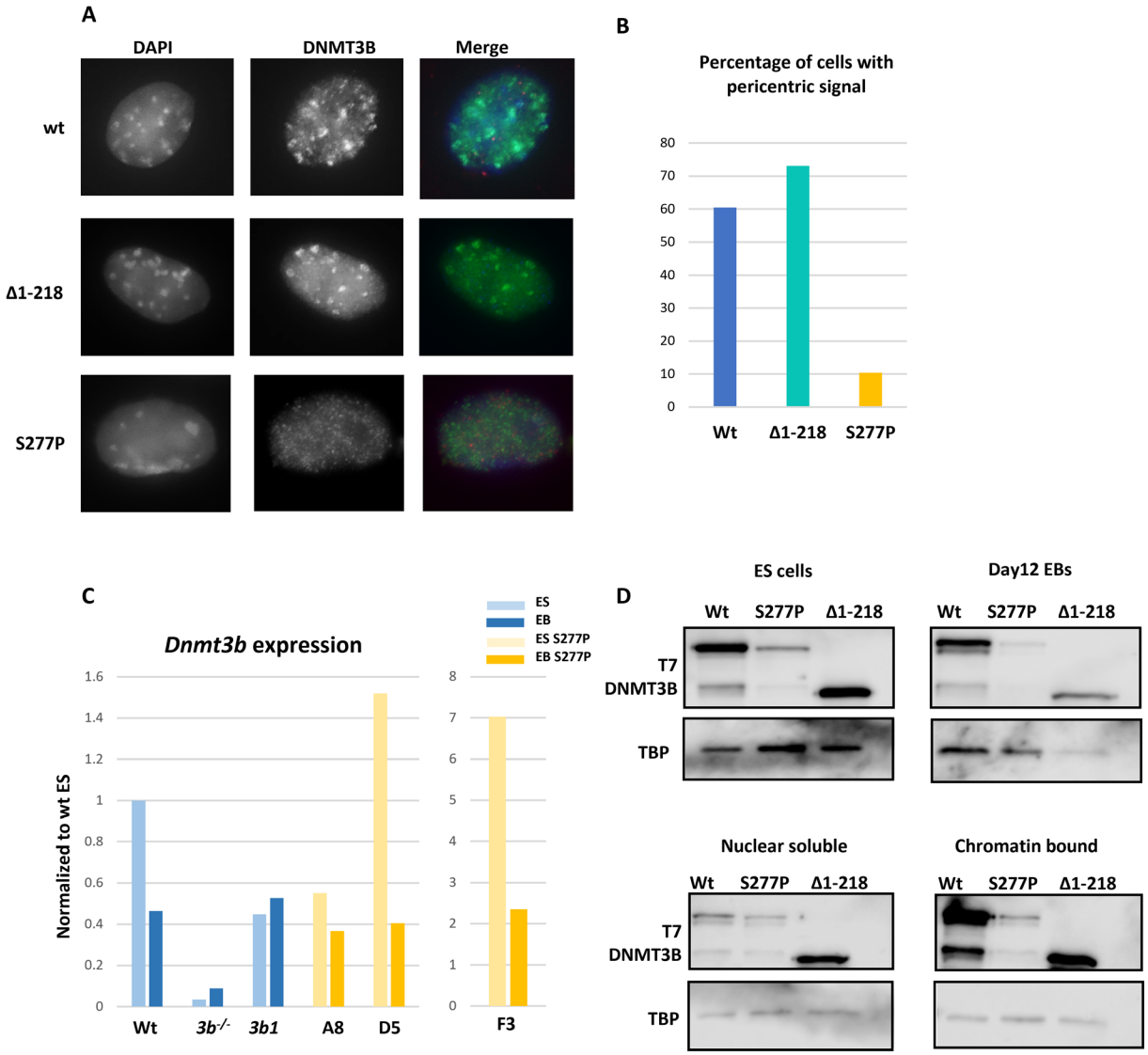


Figure 4.15: Pericentric localization of DNMT3B1, Δ1-218 DNMT3B1 and S277P DNMT3B1 (shown on the previous page).

A) Localization of DNMT3B1 on pericentric chromatin is visible as punctate signal overlapping DAPI dense regions. This localization is compromised by the S277P mutation in DNMT3B1. This IF was performed in day 12 differentiated EBs using anti-T7 tag antibody.
B) The graph shows the percentage of cells with pericentromeric DNMT3B1 signal in cell lines expressing DNMT3B1, Δ1-218 DNMT3B1 or S277P DNMT3B1. Pericentric localization was more obvious in Δ1-218 DNMT3B1 than wt. S277P DNMT3B1 signal was largely undetectable on pericentric chromatin. A minimum of 200 cells were scored for each line.
C) Expression of *Dnmt3b* in wt, *Dnmt3b*$^{-/-}$ (denoted as *3b*$^{-/-}$), *T7-Dnmt3b1-KI* (denoted as *3b1*) and, *S277P Dnmt3b1-KI* (A8, D5 and F8) in ES and EBs. *S277P Dnmt3b1-KI* line show comparable or higher levels of Dnmt3b transcripts.
D) (Upper panel) Western blot showing DNMT3B levels in nuclear extracts form *T7-Dnmt3b1-KI* (denoted as wt) and, *T7-S277P Dnmt3b1-KI* (denoted as S277P) and*, T7- Δ1-218 Dnmt3b1-KI* (denoted as Δ1-218) in ES cells and EBs. S277P DNMT3B1 has lower protein levels in comparison to both wt and Δ1-218 DNMT3B1. (Lower panel) Western blot after cell fraction for *T7-Dnmt3b1-KI* (denoted as wt) and, *T7-S277P Dnmt3b1-KI* (denoted as S277P) and, T7-Δ1-218 Dnmt3b1-KI (denoted as Δ1-218). Chromatin-bound and nuclear soluble fractions are shown.

4.12 Summary and conclusions

Although the specificity of DNMT3B and its role in CGI methylation on Xi is well established, there has been no thorough investigation of what governs this specificity. Some studies have suggested different preferences for DNMT3A and DNMT3B for defined DNA sequence motifs, but this cannot account for scenarios involving allele-specific methylation. In this chapter, I explored different models that could explain the specific recruitment of DNMT3B for Xi CGI methylation. I have analysed the contribution of two isoforms of DNMT3B, catalytically inactive DNMT3B3, and active DNMT3B1, in the methylation of X-linked CGIs. For this, I used a complementation approach where I introduced one DNMT3B isoform at a time in the *Dnmt3b*$^{-/-}$ line. I have shown that the complementation of *Dnmt3b*$^{-/-}$ line with DNMT3B1 fully rescued CGI methylation on Xi (as well as DNMT3B-dependent CGIs on autosomes), indicating that all other isoforms of DNMT3B are dispensable. To further test if DNMT3B3 has a role in rescuing CGI methylation on Xi, potentially in combination with DNMT3A, I complemented the *Dnmt3b*$^{-/-}$ line with DNMT3B3. DNMT3B3 was unable to rescue methylation of X-linked CGIs. The finding that DNMT3B3 is dispensable for CGI methylation on Xi was unexpected since a recent study reported DNMT3B3 to shows a preference for associating with DNMT3B1 (Zeng et al., 2020), and

hence one could predict/expect a scenario where expression of either isoform alone would not fully rescue CGI methylation.

Having ruled out a role for the non-catalytic DMT3B3 isoform in Xi CGI methylation, I investigated the role of the N-terminal domain as a DNMT3B-specificity imparting region. I present clear evidence that the N-terminal domain is crucial for DNMT3B-mediated CGI methylation, not only on the X chromosome but also at DNMT3B target CGIs genome-wide. This is the first report indicating the essential role of the DNMT3B N-terminal domain in CGI methylation. It has been suggested earlier that N-terminal domain of DNMT3B could serve as a non-specific/general chromatin-binding domain (Jeong et al, 2009). To test if the lack of CGI rescue was due to the inability of Δ1-218 DNMT3B1 to associate with chromatin, I employed several approaches to analyse its chromatin binding strength. I did not observe any difference in chromatin-binding strength between DNMT3B1 and Δ1-218 DNMT3B1, contrary to a previous publication (Jeong et al 2009). Intriguingly, my ChIP-seq analysis reveals a compromised localization of Δ1-218 DNMT3B1 on CGIs in comparison to full-length DNMT3B1. Furthermore, I have shown that the chromosome-wide occupancy of Δ1-218 DNMT3B1 on Xi is reduced in comparison to full-length DNMT3B1. This could indicate the role of DNMT3B1 N-terminus in mediating access of DNMT3B1 to chromatin state that occurs on Xi such as the accumulation of chromatin modifications established by PRC1/2, H3K9me3, H4K20me1 or compact chromatin structure. The IF demonstrated that Δ1-218 DNMT3B1 retains the ability to localize to pericentric heterochromatin. Together these observations suggest that the N-terminal domain of DNTM3B1 could function as a context-specific chromatin-interacting domain involved in the retention/targeting of DNMT3B1 on CGIs and Xi.

The role of DNMT3B1 N-terminal domain in determining its targeting to CGIs shown in this work is exciting given the recent studies highlighting a similar role of N-terminal domain of DNMT3A. Two recent publications describe that N-terminal domain of DNMT3A determines its localization to H2AK119ub1 modified chromatin and CGIs (Gu et al., 2022; Weinberg et al., 2021). According to these studies, the H2AK119ub1-recognizing domain is a small region within the disordered N-terminal domain of DNMT3A. Furthermore, an earlier study by (Suetake et al., 2011) showed that subregions of the DNMT3A N-terminus are associated with different DNA binding strengths. I therefore set out to explore if similar functional subdomains are present in the DNMT3B1 N-terminal domain which mediate localization of DNMT3B1 to CGIs. Lack of CGI methylation rescue in complementation lines expressing DNMT3B1 with smaller deletions in the N-terminal domain indicate that the entirety of the N-terminal domain is required for DNMT3B-mediated CGI methylation. Without further knowledge of the structure of DNMT3B1 N-terminus, the impact that these smaller deletions may have, or the

outcome of bringing together sequences that are normally spaced farther apart, is largely unpredictable.

A few curious observations came from the differential methylation analysis of complementation lines expressing DNMT3B1 with deletions in N-terminal in comparison to *Dnmt3b*$^{-/-}$. I found 2378 regions hypermethylated in *Δ1-218 Dnmt3b1-KI* complementation compared to *Dnmt3b*$^{-/-}$ ES cells, mostly residing within exons or intergenic regions. As shown in Figure 4.8 G, these regions are methylated in wt cells albeit to low levels, and thus Δ1-218 DNMT3B1 behaves like a hypermorphic mutant at these regions in ES cells. In EBs however, Δ1-218 DNMT3B1 does not rescue DNMT3B-dependent CGIs. Most of the regions found hypermethylated in *Δ1-218 Dnmt3b1-KI* line at ES stage (1873/2387) gain methylation in *Dnmt3b*$^{-/-}$ cells during differentiation. Approaches with genomewide coverage such as WGBS would reveal the CpGs targeted by Δ1-218 DNMT3B1 and tell if targeting of Δ1-218 DNMT3B1 becomes biased towards certain genomic regions/features. Since delta variants of DNMT3B are expressed in non-small lung cancer where they are thought to lead to aberrant DNA methylation patterns, this analysis would have implications in advancing our understanding of cancer.

Given the results that a full N-terminal domain is required for CGI methylation, I attempted to switch the entire N-terminal domain of DNMT3A2 with that of DNMT3B1, to determine if this would direct the chimeric protein to DNMT3B-target regions. However, the chimeric protein did not rescue Xi CGI methylation. Although the fusion of DNMT3A and DNMT3B was theoretically designed to avoid disruption of protein structure, the possibility of conformation changes that render the protein catalytically inactive cannot be excluded. To address this caveat, I tried to test the catalytic activity of the complementation constructs using two strategies, but the results of these experiments were inconclusive (see Appendix 3).

I also generated an ICF model cell line expressing a point mutation in PWWP domain of DNMT3B1 (S277P DNMT3B1). A previous study showed that ICF patients bearing mutations in DNMT3B catalytic domain show hypomethylation of Xi CGIs (Hansen et al., 2000). The impact of S282P mutation (S277P in mouse) on CGI methylation was not explored. I have shown that similar to ICF mutations in DNMT3B catalytic domain, S277P DNMT3B1 also fails to rescue CGI methylation. However, I have discovered that S277P mutation causes mouse DNMT3B1 to become unstable which could be an explanation for the inability of S277P DNMT3B1 to rescue CGI methylation on Xi and possibly the development of ICF symptoms among patients. The finding that S277P mutation causes instability of DNMT3B1 could also explain the loss of DNMT3B1 staining on the pericentric heterochromatin in IF experiments reported before (Chen et al., 2004) and in this work. The instability of S277P DNMT3B1 must also be factored

in when considering the conclusions drawn by other published ChIP-seq studies such as by Baubec et al., 2015, which have reported the loss of localization on gene bodies in this mutant.

Given the noisy nature of DNMT3B1 ChIP, I am cautious in drawing solid conclusions from the DNMT3B1 ChIP-seq. However, an interesting observation regarding DNMT3B1 came from ChIP-seq experiment. Contrary to expectation, DNMT3B1 ChIP signal on X-linked CGIs does not change between ES cells, which are unmethylated, and EBs; rather it seems to reside on X-linked CGIs in both. Since CGI methylation on Xi only takes place in EBs, DNMT3B1 must get a cue to carry out methylation only in EBs. This cue could be (1) a chromatin modification that flags/marks the target CGIs or render them permissive for methylation, (2) allosteric control of DNMT3B1 by interacting protein, or (3) PTMs on DNMT3B1. I explore the contribution of these factors in the next chapter.

Chapter 5

Investigating the role of H3K9me3 and protein interactors in influencing DNMT3A and DNMT3B specificity

5.1 Introduction

An alternative approach to the investigation of the underlying mechanisms that govern DNMT3B specificity for CGI methylation on Xi is to define any chromatin modifications or DNMT3B interacting proteins that could confer Xi CGI specificity. It may then be possible to link these factors to information concerning critical domains of DNMT3B described in chapter 4.

5.1.1 SMCHD1-linked H3K9me3 accumulation.

In chapter 1, I have detailed that XCI is a multi-step process, involving a cascade of events, which lay the ground for the steps to follow. CGI methylation is a late-stage process preceded by a number of crucial upstream steps such as transcriptional silencing, accumulation of histone modifications (notably H3K27me3, H2AK119ub1, H3K9me3 and H4K20me1), all of which may make Xi CGIs distinct targets for DNMT3B. Possibly linked to this is the role of SMCHD1. Despite the clear evidence for the role of SMCHD1 in targeting Xi CGI methylation, the underlying mechanism is unknown. A recent study suggests the role of SMCHD1 in the formation of H3K9me3 blocks on Xi. According to this study, the block-like signal of H3K9me3 on Xi shows at least partial dependence on SMCHD1 in MEFs and epiblast stem cells (EpiSCs, derived from late stage epiblast in which XCI has started) (Ichihara et al., 2022). A similar observation showing the reduction in H3K9me3 accumulation on Xi in *Smchd1* mutant MEFs was also made in our lab (Dr. Tatyana Nesterova, unpublished data). SMCHD1-mediated enrichment of H3K9me3 in turn could influence/ determine CGI methylation on Xi.

H3K9me3 is established on Xi earlier than CGI methylation (EB day 3, Dr. Tatyana Nesterova unpublished data), hence placing it upstream of CGI methylation. Support for the role of H3K9me3 in CGI methylation comes from the study by Keniry et al., which shows that depletion of H3K9me3 methyltransferase, SETDB1, leads to a reduction in CGI methylation on Xi (Keniry et al., 2016). Furthermore, a recent publication reported that allele-specific DNA methylation of ICRs is partially determined by H3K9me3 levels on these regions, adding further support to the role of H3K9me3 in determining DNA methylation (Yang et al., 2022). In view of these studies, it is interesting to investigate the role of H3K9me3 in DNMT3B-mediated CGI methylation on Xi.

5.1.2 DNMT3B-specific interactors

Knowledge of the interactome of a protein can provide useful insights into its function and regulation in the cell. In chapter 4, I have presented a hypothesis that the specificity of DNMT3B for Xi is governed by its interactors which are distinct from those of DNMT3A. Identification of protein interactors of DNMT3A and DNMT3B can provide useful insight into the regulation of CGI methylation and DNMT3B specificity for Xi. Importantly, a comparison of DNMT3B interactome in ES cells with that of differentiated EBs where XCI and CGI methylation has occurred, is a key experiment in this investigation.

A combination of affinity purification and mass spectrometry (AP-MS) is frequently used to study protein-protein interactions. AP-MS methods allow the isolation of protein complexes from cells or nuclear lysates through affinity purification of epitope-tagged 'bait' protein. The purified complex is then analysed using mass spectrometry which is a sensitive technique for identifying the proteins co-purified with the bait protein. It is also possible to analyse any differential PTMs present on the bait protein. The advantage of AP-MS approach is that protein complexes can be studied under physiological conditions. Many modifications of the AP-MS approach have now been adapted which allow quantitative comparison of the interacting proteins and the analysis of changes in the interactors in different conditions. Such quantitative MS methods involve the labelling of proteins with heavy isotopes incorporated by culturing cells in a growth medium containing amino acids with heavy isotopes. MS provides a list of all the interactors of a given protein in the cell without information regarding the composition of individual complexes; and thus, requires downstream validation, for instance, the AP-MS analysis of candidate interactors as a reciprocal experiment/validation of the interaction with the protein of interest.

A limitation of AP-MS is the detection of protein contaminants which due to their high abundance or non-specific interaction with the epitope, elute with the true protein interactors and add noise to the data. To reduce the background, approaches such as fusion of the protein of interest with tandem affinity purification tag (TAP-tag) are used. The TAP-tag consists of two affinity tags separated by a cleavage site for the TEV protease, allowing two rounds of purification steps. A disadvantage of this approach is that the weak interactors of the bait protein may be lost during the tandem purification steps (reviewed in (Berggård et al., 2007; Gingras et al., 2007).

AP-based methods are biased towards stable high affinity interactions that also show slow dissociation kinetics. The use of cross-linking agents such as formaldehyde and di-thiobis-succinimidyl-propionate (DSP) has been adapted to capture/preserve transient interactors while also allowing stringent washes to reduce non-specific noise. Rapid immunoprecipitation mass spectrometry of endogenous proteins (RIME) is such a recently developed AP-MS approach that uses *in vivo* cross-linked protein complexes (Mohammed et al., 2016). Proximity labelling techniques such as BioID present an alternative method to capture weak or transient interactions. The protein of interest is fused with bacterial biotin ligase BirA, which promiscuously biotinylates proteins present in proximity (reviewed in (Varnaitė and MacNeill, 2016)). A disadvantage to cross-linking and proximity labelling methods is the false positive protein hits which must be followed by experimental validation. Thus, these proteomics approaches offer different strengths and the choice of the method depends on the dynamics and abundance of the protein of interest.

5.2 H3K9me3 is enriched on CGIs

To study the correlation between CGI methylation and H3K9me3 on Xi, I performed native ChIP-seq for H3K9me3 in the iXist-ChrX line before and after 12-day EB differentiation. In agreement with prior studies, H3K9me3 forms a broad “blanket” like signal across the X chromosome (Figure 5.1 A). The H3K9me3 levels increase during differentiation and are skewed towards Xi causing the allelic ratio to rise (Figure 5.1 C). Although the exact allelic ratio varied from one differentiation to another, as a trend, H3K9me3 is seen to be weakly enriched over Xi in comparison to Xa. On average my data (and that from others in the lab) has shown that the allelic ratio rises to 0.6-0.7 during differentiation (Figure 5.1 C). This enrichment of H3K9me3 on Xi is, however, low as compared to other histone marks known to enrich on Xi like H2AK119ub1 and H3K27me3 for instance (Bowness et al., 2022).

I reasoned that if H3K9me3 marks the CGIs for methylation, then H3K9me3 would show higher enrichment over CGIs than Xi on average. H3K9me3 domains can be seen to overlap the MBD-seq signal at various loci. In Figure 5.1 B genome browser tracks show the H3K9me3 signal and DNA methylation in day 12 EBs and the co-occurrence of the two. Figure 5.1 D shows that this correlation is allele-specific. H3K9me3 signal is higher on Xi (Cast) allele which gains methylation. I find that overall, H3K9me3 enrichment is higher on CGIs than Xi on average (Figure 5.1 E, allelic ratio > 0.75 on CGIs). It is therefore plausible that H3K9me3 marks the CGIs that recruit DNMT3B, and then gain methylation.

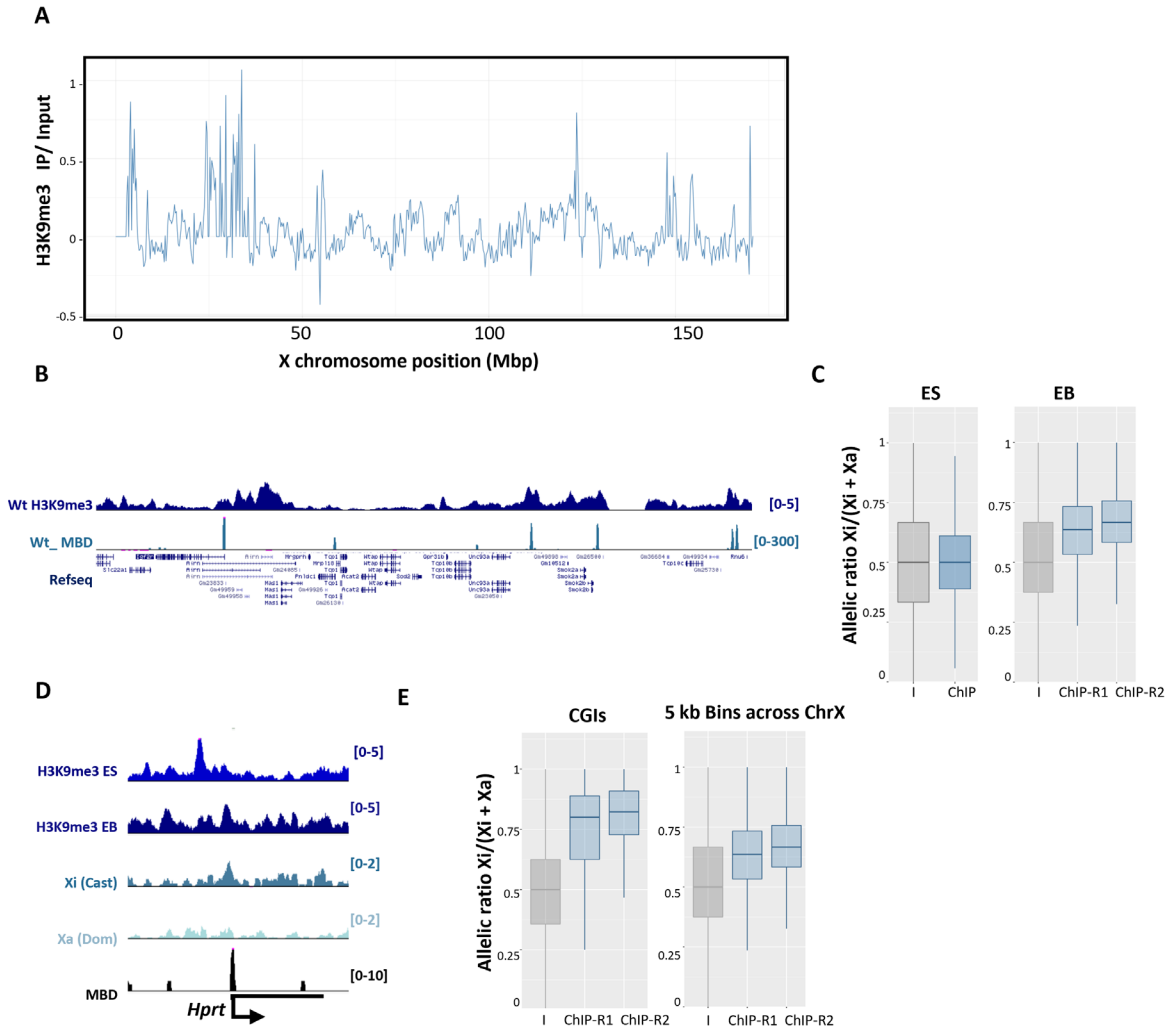


Figure 5.1: H3K9me3 accumulation on Xi (shown on the previous page).

A) The general pattern of H3K9me3 signal on the X chromosome. H3K9me3 forms a broad mat-like accumulation on the entire chromosome. H3K9me3 signal/input has been plotted for 250 kb bins across the X chromosome. The figure is representative of two well correlated replicates (see Appendix 2).
B) Genome browser tracks showing the overlap between H3K9me3 ChIP- and MBD-seq signal in wt iXist-ChrX line at 12-day differentiated EBs.
C) Boxplots showing the relative H3K9me3 signal on the Xi in comparison to the Xa in ES cells and day 12 EBs. H3K9me3 enrichment can be visualized by the rise in allelic ratio from 0.5 in ES cells to 0.6-0.7 in EBs. (I = Input, R1 = replicate 1, R2 = replicate2).
D) Genome browser tracks showing H3K9me3 gain on region overlapping CGI. The gain in H3K9me3 can only be seen in Xi (Cast) allele indicating the close correlation between H3K9me3 and CGI methylation on the Xi.
E) Boxplots showing allelic ratio at CGIs (left). CGIs show a higher accumulation of H3K9me3 than across the X chromosome on average in EBs (shown on right). (as above, I = Input, R1 = replicate 1, R2 = replicate2).

5.3 H3K9me3 correlates with DNMT3B localization on the Xi chromosome

I set out to test if H3K9me3 accumulation correlates with DNMT3B localization on the Xi chromosome. Genomic localization of DNMT3B was determined by ChIP-seq (introduced first in chapter 4 section 4.9), using the *T7-Dnmt3b1-KI* line on ES cells and day 12 EBs. To serve as a comparison and test if H3K9me3 specifically correlates with DNMT3B localization, I also performed ChIP-seq for DNMT3A in parallel with DNMT3B ChIP (Figure A and B). To avoid variability in pulldown efficiency owing to different antibodies, I generated a cell line with N-terminal 3X-T7 epitope-tagged *Dnmt3a2* targeted to the TIGRE locus (*T7-Dnmt3a2-KI* line) (Figure 5.2 A). DNMT3A ChIP was validated using a western blot for the histone modifications. Figure 5.2 C shows that DNMT3B is more enriched over H3K36me3 modified chromatin as compared to DNMT3A which agrees with the literature (Baubec et al., 2015). Similarly, H3K4me3 which is known to be antagonistic by both DNMT3A and DNMT3B did not show any enrichment in the IP (Figure 5.2 C, lower panel). Overall, DNMT3A and DNMT3B ChIP-seq were reproducible with published datasets (Baubec et al., 2015) (Figure 5.2 D).

I next searched for genomic regions showing H3K9me3 enrichment. For this, I used a peak calling software, MACS2, which captured the broad domains of the H3K9me3 ChIP-seq signal on ES and day 12 EBs. I then formulated a list of regions that gain H3K9me3 during differentiation genome-wide. Most of these regions are intergenic (Figure 5.2 E). The heatmap in Figure 5.2 F (left) shows that these

regions have a modest increase in H3K9me3 levels in day 12 EBs. DNMT3B localizes on these H3K9me3 enriched regions (Figure 5.2 F middle panel). DNMT3A shows a weaker correlation with H3K9me3-modified regions as compared to DNMT3B.

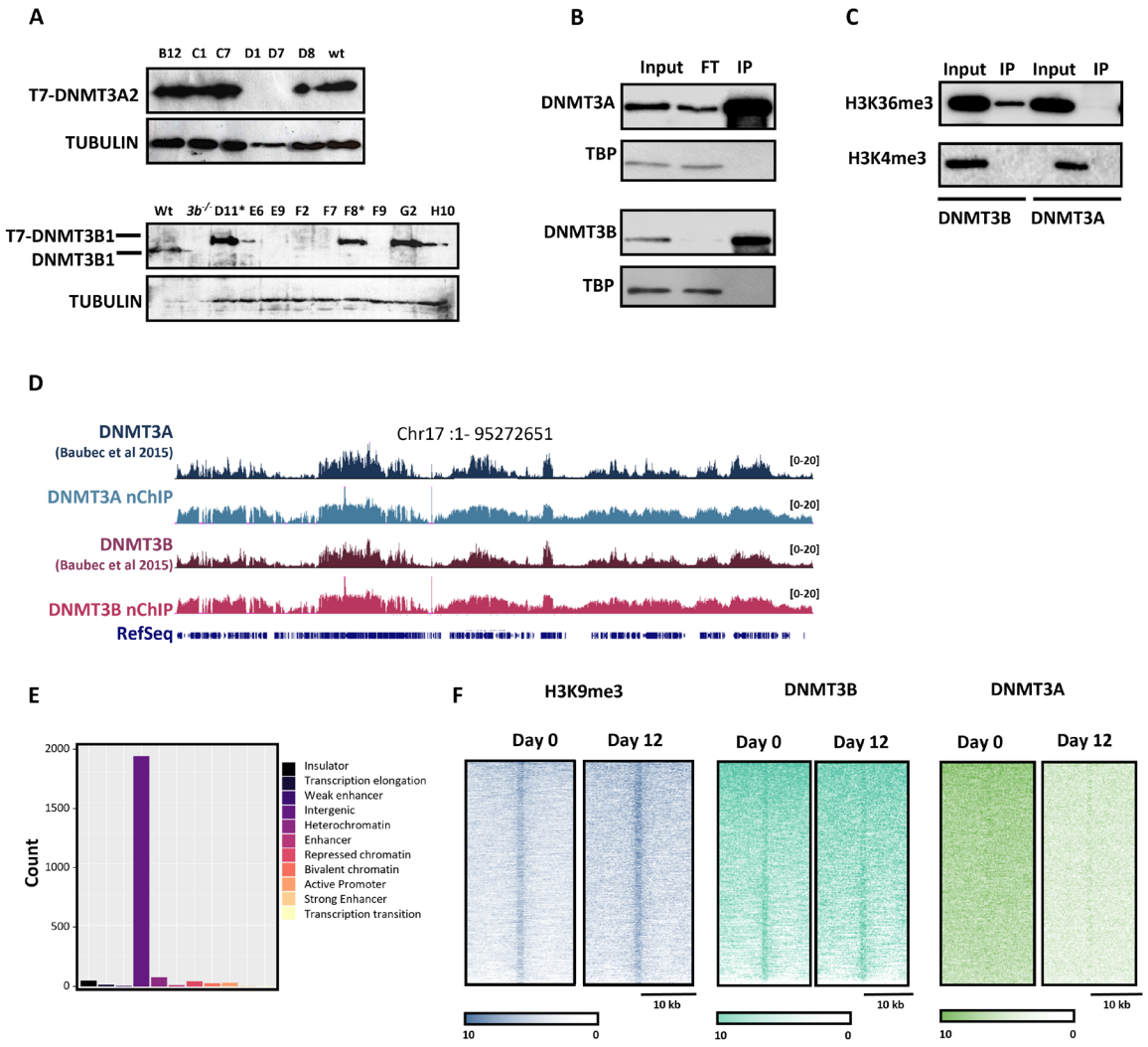


Figure 5.2: H3K9me3 enrichment correlates with DNMT3B localization on the X chromosome (continued to next page).

A) Western blot of clones showing successful targeting of *T7-Dnmt3a2* (upper panel) and *T7-Dnmt3b1* (lower panel) in *Dnmt3b*$^{-/-}$ iXist-ChrX cells. F8 and D11 lines for *T7-Dnmt3b1* targeting were used for ChIP-seq experiments. B12 was used for DNMT3A ChIP-seq experiments.

B) Western blot showing the efficient pulldown from DNMT3A and DNMT3B ChIP-seq. Both the proteins can be seen enriched in immunoprecipitated fraction (denoted as IP). For DNMT3B, there was also a clear depletion in the flowthrough (denoted as FT) which represents the unbound fraction after the pulldown. Both DNMT3A and DNMT3B show enrichment in IP.

Figure 5.2: H3K9me3 enrichment correlates with DNMT3B localization on the X chromosome (continued).

C) Western blot comparing H3K36me3 (top) and H3K4me3 (bottom) in IP fraction in DNMT3A and DNMT3B ChIP.

D) Genome browser tracks showing DNMT3A and DNMT3B native ChIP-seq (nChIP-seq) in comparison to published dataset (Baubec et al., 2015). The native ChIP-seq signal is reproducible with the published ChIP-seq. DNMT3A and DNMT3B enrich on gene-rich regions. Positions of the genes are indicated by the blue lines in the bottom track (labelled RefSeq).

E) Distribution of H3K9me3 peaks on the X chromosome among genomic features as determined by chromHMM. Most of these H3K9me3 peaks are intergenic.

F) (left) Heatmap for H3K9me3, in day 0 ES cells and day 12 EBs ordered from top to bottom according to H3K9me3 levels in day 12 EBs. (Middle and right) DNMT3A and DNMT3B ChIP-seq-seq in ES cells and day 12 EBs plotted on the H3K9me3 ordered according to H3K9me3 levels in the heatmap shown on left. DNMT3B shows a correlation with H3K9me3 peaks on the X chromosome. Panels (D), (E) and (F) show representative data from two highly correlated replicates for DNMT3A, DNMT3B and H3K9me3 ChIP (see Appendix 2).

In section 4.9 I have shown that DNMT3B localizes over X-linked CGIs in ES cells where methylation has not been established. Hence DNMT3B ChIP-seq does not reflect the loci where DNMT3B-mediated CpG methylation occurs. MBD-seq data, therefore, provides a functional definition of DNMT3B target sites in comparison to ChIP-seq. In chapter 3 (section 3.9), I have shown that "CGI-like elements" are dependent upon DNMT3A for their methylation. If indeed H3K9me3 marks only the DNMT3B-target CGIs, then "CGI-like elements" should be devoid of H3K9me3. Figure 5.3 compares the H3K9me3 enrichment on CGIs and "CGI-like elements". Since H3K9me3 is accumulated on the entire Xi, the "CGI-like elements" also showed a gain in H3K9me3 upon differentiation (allelic ratio increased from 0.5 in ES cells to ~0.6 in EBs). However, the gain of H3K9me3 was much higher on CGIs as evident from the allelic ratio > 0.75 in EBs.

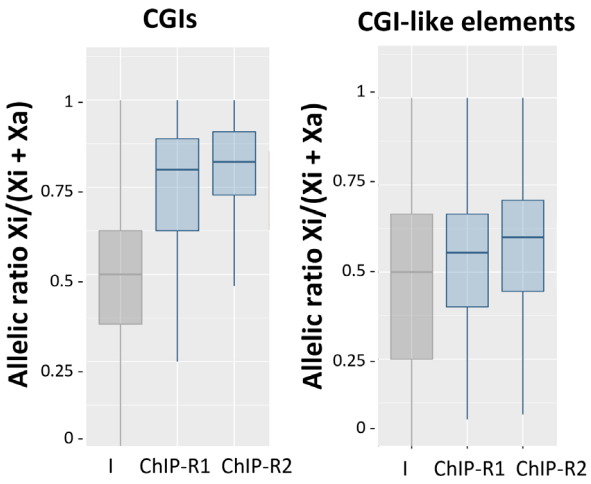


Figure 5.3: H3K9me3 is more enriched at CGIs in comparison to DNMT3B-independent "CGI-like elements" on the Xi in EBs (shown on the previous page).

Boxplots compare the relative gain of H3K9me3 on CGIs compared to "CGI-like elements" on the Xi. The gain of H3K9me3 on CGIs is larger as compared to "CGI-like elements" which gain methylation independently of DNMT3B.

5.4 H3K9me3 accumulation at CGI is upstream of their DNA methylation but shows partial loss in *Dnmt3b*$^{-/-}$ cells

Key evidence for the role H3K9me3 in establishment of CGI methylation on Xi would come from experiments which show that loss of H3K9me3 results in lack of CGI methylation on Xi. Ongoing work in the lab has to date failed to pinpoint a specific H3K9me3 methyltransferases responsible for the establishment of H3K9me3 on Xi (see discussion). Therefore, the analysis of CGI methylation in cells that lack H3K9me3 enrichment on Xi has not been possible. If H3K9me3 determines CGI methylation on Xi, then by extension, one could assume that H3K9me3 enrichment on Xi would remain unaffected in the absence of CGI methylation (i.e., in *Dnmt3b*$^{-/-}$ cells). To test this, I performed H3K9me3 ChIP-seq on *Dnmt3b*$^{-/-}$ line. As expected H3K9me3 was established in *Dnmt3b*$^{-/-}$ cells (Figure 5.4 A). Intriguingly, this gain of H3K9me3 is less efficient in *Dnmt3b*$^{-/-}$ EBs in comparison to wt (Figure 5.4 A, right panel). The decrease in H3K9me3 is more pronounced at CGIs which show higher gain of H3K9me3 than Xi on average (Figure 5.4 B, left). Figure 5.4 C and D show that H3K9me3 levels are compromised in *Dnmt3b*$^{-/-}$ chromosome wide. Thus, on Xi H3K9me3 is established but attains a lower level in the absence of DNMT3B. This data suggests that DNMT3B is important for maintaining H3K9me3 levels in wt cells. The mechanisms at play cannot be inferred from this experiment alone.

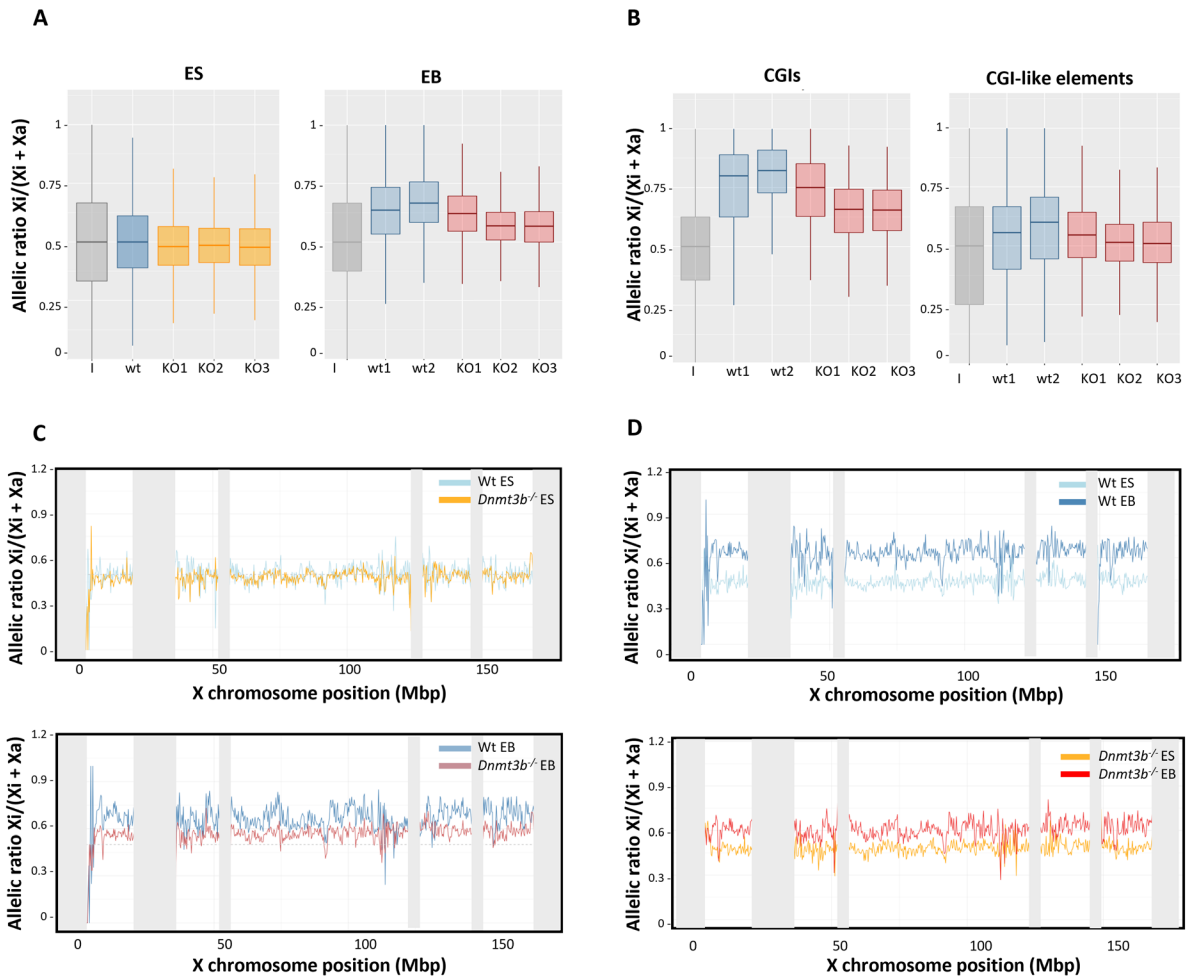


Figure 5.4: Enrichment of H3K9me3 on the Xi is lower in *Dnmt3b*$^{-/-}$ cells as compared to wt.

A) Boxplots showing the relative H3K9me3 signal on Xi in comparison to Xa in ES cells and day-12 EBs. H3K9me3 enrichment can be visualized by the rise in allelic ratio from 0.5 in ES cells to 0.6-0.7 in EBs. *Dnmt3b*$^{-/-}$ cells (denoted here as KO for knockout) show less enrichment in comparison to wt. Boxplots compare the H3K9me3 signal on 5 kb bins on X chromosome. (wt1 and wt2 show the two replicates for H3K9me3 ChIP-seq performed in wt iXist-ChrX line, KO1-KO3 show three replicates for H3K9me3 ChIP-seq in *Dnmt3b*$^{-/-}$ line)

B) Boxplots showing allelic ratio in wt and *Dnmt3b*$^{-/-}$ cells at CGIs (left) and "CGI-like elements" (right).

C) Comparison of the H3K9me3 on Xi (relative to Xa) in wt and *Dnmt3b*$^{-/-}$ cells before (top) and after differentiation (bottom). *Dnmt3b*$^{-/-}$ cells have lower H3K9me3 enrichment than wt on Xi in EBs. H3K9me3 signal per 250 kb bin along the X chromosome is plotted. The repetitive regions that show poor mappability have been masked.

D) Data from (C) with pre- and post-differentiation H3K9me3 signal plotted together for wt and *Dnmt3b*$^{-/-}$ cells. H3K9me3 signal is gained on the Xi in both the cell lines but the gain is larger in wt cells. Panels C and D show representative data from two and three highly reproducible H3K9me3 ChIP replicates for wt and *Dnmt3b*$^{-/-}$ cells respectively.

5.5 H3K9me3 is not enriched on autosomal CGIs dependent upon DNMT3B for methylation

In chapter 4, the mutations in DNMT3B which did not rescue CGI methylation on Xi, also failed to rescue methylation on autosomal CGIs dependent upon DNMT3B. Thus, it is tempting to speculate that a unified mechanism is at play that marks/flags all the DNMT3B target CGIs. Since H3K9me3 shows specific enrichment on X-linked CGIs, I tested if the autosomal CGIs that are dependent upon DNMT3B for methylation also show a preferential gain in H3K9me3 during differentiation.

To see if H3K9me3 levels do indeed govern/determine DNMT3B localization and methylation, I quantified H3K9me3 enrichment over autosomal CGIs that get methylated during differentiation in DNMT3B-dependent or independent manner. On autosomal CGIs, I could not see a clear trend between DNMT3B-dependent and independent CGIs, both of which showed a range of H3K9me3 levels and no significant gain in H3K9me3 levels in EBs (Figure 5.5 A and B, middle and right panel). As expected, H3K9me3 levels on X-linked CGIs are higher in differentiated cells (Figure 5.5 A and B, left). Overall, there is a modest increase in H3K9me3 levels on DNMT3B-dependent autosomal CGIs but it is not convincing enough to pursue H3K9me3 as a DNMT3B recruiting chromatin modification genome-wide (data not shown).

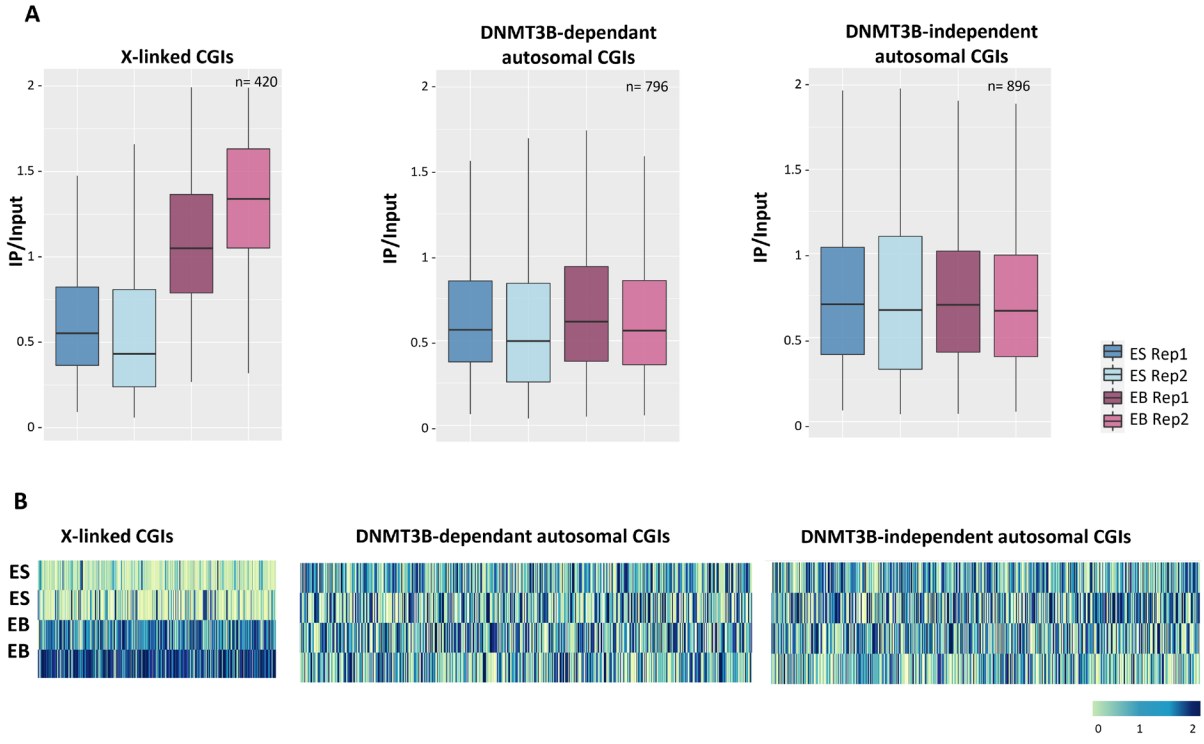


Figure 5.5: H3K9me3 enrichment on autosomal CGIs (shown on the previous page).

A) Boxplot showing H3K9me3 enrichment (represented as ChIP-seq signal/input, IP/Input) on X-linked (left), DNMT3B-dependant and (middle) DNMT3B-independent (right) CGIs on autosomes. As expected, H3K9me3 is enriched in EBs on the Xi. Unlike the X-linked CGIs, H3K9me3 is not enriched on autosomal CGIs. (n = number of CGIs in category).

B) Heatmaps represent the same regions as in (A) and show H3K9me3 enrichment (calculated by IP/Input) on individual CGIs on day 0 (denoted as ES) and day 12 EBs in each category. No obvious trend in the H3K9me3 signal is visible on autosomal CGIs. Each tile/line is one CGI.

5.6 A strategy to determine the DNMT3A and DNMT3B interactome in ES cells and EBs

An alternative approach to understanding the mechanisms governing DNMT3B specificity is to identify protein interactors specific to DNMT3B compared to DNMT3A. These protein interactors could in turn regulate/influence the activity and function of DNMT3A and DNMT3B. In order to identify DNMT3A- or DNMT3B-specific interactors, I performed an IP on nuclear lysates of the cell lines expressing T7-tagged DNMT3A2 and DNMT3B1 in ES cells and 12-day differentiated EBs under native conditions. IP was also performed in the parental iXist-ChrX line (on ES and EBs) that did not express T7-tagged DNMT3A/3B proteins to serve as a negative control. This strategy would allow me to (1) identify interactors of DNMT3A and DNMT3B in ES cells, (2) identify interactors of DNMT3A and DNMT3B in differentiated EB, (3) analyse if the interacting partners for DNMT3A and DNMT3B change upon differentiation and (4) see if DNMT3A and DNMT3B have unique protein interactors which could potentially allow their specific recruitment and in turn explain the specificity for their target loci (Figure 5.6).

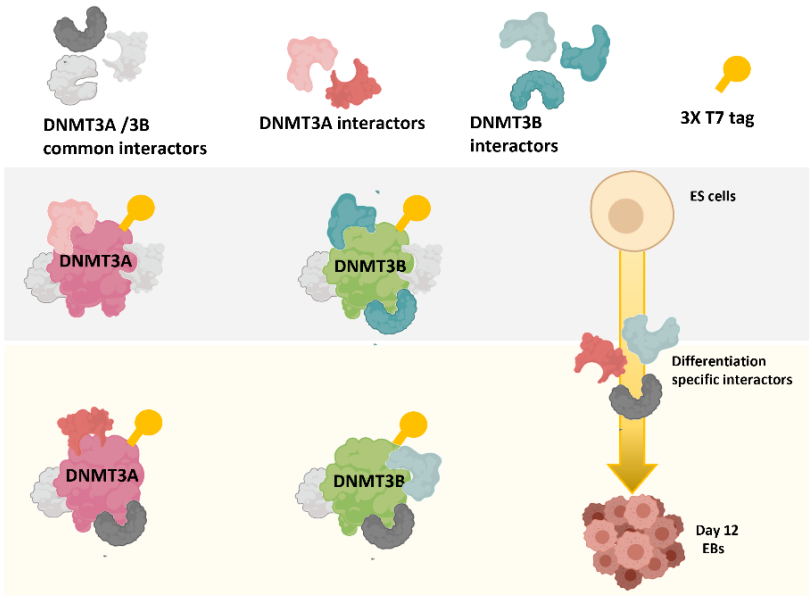


Figure 5.6: Schematic showing the rationale behind performing IP experiments to identify DNMT3A and DNMT3B interactors in undifferentiated and differentiated cells (shown on the previous page).

T7-tagged DNMT3A2 and DNMT3B1 expressing cell lines were generated for performing IP on nuclear lysates with the aim of identifying DNMT3A and DNMT3B interactors. DNMT3A and DNMT3B, being homologues are expected to have some common interactors. This screen will show if DNMT3A and DNMT3B also have specific interactors. The interactome of both DNMT3A and DNMT3B might undergo changes during differentiation which could explain the different functions of DNMT3A and DNMT3B on CGI. A comparative analysis of the interactome of DNMT3A and DNMT3B in ES cells and differentiated EBs would give some insight into the mechanism for driving DNMT3B specificity.

The efficiency of IPs was tested using western blot analysis. Figure 5.7 A shows a representative western blot validating the efficient pulldown for DNMT3A2 and DNMT3B1 from ES cells. For simplification, I have referred to both the bait proteins henceforth as T7-DNTM3A and T7-DNMT3B. Both the bait proteins can be seen to be enriched in the IP. DNMT3B is visibly depleted from the unbound flowthrough (FT) fraction. As expected, the untagged control iXist-ChrX line does not show any IP'ed protein (detectable by T7 antibody) on the western blot. To ensure that the IP pulled down additional/interacting proteins with the bait, I performed silver staining. Figure 5.7 B shows silver staining for IP of DNMT3A and DNMT3B (two replicates each) in ES cells. The control lane shows faint staining indicating a low abundance of proteins as expected. Figure 5.7 C shows the abundance of proteins in all the IP samples from ES cells after mass spectrometry: three replicates for DNMT3A and DNMT3B IP and two replicates for the control. The data obtained from mass spectrometry was analysed via three methods which largely agreed on the proteins significantly enriched in a given IP as compared to the negative control (as detailed in chapter 2, section 2.2.10). The proteins enriched in IP for DNMT3A and DNMT3B showed clustering among the replicates, validating that the replicates were reproducible (Figure 5.7 D). DNMT3A and DNMT3B IP resemble each other more (separated on PC2) as compared to the control (separated on PC1). Figure 5.7 E-H similarly shows the data for IP performed in day 12 EBs. Western blot validation of IP and silver staining are shown in (Figure 5.7 E and F). PCA plot and correlation plot in Figure 5.7 G and H indicate reproducibility among the technical replicates collected form independent differentiation experiments.

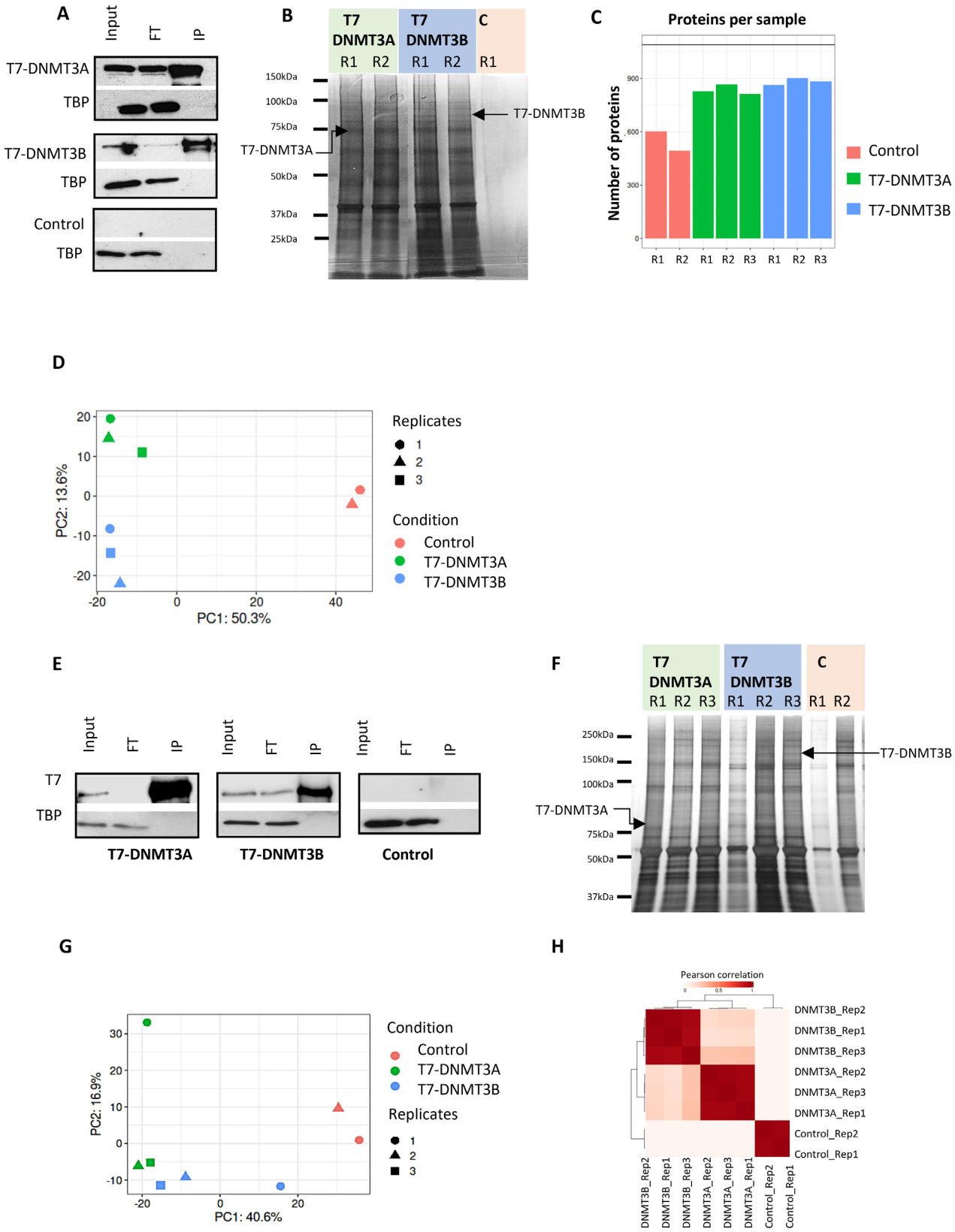

A
Input
FT
IP
T7-DNMT3A
TBP
T7-DNMT3B
TBP
Control
TBP
B
T7 DNMT3A
R1 R2
T7 DNMT3B
R1 R2
C
R1
150kDa
100kDa
75kDa
50kDa
37kDa
25kDa
T7-DNMT3B
T7-DNMT3A
C
Proteins per sample
Number of proteins
900
600
300
0
R1 R2 R1 R2 R3 R1 R2 R3
Control
T7-DNMT3A
T7-DNMT3B
D
PC2: 13.6%
PC1: 50.3%
Replicates
1
2
3
Condition
Control
T7-DNMT3A
T7-DNMT3B
E
Input
FT
IP
T7
TBP
T7-DNMT3A
T7-DNMT3B
Control
F
T7 DNMT3A
R1 R2 R3
T7 DNMT3B
R1 R2 R3
C
R1 R2
250kDa
150kDa
100kDa
75kDa
50kDa
37kDa
T7-DNMT3B
T7-DNMT3A
G
PC2: 16.9%
PC1: 40.6%
Condition
Control
T7-DNMT3A
T7-DNMT3B
Replicates
1
2
3
H
Pearson correlation
DNMT3B_Rep2
DNMT3B_Rep1
DNMT3B_Rep3
DNMT3A_Rep2
DNMT3A_Rep3
DNMT3A_Rep1
Control_Rep2
Control_Rep1

Figure 5.7: Validation and analysis steps for IP from nuclear lysates (shown on the previous page)

A) Western blot validation for the proper functioning of IP for T7-DNMT3A2 and T7-DNMT3B1 bait proteins from nuclear lysates of ES cells. For simplification, the bait proteins have been denoted T7-DNMT3A and T7-DNMT3B. The bait proteins are enriched in the IP. FT is unbound flowthrough.
B) Silver staining of the IP samples from ES cells to test efficient pulldown of bait interacting proteins. (R = replicate).
C) The abundance of proteins in all IP samples from ES cells as detected by the mass spectrometer. (as before R = replicate).
D) PCA plot showing the clustering among replicates of IP on ES cell samples. DNMT3A and DNMT3B IP resemble each other more compared to the control.
E) Western blot validation for the IP performed in EBs. The figure shows successful pulldown for T7-DNMT3A (left), and T7-DNMT3B (middle) bait proteins from nuclear lysates of EBs as indicated by enrichment of bait proteins in the IP fraction relative to input and flowthrough (FT).
F) Silver staining of the IP samples from EBs to test efficient pulldown of bait interacting proteins. Three replicates for T7-DNMT3A and T7-DNMT3B, and two control samples (iXist-ChrX line differentiated for 12 days into EBs) are shown. (R = replicate).
G) PCA plot showing the clustering among replicates. Two replicates for DNMT3A and DNMT3B IP cluster closer than a third replicate. Overall, the IP with bait proteins is well separated from the control.
H) Correlation plot clustered by the protein hits found in IP samples from EBs. All the replicates per condition cluster together. DNMT3A and DNMT3B IP cluster closer to each other compared to the control. (Rep = replicate).

5.7 DNMT3A and DNMT3B associated proteins in ES cells

Reassuringly, a number previously reported interactors of DNMT3A and DNMT3B were also identified in this experiment (shown in Figure 5.8, reviewed in (Saravanaraman et al., 2020)). The proteins enriched in DNMT3A and DNMT3B IPs are shown in Figure 5.9. The majority of these proteins are depleted in control and reproducibly found among replicates (Figure 5.9 A). I performed a pairwise comparison of DNMT3A or DNMT3B IP with control. Figure 5.9 B and C show volcano plots for the interactors found for DNMT3A and DNMT3B respectively as a function of the fold change enrichment over control and p-values. DNMT3B showed very few protein hits that qualified the FDR < 0.05 (Figure 5.9 C). This could be due to the relatively low expression of DNMT3B in comparison to DNMT3A. Since the number of significant protein hits in the DNMT3B IP using FDR < 0.05 threshold is small, I have used a p-value cut-off of < 0.01 to filter significant DNMT3B-associated proteins. In accordance with

the literature, DNMT3L was the most significant and abundant interactor for both DNMT3A and DNMT3B. The log fold change for DNMT3L was higher for DNMT3A in comparison to DNMT3L. This slight preference of DNMT3L for DNMT3A has been observed previously (Veland et al., 2019). Figure 5.9 D shows the overlap between the DNMT3A and DNMT3B-associated proteins.

In Figure 5.9 E, I have highlighted some DNMT3A-, DNMT3B-specific and commonly associated proteins. HELLS was found as a common interactor of DNMT3A and DNMT3B. As described in chapter 1, HELLS is a member of SNF2 family of ATPases that possess chromatin remodelling activity which facilitates DNMT3A and DNMT3B to carry out methylation in the linker DNA. Mutations in HELLS have been reported to lead to defects in DNA methylation (Myant et al., 2011; Zhu et al., 2006). Two other SWI/SNF family ATPases, SMARCA1 and SMARCA5 were also among this group. Another common interactor identified was EHMT2. Some literature suggests a functional link between EHMT2 and establishment of promoter methylation of germline and developmental genes in mouse embryos (Auclair et al., 2016) and ES cells (Dong et al., 2008; Myant et al., 2011). The direct interaction of EHMT2 with catalytic domains of DNMT3A and DNMT3B has previously been reported (Epsztejn-Litman et al., 2008). The screen also picked up subunits of PRC2 (EED, EZH2, SUZ12). The interaction of DNMT3A and DNMT3B with EED and EZH2 has also been reported in pulldown assay and was suggested to ensure the silencing of selective gene promoters (Viré et al., 2006). Some proteins associated with H3K9me1/2/3 were identified among commonly associated proteins. CBX1, 3 and 5 (also known as HP1 β, γ and α respectively) are H3K9me2/3 binding chromobox domain proteins that associate with heterochromatin (Kaustov et al., 2011). TRIM28 (KAP1) is involved in the recruitment of H3K9me3 methyltransferase SETDB1. Mphosph8 (MPP8) a subunit of the human silencing hub (HUSH) complex which recruits SETDB1 was also found in this group. Another member of the HUSH complex, TASOR was identified as interacting with DNMT3B but not DNMT3A. Developmental pluripotency associated (DPPA)2 and 4 have been shown to regulate zygotic gene activation (Eckersley-Maslin et al., 2019) and counteract DNA methylation (Gretarsson and Hackett, 2020). Paradoxically, histone acetyltransferase KAT6B (KAT7 for only DNMT3B-associated) and members of histone deacetylase NuRD complex, MTA1 together with RBBP7 and 6 were found among the common associated proteins of DNMT3A and DNMT3B. DNA demethylase TET1 was also identified as a common interactor and has been known to compete with DNMT3A and DNMT3B on regulatory sequences in pluripotent cells (Charlton et al., 2020).

Among DNMT3A-only associated proteins were the H3K36me2 methyltransferase NSD1, and MTA3, which has been shown to work in concert with histone deacetylases (Chen et al., 2013), Zinc finger protein ZNF462 and WDR5 known to be associated with H3K4me depositing MLL complexes.Centromeric protein CENP-C has previously been shown to associate with DNMT3B and

was also seen as DNMT3B-associated protein in this screen (Gopalakrishnan et al., 2009). Other DNMT3B-specific associated proteins include histone (H3K4) demethylases KDM2B and 5B, and H3K4me3 reader protein BAP18. KDM2B binds CGIs through a CXXC domain to recruit PRC1 (Farcas et al., 2012). SAP18 and PNN which associate with SIN3A and HDAC1, were also found associated with DNMT3B alongside FUS, TARDBP and MATR3 that function in RNA splicing and processing.

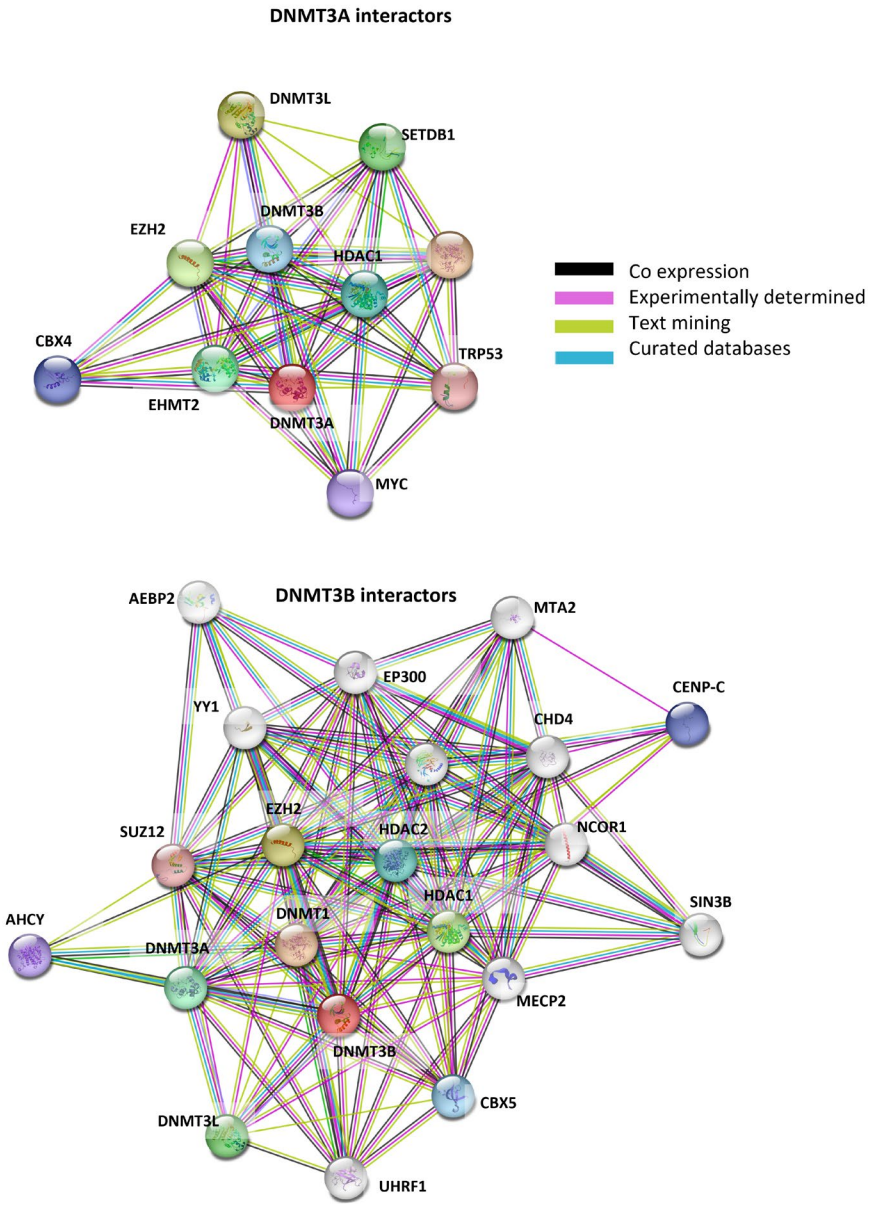


Figure 5.8: DNMT3A and DNMT3B interactors

Figure generated using Strings database (Szklarczyk et al., 2015), showing well characterized protein interactors for DNMT3A (top) and DNMT3B (bottom). Strings database uses coloured lines to show the criteria based on which any given protein is suggested to be associated with another; a key for the criteria used is shown on right.

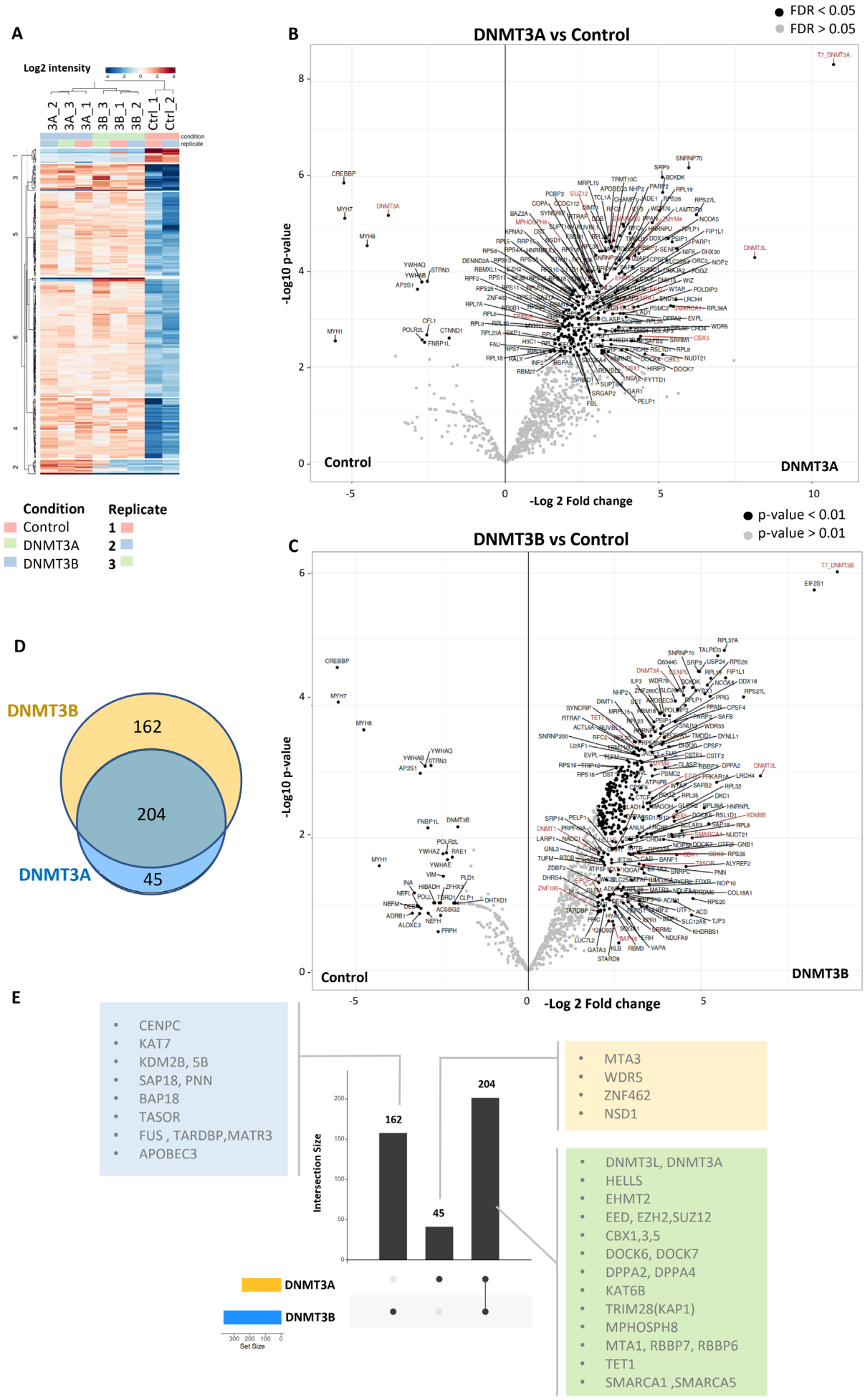

A
Log2 intensity
3A_2
3A_3
3A_1
3B_3
3B_1
3B_2
Ctrl_1
Ctrl_2
condition
replicate
Condition
Control
DNMT3A
DNMT3B
Replicate
1
2
3
B
DNMT3A vs Control
FDR < 0.05
FDR > 0.05
-Log10 p-value
-Log 2 Fold change
Control
DNMT3A
C
DNMT3B vs Control
p-value < 0.01
p-value > 0.01
-Log10 p-value
-Log 2 Fold change
Control
DNMT3B
D
DNMT3B
162
204
DNMT3A
45
E
CENPC
KAT7
KDM2B, 5B
SAP18, PNN
BAP18
TASOR
FUS , TARDBP,MATR3
APOBEC3
MTA3
WDR5
ZNF462
NSD1
Intersection Size
162
204
45
DNMT3A
DNMT3B
Set Size
DNMT3L, DNMT3A
HELLS
EHMT2
EED, EZH2,SUZ12
CBX1,3,5
DOCK6, DOCK7
DPPA2, DPPA4
KAT6B
TRIM28(KAP1)
MPHOSPH8
MTA1, RBBP7, RBBP6
TET1
SMARCA1 ,SMARCA5

Figure 5.9: DNMT3A- and DNMT3B-associated proteins immunoprecipitated from nuclear lysates in ES cells (shown on the previous page).

A) Heatmap showing the enrichment of proteins identified among samples. These are mostly depleted from the negative control as per expectation. Every row represents an individual protein found enriched in DNMT3A IP. The enrichment value for a given protein (represented here as intensity) is reproducible among replicates.
B) Volcano plot showing proteins enriched in the DNMT3A IP as compared to control. The proteins labelled have a FDR < 0.05. Some associated proteins discussed in text are labelled red.
C) Volcano plot showing proteins enriched in the DNMT3B IP as compared to control. The proteins have been labelled have a p-value < 0.01 cut off.
D) Venn diagram showing the overlap between DNMT3A- and DNMT3B-associated proteins.
E) Plot depicting the number of proteins recognized as DNMT3A-, DNMT3B- or common associated proteins. A few examples have been shown.

5.8 DNMT3A and DNMT3B associated proteins in EBs

The proteins associated with DNMT3A and DNMT3B in EBs are shown in Figure 5.10. The heatmaps show the intensity for the enriched proteins in DNMT3A (Figure 5.10 A) and DNMT3B IP (Figure 5.10 C) among replicates versus control. The heatmaps also show that the proteins enriched in the IPs were found among all replicates. Volcano plots in Figure 5.10 B and D show the proteins significantly enriched in DNMT3A and DNMT3B IP respectively. Only 13 proteins found enriched in DNMT3B IP qualified the significance cut-off of FDR < 0.05 (Figure 5.10 C and D). As for DNMT3B IP in ES cells, I relaxed the cut-off to p-value < 0.01 (Figure 5.10 E). DNMT3L was again the most significantly enriched protein in DNMT3A and DNMT3B IP. The total number of associated proteins found in the IP was smaller in EB samples for both DNMT3A and DNMT3B in comparison to ES cells. The subunits of PRC2 complex, CBX proteins and TET1 which were found as DNMT3A and DNMT3B common associated proteins, were absent in IP conducted on EBs. A few proteins like SAP18, BAP18, KAT6B, KAT7, HELLS, DPPA2 / DPPA4, DOCK6 and DOCK7, found as DNMT3A/3B common associated proteins can be seen enriched in EBs as well. Among proteins enriched in DNMT3B IP were MARK1, MARK2, KIF14, KIF21B, KIFC3 all of which are associated with microtubule dynamics. Enrichment of these proteins could be the result of the association of DNMT3B with centromeres (and microtubules) during cell mitosis (Geiman et al., 2004). Aurora kinase B, a key regulator of mitosis and chromosome alignment /segregation was also enriched in DNMT3A and DNMT3B IP. Interestingly MPP8 and TASOR were found enriched in DNMT3B IP suggesting an association of DNMT3B with H3K9me3 establishing proteins.

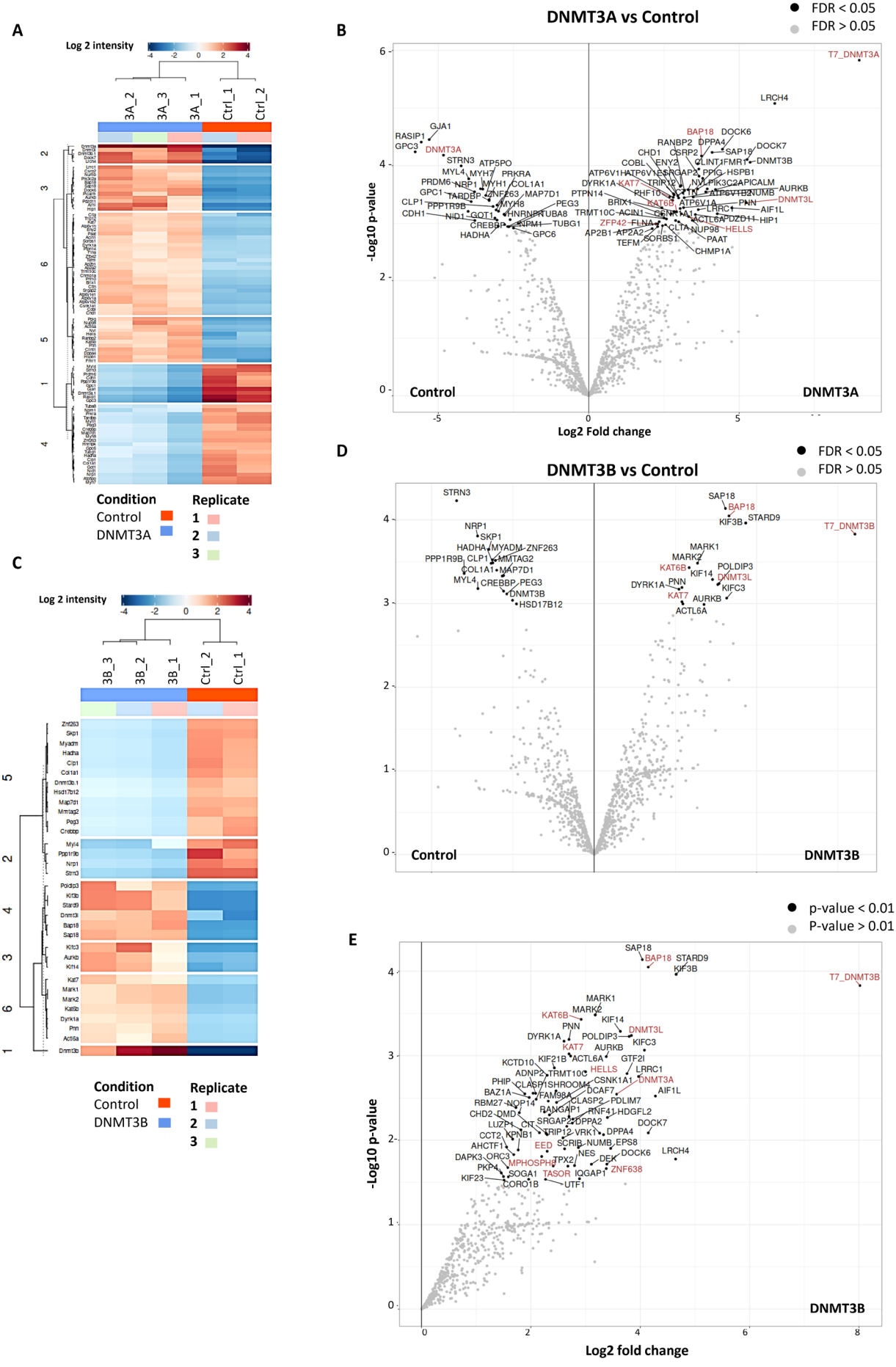

A
Log 2 intensity
-4 -2 0 2 4
3A_2
3A_3
3A_1
Ctrl_1
Ctrl_2
Condition
Control
DNMT3A
Replicate
1
2
3
B
DNMT3A vs Control
FDR < 0.05
FDR > 0.05
-Log10 p-value
Log2 Fold change
Control
DNMT3A
T7_DNMT3A
DNMT3A
DNMT3L
C
Log 2 intensity
-4 -2 0 2 4
3B_3
3B_2
3B_1
Ctrl_2
Ctrl_1
Condition
Control
DNMT3B
Replicate
1
2
3
D
DNMT3B vs Control
FDR < 0.05
FDR > 0.05
Control
DNMT3B
T7_DNMT3B
E
p-value < 0.01
P-value > 0.01
-Log10 p-value
Log2 fold change
DNMT3B
T7_DNMT3B
DNMT3A
DNMT3L

Figure 5.10: DNMT3A- and DNMT3B-associated proteins in EBs (shown on the previous page).

A) Heatmap showing the intensity values of proteins found significantly enriched in DNMT3A IP in comparison to control and clustering among replicates (cut off = FDR < 0.05). Each row represents an individually associated protein found enriched in the DNMT3A IP. The enrichment value for a given protein (represented here as intensity) is reproducible among replicates.
B) Volcano plot showing DNMT3A associated proteins. Some of the proteins enriched in IP and described in section 5.7 have been labelled in red.
C) Heatmap as in (A) for the DNMT3B IP.
D) Volcano plot showing DNMT3B associated proteins. The proteins labelled have a FDR <0.05
E) Volcano plot as in (D) calculated using a less stringent significance cut off (p-value < 0.01)).

As mentioned in section 5.1.2 AP-MS based methods are susceptible to noise/background contributed by highly abundant proteins or non-specific interactors. It is therefore important that the abundant proteins identified in AP-MS analysis are experimentally validated in order to derive confident conclusions. It is reassuring that a number of well-characterized interactors of DNMT3A and DNMT3B were found enriched in the IP in this screen. I was not able to experimentally test the interaction between DNMT3A and DNMT3B and their associated proteins in the time I had available for my DPhil studies, but this will be pursued by the lab in the future. Given the possible link between CGI methylation and H3K9me3, TASOR and MPP8 in particular provide interesting targets for further investigation.

5.9 PTMs on DNMT3A and DNMT3B

PTMs can regulate protein function. Some PTMs have been reported to occur on DNMT3A and DNMT3B in prior studies. For instance sumolyation of DNMT3A has been reported (Li et al., 2007). Sumolyation of DNMT3A on the N-terminus has been shown to result in loss of its interaction with HDAC1/2 and ability to repress transcription (Ling et al., 2004). Similarly phosphorylation of DNMT3A (serine 386 and 389) enables DNMT3A to be targeted to heterochromatin and methylate heterochromatic repeats (Deplus et al., 2014). Additionally, software such as PhoshoSitePlus (Hornbeck et al., 2015) predicts many more residues in both DNMT3A and DNMT3B as undergoing PTMs, particularly phosphorylation, on the N-terminus. Given that CGI methylation by DNMT3B occurs only upon differentiation, PTMs in DNMT3B could explain this differentiation-specific function. On the western blot, DNMT3B forms a “ladder” pattern of three bands which could suggest a differentially

modified form of DNMT3B. By further analysing the IP samples for DNMT3A and DNMT3B in ES cells and EBs, the mass spectrometry facility provided me with analysis for the PTMs found on DNMT3A and DNMT3B. No PTMs were reliably found in DNMT3B IP samples, which could be due to its low abundance, especially in EBs where DNMT3B is downregulated (as described in chapter 4). DNMT3A on the other hand did exhibit differentiation-specific PTMs. In ES cells DNMT3A2 was found to be acetylated on K398 (K617 in DNMT3A1, mapping to linker between ADD and catalytic domain) and methylated on R143 (equivalent residue on DNMT3A1 362, the linker between PWWP domain and ADD domain). Upon differentiation, residue T422 (T641 in DNMT3A1, catalytic domain) was phosphorylated and K407 and K409 (K627 and K629 in DNMT3A1, start of the catalytic domain) were methylated (Figure 5.11, schematic in top panel). The residues that are modified in ES cells (R143 and K398) and the phospho site T422 are conserved among DNMT3A and DNMT3B in mammals. K407 and K409 are conserved in DNMT3A in mammals and replaced with R in DNMT3B (Figure 5.11, bottom panel). These residues, due to their position in the catalytic domain, could function to regulate DNMT3A catalytic activity in EBs and possibly prevent it from methylating CGIs. Experimental validation of these PTMs on DNMT3A and studying the impact of point mutations/substitutions of the modified residues would be interesting topics for further investigation.

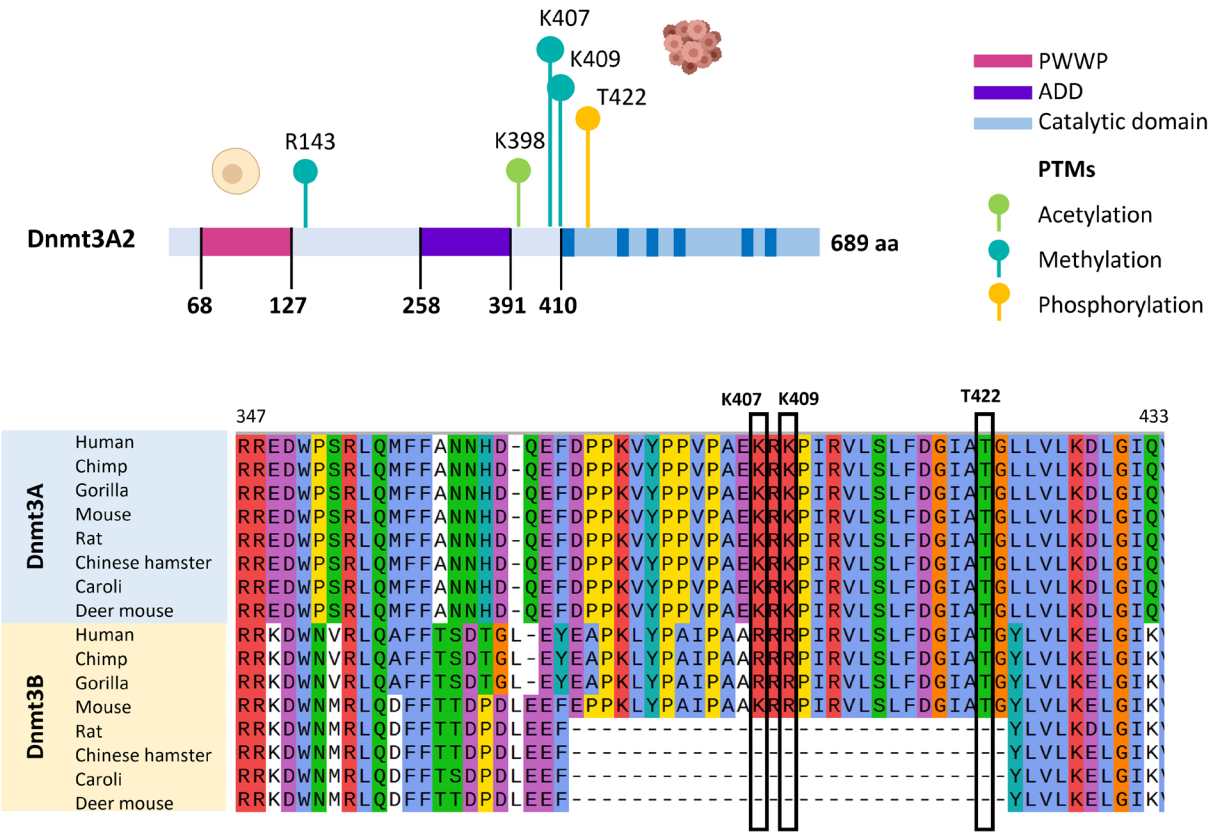


Figure 5.11: PTMs on DNMT3A2 (shown on the previous page).

(Top) A schematic showing the PTMs found on DNMT3A in ES cells (R143, K398) and EBs (K407, K409 and T422) detected in mass spectrometry analysis of IPs perfumed in this work.

(Bottom) Protein sequence alignment of DNMT3A and DNMT3B for selective mammals, highlighting residues that exhibit PTMs in DNMT3A2 in EBs. K407 and K409 are replaced by R in DNMT3B for most species. Amino acids have been coloured as per their biochemical properties. Protein sequence alignment was performed using Clustal Omega.

5.10 Summary and conclusions

Experiments in this chapter explore the role of factors outside DNMT3B itself that could govern/impact DNMT3B-mediated CGIs methylation on Xi. Keeping in mind the recent literature suggesting the role of SMCHD1 in the establishment of H3K9me3 on Xi, the first part of this chapter investigates H3K9me3 serving as a potential recruiting mechanism for DNMT3B on CGIs. Interestingly CGIs are among the most H3K9me3-enriched regions on Xi. DNMT3B-independent "CGI-like elements" (introduced in chapter 3, section 3.9) show lower enrichment of H3K9me3. The intergenic regions on Xi that show a high accumulation of H3K9me3 also correlate with DNMT3B ChIP-seq, providing support for H3K9me3 as a putative chromatin mark for targeting DNMT3B and possibly, CGI methylation on Xi. Clear evidence for the contribution of H3K9me3 in the establishment of CGI methylation on Xi would come from MBD-seq analysis of a cell line which lacks H3K9me3 establishment on Xi. In a reporter assay, Keniry et al., identified SETDB1 as the predominant histone methyltransferase responsible for establishment of H3K9me3 on Xi (Keniry et al., 2016). However, ongoing analysis in our lab indicates at most a contributory role of SETDB1 (Dr. Mafalda Almeida unpublished data). The role of H3K9me3 in DNA methylation is well studied and reported in plants and filamentous fungi *Neurospora crassa* (reviewed in (Rountree and Selker, 2010). In the mouse model, the role of H3K9me3 in the recruitment of DNMT3A and DNMT3B to major satellite repeats via HP1 has been described (Lehnertz et al., 2003). H3K9me2 methyltransferase G9A (also known as EHMT2), has also been implicated in regulating DNA methylation at ICRs in ES cells (Zhang et al., 2016). Thus, in view of the published literature, the upstream role of H3K9me3 in CGI methylation on Xi is plausible. Intriguingly, I have discovered that although H3K9me3 is established on Xi in *Dnmt3b*$^{-/-}$ cells, H3K9me3 accumulation on Xi is compromised in *Dnmt3b*$^{-/-}$ cells. Arguing against the importance of H3K9me3 for CGI methylation is that autosomal CGIs that are methylated by DNMT3B do not show a

clear trend of elevated H3K9me3. The underlying mechanisms for crosstalk between DNMT3B and H3K9me3 present an important subject for further investigation.

The second part of this chapter focuses on the identification of DNMT3B-specific interactors as a means to gaining insight into mechanisms governing DNMT3B specificity. Several previously reported interactors of DNMT3A and DNMT3B, such as DNMT3L, HELLS, SUZ12, EZH2 and EHMT2 were identified in this experiment, which together with high reproducibility among the replicates, validate its efficiency. DNMT3B was also found to be associated with protein complexes implicated in the recruitment of H3K9me3 methyltransferase SETDB1 such as TASOR, MPP8 and TRIM28. The interactions of these proteins with DNMT3B needs to be experimentally validated which was not possible in the available time but will be pursued in future. The interaction of these proteins with DNMT3B, if proved, could provide hints into the crosstalk between CGI methylation and H3K9me3.

A potential issue of the AP-MS analysis was the low abundance of DNMT3B and its further downregulation during differentiation. The ChIP-seq for DNMT3A and DNMT3B hint that the engagement of these enzymes with their target site is transient and depicts fast kinetics (evident by lack of clear enrichment/peaks on DNMT3B-target CGIs). It is possible that weak/low abundance interactors of DNMT3B were masked in this analysis due to this reason. Proteomics approaches that involve cross-linking such as RIME or proximity labelling could therefore improve the detection of such transient DNMT3B-specific interactors that might have escaped detection in AP-MS experiment used here.

AP-MS analysis could also benefit from additional controls. In chapter 4, I have shown that Δ1-218 DNMT3B1 fails to localize to and rescue methylation of CGI methylation. AP-MS analysis of Δ1-218 DNMT3B1 could serve as a negative control to identify the interactors of DNMT3B required for methylation of CGIs. Furthermore, a comparison of the interactome of DNMT3B between differentiated male and female cells could provide information regarding the presence of Xi-specific complexes formed by DNMT3B. However, it is tempting to assume that a unified global mechanism for DNMT3B-mediated CGI methylation is at play, a notion supported by a similar differentiation window for DNMT3B-mediated CGI methylation on autosomes and Xi, and mutations in DNMT3B failing to rescue autosomal CGI methylation as well as those on Xi. If distinct mechanisms govern CGI methylation on Xi, then Xi-specific complexes of DNMT3B would be a small fraction of other protein complexes that DNMT3B is a part of, making the detection very challenging.

Finally, I have briefly looked at PTMs as a regulatory mechanism for controlling the methylation activity of DNMT3A and DNMT3B. Although no PTMs could be reliably detected for DNMT3B, residues in the DNMT3A catalytic domain undergo differentiation-specific phosphorylation (T422) and methylation (K407 and 409) which could modulate its catalytic activity or its interaction with the DNA substrate, particularly X-linked CGIs. As for other findings from the proteomics data, these PTMs need to be experimentally validated and serve as an exciting subject for future investigation.

Chapter 6

Discussion



DNA methylation is crucial for proper development and survival in mammals (Okano et al., 1999). Studies on the misregulation of DNA methylation machinery in cancers, notably the *de novo* methyltransferases, DNMT3A and DNMT3B, further consolidates the significance of the establishment of accurate DNA methylation patterns in development and disease (reviewed in (Gagliardi et al., 2018)). Although DNA methylation is one of the best characterised epigenetic modifications with much research conducted on understanding its establishment and determinants, there remain many important questions. One example is how normally unmethylated CGIs sometimes acquire DNA methylation, XCI being a model example, and the dependence of this process for the most part on DNMT3B but not DNMT3A. This specificity is particularly enigmatic as DNMT3A and DNMT3B are highly similar and function redundantly in methylation at most other sites in the genome (Auclair et al., 2014). In this work, I have described experiments aimed at shedding light on the question. In this chapter, I will discuss my development of a robust experimental system to study CGI methylation on Xi. I will then discuss the insights gained from the complementation experiments, ChIP experiments, and mass spectrometry aimed to draw comparisons between DNMT3A and DNMT3B. Finally, I describe some testable models of DNMT3B specificity by taking into consideration the key findings from my work.

6.1 Differentiation of iXist-ChrX ES cells coupled to calibrated MBD-seq - a versatile model system to study CGI methylation on Xi

One of the main strengths of this study is the optimization of a robust and sensitive experimental system to study CGI methylation. The iXist-ChrX line developed by Dr. Tatyana Nesterova in the lab allows allelic analysis of individual CGIs. Doxycycline-inducible Xist causes the iXist-ChrX cells to undergo synchronous XCI of one of two X chromosomes when grown as ES cells following cell-culture-based EB or NPC differentiation protocols. Most of my experiments used EB differentiation which was

reproducible and ideal for my purposes (section 3.6). EB differentiation is simpler, cheaper, and as revealed in this work, a faster system (in comparison to NPC differentiation) for studying the establishment and dynamics of CGI methylation on Xi. Calibrated MBD-seq, due to its high signal-to-noise ratio provided a sensitive assay to capture small changes in CGI methylation. The sensitivity of MBD-seq is facilitated by the absence of a signal on CGIs on Xa; resulting in large shifts in allelic ratio when compared to alternative assays like BioCAP-seq. The calibration of MBD-seq with exogenous human DNA allows quantitative analysis between different timepoints during differentiation.

I was able to perform for the first time a chromosome-wide analysis of the dynamics of X-linked CGI methylation during XCI. Allelic analysis confirmed that CGI methylation on Xi occurs in two phases: fast and slow, with the majority of CGIs on Xi falling in the slow methylating category. The exact timing of CGI methylation varies depending on the differentiation system used (day 4-8 in EBs versus day 12-15 in NPCs). This finding agrees with the previous report from our lab (Gendrel et al., 2012). Gendrel et al., also found that methylation of fast CGIs occurs before the recruitment of SMCHD1. In this study I have observed similarly that the majority of the SMCHD1-dependent CGIs show slow methylation dynamics. A minority of SMCHD1 dependent CGIs showed fast dynamics in my analysis, which could be due to the faster dynamics of XCI in the iXist-ChrX line in comparison to MEFs derived from embryos which use the native Xist promoter, used to define SMCHD1-dependent CGIs (Gdula et al.,2019). Fast CGIs are mostly associated with genes that have low gene expression in ES cells. These genes tend to get silenced earlier, which could in turn make them permissive for establishment of CGI methylation earlier than late silencing genes with high expression levels (Bowness et al., 2022). In agreement with this model, I found a correlation between the gene silencing dynamics and onset of CGI methylation during XCI, i.e. the genes that have been reported to be silenced late (Bowness et al., 2022) are linked to slow methylating CGIs. However, the opposite trend i.e., correlation between fast CGIs with fast silencing gene was not very obvious.

MBD-seq analysis of autosomal CGIs in wt iXist-ChrX was consistent with an earlier study conducted in mouse embryos. For instance, the majority of CGIs in the genome remained hypomethylated after differentiation (Figure 6.1 A). CGIs reported to be methylated in post implantation embryos such as those associated with germline genes (*Brdt*, *Rhox13, Taf7l, AU022751*), meiosis (*Spo11, Sycp1/2/3, Syce1/2/3)* , piRNA pathway (*Piwi1/2/3, Ddx4, Asz1*) and key developmental transcription factors (*Klf3, Daxx, Foxo3, Bcl11a/b, Zfp64*) and several genes of Wnt/β-catenin pathway, were also found methylated in differentiated iXist-ChrX cells (Auclair et al., 2014). Auclair et al, observed that the proportion of methylated CGIs associated with TSS is lower than those residing within intergenic and intragenic regions in post-implantation embryos. My data agrees with this finding (Figure 6.1 B). In

agreement with Auclair et al, I observed that the level of methylation observed at CGIs associated with promoters was lower than those found intergenically (Figure 6.1 C).

Further improvement in this experimental system could be achieved by using WGBS which provides a more accurate quantitative measure of the methylation levels of all cytosines in the genome. RRBS also provides a larger coverage of CpGs in comparison to MBD-seq, making it possible to analyse the methylation status of some gene bodies in addition to CGIs. The latter could be of particular interest for gaining more details on the possible redistribution of CpG methylation by DNMT3B mutants. Since the main utility of this experimental system was to understand the role of various isoforms and domains of DNMT3B in X-linked CGIs, I opted for MBD-seq for this study.

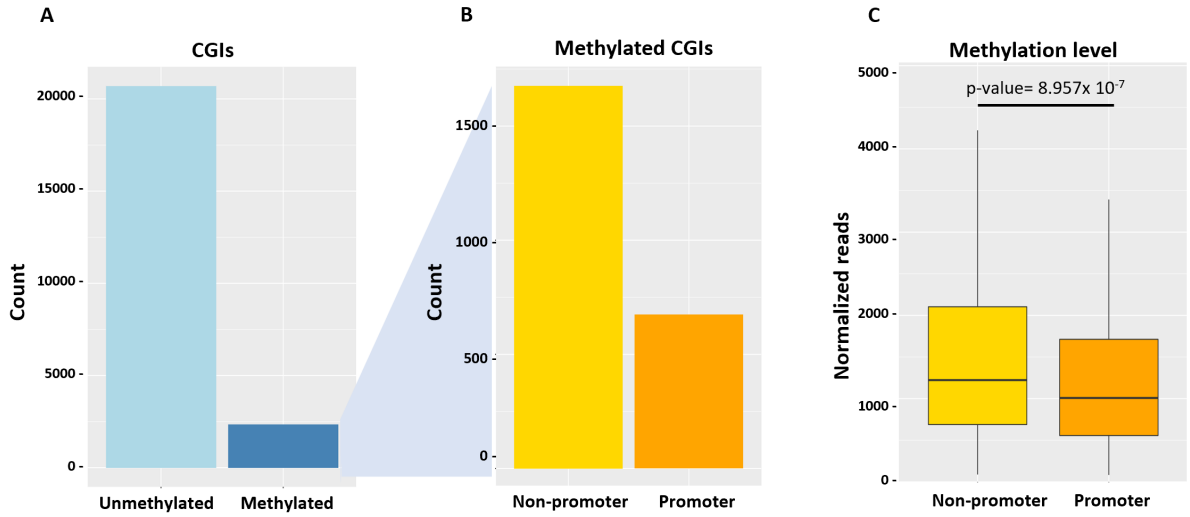


Figure 6.1: The proportion of promoter-associated and inter/intragenic CGIs among those gaining methylation during EB differentiation.

A) Bar plot shows the proportion of CGIs that get methylated during differentiation. Only a small fraction of CGIs (10.2 % of the total) gain methylation.
B) Out of the methylated CGIs, the proportion of promoter-associated CGIs is smaller than those found intergenically and intragenically (denoted here as non-promoter category).
C) The methylation level in EBs for promoter-associated CGIs is generally lower than non-promoter CGIs.

6.2 Characterization of DNMT3B-dependent CGIs on Xi and genomewide

Gendrel et al., using systematic deletions of *Dnmt3a, Dnmt3b* and *Dnmt3l* in E9.5 mouse embryos reported that DNMT3B is responsible for CGI methylation on Xi (Gendrel et al., 2012). This methylation analysis was performed on 7 selected X-linked CGIs using Sequenom EpiTyper analysis. Later Auclair et al., validated that the specificity of DNMT3B was true for the majority of X-linked CGIs using RRBS analysis of female *Dnmt3b* null E8.5 embryos (Auclair et al., 2014). My data agrees with previous reports. Apart from a few exceptions (15 CGIs), the majority of X-linked CGIs are fully dependent on DNMT3B for their methylation. Table A1 in Appendix 5 summarizes the key information regarding X-linked CGIs, their associated genes and dependence on DNMT3B and SMCHD1.

On the autosomes, well-characterized targets of DNMT3B such as several germline expressed genes, protocadherin genes, genes associated with Wnt/β-catenin pathway and *Dppa3* were hypomethylated in my *Dnmt3b*$^{-/-}$ line (Auclair et al., 2014; Yagi et al., 2020). CGIs hypomethylated in ICF patients such as *Zbtb22, Slc25a31, Slc8a2, Trim42, Ly6k, Tdrd1/6, Rem1, Meioc, Srrm2/4, Znrf4, Jph4, Neurl1b, Mapk8, Kcng4* and *Tnxb*) were also identified as DNMT3B-dependent CGIs in this study (Yagi et al., 2020). Thus, these findings further validate the iXist-ChrX EB differentiation system as a model for the developmental regulation of DNA methylation *in vivo*.

6.3 Expansion of the current definition of CGIs

Originally, CGIs were defined based on predictive algorithms that identified regions with elevated CpG content as CGIs (i.e., based on high observed/expected ratio for CpGs) (Chinwalla et al., 2002). Using this criterion, CGIs detected in the mouse genome were reported to be frequently associated with gene promoters. Using the predictive algorithm Chinwalla et al., study reported that mouse genome harboured half the number of CGIs as compared to the human genome (15500 CGIs in the mouse genome compared with 27000 CGIs in the human genome), suggesting a lack of conservation of CGIs and, by extension, questioning the significance of CGIs. This difference in the number of mouse and human CGIs was previously also noted by Adrian Bird (Antequera and Bird, 1993). Illingworth et al., then experimentally determined CGIs using CAP-seq and reported many more CGIs in the mouse genome (23044 experimentally determined CGIs in comparison to 15500 predicted CGIs) and showed that the number of CGIs in mouse was not only comparable to those in humans but also showed high conservation among syntenic regions (Illingworth et al., 2010). One-fifth of the experimentally

determined CGIs failed to qualify the computational criteria of CpG observed/expected ratio being 0.6 or greater, proving that the computational predictions were too stringent.

The total number of CGIs defined in the genome are prone to high variability based on the method used. Illingworth et al., for instance, comment that they were able to capture > 7000 additional CGIs by making the elution conditions for CAP-seq less stringent (Illingworth et al., 2010). Similarly, Long et al., using a similar but improved BioCAP-seq method expanded the previous definition by Illingworth et al., and reported 27699 CGIs in the mouse genome (termed as NMIs in the publication). In both publications, about half of these CGIs were found to be residing within intragenic and intergenic regions in both human and mouse genomes. These regions/non-promoter CGIs were shown to either correlate with chromatin modifications like H3K4me1 and H3K4me3, or RNA Pol II and low-level transcripts, and were frequently methylated during development and tissue specification.

My MBD-seq experiments reproducibly revealed a number of regions that have previously not been annotated as CGIs. One class of these methylated CpG-rich regions resided within 2 kb of TSS (1922 regions, 120 on X chromosome), therefore resembling “classical” or “canonical” CGIs. Another class of the MBD-seq signal was > 2 kb away from TSS but overlapped regions of high chromatin accessibility (defined by ATAC-seq dataset from Dr. Joseph Bowness performed on iXist-ChrX line. 4890 regions, 374 on the X chromosome) (Figure 6.2 A). Owing to the high GC content and correlation with regions with high chromatin accessibility, these regions could correspond to CREs. Indeed, 25% of the regions in this intergenic/intragenic MBD-seq signal class also correlated with known CREs as defined by ENCODE project based on H3K4me3, H3K27ac, and ChIP-seq data for TFs (Figure 6.2 B). In this study, I, therefore, propose that these are CGIs, corresponding with regions potentially involved in gene regulation, that have not been included in previous annotations. In my experiments, these regions are hypomethylated in ES cells and gain methylation upon differentiation. It is possible that these regions escaped detection by CAP-seq based methods used by (Illingworth et al., 2010; Long et al., 2013) due to low signal, but were brought to attention in this study by a more sensitive MBD-seq assay upon gaining methylation in EBs. A comparison of ATAC-seq, H3K4me1, H3K4me3, and H3K27ac ChIP-seq in ES cells with EBs would further validate if these regions truly represent enhancers that undergo silencing /inactivation upon differentiation. It would also be interesting to test if these regions are associated with lowly expressed RNAs (such as eRNAs reported to be expressed from enhancers) which are then silenced in EB using assays like CAGE-seq (cap analysis of gene expression).

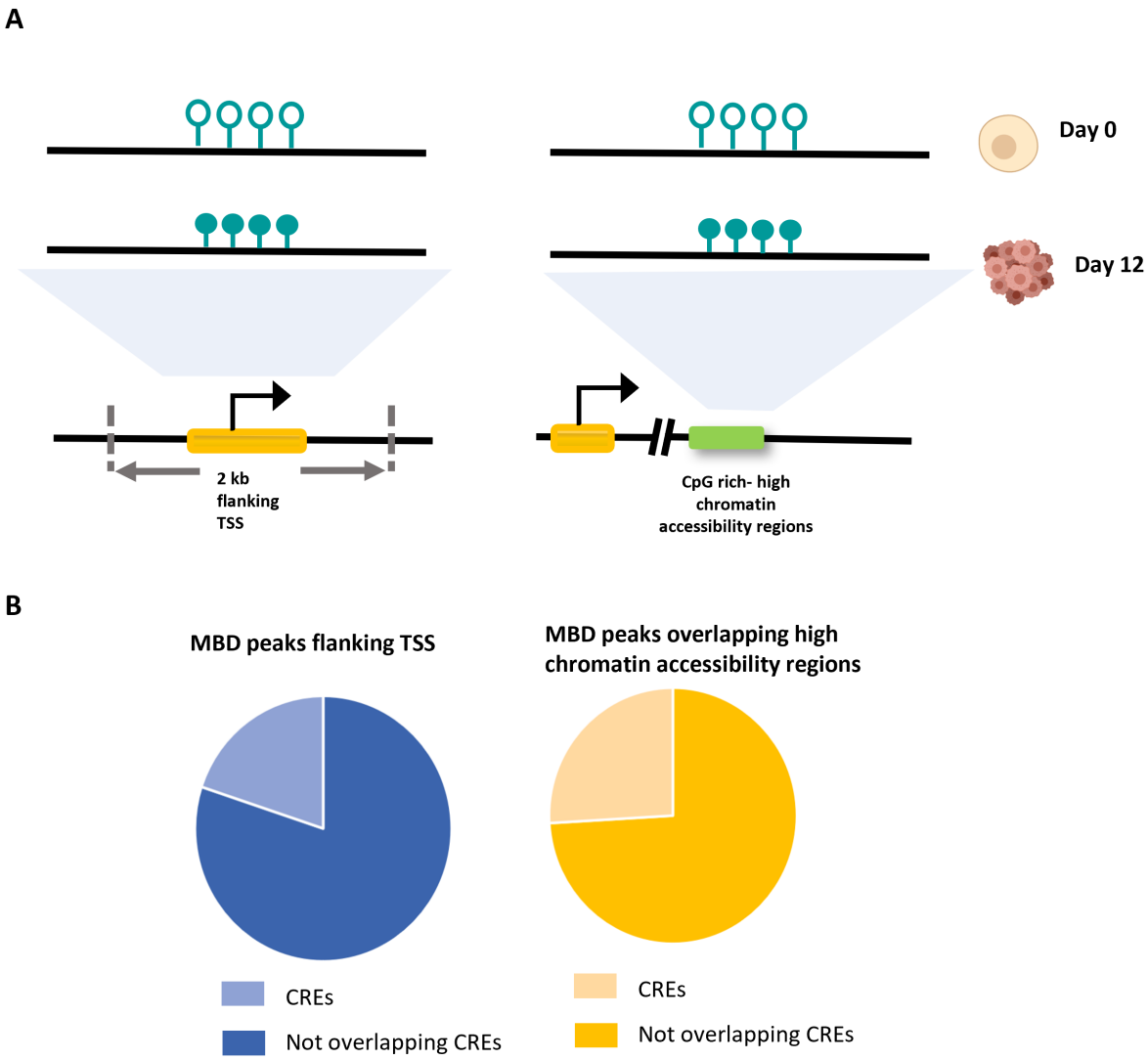


Figure 6.2: Discovery of additional CGIs by MBD-seq.

A) MBD-seq reproducibly captures two classes of regions (left) TSS-associated regions are found within 2 kb flanking TSS and (right) regions > 2 kb away from TSS that correspond to high chromatin accessibility. These regions are likely to represent gene regulatory regions and therefore likely to be CGIs. Both of these regions gain methylation upon differentiation. This figure summarizes the results described in section in 3.9.

B) Pie chart showing the proportion of CREs among the TSS- and high chromatin accessibility regions as defined by ENCODE project.

6.4 Methylation of "CGI-like elements" on Xi is DNMT3B-independent

Further investigation of reproducible MBD-seq peaks from intergenic regions revealed the presence of "CGI-like elements" on the X chromosome. "CGI-like elements" reside more than 2 kb away from TSS and correlate with regions showing H3K4me1 enrichment and high chromatin accessibility; suggesting these could be CREs. Allelic analysis showed that methylation gained upon differentiation on these regions is skewed towards Xi. These regions could correspond to enhancer elements that, after the establishment of XCI, are not being used and passively gain DNA methylation. DNA methylation at a subset of enhancers during mouse and human embryogenesis has previously been reported (Li et al., 2018).

A subset of these "CGI-like elements" overlapped with transposons. The Xi-selective gain of DNA methylation on transposons is unexpected. One possibility is that chromosome-wide gene silencing and heterochromatinization leads to the selective gain of DNA methylation at these regions. I have also discovered Xi specific methylation of another group of regions called "rising peaks" which mainly reside within transposons and heterochromatin. Unlike "CGI-like elements", "rising peaks" are partially methylated in ES cells. "Rising peaks" also gain methylation during differentiation and show bias towards Xi in comparison to Xa. "CGI-like elements" and "rising peaks" share some common features. (1) They both show an early gain of methylation (i.e., between day 0-day 4 of EB differentiation) in comparison to CGIs (that gain methylation between day 4-day 8 of EB differentiation, Figure 6.3 A), (2) they have lower CpG density in comparison to annotated CGIs, and (3) these regions gain methylation independent of DNMT3B (Figure 6.3 B and C). In their study on mouse embryos, Auclair et al., also observed that CGI methylation was slightly delayed in comparison to bulk genome such as gene bodies, transposable elements, and intergenic regions (Auclair et al., 2014). They also reported that this delay in methylation was most prominent for the CGIs associated with TSS. Similarly, Auclair et al., mention that the methylation of CGIs showing fast dynamics is less affected in *Dnmt3b*$^{-/-}$ embryos and can be compensated by DNMT3A (Auclair et al., 2014).

Whether or not the DNA methylation on "CGI-like elements" achieves a purpose in XCI is an open question. ChrRNA-seq shows that "CGI-like elements" overlap with low expression transcripts which are silenced on Xi in EBs. Methods like CAGE-seq will provide further insight into the nascent lowly transcripts expressed from "CGI-like elements". As for other yet-unannotated CGIs observed in this study, analysis by ATAC-seq, H3K4me1/3 ChIP-seq during the differentiation timecourse would further validate the existence of gene regulatory regions corresponding to "CGI-like elements".

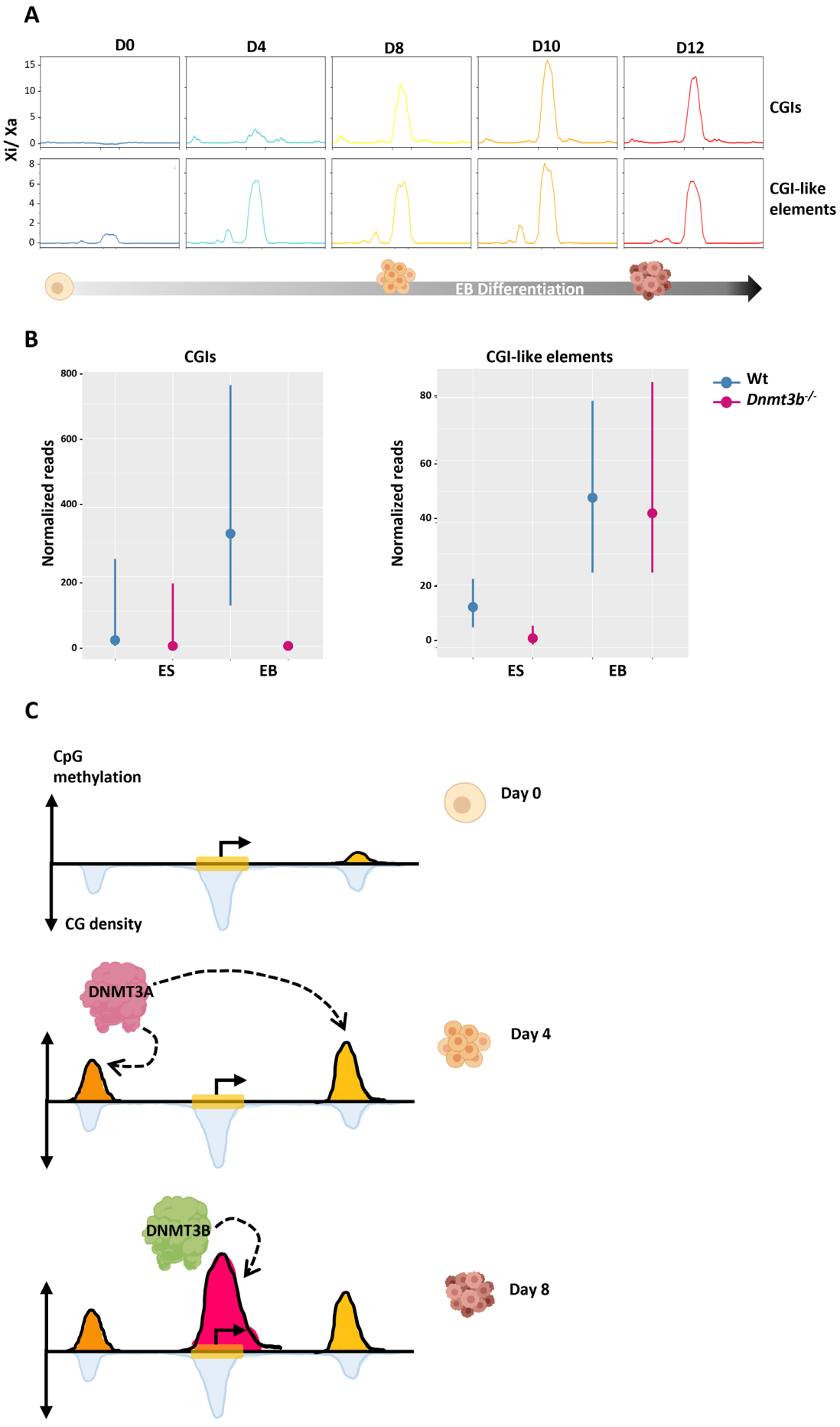

A
D0
D4
D8
D10
D12
Xi/ Xa
CGIs
CGI-like elements
EB Differentiation
B
CGIs
CGI-like elements
Normalized reads
Wt
Dnmt3b-/-
ES
EB
C
CpG methylation
CG density
Day 0
DNMT3A
Day 4
DNMT3B
Day 8

Figure 6.3: "CGI-like elements" gain methylation on Xi independent of DNMT3B (shown on the previous page).

A) Plots show that gain of methylation (MBD-seq signal) during EB timecourse on CGIs and "CGI-like elements". "CGI-like elements" gain methylation between day 0 and day 4. Days of the timecourse are denoted by D0-D12.
B) Plots show methylation levels on "CGI-like elements" and annotated CGIs in wt (blue) and *Dnmt3b*$^{-/-}$ (pink). Unlike CGIs, DNA methylation at "CGI-like elements" remain unaffected in *Dnmt3b*$^{-/-}$ cells. In these plots, median value is represented by dot and the line shows interquartile range.
C) Schematic showing gain of DNA methylation at "CGI-like elements" (left), and annotated CGIs (middle) on Xi. "CGI-like elements" have lower CpG content and get methylated earlier than annotated CGIs. Repeat-associated "rising peaks" also partially methylated in ES cells and show earlier/faster gain of methylation in comparison to annotated CGIs (right).

6.5 DNMT3B isoforms and specificity

Confirmation of the dependence of X-linked CGI methylation on DNMT3B, together with the development of a tractable cell culture-based assay laid a foundation for complementation assays to define DNMT3B isoforms/domains that confer specificity towards Xi CGI methylation, described in chapter 4. In this work, I made use of a complementation strategy to test the contribution of individual isoforms of DNMT3B in CGI methylation on Xi. I have shown that the major catalytic form, DNMT3B1, is able to fully complement Xi CGI methylation and that the catalytically inactive DNMT3B3 isoform is completely dispensable. The full rescue of CGI methylation on Xi in the *Dnmt3b1* complementation cell line (*Dnmt3b1-KI*) also indicates that all other isoforms of DNMT3B are dispensable for this process. Additionally, I found that methylation of DNMT3B-dependent autosomal CGIs was rescued by DNMT3B1 and not DNMT3B3, suggesting common mechanisms underlying DNMT3B-mediated methylation on both Xi and autosomal CGIs. A caveat of the complementation experiment is variable expression levels of rescue constructs despite being expressed from the same constitutive pCAG promoter. The protein levels of DNMT3B1 were much lower than DNMT3B3 as apparent by western blots. However, the protein levels of DNMT3B1, although lower than other rescue constructs used in this work, were comparable to those in the wt iXist-ChrX line, and as evident from MBD-seq, sufficient to rescue DNMT3B-dependent methylation genomewide.

My findings are important in light of recent work that reported that catalytically inactive DNMT3B (denoted as DNMT3B^{CI}) rescues developmentally regulated CpG methylation at many sites in the genome in *Dnmt3b* mutant embryos (Nowialis et al., 2019). It was suggested that there are accessory roles of DNM3B, other than its catalytic functions, which were responsible for the observed rescue of

CpG methylation. I analysed CpGs reported to have been rescued by DNMT3B^{CI}, in the *Dnmt3b3-KI* complementation line, and I was not able to find an equivalent rescue/accessory function of DNMT3B in my experimental system. A few limitations of my analysis are that Nowialis et al., used RRBS for profiling CpG methylation, which has much larger genomic coverage than MBD-seq. I was therefore able to capture only 20% of the CpGs claimed to have been rescued by DNMT3B^{CI} in my MBD-seq. Secondly, Nowialis et al., analysed the rescue of CpG methylation by DNMT3B^{CI} in E11.5 mouse embryos unlike the cell-culture based system used in this study, which is expected to have higher/complete CpG methylation in comparison to day 12 EBs.

It is also important to consider that DNMT3B3 and DNMT3B^{CI} used by Nowialis et al., are not directly comparable. DNMT3B^{CI} is full-length DNMT3B1 with two-point mutations (P656V/C657D) which were claimed to disrupt the catalytic activity of DNMT3B. DNMT3B3, although catalytically inactive, is structurally distinct from DNMT3B1. Although Nowialis et al., inferred that DNMT3B^{CI} is catalytically inactive, the expression of DNMT3B^{CI} in a *Dnmt3a/3b* double knockout lymphoma cell line led to methylation of some CpGs (albeit 10-fold lower than wt DNMT3B), suggesting that some catalytic activity of DNMT3B^{CI} was intact. Taken together these considerations indicated that DNMT3B^{CI} is a hypomorphic mutation that retains some catalytic activity. It would be interesting in the future to directly test the effect of expressing DNMT3B^{CI} in my cell culture model system.

6.6 A role for the DNMT3B N-terminal domain in CGI methylation

The N-terminal domain of DNMT3B was a good candidate as a specificity-imparting region given the poor conservation among the N-terminal domains of DNMT3A and DNMT3B. Complementation experiments with cell lines expressing complete or partial deletions of the DNMT3B1 N-terminal domain revealed that the N-terminal domain is crucial for CGI methylation on Xi and DNMT3B-dependent autosomal CGIs. Prior analysis of the N-terminal domain of DNMT3B is limited so the function of this region is largely unknown. A complete structure of DNMT3B or DNMT3A has not been solved. Based on the amino acid sequence, the structure prediction suggests that the N-terminal domain of both DNMT3A and DNMT3B forms a flexible unstructured loop. Collectively, the literature regarding the N-terminal domain of DNMT3A and DNMT3B suggests two potential modes of action of the N-terminal domain. Jeong et al., in 2009 first published that the N-terminal domain of DNMT3B serves as a general chromatin binding domain (Figure 6.4). Considering this information, one can assume a model that the N-terminus is required for the association of DNMT3B with chromatin,

serving as general chromatin binding hook and Δ1-218 DNMT3B1, therefore, fails to rescue CGI methylation despite the evidence of fully intact catalytic activity (Zeng et al., 2020). The smaller deletions in the N-terminal domain could disturb the structure of this hook and result in the failure of DNMT3B to engage with the chromatin. I have presented strong evidence that complete or partial deletions of the N-terminus do not weaken the chromatin association of DNMT3B.

A second model of the N-terminal domain function comes from the literature on DNMT3A, which suggests the N-terminal domain to be composed of functional subdomains. The evidence first came from Suetake et al., in 2011 who reported that the N-terminal domain of DNMT3A mediated its binding to DNA by showing that DNMT3A1, the isoform possessing a longer N-terminal domain, bound to DNA with greater strength than DNMT3A2 which has a shorter N-terminus. This study showed that the N-terminal domain alone was sufficient to bind DNA. By testing the DNA binding strength of various N-terminal domain fragments, Suetake et al., showed that a fragment composed of residues 1-57 showed the highest DNA-binding strength while fragment spanning 161-211 residues possessed a weak DNA binding ability. These fragments harboured basic residues K51, K53, R177 and R179 that were shown to have a major role in mediating DNA binding. Interestingly, residues 58-146 were suggested to antagonize DNA binding since N-terminal domain fragments of 1-161 or 1-146 did not have any DNA binding ability. It was not tested if, in the full protein context, deletion of residues 1-57 or 161-21 or mutation of 4 basic residues within these regions, would result in a substantial impact on the chromatin binding strength of DNMT3A. The indication of a similar subdomain in DNMT3B N-terminus has been presented in a study by (Molaro et al., 2020), which points out the presence of a conserved RSRRRR motif in the mammalian DNMT3B N-terminus (spanning residues 160-165 in the mouse protein) suggesting an unknown and critical role. Recently, Weinberg et al., 2021 suggested the presence of a ubiquitin-dependent recruitment region (UDR) between residues 160-219 in the N-terminal domain of mouse DNMT3A which mediates the interaction of DNMT3A with H2AK119ub1 modified nucleosomes *in vitro* (Figure 6.4). This paper suggests two competing mechanisms of DNMT3A recruitment. In mesenchymal stem-like cells (MSCs, used in this study) the UDR-mediated recruitment of DNMT3A to H2AK119ub1 modified CGIs only became apparent in the absence of or with mutations of the PWWP domain, which predominantly mediates DNMT3A localization to H3K36me3 modified genic regions. The presence of this H2AK119ub1 binding domain (spanning residues 189-206 in the mouse DNMT3A1) was also later reported by (Gu et al., 2022). Studies by (Gu et al., 2022; Weinberg et al., 2021) together with work from (Manzo et al., 2017) suggest that the N-terminal domain of DNMT3A is composed of functional subdomains which govern its recruitment to

H2AK119ub1 modified chromatin rather than serving as a general chromatin/DNA binding region as suggested earlier by Jeong et al., 2009 or Suetake et al., 2011.

Complementation cell lines expressing smaller deletions in the N-terminal domain of DNMT3B1 in this work, were aimed at revealing subregions of DNMT3B1 N-terminus that are critical for methylation of X-linked CGIs. However, none of the smaller N-terminal deletions in DNMT3B1 rescued CGI methylation mediated by DNMT3B. Thus, my data argues against the model of DNMT3B1 N-terminus being composed of submodular domains/functional motifs that could influence DNMT3B mediated CGI methylation. One deletion in the N-terminus, (*ΔC Dnmt3b1-KI, Δ98-147 Dnmt3b1-KI*), rendered the DNMT3B1 protein unstable.

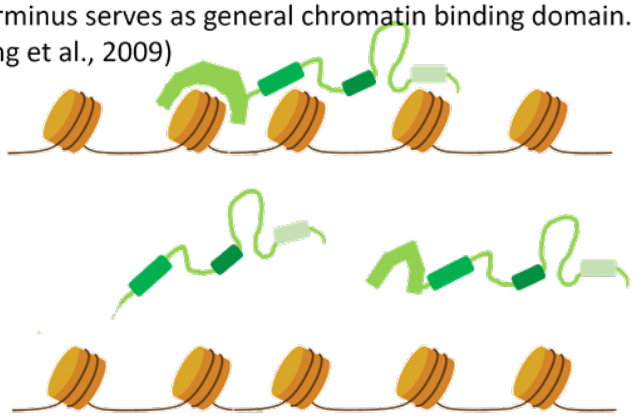
I have also shown that chromatin binding strength largely remained unaffected by the small deletions in the N-terminal domain. It is possible that deletions in the N-terminal domain led to the redistribution of DNMT3B1 to other genomic locations causing ectopic/hypermethylation which was not captured with MBD-seq. ChIP-seq and methods like WGBS or RRBS which have a larger genomic coverage could reveal the differential localization and methylation performed by N-terminal domain deletion mutants of DNMT3B1.

Some of my observations hint towards the distinct contribution/influence of individual N-terminal domain regions on DNMT3B1 function. For example, a common trend observed among N-terminal domain mutants of DNMT3B1 was the hypermethylation of intergenic regions and exons in comparison to the *Dnmt3b*$^{-/-}$ line. This hypermethylation was 5-6-fold reduced in *Δ1-50* and *Δ51-98 Dnmt3b1-KI* (*ΔA* and *ΔB Dnmt3b-KI*) in comparison to *Δ146-177* and *Δ178-229 Dnmt3b1-KI (ΔD and ΔE Dnmt3b1-KI* lines, see Appendix 4). Similarly, salt extraction experiments suggest a weaker chromatin association of Δ1-50 and Δ51-98 DNMT3B1. However overall, these observations present weak evidence for a model that DNMT3B1 N-terminus has functional subdomains. The exons found hypermethylated by N-terminal domain mutants of DNMT3B1 make only a small fraction of MBD-seq signal (< 1200 out of ~25000 CpGs analysed genomewide which include CGIs and unannotated CpG rich regions found both inter and intragenically). Similarly in all the N-terminal deletions of DNMT3B1, ~90% of the protein was found strongly associated with chromatin; and so irrespective of the region deleted, all the N-terminal deletion mutants of DNMT3B1 largely functioned similarly. Further investigation, particularly a structure of the DNMT3B1 N-terminal would help in understanding the impact that complete or partial deletions have on DNMT3B1 function.



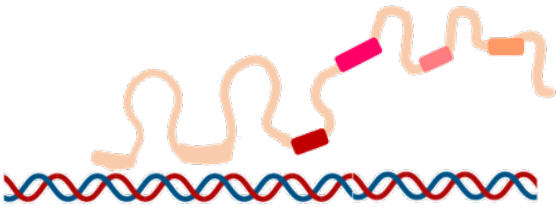
1

N-terminus serves as general chromatin binding domain.
(Jeong et al., 2009)



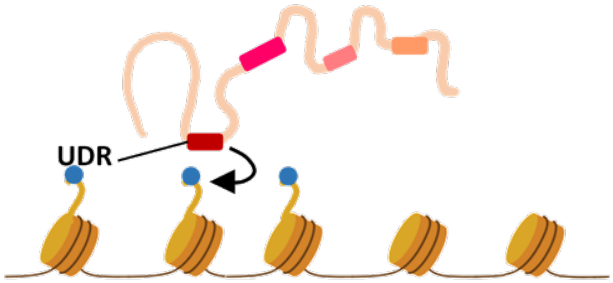
2

N-terminus harbours DNA binding domains rich in basic residues.
(Suetake et al., 2011)



3

N-terminus contains a ubiquitin-dependent recruitment region (UDR) to bind H2AK119ub1 modified CGIs
(Weinberg et al., 2021 and Gu et al., 2022)



4

This work | N-terminus mediates localization of DNMT3B to CGIs

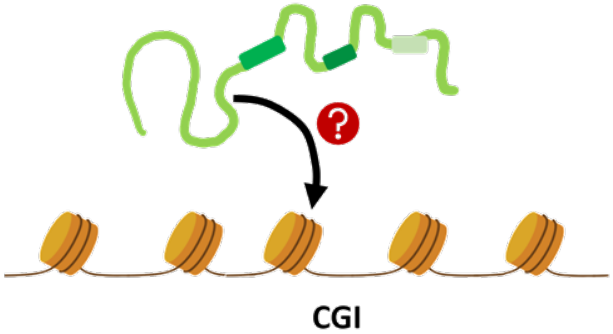


Figure 6.4: Suggested models for the role of the N-terminus in DNMT3A and DNMT3B (shown on the previous page).

1) Jeong et al., showed that the N-terminal domain of DNMT3B is important for its association with chromatin. In the schematic, the N-terminal domain of DNMT3B has been shown as a chromatin-binding hook. According to this study removal of the N-terminal domain or smaller deletion in this hook would result in dissociation of DNMT3B from chromatin binding.
2) Suetake et al., suggest that the N-terminal domain of DNMT3A act as a DNA binding domain. Some regions of the DNMT3A N-terminus have a stronger affinity for DNA than others.
3) Recent work by Weinberg et al., and Gu et al., suggest context-specific role N-terminal domain of DNMT3A. They show that N-terminus of DNMT3A mediates its recruitment to H2AK119ub1 enriched CGIs through a small, ordered region called ubiquitin-dependent recruitment domain (UDR).
4) ChIP-seq data in this work shows that deletion of N-terminus of DNMT3B results in loss of its residence on CGIs. The underlying mechanism or the region within the N-terminus determining this function was not found.

Since the results from smaller N-terminal domain deletions support that the entirety of DNMT3B N-terminus is crucial for CGI methylation, theoretically switching the N-terminal domain of DNMT3A with that of DNMT3B (i.e *Dnmt3b(N)3a* complementation cell line) should have rescued CGI methylation mediated by DNMT3B. Similar experiments have been conducted in the past studies, for example, Weinberg et al., 2021 have shown that replacing the N-terminal domain of DNMT3B with DNMT3A causes it to be recruited to H2AK119ub1 enriched CGIs. Although theoretically the chimeric DNMT3B(N)3A construct was carefully designed so that based on amino acid sequence, the transition of DNMT3B N-terminus into DNMT3A C-terminus could not be told, loss/disruption of catalytic activity of resulting protein due to this fusion cannot be ruled out.

In the future, additional rescue constructs can be tested using the complementation strategy. For instance, given the evidence of specific localization of the longer isoform DNMT3A1 on H2AK119ub1 enriched CGIs, it would be interesting to perform MBD-seq on complementation lines expressing chimeric DNMT3B where DNMT3B N-terminal is switched with that of DNMT3A1. If this chimeric DNMT3B rescues CGI methylation on Xi, this could allude towards the role of H2AK119ub1 enriched on Xi in methylation of X-linked CGIs.

6.7 Contribution of N-terminal domain to DNMT3B nuclear localization

Since the N-terminal domain of DNMT3B has been shown to be dispensable for catalysis *in vitro* (Zeng et al., 2020), the inability of the N-terminal deletion mutant of DNMT3B1 to rescue CGI methylation indicates a localization/targeting defect in Δ1-218 DNMT3B1. Since the overall chromatin association of Δ1-218 DNMT3B1 was not compromised, I performed ChIP-seq to test the differential localization of Δ1-218 DNMT3B1 as compared to wt. I have shown that deletion of the N-terminus resulted in compromised localization of DNMT3B1 on CGIs on the Xi. IF experiments revealed that localization of Δ1-218 DNMT3B1 is retained on the pericentric heterochromatin, suggesting that the role of the N-terminus of DNMT3B1 is context specific. A qualitative analysis of IF images showed that the punctate signal of Δ1-218 DNMT3B1 on pericentric heterochromatin became sharper than wt which could be due to the redistribution of Δ1-218 DNMT3B1 towards pericentric heterochromatin. Higher expression Δ1-218 DNMT3B1 in comparison to wt could be another explanation for the relatively clear punctate signal on pericentric heterochromatin.

In agreement with a previous report (Baubec et al., 2015), I observed that wt DNMT3B1 ChIP-seq signal gradually rises from the TSS towards the TTS, mirroring/correlating with the co-transcriptional accumulation of H3K36me3 on gene bodies. Unlike wt DNMT3B1, Δ1-218 DNMT3B1 did not show higher localization around the TTS. Similar aberrant localization of DNMT3B on gene bodies is associated with mutations in the PWWP domain which mediates the recruitment of DNMT3B to H3K36me3 enriched regions (Baubec et al., 2015). This latter observation suggested that the deletion of the N-terminal domain might have led to the loss of PWWP domain function or integrity. Support for this comes from the study by Baubec et al., which showed that the N-terminal domain is required for the PWWP domain fragment of DNMT3B (amino acids 227 to 343 of the mouse DNMT3B1) to bind H3K36me3 (Baubec et al., 2015). This dependence of the PWWP domain on the N-terminus has not been validated in the context of full-length DNMT3B. Since the mutations in the DNMT3B PWWP domain lead to loss of punctate signal on pericentric heterochromatin (Chen et al., 2004) which remains intact upon deletion of DNMT3B1 N-terminus, I assume that the PWWP domain is functional in Δ1-218 DNMT3B1 and that the N-terminal domain deletion does not phenocopy PWWP domain mutations. Taken together, my experiments show that the N-terminal domain of DNMT3B1 is crucial for CGI methylation mediated by DNMT3B. The N-terminal is necessary for the localization of DNMT3B1 on CGIs and gene bodies but not for localization on pericentric heterochromatin. The

mechanisms via which the N-terminus mediates the localization of DNMT3B1 on CGIs are not clear. An attempt to narrow down the functionally critical region of the DNMT3B1 N-terminal domain has not yielded informative results. Analysis of the effect of point mutations in the DNMT3B N-terminus on CGI methylation may overcome the confounding factor of conformational changes that smaller deletions in DNMT3B1 N-terminus could have caused. Mutation of the conserved RSRRRR motif (spanning residues 160-165 in mouse DNMT3B1) to hydrophobic residues such as alanine, for instance, could serve as a good future experiment.

6.8 Analysis of the impact of DNMT3B1 PWWP point mutant reported in ICF

In my search for the specificity-imparting domain of DNMT3B1, I explored mutations reported in ICF. Hansen et al., reported that ICF leads to hypomethylation on X-linked CGIs in female patients (Hansen et al., 2000), which makes the causative mutations of ICF of interest for this study. Most of the causative mutations of ICF either lead to loss of DNMT3B (due to nonsense mutations in the start of the protein) or missense mutations in the catalytic domain which are thought to result in hypomorphic enzyme (Hamidi et al., 2015). My observation that the catalytically active isoform DNMT3B1 is crucial for CGI methylation on Xi makes good sense in light of these studies. An exceptional case however is the point mutation in the PWWP domain (S277P) reported in three ICF patients (Shirohzu et al., 2002). Since these patients showed diagnostic features typical of ICF (i.e., hypomethylation pericentric repeats, the fusion of chromosomes 1, 9, and 16, alongside mental retardation and high vulnerability to infections), it is expected that X-linked CGIs would also depict hypomethylation in these patients; this, however, had not been tested. Furthermore, unlike other missense mutations responsible for ICF, S277P is not expected to disturb the catalytic activity of DNMT3B but is clearly crucial for DNMT3B function. To gain insight into the effect of S277P DNMT3B1, I generated a complementation line expressing S277P DNMT3B1. As expected, this point mutation did not rescue CGI methylation on Xi nor on other DNMT3B-dependent CGIs on autosomes. My experiments reveal that S277P mutation causes loss of DNMT3B1 protein stability which could in turn be due to misfolding of the S277P DNMT3B1. I also performed ChIP-seq for S277P DNMT3B1 which revealed that (1) unlike Δ1-218 DNMT3B1, S277P DNMT3B1 showed comparable localization on CGIs on Xi and (2) as reported by Baubec et al., 2015, S277P DNMT3B1 showed poor localization on gene bodies. However, I have not presented this data here since the low expression levels of S277P DNMT3B1 add more noise to DNMT3B ChIP-seq which already has a high background.

6.9 Contribution of H3K9me3 in CGI methylation

One consideration in investigating the mechanisms governing DNMT3B specificity for CGI methylation on Xi is the possible role of chromatin modifications specific to DNMT3B compared to DNMT3A target sites. I reasoned that a differential ChIP-seq analysis could indicate the binding sites for DNMT3A and DNMT3B and I could then search for the chromatin modifications that mark their target binding sites. However due to the broad distribution of DNMT3A and DNMT3B binding determined by ChIP-seq, the definition of DNMT3A- and DNMT3B-specific target sites was not straightforward. For this analysis I measured the ChIP-seq signal over 1 kb windows across the genome and used DESeq2 to calculate the differentially occupied regions between DNMT3A and DNMT3B. As a reference, I processed the published dataset by Baubec et al., using the same pipeline (Baubec et al., 2015). Although I was able to find differential regions, the FDR values were well above ideal (FDR < 0.1) for this ChIP-seq (both for my data and for the dataset from Baubec et al., 2015). Contrary to expectation, DNMT3B ChIP-seq signal did not show enrichment on X-linked CGIs in EBs and the ChIP-seq signal did not vary significantly at X-linked CGIs between ES and EBs. One explanation could be that DNMT3A and DNMT3B occupation on chromatin has fast kinetics which are difficult to capture by IP-based methods. In ES cells DNMT3A and DNMT3B are not excluded from CGIs but rather appear to sample all sites across the genome. This observation begs the question as to why CGIs are normally hypomethylated. One hypothesis is that CGIs overlap the promoters of actively transcribing genes and therefore regions enriched in H3K4me3 which inhibits DNMT3 proteins from carrying out methylation due to the inhibition of the catalytic domain with the ADD domain. Thus, during differentiation, these regions must become silenced, lose H3K4me3 and therefore allow DNMT3B to then methylate CGIs. This makes sense especially considering XCI (Figure 6.5).

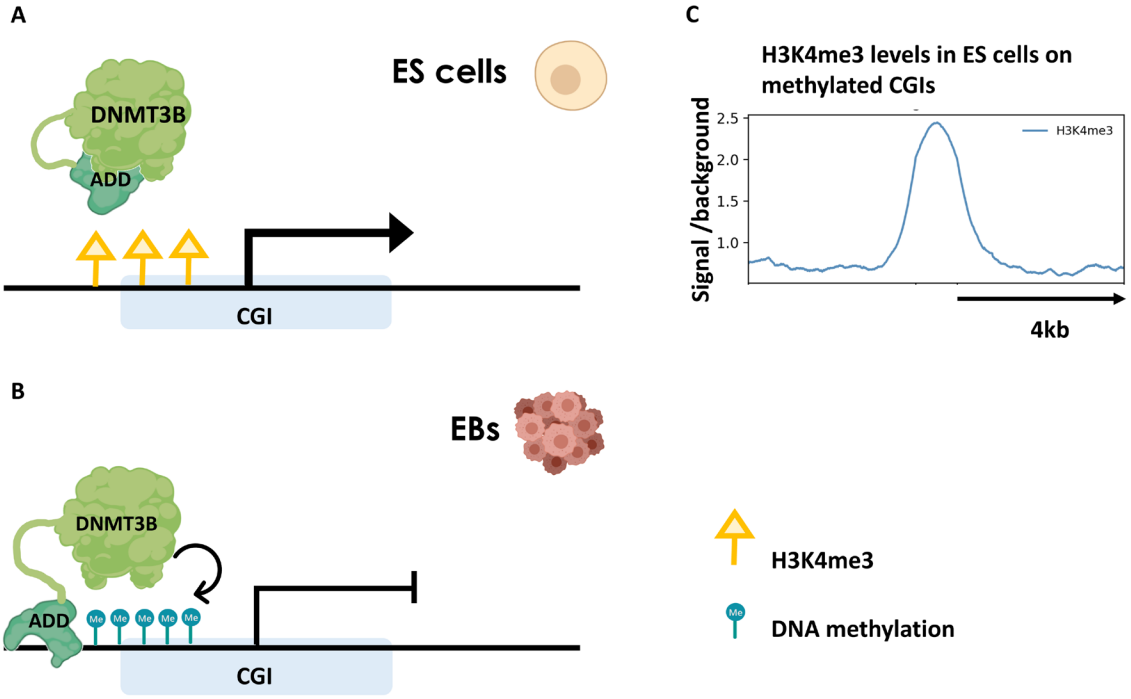


Figure 6.5: Models for maintenance of a hypomethylated state of CGIs in ES cells.

A) In ES cells, due to ongoing transcription, CGIs are modified with H3K4me3 which inhibits DNMT3B to carry out methylation due to autoinhibition via the ADD domain.
B) In EBs, transcription on the Xi is silenced and H3K4me3 is removed which creates permissive conditions for DNMT3B to methylate CGIs.
C) H3K4me3 is enriched at CGIs in ES cells that gain methylation during differentiation. H3K4me3 ChIP-seq data was taken from the ENCODE project, accession ENCBS043ENC.

However, transcriptional silencing is not sufficient to lead to CGI methylation since the majority of CGIs tend to stay hypomethylated despite the silencing of associated genes. On Xi, a differentiation-dependent cue is necessary. Counterintuitively, DNMT3B gets downregulated during this time window (shown in this work and (Gendrel et al., 2012)). The clear evidence for the differentiation-dependent recruitment of SMCHD1 on Xi and its indispensable role in determining CGI methylation on Xi, makes it a prime factor to consider in this investigation (Gdula et al., 2019; Gendrel et al., 2012). However, the mechanism via which SMCHD1 regulates CGI methylation on Xi remains unknown. A recent study has shown that loss of SMCHD1 results in narrowing to H3K9me3 blocks/enrichment on Xi (Ichihara et al., 2022), indicating a network between SMCHD1 and H3K9me3 that could in turn affect CGI methylation. I have shown that H3K9me3 enrichment on Xi occurs upstream of CGI methylation and

that the X-linked CGIs show higher accumulation of H3K9me3 as compared to Xi in general. The H3K9me3 levels on DNMT3B-independent "CGI-like elements", although gained during differentiation, are much lower than CGIs and resemble the general/average levels on Xi. Together, these findings support a role for H3K9me3 in governing CGI methylation on Xi.

Mammals have six H3K9 methyltransferases. G9A and GLP (also called EHMT2 and EHMT1 respectively) deposit mono and di methylation of H3K9 on silent genes in euchromatic regions. SUV39H1 and SUV39H2 are responsible for the enrichment of H3K9me3 on pericentromeric and centromeric regions (reviewed in (Du et al., 2015). SETDB1 is histone mono- di- and tri-methyltransferase known for H3K9me3 establishment on retrotransposons. A paralogue of SETDB1 called SETDB2 is known to catalyse H3K9me3 from H3K9me1/2 (reviewed (Padeken et al., 2022)). A role for H3K9me3 in governing DNA methylation has been observed in different organisms. For instance, in filamentous fungi *Neurospora crassa*, H3K9me3 deposited by DIM5, leads to the recruitment of HP1 via its chromodomain. HP1 in turn acts as a scaffold to recruit DNA cytosine methyltransferase DIM2 via its chromoshadow domain. This link between H3K9me3 and cytosine methylation is unidirectional as loss of HP1 abrogates cytosine methylation but has little effect on H3K9me3 levels (reviewed in (Du et al., 2015)). In *Arabidopsis thaliana* cytosine methyltransferases, CMT2 and CMT3 (which prefer unmethylated CHG and CHH or hemimethylated CHG respectively) contain a chromodomain which allows direct binding to H3K9me3-modified chromatin. Mutations in SUV39 homologue in *A.thaliana,* KYP, leads to a reduction in CHG methylation. However, in Arabidopsis loss of CMT3 also leads to loss of H3K9me3 levels as KYP in turn binds methylated CHG and CHH via SRA domain, forming a feedforward loop. Cytosine methylation by the RdDM pathway in *A.thaliana* is also dependent upon H3K9me2 binding SHH1 protein (reviewed in (Du et al., 2015)).

The link between CpG methylation and H3K9me3 is not as well characterized in mammals, but some evidence has been presented for the dependence of CpG methylation on H3K9me3. At the pericentric chromatin, the recruitment of DNMT3B has been shown to be dependent on SUV39H1/H2 (Lehnertz et al., 2003). In mouse ES cells the loss of G9A results in the hypomethylation of retrotransposons and selected CGIs and ICRs (Dong et al., 2008; Zhang et al., 2016). Dong et al., also suggest that the recruitment of DNMT3A at reterotransposons is compromised in *G9a* null ES cells. On the other hand, some studies show that H3K9me2 itself does not determine DNA methylation. For instance, a study suggests that G9A catalytic mutants can rescue DNA methylation at selected CGIs (associated with *Mage-a2* and *Wfdc15*) in mouse ES cells indicating that H3K9me2 itself is not required for methylation of these CGIs (Tachibana et al., 2008). Similarly, another study showed that G9A protein, rather than H3K9me2, is required for the methylation at ICRs (Zhang et al., 2016).

Keniry et al., have shown that the depletion of SETDB1 in MEFs leads to partial loss of H3K9me3 on Xi. They have also shown that loss of SETDB1 leads to a small loss of CGI methylation. Another publication suggests that depletion of a SETDB1 interacting protein ATF7IP, leads to reactivation of a reporter gene on Xi when combined with loss of DNA methylation on Xi (by 5-azaC treatement or DNMT1 depletion) (Minkovsky et al., 2014). Minkovsky et al., propose that ATF7IP bridges two of its interacting proteins SETDB1 and MBD1, and by extension mediates a link between DNA methylation and H3K9me3 accumulation on Xi. Further investigation is required to build a strong argument for the MBD1-ATF7IP-SETDB1 pathway operating on Xi. Thus, in light of the published literature, the accumulation of H3K9me3 on CGIs on Xi could be suggestive of its role in mediating CGI methylation. Proof of a role for H3K9me3 on Xi would require analysis of CGI methylation in a mutant cell line which lacks the H3K9me3 establishment on Xi. As mentioned earlier, contrary to the published results, loss of SETDB1 does not result in loss of H3K9me3 on Xi in iXist-ChrX cells and the search for the H3K9me3 methyltransferase functioning on Xi is an active area of research in our lab (Dr. Mafalda Almeida , unpublished data).

Assuming that H3K9me3 does govern/mediate CGI methylation, the question is, how does this occur mechanistically since there is no recognized H3K9me3-reading domain in DNMT3 proteins. Interestingly the ADD domain of the protein ATRX, has been reported to bind H3K9me3 and lead to its recruitment to pericentromeric heterochromatin (Iwase et al., 2011). Another study reported that DNMT3A can catalyse methylation on H3K9me3-modified nucleosomes reconstituted *in vitro* (Zhang et al., 2010), suggesting that H3K9me3 may release DNMT3A from autoinhibition. It has been suggested that despite the high conservation between the ADD domains of DNMT3A and DNMT3B, the ADD domain of DNMT3B has more flexibility to accommodate/interact with H3K9me3 than DNMT3A (Boyko et al., 2022; Iwase et al., 2011). Although the residues critical for creating the interacting interface of ATRX ADD domain with H3K9me3 were not found conserved in DNMT3A and DNMT3B, it is worth investigating if DNMT3B does indeed have the ability to bind/recognize H3K9me3 modified histone/chromatin. To this end previous biochemical analysis in the lab showed using microscale thermophoresis that a fragment of DNMT3B spanning the PWWP and ADD domains could interact with $H3_{1-20}$K9me3 (Dr. Burcu Anil, unpublished data). Building on this work is an important direction for future studies to investigate the possible link between H3K9me3 and DNMT3B.

6.10 Crosstalk between CGI methylation and H3K9me3 accumulation on Xi

I have shown that knockout of *Dnmt3b* results in partial loss of H3K9me3 accumulation on Xi. The defect in H3K9me3 accumulation on Xi varied between individual EB differentiation experiments, however, this trend was constant among all replicates. To gain further confidence in these findings and to exclude any unknown/undiscovered confounding clonal factors, H3K9me3 ChIP-seq should be repeated in other *Dnmt3b*$^{-/-}$ clones/cell lines generated for this study. Theoretically, H3K9me3 levels on Xi should be rescued in the *Dnmt3b1-KI* line. The data obtained so far suggests that accumulation of H3K9me3 lies upstream of CGI methylation, since in *Dnmt3b*$^{-/-}$ lines H3K9me3 is gained on Xi during differentiation. The partial H3K9me3 accumulation defect might therefore indicate that DNMT3B (or CGI methylation) is required for the maintenance of H3K9me3 enrichment on Xi. A crosstalk of CGI methylation and H3K9me3 enrichment has been suggested to occur on ICRs (Quenneville et al., 2011).

KRAB-ZFPS, ZFP57 and ZFP445 are known to bind ICRs and protect their allele-specific methylation during the global wave of demethylation in preimplantation embryos (Li et al., 2008; Takahashi et al., 2019). ZFP57 is known to recruit KAP1. Deletion of *Zfp57* and *Kap1* results in the loss of almost all imprints and embryonic lethality (Li et al., 2008; Messerschmidt et al., 2012). In ES cells, loss of ZFP57 and lack of ICR methylation has been shown to result in loss of KAP1 binding and reduction in H3K9me3 on ICRs (Quenneville et al., 2011). ChIP-seq shows that SETDB1 localizes to ICRs and has therefore been proposed to be recruited to ICRs via KAP1 as a mechanism for the deposition H3K9me3 (Messerschmidt et al., 2012; Quenneville et al., 2011). A similar KRAB-ZFP-KAP1-SETDB1 pathway has been proposed for silencing certain subfamilies of retrotransposons (Ondičová et al., 2020). It is possible that this pathway is also responsible for the maintenance of H3K9me3 levels on Xi and therefore in the absence of CGI methylation in *Dnmt3b*$^{-/-}$ line, H3K9me3 enrichment is also depleted. On ICRs, ChIP-seq shows a high correlation between the localization of ZFP57, KAP1, SETDB1 and H3K9me3 (Quenneville et al., 2011). It would therefore be interesting to perform ChIP-seq for KAP1 and SETDB1 in EBs which would give an indication if such a pathway could be involved on CGIs on Xi.

Intriguingly among the proteins found associated with DNMT3B in my proteomic screen were TASOR, MPP8, and KAP1. MPP8 and KAP1 were detected among DNMT3A and DNMT3B common interactors in ES cells, and in EBs these were found associated only with DNMT3B. If the interaction of DNMT3B with SETDB1 recruiting KAP1 and MPP8/TASOR can be experimentally validated, this could indicate a mechanism via which DNMT3B (the protein itself or the CGI methylation) maintains H3K9me3 on Xi. To tease apart the contribution of CGI methylation or DNMT3B protein in the efficient enrichment of

H3K9me3 on Xi, H3K9me3 ChIP-seq can be performed in a *Dnmt3b*$^{-/-}$ line complemented with a catalytic inactive mutant of DNMT3B.

6.11 Proteomic analysis for discovery of DNMT3A and DNMT3B interactome dynamic during differentiation

I employed an AP-MS based approach to identify the changes in the interacting partners of DNMT3A and DNMT3B between the ES and EB state. Theoretically, this experiment would inform about the potential contributors/determinants of DNMT3B specificity. Reassuringly, several reported/known interactors of DNMT3A and DNMT3B such as DNMT3L, G9A, EZH2, SUZ12, EED, HELLS and HP1α (CBX5), were also identified in this study (reviewed in (Saravanaraman et al., 2020)). Since both DNMT3A and DNMT3B associate broadly with chromatin (as evident from ChIP-seq) it is important to experimentally validate the potential interactors suggested by AP-MS experiment since many chromatin binding proteins captured by IP can be false positives.

TET1 was found associated with both DNMT3A and DNMT3B in ES cells suggesting that it works in proximity to DNMT3 proteins. TET proteins are reported to bind CGIs and guard the promoters against methylation (Wu et al., 2011). Using ChIP-seq, Manzo et al., have shown that TET1 binds hypomethylated CpGs and forms a contrasting binding pattern to methylated CpGs (Manzo et al., 2017). Furthermore, 5hmC was enriched at the borders of TET1 binding sites on hypomethylated CpGs suggesting that 5mC was established on these regions before (Manzo et al., 2017). It is possible that TET1 is responsible for constantly demethylating any methylation brought about by DNMT3A and DNMT3B in ES cells (Figure 6.6).

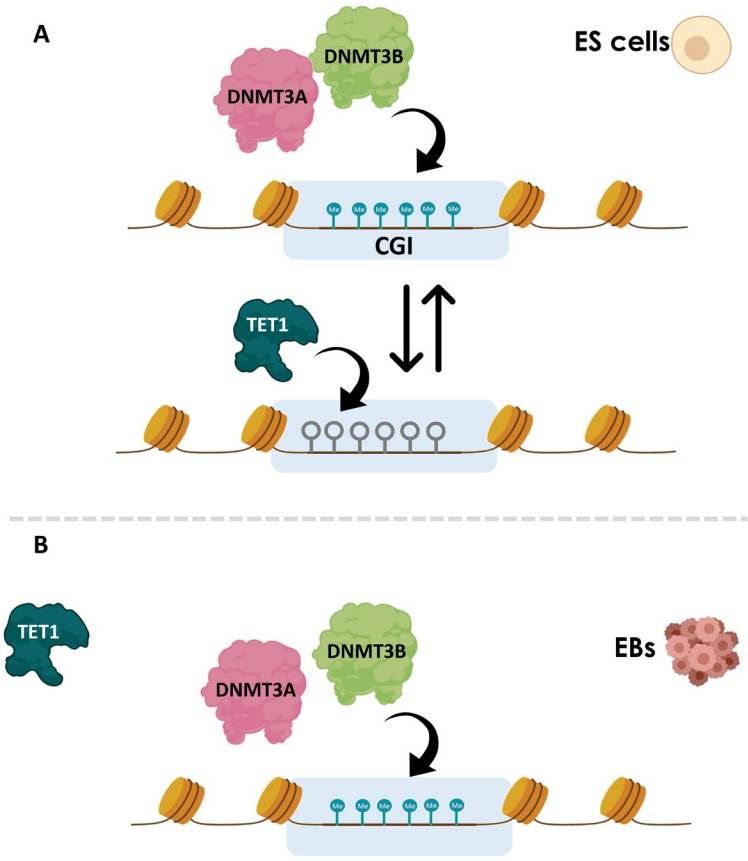


Figure 6.6: TET1-mediated hypomethylation on CGIs in ES cells and proposed model for CGI methylation in EBs.

A) In ES cells both DNMT3 proteins methylate CGIs but these are demethylated by TET1 found associated with CGIs.
B) In EBs, TET1 could be removed from CGIs allowing DNMT3A and 3B to establish CGI methylation.

As mentioned above an interesting finding was the presence of SETDB1 recruiting proteins such as TASOR, MPP8, and KAP1 among the DNMT3B interactors in EBs. Although this observation awaits experimental validation, there is already some evidence of the interaction of KAP1 with DNMT3A and DNMT3B in mouse ES cells (Quenneville et al., 2011). Furthermore, SETDB1 has been shown to associate with DNMT3A and DNMT3B in Cos-7 cells (an African monkey kidney fibroblast cell line) (Li et al., 2006). An important control to add to the AP-MS analysis is an IP for Δ1-218 DNMT3B1. The differential analysis of the interacting proteins found in DNMT3B1 and Δ1-218 DNMT3B1 could provide insights into the mechanisms underlying the function of the DNMT3B1 N-terminus.

There is also evidence that PTMs can regulate the activity of DNMT3 proteins. For instance, Casein kinase 2, CK2 mediates phosphorylation of DNMT3A on S386 and S389 which increases heterochromatin targeting of DNMT3A and reduces its DNA methyltransferase activity. The latter mechanism could explain hypomethylation associated with cancers showing upregulation of CK2 (Jeltsch and Jurkowska, 2016). Both DNMT3A and DNMT3B have been suggested to be sumoylated (Li et al., 2007)(Eun Sung Kang et al., 2001). Many more PTMs have been predicted for both DNMT3A and DNMT3B, most of which have not been investigated (Jeltsch and Jurkowska, 2016). In my experiments, I was able to find three PTMs specifically on DNMT3A2 in EBs, phosphorylation on T422 (T641 in DNMT3A1), and methylation on K407 and K409 (corresponding to K627 and K629 in DNMT3A1) all of which resided within or near the catalytic domain). This methylation could render the hydrophilic lysine residues hydrophobic. Interestingly these residues are well conserved in DNMT3A among mammals and differ between the two homologues, DNMT3A and DNMT3B. In ES cells DNMT3A2 was found to be acetylated on K398 (K617 in DNMT3A1, linker between ADD and catalytic domain) and methylated on R143 (equivalent residue on DNMT3A1 362, linker between PWWP domain and ADD domain). Further investigation, particularly for the DNMT3A PTMs detected in EBs could indicate if these regulate DNMT3A activity. Alternatively, these PTMs could modulate the interaction of DNMT3A with histone modification or modify its targeting for instance away from X-linked CGIs.

Taking into consideration the findings from proteomics experiments, I present the following speculative model. In ES cells the ongoing transcription and the resulting H3K4me3 accumulation on the X chromosome prevents CGI methylation (Figure 6.7). Once XCI is initiated, transcriptional silencing takes place and H2AK119ub1 and other histone modifications are established on Xi. Later during differentiation, SMCHD1 is recruited to XCI via H2AK119ub1 (Jansz et al., 2018) which in turn stimulates the accumulation of H3K9me3 on Xi (Ichihara et al., 2022). The high levels of H3K9me3 on CGIs could act to facilitate DNMT3B recruitment via recognition by the DNMT3B ADD domain. DNMT3A on the other hand might lose its interaction with chromatin due to PTMs detected only in EBs. The association of DNMT3B with SETDB1 recruiting MPP8, TASOR or KAP1 could indicate DNMT3B having a role in H3K9me3 maintenance and explain the lower levels of H3K9me3 in *Dnmt3b*$^{-/-}$ cells (Figure 6.7). Many aspects of this model require further investigation in future studies.

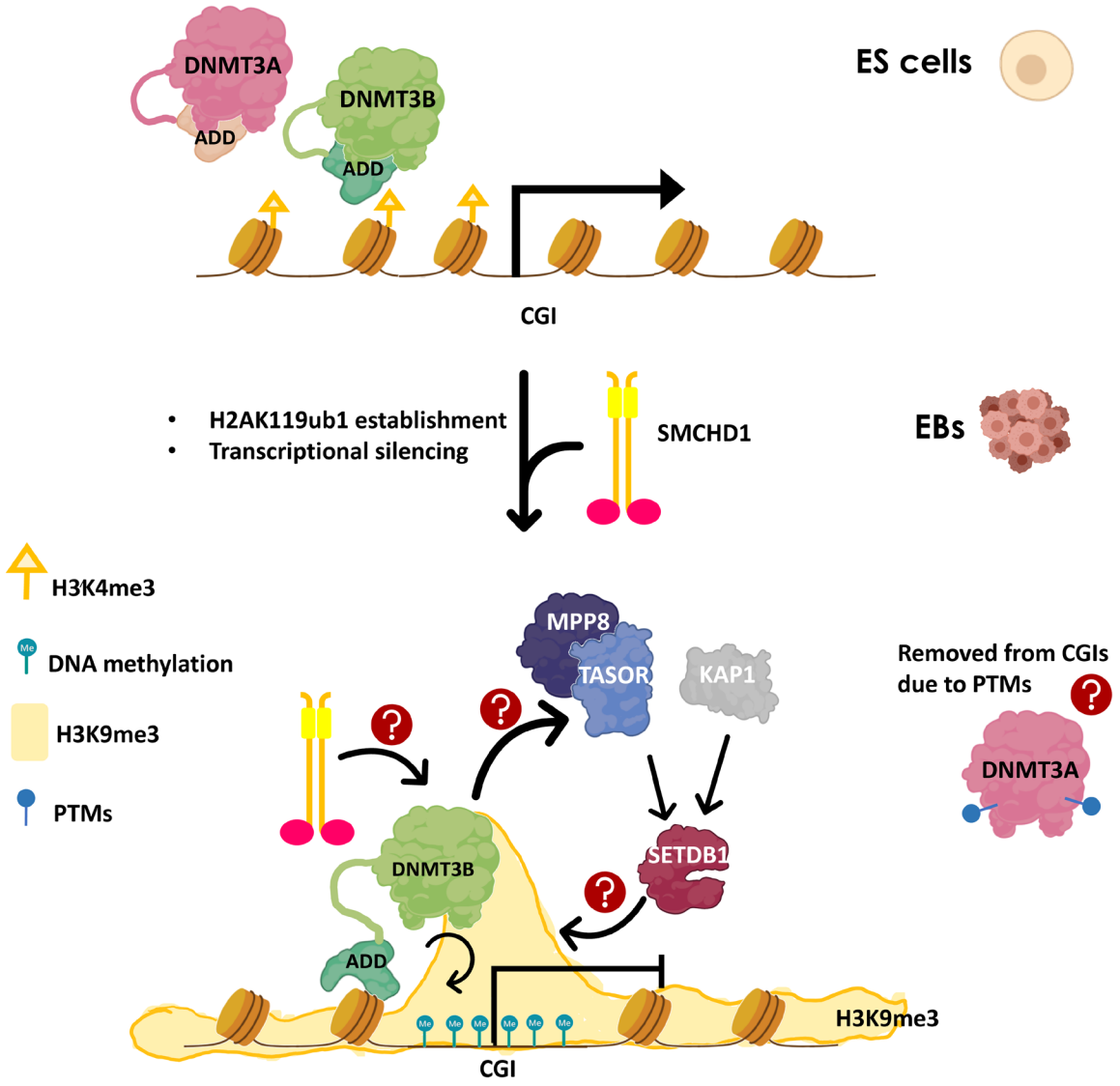


Figure 6.7: Model for DNMT3B mediated methylation of X-linked CGIs based on findings in this study and preliminary data.

DNMT3A and 3B both localize to CGIs on Xi in ES cells as shown by ChIP-seq but could be inhibited by the enrichment of H3K4me3 at these regions. During differentiation and XCI, transcriptional silencing occurs and H3K9me3 is established (with the involvement of SMCHD1, the underlying mechanisms to be discovered) which could then free DNMT3B from autoinhibition and enable it to carry out CGI methylation. H3K9me3 could act as DNMT3B- recruiting modification which would be recognized by the DNMT3B ADD domain as suggested by data from Dr. Burcu Anil. DNMT3A on the other hand might lose its interaction with chromatin due to PTMs specific to EBs. The association of DNMT3B with SETDB1 recruiting MPPB, TASOR or KAP1 could indicate a role of DNMT3B in H3K9me3 maintenance and the lower levels of H3K9me3 in *Dnmt3b*$^{-/-}$ cells.

Bibliography

Ainscough, J.F.-X., Rahman, F.A., Sercombe, H., Sedo, A., Gerlach, B., and Coverley, D. (2007). C-terminal domains deliver the DNA replication factor Ciz1 to the nuclear matrix. Journal of cell science *120*, 115-124.

Allshire, R.C., and Madhani, H.D. (2018). Ten principles of heterochromatin formation and function. Nature reviews Molecular cell biology *19*, 229-244.

Amouroux, R., Nashun, B., Shirane, K., Nakagawa, S., Hill, P.W., D'Souza, Z., Nakayama, M., Matsuda, M., Turp, A., Ndjetehe, E., Encheva, V.,Kudo, R.N., Koseki, H., Sasaki, H., and Hajkova, P. (2016). De novo DNA methylation drives 5hmC accumulation in mouse zygotes. Nature cell biology *18*, 225-233.

Andergassen, D., Muckenhuber, M., Bammer, P.C., Kulinski, T.M., Theussl, H.-C., Shimizu, T., Penninger, J.M., Pauler, F.M., and Hudson, Q.J. (2019a). The Airn lncRNA does not require any DNA elements within its locus to silence distant imprinted genes. PLoS Genetics *15*, e1008268.

Andergassen, D., Smith, Z.D., Lewandowski, J.P., Gerhardinger, C., Meissner, A., and Rinn, J.L. (2019b). In vivo Firre and Dxz4 deletion elucidates roles for autosomal gene regulation. Elife *8*, e47214.

Anguera, M.C., Ma, W., Clift, D., Namekawa, S., Kelleher III, R.J., and Lee, J.T. (2011). Tsx produces a long noncoding RNA and has general functions in the germline, stem cells, and brain. PLoS genetics *7*, e1002248.

Arber, W. (1965). Host-controlled modification of bacteriophage. Annual Reviews in Microbiology *19*, 365-378.

Ariyoshi, M., and Schwabe, J.W. (2003). A conserved structural motif reveals the essential transcriptional repression function of Spen proteins and their role in developmental signaling. Genes & development *17*, 1909-1920.

Auclair, G., Borgel, J., Sanz, L.A., Vallet, J., Guibert, S., Dumas, M., Cavelier, P., Girardot, M., Forné, T., and Feil, R. (2016). EHMT2 directs DNA methylation for efficient gene silencing in mouse embryos. Genome research *26*, 192-202.

Auclair, G., Guibert, S., Bender, A., and Weber, M. (2014). Ontogeny of CpG island methylation and specificity of DNMT3 methyltransferases during embryonic development in the mouse. Genome biology *15*, 1-16.

Banigan, E.J., and Mirny, L.A. (2020). Loop extrusion: theory meets single-molecule experiments. Current opinion in cell biology *64*, 124-138.

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*, 17732-17736.

Bannister, A.J., Zegerman, P., Partridge, J.F., Miska, E.A., Thomas, J.O., Allshire, R.C., and Kouzarides, T. (2001). Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. Nature *410*, 120-124.

Barau, J., Teissandier, A., Zamudio, N., Roy, S., Nalesso, V., Hérault, Y., Guillou, F., and Bourc'his, D. (2016). The DNA methyltransferase DNMT3C protects male germ cells from transposon activity. Science *354*, 909-912.

Barr, M.L., and Bertram, E.G. (1949). A morphological distinction between neurones of the male and female, and the behaviour of the nucleolar satellite during accelerated nucleoprotein synthesis. Nature *163*, 676-677.

Baubec, T., Colombo, D.F., Wirbelauer, C., Schmidt, J., Burger, L., Krebs, A.R., Akalin, A., and Schübeler, D. (2015). Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. Nature *520*, 243-247.

Berggård, T., Linse, S., and James, P. (2007). Methods for the detection and analysis of protein–protein interactions. Proteomics *7*, 2833-2842.

Bestor, T., Laudano, A., Mattaliano, R., and Ingram, V. (1988). Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells: the carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. Journal of molecular biology *203*, 971-983.

Bibikova, M., and Fan, J.-B. (2009). GoldenGate® assay for DNA methylation profiling. In DNA Methylation (Springer), pp. 149-163.

Bibikova, M., Le, J., Barnes, B., Saedinia-Melnyk, S., Zhou, L., Shen, R., and Gunderson, K.L. (2009). Genome-wide DNA methylation profiling using Infinium® assay. Epigenomics *1*, 177-200.

Billen, D. (1968). Methylation of the bacterial chromosome: an event at the "replication point"? Journal of Molecular Biology *31*, 477-486.

Bird, A., Taggart, M., Frommer, M., Miller, O.J., and Macleod, D. (1985). A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell *40*, 91-99.

Bird, A.P. (1986). CpG-rich islands and the function of DNA methylation. Nature *321*, 209-213.

Bishop, K.S., and Ferguson, L.R. (2015). The interaction between epigenetics, nutrition and the development of cancer. Nutrients *7*, 922-947.

Blackledge, N.P., Farcas, A.M., Kondo, T., King, H.W., McGouran, J.F., Hanssen, L.L., Ito, S., Cooper, S., Kondo, K., and Koseki, Y. (2014). Variant PRC1 complex-dependent H2A ubiquitylation drives PRC2 recruitment and polycomb domain formation. Cell *157*, 1445-1459.

Blackledge, N.P., and Klose, R.J. (2021). The molecular principles of gene regulation by Polycomb repressive complexes. Nature Reviews Molecular Cell Biology *22*, 815-833.

Blackledge, N.P., Long, H.K., Zhou, J.C., Kriaucionis, S., Patient, R., and Klose, R.J. (2012). Bio-CAP: a versatile and highly sensitive technique to purify and characterise regions of non-methylated DNA. Nucleic acids research *40*, e32-e32.

Blewitt, M.E., Gendrel, A.-V., Pang, Z., Sparrow, D.B., Whitelaw, N., Craig, J.M., Apedaile, A., Hilton, D.J., Dunwoodie, S.L., and Brockdorff, N. (2008). SmcHD1, containing a structural-maintenance-of-chromosomes hinge domain, has a critical role in X inactivation. Nature genetics *40*, 663-669.

Blewitt, M.E., Vickaryous, N.K., Hemley, S.J., Ashe, A., Bruxner, T.J., Preis, J.I., Arkell, R., and Whitelaw, E. (2005). An N-ethyl-N-nitrosourea screen for genes involved in variegation in the mouse. Proceedings of the National Academy of Sciences *102*, 7629-7634.

Bogutz, A.B., Brind'Amour, J., Kobayashi, H., Jensen, K.N., Nakabayashi, K., Imai, H., Lorincz, M.C., and Lefebvre, L. (2019). Evolution of imprinting via lineage-specific insertion of retroviral promoters. Nature communications *10*, 1-14.

Borgel, J., Guibert, S., Li, Y., Chiba, H., Schübeler, D., Sasaki, H., Forné, T., and Weber, M. (2010). Targets and dynamics of promoter DNA methylation during early mouse development. Nature genetics *42*, 1093-1100.

Bostick, M., Kim, J.K., Estève, P.-O., Clark, A., Pradhan, S., and Jacobsen, S.E. (2007). UHRF1 plays a role in maintaining DNA methylation in mammalian cells. Science *317*, 1760-1764.

Bourc'his, D., Xu, G.-L., Lin, C.-S., Bollman, B., and Bestor, T.H. (2001). Dnmt3L and the establishment of maternal genomic imprints. Science *294*, 2536-2539.

Bowness, J.S., Nesterova, T.B., Wei, G., Rodermund, L., Almeida, M., Coker, H., Carter, E.J., Kadaster, A., and Brockdorff, N. (2022). Xist-mediated silencing requires additive functions of SPEN and Polycomb together with differentiation-dependent recruitment of SmcHD1. Cell Reports *39*, 110830.

Boyes, J., and Bird, A. (1991). DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein. Cell *64*, 1123-1134.

Boyko, K., Arkova, O., Nikolaeva, A., Popov, V.O., Georgiev, P., and Bonchuk, A. (2022). Structure of the DNMT3B ADD domain suggests the absence of a DNMT3A-like autoinhibitory mechanism. Biochemical and Biophysical Research Communications *619*, 124-129.

Brambert, P.R., Kelpsch, D.J., Hameed, R., Desai, C.V., Calafiore, G., Godley, L.A., and Raimondi, S.L. (2015). DNMT3B7 expression promotes tumor progression to a more aggressive phenotype in breast cancer cells. PLoS One *10*, e0117310.

Brandeis, M., Frank, D., Keshet, I., Siegfried, Z., Mendelsohn, M., Temper, V., Razin, A., and Cedar, H. (1994). Spl elements protect a CpG island from de novo methylation. Nature *371*, 435-438.

Brind'Amour, J., Kobayashi, H., Richard Albert, J., Shirane, K., Sakashita, A., Kamio, A., Bogutz, A., Koike, T., Karimi, M.M., and Lefebvre, L. (2018). LTR retrotransposons transcribed in oocytes drive species-specific and heritable changes in DNA methylation. Nature communications *9*, 1-14.

Brockdorff, N., Ashworth, A., Kay, G.F., Cooper, P., Smith, S., McCabe, V.M., Norris, D.P., Penny, G.D., Patel, D., and Rastan, S. (1991). Conservation of position and exclusive expression of mouse Xist from the inactive X chromosome. Nature *351*, 329-331.

Brockdorff, N., Ashworth, A., Kay, G.F., McCabe, V.M., Norris, D.P., Cooper, P.J., Swift, S., and Rastan, S. (1992). The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. Cell *71*, 515-526.

Brockdorff, N., Bowness, J.S., and Wei, G. (2020). Progress toward understanding chromosome silencing by Xist RNA. Genes & development *34*, 733-744.

Brown, C.J., Lafreniere, R.G., Powers, V.E., Sebastio, G., Ballabio, A., Pettigrew, A.L., Ledbetter, D.H., Levy, E., Craig, I.W., and Willard, H.F. (1991). Localization of the X inactivation centre on the human X chromosome in Xq13. Nature *349*, 82-84.

Brown, D.A., Di Cerbo, V., Feldmann, A., Ahn, J., Ito, S., Blackledge, N.P., Nakayama, M., McClellan, M., Dimitrova, E., and Turberfield, A.H. (2017). The SET1 complex selects actively transcribed target genes via multivalent interaction with CpG island chromatin. Cell reports *20*, 2313-2327.

Butler, M.G. (2009). Genomic imprinting disorders in humans: a mini-review. Journal of assisted reproduction and genetics *26*, 477-486.

Cai, Y., Geutjes, E.-J., De Lint, K., Roepman, P., Bruurs, L., Yu, L., Wang, W., Van Blijswijk, J., Mohammad, H., and De Rink, I. (2014). The NuRD complex cooperates with DNMTs to maintain silencing of key colorectal tumor suppressor genes. Oncogene *33*, 2157-2168.

Carrozza, M.J., Li, B., Florens, L., Suganuma, T., Swanson, S.K., Lee, K.K., Shia, W.-J., Anderson, S., Yates, J., and Washburn, M.P. (2005). Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. Cell *123*, 581-592.

Cate, R., Chick, W., and Gilbert, W. (1983). Comparison of the methylation patterns of the two rat insulin genes. Journal of Biological Chemistry *258*, 6645-6652.

Cattanach, B. (1983). Location of Xce using XceaXcec heterozy-gates. Mouse News Letter *69*, 24.

Cattanach, B., Rasberry, C., Evans, E., and Burtenshaw, M. (1989). Further Xce linkage data. Mouse News Lett *83*, 165.

Charlton, J., Jung, E.J., Mattei, A.L., Bailly, N., Liao, J., Martin, E.J., Giesselmann, P., Brändl, B., Stamenova, E.K., and Müller, F.-J. (2020). TETs compete with DNMT3 activity in pluripotent cells at thousands of methylated somatic enhancers. Nature genetics *52*, 819-827.

Chedin, F., Lieber, M.R., and Hsieh, C.-L. (2002). The DNA methyltransferase-like protein DNMT3L stimulates de novo methylation by Dnmt3a. Proceedings of the National Academy of Sciences *99*, 16916-16921.

Chen, C.-K., Blanco, M., Jackson, C., Aznauryan, E., Ollikainen, N., Surka, C., Chow, A., Cerase, A., McDonel, P., and Guttman, M. (2016). Xist recruits the X chromosome to the nuclear lamina to enable chromosome-wide silencing. Science *354*, 468-472.

Chen, T., Tsujimoto, N., and Li, E. (2004). The PWWP domain of Dnmt3a and Dnmt3b is required for directing DNA methylation to the major satellite repeats at pericentric heterochromatin. Molecular and cellular biology *24*, 9048-9058.

Chen, T., Ueda, Y., Xie, S., and Li, E. (2002). A novel Dnmt3a isoform produced from an alternative promoter localizes to euchromatin and its expression correlates with activede novo methylation. Journal of Biological Chemistry *277*, 38746-38754.

Chen, Y., Miyazaki, J., Nishizawa, H., Kurahashi, H., Leach, R., and Wang, K. (2013). MTA3 regulates CGB5 and Snail genes in trophoblast. Biochemical and biophysical research communications *433*, 379-384.

Chen, Z., and Zhang, Y. (2020). Maternal H3K27me3-dependent autosomal and X chromosome imprinting. Nature Reviews Genetics *21*, 555-571.

Cheng, X., Kumar, S., Posfai, J., Pflugrath, J.W., and Roberts, R.J. (1993). Crystal structure of the Hhal DNA methyltransferase complexed with S-adenosyl-L-methionine. Cell *74*, 299-307.

Chinwalla, A.T., Cook, L.L., Delehaunty, K.D., Fewell, G.A., Fulton, L.A., Fulton, R.S., Graves, T.A., Hillier, L.W., Mardis, E.R., McPherson, J.D*., et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-562.

Chotalia, M., Smallwood, S.A., Ruf, N., Dawson, C., Lucifero, D., Frontera, M., James, K., Dean, W., and Kelsey, G. (2009). Transcription is required for establishment of germline methylation marks at imprinted genes. Genes & development *23*, 105-117.

Choy, J.S., Wei, S., Lee, J.Y., Tan, S., Chu, S., and Lee, T.-H. (2010). DNA methylation increases nucleosome compaction and rigidity. Journal of the American Chemical Society *132*, 1782-1783.

Chu, C., Zhang, Q.C., Da Rocha, S.T., Flynn, R.A., Bharadwaj, M., Calabrese, J.M., Magnuson, T., Heard, E., and Chang, H.Y. (2015). Systematic discovery of Xist RNA binding proteins. Cell *161*, 404-416.

Chuang, L.S.-H., Ian, H.-I., Koh, T.-W., Ng, H.-H., Xu, G., and Li, B.F. (1997). Human DNA-(cytosine-5) methyltransferase-PCNA complex as a target for p21WAF1. Science *277*, 1996-2000.

Chureau, C., Chantalat, S., Romito, A., Galvani, A., Duret, L., Avner, P., and Rougeulle, C. (2011). Ftx is a non-coding RNA which affects Xist expression and chromatin structure within the X-inactivation center region. Human molecular genetics *20*, 705-718.

Cierpicki, T., Risner, L.E., Grembecka, J., Lukasik, S.M., Popovic, R., Omonkowska, M., Shultis, D.D., Zeleznik-Le, N.J., and Bushweller, J.H. (2010). Structure of the MLL CXXC domain–DNA complex and its functional role in MLL-AF9 leukemia. Nature structural & molecular biology *17*, 62-68.

Cirillo, D., Blanco, M., Armaos, A., Buness, A., Avner, P., Guttman, M., Cerase, A., and Tartaglia, G.G. (2017). Quantitative predictions of protein interactions with long noncoding RNAs. Nature methods *14*, 5-6.

Clapier, C.R., and Cairns, B.R. (2009). The biology of chromatin remodeling complexes. Annual review of biochemistry *78*, 273-304.

Clouaire, T., Webb, S., and Bird, A. (2014). Cfp1 is required for gene expression-dependent H3K4 trimethylation and H3K9 acetylation in embryonic stem cells. Genome biology *15*, 1-16.

Coker, H., Wei, G., and Brockdorff, N. (2019). m6A modification of non-coding RNA and the control of mammalian gene expression. Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms *1862*, 310-318.

Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., and Jacobsen, S.E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature *452*, 215-219.

Collombet, S., Ranisavljevic, N., Nagano, T., Varnai, C., Shisode, T., Leung, W., Piolot, T., Galupa, R., Borensztein, M., and Servant, N. (2020). Parental-to-embryo switch of chromosome organization in early embryogenesis. Nature *580*, 142-146.

Conrad, T., and Akhtar, A. (2012). Dosage compensation in Drosophila melanogaster: epigenetic fine-tuning of chromosome-wide transcription. Nature Reviews Genetics *13*, 123-134.

Constantinides, P.G., Jones, P.A., and Gevers, W. (1977). Functional striated muscle cells from non-myoblast precursors following 5-azacytidine treatment. Nature *267*, 364-366.

Conti, L., Pollard, S.M., Gorba, T., Reitano, E., Toselli, M., Biella, G., Sun, Y., Sanzone, S., Ying, Q.-L., and Cattaneo, E. (2005). Niche-independent symmetrical self-renewal of a mammalian tissue stem cell. PLoS biology *3*, e283.

Cooper, S., Grijzenhout, A., Underwood, E., Ancelin, K., Zhang, T., Nesterova, T.B., Anil-Kirmizitas, B., Bassett, A., Kooistra, S.M., and Agger, K. (2016). Jarid2 binds mono-ubiquitylated H2A lysine 119 to mediate crosstalk between Polycomb complexes PRC1 and PRC2. Nature communications *7*, 1-8.

Crick, F. (1970). Central dogma of molecular biology. Nature *227*, 561-563.

Csankovszki, G., Nagy, A., and Jaenisch, R. (2001). Synergism of Xist RNA, DNA methylation, and histone hypoacetylation in maintaining X chromosome inactivation. The Journal of cell biology *153*, 773-784.

Dahl, C., and Guldberg, P. (2003). DNA methylation analysis techniques. Biogerontology *4*, 233-250.

Daniel, J.M., Spring, C.M., Crawford, H.C., Reynolds, A.B., and Baig, A. (2002). The p120 ctn-binding partner Kaiso is a bi-modal DNA-binding protein that recognizes both a sequence-specific consensus and methylated CpG dinucleotides. Nucleic acids research *30*, 2911-2919.

Darrow, E.M., Huntley, M.H., Dudchenko, O., Stamenova, E.K., Durand, N.C., Sun, Z., Huang, S.-C., Sanborn, A.L., Machol, I., and Shamim, M. (2016). Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. Proceedings of the National Academy of Sciences *113*, E4504-E4512.

Deakin, J.E., Chaumeil, J., Hore, T.A., and Marshall Graves, J.A. (2009). Unravelling the evolutionary origins of X chromosome inactivation in mammals: insights from marsupials and monotremes. Chromosome Research *17*, 671-685.

Dedeurwaerder, S., Defrance, M., Calonne, E., Denis, H., Sotiriou, C., and Fuks, F. (2011). Evaluation of the Infinium Methylation 450K technology. Epigenomics *3*, 771-784.

Dennis, K., Fan, T., Geiman, T., Yan, Q., and Muegge, K. (2001). Lsh, a member of the SNF2 family, is required for genome-wide methylation. Genes & development *15*, 2940-2944.

Deplus, R., Blanchon, L., Rajavelu, A., Boukaba, A., Defrance, M., Luciani, J., Rothé, F., Dedeurwaerder, S., Denis, H., and Brinkman, A.B. (2014). Regulation of DNA methylation patterns by CK2-mediated phosphorylation of Dnmt3a. Cell reports *8*, 743-753.

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.

Dobrinić, P., Szczurek, A.T., and Klose, R.J. (2021). PRC1 drives Polycomb-mediated gene repression by controlling transcription initiation and burst frequency. Nature structural & molecular biology *28*, 811-824.

Dong, K.B., Maksakova, I.A., Mohn, F., Leung, D., Appanah, R., Lee, S., Yang, H.W., Lam, L.L., Mager, D.L., and Schübeler, D. (2008). DNA methylation in ES cells requires the lysine methyltransferase G9a but not its catalytic activity. The EMBO journal *27*, 2691-2701.

Dossin, F., Pinheiro, I., Żylicz, J.J., Roensch, J., Collombet, S., Le Saux, A., Chelmicki, T., Attia, M., Kapoor, V., and Zhan, Y. (2020). SPEN integrates transcriptional and epigenetic control of X-inactivation. Nature *578*, 455-460.

Down, T.A., Rakyan, V.K., Turner, D.J., Flicek, P., Li, H., Kulesha, E., Graef, S., Johnson, N., Herrero, J., and Tomazou, E.M. (2008). A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Nature biotechnology *26*, 779-785.

Du, J., Johnson, L.M., Jacobsen, S.E., and Patel, D.J. (2015a). DNA methylation pathways and their crosstalk with histone methylation. Nature reviews Molecular cell biology *16*, 519-532.

Du, Q., Luu, P.-L., Stirzaker, C., and Clark, S.J. (2015b). Methyl-CpG-binding domain proteins: readers of the epigenome. Epigenomics *7*, 1051-1073.

Duymich, C.E., Charlet, J., Yang, X., Jones, P.A., and Liang, G. (2016). DNMT3B isoforms without catalytic activity stimulate gene body methylation as accessory proteins in somatic cells. Nature communications *7*, 1-9.

Eckersley-Maslin, M., Alda-Catalinas, C., Blotenburg, M., Kreibich, E., Krueger, C., and Reik, W. (2019). Dppa2 and Dppa4 directly regulate the Dux-driven zygotic transcriptional program. Genes & development *33*, 194-208.

Ehrlich, M. (2003). The ICF syndrome, a DNA methyltransferase 3B deficiency and immunodeficiency disease. Clinical immunology *109*, 17-28.

Eidahl, J.O., Crowe, B.L., North, J.A., McKee, C.J., Shkriabai, N., Feng, L., Plumb, M., Graham, R.L., Gorelick, R.J., and Hess, S. (2013). Structural basis for high-affinity binding of LEDGF PWWP to mononucleosomes. Nucleic acids research *41*, 3924-3936.

Elgin, S.C., and Reuter, G. (2013). Position-effect variegation, heterochromatin formation, and gene silencing in Drosophila. Cold Spring Harbor perspectives in biology *5*, a017780.

Emperle, M., Adam, S., Kunert, S., Dukatz, M., Baude, A., Plass, C., Rathert, P., Bashtrykov, P., and Jeltsch, A. (2019). Mutations of R882 change flanking sequence preferences of the DNA methyltransferase DNMT3A and cellular methylation patterns. Nucleic acids research *47*, 11355-11367.

Emperle, M., Rajavelu, A., Reinhardt, R., Jurkowska, R.Z., and Jeltsch, A. (2014). Cooperative DNA binding and protein/DNA fiber formation increases the activity of the Dnmt3a DNA methyltransferase. Journal of Biological Chemistry *289*, 29602-29613.

Endoh, M., Endo, T.A., Shinga, J., Hayashi, K., Farcas, A., Ma, K.-W., Ito, S., Sharif, J., Endoh, T., and Onaga, N. (2017). PCGF6-PRC1 suppresses premature differentiation of mouse embryonic stem cells by regulating germ cell-related genes. Elife *6*, e21064.

Engreitz, J.M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., and Lander, E.S. (2013). The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. Science *341*, 1237973.

Epsztejn-Litman, S., Feldman, N., Abu-Remaileh, M., Shufaro, Y., Gerson, A., Ueda, J., Deplus, R., Fuks, F., Shinkai, Y., and Cedar, H. (2008). De novo DNA methylation promoted by G9a prevents reprogramming of embryonically silenced genes. Nature structural & molecular biology *15*, 1176-1183.

Erwin, J.A., del Rosario, B., Payer, B., and Lee, J.T. (2012). An ex vivo model for imprinting: mutually exclusive binding of Cdx2 and Oct4 as a switch for imprinted and random X-inactivation. Genetics *192*, 857-868.

Farcas, A.M., Blackledge, N.P., Sudbery, I., Long, H.K., McGouran, J.F., Rose, N.R., Lee, S., Sims, D., Cerase, A., and Sheahan, T.W. (2012). KDM2B links the Polycomb Repressive Complex 1 (PRC1) to recognition of CpG islands. elife *1*, e00205.

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*, 1827-1831.

Fujita, N., Watanabe, S., Ichimura, T., Tsuruzoe, S., Shinkai, Y., Tachibana, M., Chiba, T., and Nakao, M. (2003). Methyl-CpG binding domain 1 (MBD1) interacts with the Suv39h1-HP1 heterochromatic complex for DNA methylation-based transcriptional repression. Journal of Biological Chemistry *278*, 24132-24138.

Gagliardi, M., Strazzullo, M., and Matarazzo, M.R. (2018). DNMT3B functions: novel insights from human disease. Frontiers in cell and developmental biology *6*, 140.

Galupa, R., and Heard, E. (2018). X-chromosome inactivation: a crossroads between chromosome architecture and gene regulation. Annual review of genetics *52*, 535-566.

Galupa, R., Nora, E.P., Worsley-Hunt, R., Picard, C., Gard, C., van Bemmel, J.G., Servant, N., Zhan, Y., El Marjou, F., and Johanneau, C. (2020). A conserved noncoding locus regulates random monoallelic Xist expression across a topological boundary. Molecular cell *77*, 352-367. e358.

Gao, L., Emperle, M., Guo, Y., Grimm, S.A., Ren, W., Adam, S., Uryu, H., Zhang, Z.-M., Chen, D., and Yin, J. (2020). Comprehensive structure-function characterization of DNMT3B and DNMT3A reveals distinctive de novo DNA methylation mechanisms. Nature communications *11*, 1-14.

Gao, L., Guo, Y., Biswal, M., Lu, J., Yin, J., Fang, J., Chen, X., Shao, Z., Huang, M., and Wang, Y. (2022). Structure of DNMT3B homo-oligomer reveals vulnerability to impairment by ICF mutations. Nature communications *13*, 1-11.

Gdula, M.R., Nesterova, T.B., Pintacuda, G., Godwin, J., Zhan, Y., Ozadam, H., McClellan, M., Moralli, D., Krueger, F., and Green, C.M. (2019). The non-canonical SMC protein SmcHD1 antagonises TAD formation and compartmentalisation on the inactive X chromosome. Nature communications *10*, 1-14.

Gebhard, C., Benner, C., Ehrich, M., Schwarzfischer, L., Schilling, E., Klug, M., Dietmaier, W., Thiede, C., Holler, E., and Andreesen, R. (2010). General Transcription Factor Binding at CpG Islands in Normal Cells Correlates with Resistance to De novo DNA Methylation in Cancer CellsTF Binding Excludes De novo DNA Methylation. Cancer research *70*, 1398-1407.

Geiman, T.M., Sankpal, U.T., Robertson, A.K., Chen, Y., Mazumdar, M., Heale, J.T., Schmiesing, J.A., Kim, W., Yokomori, K., and Zhao, Y. (2004). Isolation and characterization of a novel DNA

methyltransferase complex linking DNMT3B with components of the mitotic chromosome condensation machinery. Nucleic acids research *32*, 2716-2729.

Gendrel, A.-V., Apedaile, A., Coker, H., Termanis, A., Zvetkova, I., Godwin, J., Tang, Y.A., Huntley, D., Montana, G., and Taylor, S. (2012). Smchd1-dependent and-independent pathways determine developmental dynamics of CpG island methylation on the inactive X chromosome. Developmental cell *23*, 265-279.

Gendrel, A.-V., Tang, Y.A., Suzuki, M., Godwin, J., Nesterova, T.B., Greally, J.M., Heard, E., and Brockdorff, N. (2013). Epigenetic functions of smchd1 repress gene clusters on the inactive X chromosome and on autosomes. Molecular and cellular biology *33*, 3150-3165.

Gingras, A.-C., Gstaiger, M., Raught, B., and Aebersold, R. (2007). Analysis of protein complexes using mass spectrometry. Nature reviews Molecular cell biology *8*, 645-654.

Gjaltema, R.A., Schwämmle, T., Kautz, P., Robson, M., Schöpflin, R., Lustig, L.R., Brandenburg, L., Dunkel, I., Vechiatto, C., and Ntini, E. (2022). Distal and proximal cis-regulatory elements sense X chromosome dosage and developmental state at the Xist locus. Molecular Cell *82*, 190-208. e117.

Goll, M.G., Kirpekar, F., Maggert, K.A., Yoder, J.A., Hsieh, C.-L., Zhang, X., Golic, K.G., Jacobsen, S.E., and Bestor, T.H. (2006). Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. Science *311*, 395-398.

Gontan, C., Achame, E.M., Demmers, J., Barakat, T.S., Rentmeester, E., van IJcken, W., Grootegoed, J.A., and Gribnau, J. (2012). RNF12 initiates X-chromosome inactivation by targeting REX1 for degradation. Nature *485*, 386-390.

Gopalakrishnan, S., Sullivan, B.A., Trazzi, S., Della Valle, G., and Robertson, K.D. (2009a). DNMT3B interacts with constitutive centromere protein CENP-C to modulate DNA methylation and the histone code at centromeric regions. Human molecular genetics *18*, 3178-3193.

Gopalakrishnan, S., Van Emburgh, B.O., Shan, J., Su, Z., Fields, C.R., Vieweg, J., Hamazaki, T., Schwartz, P.H., Terada, N., and Robertson, K.D. (2009b). A novel DNMT3B splice variant expressed in tumor and pluripotent cells modulates genomic DNA methylation patterns and displays altered DNA binding. Molecular Cancer Research *7*, 1622-1634.

Gordon, C.A., Hartono, S.R., and Chédin, F. (2013). Inactive DNMT3B splice variants modulate de novo DNA methylation. PloS one *8*, e69486.

Graves, J.A.M. (1982). 5-azacytidine-induced re-expression of alleles on the inactive X chromosome in a hybrid mouse cell line. Experimental cell research *141*, 99-105.

Greenberg, M.V., and Bourc'his, D. (2019). The diverse roles of DNA methylation in mammalian development and disease. Nature reviews Molecular cell biology *20*, 590-607.

Gretarsson, K.H., and Hackett, J.A. (2020). Dppa2 and Dppa4 counteract de novo methylation to establish a permissive epigenome for development. Nature structural & molecular biology *27*, 706-716.

Grimaud, C., Nègre, N., and Cavalli, G. (2006). From genetics to epigenetics: the tale of Polycomb group and trithorax group genes. Chromosome research *14*, 363-375.

Gu, T., Hao, D., Woo, J., Huang, T.-W., Guo, L., Lin, X., Guzman, A.G., Tovy, A., Rosas, C., and Jeong, M. (2022). The disordered N-terminal domain of DNMT3A recognizes H2AK119ub and is required for postnatal development. Nature Genetics *54*, 625-636.

Guibert, S., Forné, T., and Weber, M. (2012). Global profiling of DNA methylation erasure in mouse primordial germ cells. Genome research *22*, 633-641.

Guo, F., Li, X., Liang, D., Li, T., Zhu, P., Guo, H., Wu, X., Wen, L., Gu, T.-P., and Hu, B. (2014). Active and passive demethylation of male and female pronuclear DNA in the mammalian zygote. Cell stem cell *15*, 447-459.

Guo, H., Zhu, P., Guo, F., Li, X., Wu, X., Fan, X., Wen, L., and Tang, F. (2015a). Profiling DNA methylome landscapes of mammalian cells with single-cell reduced-representation bisulfite sequencing. Nature protocols *10*, 645-659.

Guo, X., Wang, L., Li, J., Ding, Z., Xiao, J., Yin, X., He, S., Shi, P., Dong, L., and Li, G. (2015b). Structural insight into autoinhibition and histone H3-induced activation of DNMT3A. Nature *517*, 640-644.

Gurdon, J.B., and Uehlinger, V. (1966). "Fertile" intestine nuclei. Nature *210*, 1240-1241.

Guseinov, V., and Vanyushin, B. (1975). Content and localisation of 5-methylcytosine in DNA of healthy and wilt-infected cotton plants. Biochimica et Biophysica Acta (BBA)-Nucleic Acids and Protein Synthesis *395*, 229-238.

Hackett, J.A., Sengupta, R., Zylicz, J.J., Murakami, K., Lee, C., Down, T.A., and Surani, M.A. (2013). Germline DNA demethylation dynamics and imprint erasure through 5-hydroxymethylcytosine. Science *339*, 448-452.

Haggerty, C., Kretzmer, H., Riemenschneider, C., Kumar, A.S., Mattei, A.L., Bailly, N., Gottfreund, J., Giesselmann, P., Weigert, R., and Brändl, B. (2021). Dnmt1 has de novo activity targeted to transposable elements. Nature Structural & Molecular Biology *28*, 594-603.

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*, 15-23.

Hamidi, T., Singh, A.K., and Chen, T. (2015). Genetic alterations of DNA methylation machinery in human diseases. Epigenomics *7*, 247-265.

Handa, V., and Jeltsch, A. (2005). Profound flanking sequence preference of Dnmt3a and Dnmt3b mammalian DNA methyltransferases shape the human epigenome. Journal of molecular biology *348*, 1103-1112.

Hansen, R.S., Stöger, R., Wijmenga, C., Stanek, A.M., Canfield, T.K., Luo, P., Matarazzo, M.R., D'Esposito, M., Feil, R., and Gimelli, G. (2000). Escape from gene silencing in ICF syndrome: evidence for advanced replication time as a major determinant. Human molecular genetics *9*, 2575-2587.

Harbers, K., Schnieke, A., Stuhlmann, H., Jähner, D., and Jaenisch, R. (1981). DNA methylation and gene expression: endogenous retroviral genome becomes infectious after molecular cloning. Proceedings of the National Academy of Sciences *78*, 7609-7613.

Hasegawa, Y., Brockdorff, N., Kawano, S., Tsutui, K., Tsutui, K., and Nakagawa, S. (2010). The matrix protein hnRNP U is required for chromosomal localization of Xist RNA. Developmental cell *19*, 469-476.

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*, 4841-4849.

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*, 576-589.

Heitz, E. (1928). Das heterochromatin der moose (Bornträger).

Hendrich, B., and Bird, A. (1998). Identification and characterization of a family of mammalian methyl CpG-binding proteins. Genetics Research *72*, 59-72.

Henikoff, S., Henikoff, J.G., Sakai, A., Loeb, G.B., and Ahmad, K. (2009). Genome-wide profiling of salt fractions maps physical properties of chromatin. Genome research *19*, 460-469.

Holliday, R., and Pugh, J.E. (1975). DNA Modification Mechanisms and Gene Activity During Development: Developmental clocks may depend on the enzymic modification of specific bases in repeated DNA sequences. Science *187*, 226-232.

Hornbeck, P.V., Zhang, B., Murray, B., Kornhauser, J.M., Latham, V., and Skrzypek, E. (2015). PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic acids research *43*, D512-D520.

Howell, C.Y., Bestor, T.H., Ding, F., Latham, K.E., Mertineit, C., Trasler, J.M., and Chaillet, J.R. (2001). Genomic imprinting disrupted by a maternal effect mutation in the Dnmt1 gene. Cell *104*, 829-838.

Hu, B., Petela, N., Kurze, A., Chan, K.-L., Chapard, C., and Nasmyth, K. (2015). Biological chromodynamics: a general method for measuring protein occupancy across the genome by calibrating ChIP-seq. Nucleic acids research *43*, e132-e132.

Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo, H.C., Davis, S., Gatto, L., and Girke, T. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. Nature methods *12*, 115-121.

Huettel, B., Kanno, T., Daxinger, L., Bucher, E., van der Winden, J., Matzke, A.J., and Matzke, M. (2007). RNA-directed DNA methylation mediated by DRD1 and Pol IVb: a versatile pathway for transcriptional gene silencing in plants. Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression *1769*, 358-374.

Ichihara, S., Nagao, K., Sakaguchi, T., Obuse, C., and Sado, T. (2022). SmcHD1 underlies the formation of H3K9me3 blocks on the inactive X chromosome in mice. Development *149*, dev200864.

Ichimura, T., Watanabe, S., Sakamoto, Y., Aoto, T., Fujita, N., and Nakao, M. (2005). Transcriptional repression and heterochromatin formation by MBD1 and MCAF/AM family proteins. Journal of Biological Chemistry *280*, 13928-13935.

Iida, T., Suetake, I., Tajima, S., Morioka, H., Ohta, S., Obuse, C., and Tsurimoto, T. (2002). PCNA clamp facilitates action of DNA cytosine methyltransferase 1 on hemimethylated DNA. Genes to Cells *7*, 997-1007.

Illingworth, R., Kerr, A., DeSousa, D., Jørgensen, H., Ellis, P., Stalker, J., Jackson, D., Clee, C., Plumb, R., and Rogers, J. (2008). A novel CpG island set identifies tissue-specific methylation at developmental gene loci. PLoS biology *6*, e22.

Illingworth, R.S., Gruenewald-Schneider, U., Webb, S., Kerr, A.R., James, K.D., Turner, D.J., Smith, C., Harrison, D.J., Andrews, R., and Bird, A.P. (2010). Orphan CpG islands identify numerous conserved promoters in the mammalian genome. PLoS genetics *6*, e1001134.

Ingham, P.W. (1983). Differential expression of bithorax complex genes in the absence of the extra sex combs and trithorax genes. Nature *306*, 591-593.

Ito, S., D'Alessio, A.C., Taranova, O.V., Hong, K., Sowers, L.C., and Zhang, Y. (2010). Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. Nature *466*, 1129-1133.

Ito, S., Shen, L., Dai, Q., Wu, S.C., Collins, L.B., Swenberg, J.A., He, C., and Zhang, Y. (2011). Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science *333*, 1300-1303.

Iwase, S., Xiang, B., Ghosh, S., Ren, T., Lewis, P.W., Cochrane, J.C., Allis, C.D., Picketts, D.J., Patel, D.J., and Li, H. (2011). ATRX ADD domain links an atypical histone methylation recognition mechanism to human mental-retardation syndrome. Nature structural & molecular biology *18*, 769-776.

Jachowicz, J.W., Strehle, M., Banerjee, A.K., Blanco, M.R., Thai, J., and Guttman, M. (2022). Xist spatially amplifies SHARP/SPEN recruitment to balance chromosome-wide silencing and specificity to the X chromosome. Nature Structural & Molecular Biology *29*, 239-249.

Jansz, N. (2019). DNA methylation dynamics at transposable elements in mammals. Essays in Biochemistry *63*, 677-689.

Jansz, N., Keniry, A., Trussart, M., Bildsoe, H., Beck, T., Tonks, I.D., Mould, A.W., Hickey, P., Breslin, K., and Iminitoff, M. (2018a). Smchd1 regulates long-range chromatin interactions on the inactive X chromosome and at Hox clusters. Nature structural & molecular biology *25*, 766-777.

Jansz, N., Nesterova, T., Keniry, A., Iminitoff, M., Hickey, P.F., Pintacuda, G., Masui, O., Kobelke, S., Geoghegan, N., and Breslin, K.A. (2018b). Smchd1 targeting to the inactive X is dependent on the Xist-HnrnpK-PRC1 pathway. Cell reports *25*, 1912-1923. e1919.

Jeltsch, A., and Jurkowska, R.Z. (2016). Allosteric control of mammalian DNA methyltransferases–a new regulatory paradigm. Nucleic acids research *44*, 8556-8575.

Jeong, S., Liang, G., Sharma, S., Lin, J.C., Choi, S.H., Han, H., Yoo, C.B., Egger, G., Yang, A.S., and Jones, P.A. (2009). Selective anchoring of DNA methyltransferases 3A and 3B to nucleosomes containing methylated DNA. Molecular and cellular biology *29*, 5366-5376.

Jia, D., Jurkowska, R.Z., Zhang, X., Jeltsch, A., and Cheng, X. (2007). Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. Nature *449*, 248-251.

Jiao, L., and Liu, X. (2015). Structural basis of histone H3K27 trimethylation by an active polycomb repressive complex 2. Science *350*, aac4383.

Jin, J., Cai, Y., Li, B., Conaway, R.C., Workman, J.L., Conaway, J.W., and Kusch, T. (2005). In and out: histone variant exchange in chromatin. Trends in biochemical sciences *30*, 680-687.

Johnson, L.M., Law, J.A., Khattar, A., Henderson, I.R., and Jacobsen, S.E. (2008). SRA-domain proteins required for DRM2-mediated de novo DNA methylation. PLoS genetics *4*, e1000280.

Jones, P.A., and Taylor, S.M. (1980). Cellular differentiation, cytidine analogs and DNA methylation. Cell *20*, 85-93.

Jonkers, I., Barakat, T.S., Achame, E.M., Monkhorst, K., Kenter, A., Rentmeester, E., Grosveld, F., Grootegoed, J.A., and Gribnau, J. (2009). RNF12 is an X-Encoded dose-dependent activator of X chromosome inactivation. Cell *139*, 999-1011.

Jurkowska, R.Z., and Jeltsch, A. (2016). Enzymology of mammalian DNA methyltransferases. DNA methyltransferases-role and function, 87-122.

Kaluscha, S., Domcke, S., Wirbelauer, C., Stadler, M.B., Durdu, S., Burger, L., and Schübeler, D. (2022). Evidence that direct inhibition of transcription factor binding is the prevailing mode of gene and repeat repression by DNA methylation. Nature Genetics, 1-12.

Kanduri, C., Pant, V., Loukinov, D., Pugacheva, E., Qi, C.-F., Wolffe, A., Ohlsson, R., and Lobanenkov, V.V. (2000). Functional association of CTCF with the insulator upstream of the H19 gene is parent of origin-specific and methylation-sensitive. Current Biology *10*, 853-856.

Kaneda, M., Okano, M., Hata, K., Sado, T., Tsujimoto, N., Li, E., and Sasaki, H. (2004). Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. Nature *429*, 900-903.

Karachentsev, D., Sarma, K., Reinberg, D., and Steward, R. (2005). PR-Set7-dependent methylation of histone H4 Lys 20 functions in repression of gene expression and is essential for mitosis. Genes & development *19*, 431-435.

Kaustov, L., Ouyang, H., Amaya, M., Lemak, A., Nady, N., Duan, S., Wasney, G.A., Li, Z., Vedadi, M., and Schapira, M. (2011). Recognition and specificity determinants of the human cbx chromodomains. Journal of Biological Chemistry *286*, 521-529.

Keer, J., Hamvas, R., Brockdorff, N., Page, D., Rastan, S., and Brown, S. (1990). Genetic mapping in the region of the mouse X-inactivation center. Genomics *7*, 566-572.

Keniry, A., Gearing, L.J., Jansz, N., Liu, J., Holik, A.Z., Hickey, P.F., Kinkel, S.A., Moore, D.L., Breslin, K., and Chen, K. (2016). Setdb1-mediated H3K9 methylation is enriched on the inactive X and plays a role in its epigenetic silencing. Epigenetics & chromatin *9*, 1-20.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome research *12*, 996-1006.

Kiaee, F., Zaki-Dizaji, M., Hafezi, N., Almasi-Hashiani, A., Hamedifar, H., Sabzevari, A., Shirkani, A., Zian, Z., Jadidi-Niaragh, F., and Aghamahdi, F. (2021). Clinical, immunologic and molecular spectrum of patients with immunodeficiency, centromeric instability, and facial anomalies (ICF) syndrome: A systematic review. Endocrine, Metabolic & Immune Disorders-Drug Targets (Formerly Current Drug Targets-Immune, Endocrine & Metabolic Disorders) *21*, 664-672.

Kidder, B.L., Hu, G., Cui, K., and Zhao, K. (2017). SMYD5 regulates H4K20me3-marked heterochromatin to safeguard ES cell self-renewal and prevent spurious differentiation. Epigenetics & chromatin *10*, 1-20.

Kizer, K.O., Phatnani, H.P., Shibata, Y., Hall, H., Greenleaf, A.L., and Strahl, B.D. (2005). A novel domain in Set2 mediates RNA polymerase II interaction and couples histone H3 K36 methylation with transcript elongation. Molecular and cellular biology *25*, 3305-3316.

Klimasauskas, S., Kumar, S., Roberts, R.J., and Cheng, X. (1994). Hhal methyltransferase flips its target base out of the DNA helix. Cell *76*, 357-369.

KN, H., Chow, M.Z., Baker, E.K., Pal, S., Bassal, S., Brasacchio, D., Wang, L., Craig, J.M., Jones, P.L., and Sif, S. (2005). Brahma links the SWI/SNF chromatin-remodeling complex with MeCP2-dependent transcriptional silencing. Nature genetics *37*, 254-264.

Kriaucionis, S., and Heintz, N. (2009). The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. Science *324*, 929-930.

Krueger, F., and Andrews, S.R. (2016). SNPsplit: Allele-specific splitting of alignments between genomes with known SNP genotypes. F1000Research *5*.

Ku, J.-L., Jeon, Y.-K., and Park, J.-G. (2011). Methylation-specific PCR. In Epigenetics Protocols (Springer), pp. 23-32.

Kuroda, M.I., Kang, H., De, S., and Kassis, J.A. (2020). Dynamic competition of polycomb and trithorax in transcriptional programming. Annual review of biochemistry *89*, 235.

La Salle, S., and Trasler, J.M. (2006). Dynamic expression of DNMT3a and DNMT3b isoforms during male germ cell development in the mouse. Developmental biology *296*, 71-82.

Lachner, M., O'Carroll, D., Rea, S., Mechtler, K., and Jenuwein, T. (2001). Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. Nature *410*, 116-120.

Laget, S., Joulie, M., Le Masson, F., Sasai, N., Christians, E., Pradhan, S., Roberts, R.J., and Defossez, P.-A. (2010). The human proteins MBD5 and MBD6 associate with heterochromatin but they do not bind methylated DNA. PloS one *5*, e11982.

Landolin, J.M., Johnson, D.S., Trinklein, N.D., Aldred, S.F., Medina, C., Shulha, H., Weng, Z., and Myers, R.M. (2010). Sequence features that drive human promoter function and tissue specificity. Genome research *20*, 890-898.

Lane, N., Dean, W., Erhardt, S., Hajkova, P., Surani, A., Walter, J., and Reik, W. (2003). Resistance of IAPs to methylation reprogramming may provide a mechanism for epigenetic inheritance in the mouse. genesis *35*, 88-93.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods *9*, 357-359.

Larsen, F., Gundersen, G., Lopez, R., and Prydz, H. (1992). CpG islands as gene markers in the human genome. Genomics *13*, 1095-1107.

Latos, P.A., Pauler, F.M., Koerner, M.V., Şenergin, H.B., Hudson, Q.J., Stocsits, R.R., Allhoff, W., Stricker, S.H., Klement, R.M., and Warczok, K.E. (2012). Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. Science *338*, 1469-1472.

Lauster, R., Trautner, T.A., and Noyer-Weidner, M. (1989). Cytosine-specific type II DNA methyltransferases: a conserved enzyme core with variable target-recognizing domains. Journal of molecular biology *206*, 305-312.

Law, J.A., and Jacobsen, S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nature Reviews Genetics *11*, 204-220.

Lee, J.T. (2000). Disruption of imprinted X inactivation by parent-of-origin effects at Tsix. Cell *103*, 17-27.

Lee, J.Y., and Lee, T.-H. (2012). Effects of DNA methylation on the structure of nucleosomes. Journal of the American Chemical Society *134*, 173-175.

Lehnertz, B., Ueda, Y., Derijck, A.A., Braunschweig, U., Perez-Burgos, L., Kubicek, S., Chen, T., Li, E., Jenuwein, T., and Peters, A.H. (2003). Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. Current Biology *13*, 1192-1200.

Lewis, E.B. (1978). A gene complex controlling segmentation in Drosophila. In Genes, development and cancer (Springer), pp. 205-217.

Lewis, P. (1947). Polycomb: PC. Drosophila Information Service *21*, 69.

Li, B., Zhou, J., Liu, P., Hu, J., Jin, H., Shimono, Y., Takahashi, M., and Xu, G. (2007). Polycomb protein Cbx4 promotes SUMO modification of de novo DNA methyltransferase Dnmt3a. Biochemical journal *405*, 369-378.

Li, C., Fan, Y., Li, G., Xu, X., Duan, J., Li, R., Kang, X., Ma, X., Chen, X., and Ke, Y. (2018). DNA methylation reprogramming of functional elements during mammalian embryonic development. Cell discovery *4*, 1-12.

Li, C.F., Henderson, I.R., Song, L., Fedoroff, N., Lagrange, T., and Jacobsen, S.E. (2008a). Dynamic regulation of ARGONAUTE4 within multiple nuclear bodies in Arabidopsis thaliana. PLoS genetics *4*, e27.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078-2079.

Li, H., Rauch, T., Chen, Z.-X., Szabó, P.E., Riggs, A.D., and Pfeifer, G.P. (2006). The histone methyltransferase SETDB1 and the DNA methyltransferase DNMT3A interact directly and localize to promoters silenced in cancer cells. Journal of Biological Chemistry *281*, 19489-19500.

Li, X., Ito, M., Zhou, F., Youngson, N., Zuo, X., Leder, P., and Ferguson-Smith, A.C. (2008b). A maternal-zygotic effect gene, Zfp57, maintains both maternal and paternal imprints. Developmental cell *15*, 547-557.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923-930.

Lin, C.-C., Chen, Y.-P., Yang, W.-Z., Shen, J.C., and Yuan, H.S. (2020). Structural insights into CpG-specific DNA methylation by human DNA methyltransferase 3B. Nucleic acids research *48*, 3949-3961.

Lin, I.G., Han, L., Taghva, A., O'Brien, L.E., and Hsieh, C.-L. (2002). Murine de novo methyltransferase Dnmt3a demonstrates strand asymmetry and site preference in the methylation of DNA in vitro. Molecular and cellular biology *22*, 704-723.

Ling, Y., Sankpal, U.T., Robertson, A.K., McNally, J.G., Karpova, T., and Robertson, K.D. (2004). Modification of de novo DNA methyltransferase 3a (Dnmt3a) by SUMO-1 modulates its interaction with histone deacetylases (HDACs) and its capacity to repress transcription. Nucleic acids research *32*, 598-610.

Liu, Y., Siejka-Zielińska, P., Velikova, G., Bi, Y., Yuan, F., Tomkova, M., Bai, C., Chen, L., Schuster-Böckler, B., and Song, C.-X. (2019). Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. Nature biotechnology *37*, 424-429.

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*, 39-46.

Loda, A., Collombet, S., and Heard, E. (2022). Gene regulation in time and space during X-chromosome inactivation. Nature Reviews Molecular Cell Biology *23*, 231-249.

Long, H.K., Sims, D., Heger, A., Blackledge, N.P., Kutter, C., Wright, M.L., Grützner, F., Odom, D.T., Patient, R., and Ponting, C.P. (2013). Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. elife *2*, e00348.

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*, 1-21.

Lu, Z., Zhang, Q.C., Lee, B., Flynn, R.A., Smith, M.A., Robinson, J.T., Davidovich, C., Gooding, A.R., Goodrich, K.J., and Mattick, J.S. (2016). RNA duplex map in living cells reveals higher-order transcriptome structure. Cell *165*, 1267-1279.

Luco, R.F., Pan, Q., Tominaga, K., Blencowe, B.J., Pereira-Smith, O.M., and Misteli, T. (2010). Regulation of alternative splicing by histone modifications. Science *327*, 996-1000.

Luger, K., Mäder, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature *389*, 251-260.

Lyon, M.F. (1961). Gene action in the X-chromosome of the mouse (Mus musculus L.). nature *190*, 372-373.

Lyons, D.B., and Zilberman, D. (2017). DDM1 and Lsh remodelers allow methylation of DNA wrapped in nucleosomes. Elife *6*, e30674.

Lyst, M.J., Ekiert, R., Ebert, D.H., Merusi, C., Nowak, J., Selfridge, J., Guy, J., Kastan, N.R., Robinson, N.D., and de Lima Alves, F. (2013). Rett syndrome mutations abolish the interaction of MeCP2 with the NCoR/SMRT co-repressor. Nature neuroscience *16*, 898-902.

Ma, Z., Swigut, T., Valouev, A., Rada-Iglesias, A., and Wysocka, J. (2011). Sequence-specific regulator Prdm14 safeguards mouse ESCs from entering extraembryonic endoderm fates. Nature structural & molecular biology *18*, 120-127.

Maclary, E., Hinten, M., Harris, C., and Kalantry, S. (2013). Long nonoding RNAs in the X-inactivation center. Chromosome research *21*, 601-614.

Macleod, D., Charlton, J., Mullins, J., and Bird, A.P. (1994). Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. Genes & development *8*, 2282-2292.

Mallona, I., Ilie, I.M., Karemaker, I.D., Butz, S., Manzo, M., Caflisch, A., and Baubec, T. (2021). Flanking sequence preference modulates de novo DNA methylation in the mouse genome. Nucleic acids research *49*, 145-157.

Manzo, M., Wirz, J., Ambrosi, C., Villaseñor, R., Roschitzki, B., and Baubec, T. (2017). Isoform-specific localization of DNMT3A regulates DNA methylation fidelity at bivalent CpG islands. The EMBO journal *36*, 3421-3434.

Margueron, R., Justin, N., Ohno, K., Sharpe, M.L., Son, J., Drury Iii, W.J., Voigt, P., Martin, S.R., Taylor, W.R., and De Marco, V. (2009). Role of the polycomb protein EED in the propagation of repressive histone marks. Nature *461*, 762-767.

Markaki, Y., Chong, J.G., Wang, Y., Jacobson, E.C., Luong, C., Tan, S.Y., Jachowicz, J.W., Strehle, M., Maestrini, D., and Banerjee, A.K. (2021). Xist nucleates local protein gradients to propagate silencing across the X chromosome. Cell *184*, 6174-6192. e6132.

Mattei, A.L., Bailly, N., and Meissner, A. (2022). DNA methylation: a historical perspective. Trends in Genetics.

Matzke, M., Kanno, T., Daxinger, L., Huettel, B., and Matzke, A.J. (2009). RNA-mediated chromatin-based silencing in plants. Current opinion in cell biology *21*, 367-376.

Maunakea, A.K., Chepelev, I., Cui, K., and Zhao, K. (2013). Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. Cell research *23*, 1256-1269.

Mayer, W., Niveleau, A., Walter, J., Fundele, R., and Haaf, T. (2000). Demethylation of the zygotic paternal genome. Nature *403*, 501-502.

Meehan, R., Lewis, J.D., and Bird, A.P. (1992). Characterization of MeCP2, a vertebrate DNA binding protein with affinity for methylated DNA. Nucleic acids research *20*, 5085-5092.

Meissner, A., Gnirke, A., Bell, G.W., Ramsahoye, B., Lander, E.S., and Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic acids research *33*, 5868-5877.

Meissner, A., Mikkelsen, T.S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B.E., Nusbaum, C., and Jaffe, D.B. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature *454*, 766-770.

Mendenhall, E.M., Koche, R.P., Truong, T., Zhou, V.W., Issac, B., Chi, A.S., Ku, M., and Bernstein, B.E. (2010). GC-rich sequence elements recruit PRC2 in mammalian ES cells. PLoS genetics *6*, e1001244.

Messerschmidt, D.M., de Vries, W., Ito, M., Solter, D., Ferguson-Smith, A., and Knowles, B.B. (2012). Trim28 is required for epigenetic stability during mouse oocyte to embryo transition. Science *335*, 1499-1502.

Millán-Zambrano, G., Burton, A., Bannister, A.J., and Schneider, R. (2022). Histone post-translational modifications—cause and consequence of genome function. Nature Reviews Genetics, 1-18.

Minkovsky, A., Barakat, T.S., Sellami, N., Chin, M.H., Gunhanlar, N., Gribnau, J., and Plath, K. (2013). The pluripotency factor-bound intron 1 of Xist is dispensable for X chromosome inactivation and reactivation in vitro and in vivo. Cell reports *3*, 905-918.

Minkovsky, A., Sahakyan, A., Rankin-Gee, E., Bonora, G., Patel, S., and Plath, K. (2014). The Mbd1-Atf7ip-Setdb1 pathway contributes to the maintenance of X chromosome inactivation. Epigenetics & chromatin *7*, 1-15.

Mohammed, H., Taylor, C., Brown, G.D., Papachristou, E.K., Carroll, J.S., and D'santos, C.S. (2016). Rapid immunoprecipitation mass spectrometry of endogenous proteins (RIME) for analysis of chromatin complexes. Nature protocols *11*, 316-326.

Mohandas, T., Sparkes, R., and Shapiro, L. (1981). Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation. Science *211*, 393-396.

Moindrot, B., Cerase, A., Coker, H., Masui, O., Grijzenhout, A., Pintacuda, G., Schermelleh, L., Nesterova, T.B., and Brockdorff, N. (2015). A pooled shRNA screen identifies Rbm15, Spen, and Wtap as factors required for Xist RNA-mediated silencing. Cell reports *12*, 562-572.

Molaro, A., Malik, H.S., and Bourc'his, D. (2020). Dynamic evolution of de novo DNA methyltransferases in rodent and primate genomes. Molecular biology and evolution *37*, 1882-1892.

Monfort, A., Di Minin, G., Postlmayr, A., Freimann, R., Arieti, F., Thore, S., and Wutz, A. (2015). Identification of Spen as a crucial factor for Xist function through forward genetic screening in haploid embryonic stem cells. Cell reports *12*, 554-561.

Monk, M., Boubelik, M., and Lehnert, S. (1987). Temporal and regional changes in DNA methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development. Development *99*, 371-382.

Mulholland, C.B., Nishiyama, A., Ryan, J., Nakamura, R., Yiğit, M., Glück, I.M., Trummer, C., Qin, W., Bartoschek, M.D., and Traube, F.R. (2020). Recent evolution of a TET-controlled and DPPA3/STELLA-driven pathway of passive DNA demethylation in mammals. Nature communications *11*, 1-24.

Myant, K., and Stancheva, I. (2008). LSH cooperates with DNA methyltransferases to repress transcription. Molecular and cellular biology *28*, 215-226.

Myant, K., Termanis, A., Sundaram, A.Y., Boe, T., Li, C., Merusi, C., Burrage, J., Jose, I., and Stancheva, I. (2011). LSH and G9a/GLP complex are required for developmentally programmed DNA methylation. Genome research *21*, 83-94.

Nady, N., Lemak, A., Walker, J.R., Avvakumov, G.V., Kareta, M.S., Achour, M., Xue, S., Duan, S., Allali-Hassani, A., and Zuo, X. (2011). Recognition of multivalent histone states associated with heterochromatin by UHRF1 protein. Journal of Biological Chemistry *286*, 24300-24311.

Nagano, T., Mitchell, J.A., Sanz, L.A., Pauler, F.M., Ferguson-Smith, A.C., Feil, R., and Fraser, P. (2008). The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. Science *322*, 1717-1720.

Nakagawa, T., Kajitani, T., Togo, S., Masuko, N., Ohdan, H., Hishikawa, Y., Koji, T., Matsuyama, T., Ikura, T., and Muramatsu, M. (2008). Deubiquitylation of histone H2A activates transcriptional initiation via trans-histone cross-talk with H3K4 di-and trimethylation. Genes & development *22*, 37-49.

Nan, X., Ng, H.-H., Johnson, C.A., Laherty, C.D., Turner, B.M., Eisenman, R.N., and Bird, A. (1998). Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature *393*, 386-389.

Navarro, P., Chambers, I., Karwacki-Neisius, V., Chureau, C., Morey, C., Rougeulle, C., and Avner, P. (2008). Molecular coupling of Xist regulation and pluripotency. Science *321*, 1693-1695.

Neri, F., Rapelli, S., Krepelova, A., Incarnato, D., Parlato, C., Basile, G., Maldotti, M., Anselmi, F., and Oliviero, S. (2017). Intragenic DNA methylation prevents spurious transcription initiation. Nature *543*, 72-77.

Nesterova, T.B., Slobodyanyuk, S.Y., Elisaphenko, E.A., Shevchenko, A.I., Johnston, C., Pavlova, M.E., Rogozin, I.B., Kolesnikov, N.N., Brockdorff, N., and Zakian, S.M. (2001). Characterization of the genomic Xist locus in rodents reveals conservation of overall gene structure and tandem repeats but rapid evolution of unique sequence. Genome research *11*, 833-849.

Nesterova, T.B., Wei, G., Coker, H., Pintacuda, G., Bowness, J.S., Zhang, T., Almeida, M., Bloechl, B., Moindrot, B., Carter, E.J., Rodrigo, I.A., Pan, Q., Bi, Y., Song, C.X., and Brockdorff, N. (2019). Systematic allelic analysis defines the interplay of key pathways in X chromosome inactivation. Nature communications *10*, 1-15.

Ng, H.-H., Zhang, Y., Hendrich, B., Johnson, C.A., Turner, B.M., Erdjument-Bromage, H., Tempst, P., Reinberg, D., and Bird, A. (1999). MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex. Nature genetics *23*, 58-61.

Nishiyama, A., Yamaguchi, L., Sharif, J., Johmura, Y., Kawamura, T., Nakanishi, K., Shimamura, S., Arita, K., Kodama, T., and Ishikawa, F. (2013). Uhrf1-dependent H3K23 ubiquitylation couples maintenance DNA methylation and replication. Nature *502*, 249-253.

Nora, E.P., Goloborodko, A., Valton, A.-L., Gibcus, J.H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L.A., and Bruneau, B.G. (2017). Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. Cell *169*, 930-944. e922.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., Van Berkum, N.L., Meisig, J., and Sedat, J. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381-385.

Nowialis, P., Lopusna, K., Opavska, J., Haney, S.L., Abraham, A., Sheng, P., Riva, A., Natarajan, A., Guryanova, O., and Simpson, M. (2019). Catalytically inactive Dnmt3b rescues mouse embryonic development by accessory and repressive functions. Nature communications *10*, 1-16.

Nozawa, R.-S., Nagao, K., Igami, K.-T., Shibata, S., Shirai, N., Nozaki, N., Sado, T., Kimura, H., and Obuse, C. (2013). Human inactive X chromosome is compacted through a PRC2-independent SMCHD1-HBiX1 pathway. Nature structural & molecular biology *20*, 566-573.

Ogawa, Y., and Lee, J.T. (2003). Xite, X-inactivation intergenic transcription elements that regulate the probability of choice. Molecular cell *11*, 731-743.

Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell *99*, 247-257.

Ondičová, M., Oakey, R.J., and Walsh, C.P. (2020). Is imprinting the result of “friendly fire” by the host defense system? PLoS Genetics *16*, e1008599.

Ono, R., Taki, T., Taketani, T., Taniwaki, M., Kobayashi, H., and Hayashi, Y. (2002). LCX, leukemia-associated protein with a CXXC domain, is fused to MLL in acute myeloid leukemia with trilineage dysplasia having t (10; 11)(q22; q23). Cancer research *62*, 4075-4080.

Ostler, K., Davis, E., Payne, S., Gosalia, B., Exposito-Cespedes, J., Le Beau, M., and Godley, L. (2007). Cancer cells express aberrant DNMT3B transcripts encoding truncated proteins. Oncogene *26*, 5553-5563.

Ostler, K.R., Yang, Q., Vasanthakumar, A., Zhang, L., Looney, T.J., Naranjo, A., Salwen, H., Kocherginsky, M., Lahn, B.T., and Cohn, S.L. (2012). A truncated DNMT3B protein (DNMT3B7) drives DNA methylation and suppresses growth in human neuroblastoma. Cancer Research *72*, 1040-1040.

Otani, J., Nankumo, T., Arita, K., Inamoto, S., Ariyoshi, M., and Shirakawa, M. (2009). Structural basis for recognition of H3K4 methylation status by the DNA methyltransferase 3A ATRX–DNMT3–DNMT3L domain. EMBO reports *10*, 1235-1241.

Padeken, J., Methot, S.P., and Gasser, S.M. (2022). Establishment of H3K9-methylated heterochromatin and its functions in tissue differentiation and maintenance. Nature Reviews Molecular Cell Biology, 1-18.

Pandya-Jones, A., Markaki, Y., Serizay, J., Chitiashvili, T., Mancia Leon, W.R., Damianov, A., Chronis, C., Papp, B., Chen, C.-K., and McKee, R. (2020). A protein assembly mediates Xist localization and gene silencing. Nature *587*, 145-151.

Pant, V., Mariano, P., Kanduri, C., Mattsson, A., Lobanenkov, V., Heuchel, R., and Ohlsson, R. (2003). The nucleotides responsible for the direct physical contact between the chromatin insulator protein CTCF and the H19 imprinting control region manifest parent of origin-specific long-distance insulation and methylation-free domains. Genes & development *17*, 586-590.

Pasque, V., Tchieu, J., Karnik, R., Uyeda, M., Dimashkie, A.S., Case, D., Papp, B., Bonora, G., Patel, S., and Ho, R. (2014). X chromosome reactivation dynamics reveal stages of reprogramming to pluripotency. Cell *159*, 1681-1697.

Pikaard, C.S., Haag, J.R., Ream, T., and Wierzbicki, A.T. (2008). Roles of RNA polymerase IV in gene silencing. Trends in plant science *13*, 390-397.

Pidsley, R., Zotenko, E., Peters, T.J., Lawrence, M.G., Risbridger, G.P., Molloy, P., Van Djik, S., Muhlhausler, B., Stirzaker, C., and Clark, S.J. (2016). Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome biology *17*, 1-17.

Pintacuda, G., Wei, G., Roustan, C., Kirmizitas, B.A., Solcan, N., Cerase, A., Castello, A., Mohammed, S., Moindrot, B., and Nesterova, T.B. (2017a). hnRNPK recruits PCGF3/5-PRC1 to the Xist RNA B-repeat to establish polycomb-mediated chromosomal silencing. Molecular cell *68*, 955-969. e910.

Pintacuda, G., Young, A.N., and Cerase, A. (2017b). Function by structure: spotlights on Xist long non-coding RNA. Frontiers in molecular biosciences *4*, 90.

Piunti, A., and Shilatifard, A. (2016). Epigenetic balance of gene expression by Polycomb and COMPASS families. Science *352*, aad9780.

Pósfai, J., Bhagwat, A.S., Pósfai, G., and Roberts, R.J. (1989). Predictive motifs derived from cytosine methyltransferases. Nucleic acids research *17*, 2421-2435.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research *33*, D501-D504.

Qin, W., Wolf, P., Liu, N., Link, S., Smets, M., Mastra, F.L., Forné, I., Pichler, G., Hörl, D., and Fellinger, K. (2015). DNA methylation requires a DNMT1 ubiquitin interacting motif (UIM) and histone ubiquitination. Cell research *25*, 911-929.

Qiu, C., Sawada, K., Zhang, X., and Cheng, X. (2002). The PWWP domain of mammalian DNA methyltransferase Dnmt3b defines a new family of DNA-binding folds. Nature structural biology *9*, 217-224.

Quenneville, S., Verde, G., Corsinotti, A., Kapopoulou, A., Jakobsson, J., Offner, S., Baglivo, I., Pedone, P.V., Grimaldi, G., and Riccio, A. (2011). In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. Molecular cell *44*, 361-372.

Quinlan, A.R. (2014). BEDTools: the Swiss-army tool for genome feature analysis. Current protocols in bioinformatics *47*, 11.12. 11-11.12. 34.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Rajavelu, A., Jurkowska, R.Z., Fritz, J., and Jeltsch, A. (2012). Function and disruption of DNA methyltransferase 3a cooperative DNA binding and nucleoprotein filament formation. Nucleic acids research *40*, 569-580.

Rand, A.C., Jain, M., Eizenga, J.M., Musselman-Brown, A., Olsen, H.E., Akeson, M., and Paten, B. (2017). Mapping DNA methylation with high-throughput nanopore sequencing. Nature methods *14*, 411-413.

Ramirez-Carrozzi, V.R., Braas, D., Bhatt, D.M., Cheng, C.S., Hong, C., Doty, K.R., Black, J.C., Hoffmann, A., Carey, M., and Smale, S.T. (2009). A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. Cell *138*, 114-128.

Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. Nucleic acids research *42*, W187-W191.

Ramírez, J., Dege, C., Kutateladze, T.G., and Hagman, J. (2012). MBD2 and multiple domains of CHD4 are required for transcriptional repression by Mi-2/NuRD complexes. Molecular and cellular biology *32*, 5078-5088.

Ran, F., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. Nature protocols *8*, 2281-2308.

Rasmussen, K.D., Berest, I., Keβler, S., Nishimura, K., Simón-Carrasco, L., Vassiliou, G.S., Pedersen, M.T., Christensen, J., Zaugg, J.B., and Helin, K. (2019). TET2 binding to enhancers facilitates transcription factor recruitment in hematopoietic cells. Genome research *29*, 564-575.

Rasmussen, K.D., and Helin, K. (2016). Role of TET enzymes in DNA methylation, development, and cancer. Genes & development *30*, 733-750.

Rastan, S. (1983). Non-random X-chromosome inactivation in mouse X-autosome translocation embryos—location of the inactivation centre.

Rastan, S., and Robertson, E.J. (1985). X-chromosome deletions in embryo-derived (EK) cell lines associated with lack of X-chromosome inactivation.

Rauch, T., and Pfeifer, G.P. (2005). Methylated-CpG island recovery assay: a new technique for the rapid detection of methylated-CpG islands in cancer. Laboratory investigation *85*, 1172-1180.

Rausch, C., Hastert, F.D., and Cardoso, M.C. (2020). DNA modification readers and writers and their interplay. Journal of molecular biology *432*, 1731-1746.

Riggs, A.D. (1975). X inactivation, differentiation, and DNA methylation. Cytogenetic and Genome Research *14*, 9-25.

Rodermund, L., Coker, H., Oldenkamp, R., Wei, G., Bowness, J., Rajkumar, B., Nesterova, T., Susano Pinto, D.M., Schermelleh, L., and Brockdorff, N. (2021). Time-resolved structured illumination microscopy reveals key principles of Xist RNA spreading. Science *372*, eabe7500.

Rothbart, S.B., Krajewski, K., Nady, N., Tempel, W., Xue, S., Badeaux, A.I., Barsyte-Lovejoy, D., Martinez, J.Y., Bedford, M.T., and Fuchs, S.M. (2012). Association of UHRF1 with methylated H3K9 directs the maintenance of DNA methylation. Nature structural & molecular biology *19*, 1155-1160.

Rountree, M., and Selker, E. (2010). DNA methylation and the formation of heterochromatin in Neurospora crassa. Heredity *105*, 38-44.

Russell, L. (1983). X-autosome translocations in the mouse: their characterization and use as tools to investigate gene inactivation and gene action. Cytogenetics of the Mammalian X chromosome, 205-250.

Sado, T., Li, E., and Sasaki, H. (2002). Effect of TSIX disruption on XIST expression in male ES cells. Cytogenetic and genome research *99*, 115-118.

Sado, T., Okano, M., Li, E., and Sasaki, H. (2004). De novo DNA methylation is dispensable for the initiation and propagation of X chromosome inactivation.

Sado, T., Wang, Z., Sasaki, H., and Li, E. (2001). Regulation of imprinted X-chromosome inactivation in mice by Tsix. Development *128*, 1275-1286.

Sager, R., and Kitchin, R. (1975). Selective Silencing of Eukaryotic DNA: A molecular basis is proposed for programmed inactivation or loss of eukaryotic DNA. Science *189*, 426-433.

Saravanaraman, P., Selvam, M., Ashok, C., Srijyothi, L., and Baluchamy, S. (2020). De novo methyltransferases: Potential players in diseases and new directions for targeted therapy. Biochimie *176*, 85-102.

Sardina, J.L., Collombet, S., Tian, T.V., Gómez, A., Di Stefano, B., Berenguer, C., Brumbaugh, J., Stadhouders, R., Segura-Morales, C., and Gut, M. (2018). Transcription factors drive Tet2-mediated enhancer demethylation to reprogram cell fate. Cell stem cell *23*, 727-741. e729.

Saxonov, S., Berg, P., and Brutlag, D.L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proceedings of the National Academy of Sciences *103*, 1412-1417.

Schertzer, M.D., Braceros, K.C., Starmer, J., Cherney, R.E., Lee, D.M., Salazar, G., Justice, M., Bischoff, S.R., Cowley, D.O., and Ariel, P. (2019). lncRNA-induced spread of polycomb controlled by genome architecture, RNA abundance, and CpG island DNA. Molecular cell *75*, 523-537. e510.

Schnetz, M.P., Bartels, C.F., Shastri, K., Balasubramanian, D., Zentner, G.E., Balaji, R., Zhang, X., Song, L., Wang, Z., and LaFramboise, T. (2009). Genomic distribution of CHD7 on chromatin tracks H3K4 methylation patterns. Genome research *19*, 590-601.

Schübeler, D. (2015). Function and information content of DNA methylation. Nature *517*, 321-326.

Schuettengruber, B., Bourbon, H.-M., Di Croce, L., and Cavalli, G. (2017). Genome regulation by polycomb and trithorax: 70 years and counting. Cell *171*, 34-57.

Schwartz, B.E., and Ahmad, K. (2005). Transcriptional activation triggers deposition and removal of the histone variant H3. 3. Genes & development *19*, 804-814.

Seisenberger, S., Andrews, S., Krueger, F., Arand, J., Walter, J., Santos, F., Popp, C., Thienpont, B., Dean, W., and Reik, W. (2012). The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells. Molecular cell *48*, 849-862.

Serre, D., Lee, B.H., and Ting, A.H. (2010). MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. Nucleic acids research *38*, 391-399.

Shah, M.Y., Vasanthakumar, A., Barnes, N.Y., Figueroa, M.E., Kamp, A., Hendrick, C., Ostler, K.R., Davis, E.M., Lin, S., and Anastasi, J. (2010). DNMT3B7, a truncated DNMT3B isoform expressed in human tumors, disrupts embryonic development and accelerates lymphomagenesis. Cancer research *70*, 5840-5850.

Sharif, J., Muto, M., Takebayashi, S.-i., Suetake, I., Iwamatsu, A., Endo, T.A., Shinga, J., Mizutani-Koseki, Y., Toyoda, T., and Okamura, K. (2007). The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. Nature *450*, 908-912.

Shayevitch, R., Askayo, D., Keydar, I., and Ast, G. (2018). The importance of DNA methylation of exons on alternative splicing. Rna *24*, 1351-1362.

Shen, H., and Laird, P.W. (2012). In epigenetic therapy, less is more. Cell stem cell *10*, 353-354.

Shen, L., Inoue, A., He, J., Liu, Y., Lu, F., and Zhang, Y. (2014). Tet3 and DNA replication mediate demethylation of both the maternal and paternal genomes in mouse zygotes. Cell stem cell *15*, 459-471.

Shirane, K., Miura, F., Ito, T., and Lorincz, M.C. (2020). NSD1-deposited H3K36me2 directs de novo methylation in the mouse male germline and counteracts Polycomb-associated silencing. Nature genetics *52*, 1088-1098.

Shirohzu, H., Kubota, T., Kumazawa, A., Sado, T., Chijiwa, T., Inagaki, K., Suetake, I., Tajima, S., Wakui, K., and Miki, Y. (2002). Three novel DNMT3B mutations in Japanese patients with ICF syndrome. American journal of medical genetics *112*, 31-37.

Shumskaya, V.S., Zhigalova, N.A., Prokhorchouk, A.V., and Prokhorchouk, E.B. (2015). Distribution of Kaiso protein in mouse tissues. Histochemistry and cell biology *143*, 29-43.

Simon, M.D., Pinter, S.F., Fang, R., Sarma, K., Rutenberg-Schoenberg, M., Bowman, S.K., Kesner, B.A., Maier, V.K., Kingston, R.E., and Lee, J.T. (2013). High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. Nature *504*, 465-469.

Sinsheimer, R.L., Koerner, J.F., Vadla, J., and Lunan, K. (1954). The action of pancreatic desoxyribonuclease. I. Isolation of mono-and dinucleotides. Journal of Biological Chemistry *208*, 445-459.

Smallwood, S.A., Tomizawa, S.-i., Krueger, F., Ruf, N., Carli, N., Segonds-Pichon, A., Sato, S., Hata, K., Andrews, S.R., and Kelsey, G. (2011). Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. Nature genetics *43*, 811-814.

Smith, H.O., and Welcox, K. (1970). A restriction enzyme from Hemophilus influenzae: I. Purification and general properties. Journal of molecular biology *51*, 379-391.

Song, J., Rechkoblit, O., Bestor, T.H., and Patel, D.J. (2011). Structure of DNMT1-DNA complex reveals a role for autoinhibition in maintenance DNA methylation. science *331*, 1036-1040.

Song, J., Teplova, M., Ishibe-Murakami, S., and Patel, D.J. (2012). Structure-based mechanistic insights into DNMT1-mediated maintenance DNA methylation. Science *335*, 709-712.

Splinter, E., de Wit, E., Nora, E.P., Klous, P., van de Werken, H.J., Zhu, Y., Kaaij, L.J., van IJcken, W., Gribnau, J., and Heard, E. (2011). The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. Genes & development *25*, 1371-1383.

Stark, R., and Brown, G. DiffBind: Differential binding analysis of ChIP.

Stark, R., and Brown, G. (2011). DiffBind: differential binding analysis of ChIP-Seq peak data. R package version *100*.

Suetake, I., Mishima, Y., Kimura, H., Lee, Y.-H., Goto, Y., Takeshima, H., Ikegami, T., and Tajima, S. (2011). Characterization of DNA-binding activity in the N-terminal domain of the DNA methyltransferase Dnmt3a. Biochemical Journal *437*, 141-148.

Sun, Z., and Bernstein, E. (2019). Histone variant macroH2A: from chromatin deposition to molecular function. Essays in biochemistry *63*, 59-74.

Sunwoo, H., Colognori, D., Froberg, J.E., Jeon, Y., and Lee, J.T. (2017). Repeat E anchors Xist RNA to the inactive X chromosomal compartment through CDKN1A-interacting protein (CIZ1). Proceedings of the National Academy of Sciences *114*, 10654-10659.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., and Tsafou, K.P. (2015). STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic acids research *43*, D447-D452.

Tachibana, M., Matsumura, Y., Fukuda, M., Kimura, H., and Shinkai, Y. (2008). G9a/GLP complexes independently mediate H3K9 and DNA methylation to silence transcription. The EMBO journal *27*, 2681-2690.

Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., and Aravind, L. (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science *324*, 930-935.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. cell *126*, 663-676.

Takahashi, N., Coluccio, A., Thorball, C.W., Planet, E., Shi, H., Offner, S., Turelli, P., Imbeault, M., Ferguson-Smith, A.C., and Trono, D. (2019). ZNF445 is a primary regulator of genomic imprinting. Genes & development *33*, 49-54.

Takeshita, K., Suetake, I., Yamashita, E., Suga, M., Narita, H., Nakagawa, A., and Tajima, S. (2011). Structural insight into maintenance methylation by mouse DNA methyltransferase 1 (Dnmt1). Proceedings of the National Academy of Sciences *108*, 9055-9059.

Tang, Z., Chen, W.-Y., Shimada, M., Nguyen, U.T., Kim, J., Sun, X.-J., Sengoku, T., McGinty, R.K., Fernandez, J.P., and Muir, T.W. (2013). SET1 and p300 act synergistically, through coupled histone modifications, in transcriptional activation by p53. Cell *154*, 297-310.

Taylor, S.M., and Jones, P.A. (1979). Multiple new phenotypes induced in 10T12 and 3T3 cells treated with 5-azacytidine. Cell *17*, 771-779.

Thomson, J.P., Skene, P.J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., Kerr, A.R., Deaton, A., Andrews, R., and James, K.D. (2010). CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature *464*, 1082-1086.

Tian, D., Sun, S., and Lee, J.T. (2010). The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. Cell *143*, 390-403.

Tucci, V., Isles, A.R., Kelsey, G., Ferguson-Smith, A.C., Bartolomei, M.S., Benvenisty, N., Bourc'his, D., Charalambous, M., Dulac, C., and Feil, R. (2019). Genomic imprinting and physiological processes in mammals. Cell *176*, 952-965.

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. Nature protocols *10*, 475-483.

Uysal, F., Ozturk, S., and Akkoyunlu, G. (2017). DNMT1, DNMT3A and DNMT3B proteins are differently expressed in mouse oocytes and early embryos. Journal of molecular histology *48*, 417-426.

Vaisvila, R., Ponnaluri, V.C., Sun, Z., Langhorst, B.W., Saleh, L., Guan, S., Dai, N., Campbell, M.A., Sexton, B.S., and Marks, K. (2021). Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. Genome research *31*, 1280-1289.

Van Bemmel, J.G., Galupa, R., Gard, C., Servant, N., Picard, C., Davies, J., Szempruch, A.J., Zhan, Y., Żylicz, J.J., and Nora, E.P. (2019). The bipartite TAD organization of the X-inactivation center ensures opposing developmental regulation of Tsix and Xist. Nature genetics *51*, 1024-1034.

van Nuland, R., van Schaik, F., Simonis, M., van Heesch, S., Cuppen, E., Boelens, R., Timmers, H., and van Ingen, H. (2013). Nucleosomal DNA binding drives the recognition of H3K36-methylated nucleosomes by the PSIP1-PWWP domain. Epigenetics & chromatin *6*, 1-12.

Vanyushin, B., Tkacheva, S., and Belozersky, A. (1970). Rare bases in animal DNA. Nature *225*, 948-949.

Varnaitė, R., and MacNeill, S.A. (2016). Meet the neighbors: Mapping local protein interactomes by proximity-dependent labeling with BioID. Proteomics *16*, 2503-2518.

Vasanthakumar, A., Lepore, J.B., Zegarek, M.H., Kocherginsky, M., Singh, M., Davis, E.M., Link, P.A., Anastasi, J., Le Beau, M.M., and Karpf, A.R. (2013). Dnmt3b is a haploinsufficient tumor suppressor

gene in Myc-induced lymphomagenesis. Blood, The Journal of the American Society of Hematology *121*, 2059-2063.

Vedel, M., Gomez-Garcia, M., Sala, M., and Sala-Trepat, J.M. (1983). Changes in methylation pattern of albumin and α-fetoprotein genes in developing rat liver and neoplasia. Nucleic Acids Research *11*, 4335-4354.

Veland, N., Lu, Y., Hardikar, S., Gaddis, S., Zeng, Y., Liu, B., Estecio, M.R., Takata, Y., Lin, K., and Tomida, M.W. (2019). DNMT3L facilitates DNA methylation partly by maintaining DNMT3A stability in mouse embryonic stem cells. Nucleic acids research *47*, 152-167.

Velasco, G., Hubé, F., Rollin, J., Neuillet, D., Philippe, C., Bouzinba-Segard, H., Galvani, A., Viegas-Péquignot, E., and Francastel, C. (2010). Dnmt3b recruitment through E2F6 transcriptional repressor mediates germ-line gene silencing in murine somatic tissues. Proceedings of the National Academy of Sciences *107*, 9281-9286.

Vermeulen, M., Mulder, K.W., Denissov, S., Pijnappel, W.P., van Schaik, F.M., Varier, R.A., Baltissen, M.P., Stunnenberg, H.G., Mann, M., and Timmers, H.T.M. (2007). Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. Cell *131*, 58-69.

Veselovska, L., Smallwood, S.A., Saadeh, H., Stewart, K.R., Krueger, F., Maupetit-Méhouas, S., Arnaud, P., Tomizawa, S.-i., Andrews, S., and Kelsey, G. (2015). Deep sequencing and de novo assembly of the mouse oocyte transcriptome define the contribution of transcription to the DNA methylation landscape. Genome biology *16*, 1-17.

Vincent, J.J., Huang, Y., Chen, P.-Y., Feng, S., Calvopiña, J.H., Nee, K., Lee, S.A., Le, T., Yoon, A.J., and Faull, K. (2013). Stage-specific roles for tet1 and tet2 in DNA demethylation in primordial germ cells. Cell stem cell *12*, 470-478.

Viré, E., Brenner, C., Deplus, R., Blanchon, L., Fraga, M., Didelot, C., Morey, L., Van Eynde, A., Bernard, D., and Vanderwinden, J.-M. (2006). The Polycomb group protein EZH2 directly controls DNA methylation. Nature *439*, 871-874.

Waalwijk, C., and Flavell, R. (1978a). DNA methylation at a CCGG sequence in the large intron of the rabbit β-globin gene: tissue-specific variations. Nucleic acids research *5*, 4631-4642.

Waalwijk, C., and Flavell, R. (1978b). MspI, an isoschizomer of hpaII which cleaves both unmethylated and methylated hpaII sites. Nucleic acids research *5*, 3231-3236.

Walsh, C.P., Chaillet, J.R., and Bestor, T.H. (1998). Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. Nature genetics *20*, 116-117.

Wang, C.-Y., Jégu, T., Chu, H.-P., Oh, H.J., and Lee, J.T. (2018). SMCHD1 merges chromosome compartments and assists formation of super-structures on the inactive X. Cell *174*, 406-421. e425.

Wang, H., Farnung, L., Dienemann, C., and Cramer, P. (2020). Structure of H3K36-methylated nucleosome–PWWP complex reveals multivalent cross-gyre binding. Nature structural & molecular biology *27*, 8-13.

Wang, H., Maurano, M.T., Qu, H., Varley, K.E., Gertz, J., Pauli, F., Lee, K., Canfield, T., Weaver, M., and Sandstrom, R. (2012). Widespread plasticity in CTCF occupancy linked to DNA methylation. Genome research *22*, 1680-1688.

Wang, J., Bhutani, M., Pathak, A.K., Lang, W., Ren, H., Jelinek, J., He, R., Shen, L., Issa, J.-P., and Mao, L. (2007). ΔDNMT3B variants regulate DNA methylation in a promoter-specific manner. Cancer Research *67*, 10647-10652.

Wang, J., Walsh, G., Liu, D.D., Lee, J.J., and Mao, L. (2006a). Expression of ΔDNMT3B Variants and Its Association with Promoter Methylation of p16 and RASSF1A in Primary Non–Small Cell Lung Cancer. Cancer research *66*, 8361-8366.

Wang, L., Wang, J., Sun, S., Rodriguez, M., Yue, P., Jang, S.J., and Mao, L. (2006b). A novel DNMT3B subfamily, ΔDNMT3B, is the predominant form of DNMT3B in non-small cell lung cancer. International journal of oncology *29*, 201-207.

Wang, L., Zhang, J., Duan, J., Gao, X., Zhu, W., Lu, X., Yang, L., Zhang, J., Li, G., and Ci, W. (2014). Programming and inheritance of parental DNA methylomes in mammals. Cell *157*, 979-991.

Warder, D.E., and Keherly, M.J. (2003). Ciz1, Cip1 interacting zinc finger protein 1 binds the consensus DNA sequence ARYSR (0–2) YYAC. Journal of biomedical science *10*, 406-417.

Wassenegger, M., Heimes, S., Riedel, L., and Sänger, H.L. (1994). RNA-directed de novo methylation of genomic sequences in plants. Cell *76*, 567-576.

Watson, J.D., and Crick, F.H. (1953). Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature *171*, 737-738.

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*, 1-13.

Weber, M., Davies, J.J., Wittig, D., Oakeley, E.J., Haase, M., Lam, W.L., and Schuebeler, D. (2005). Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nature genetics *37*, 853-862.

Wei, G., Almeida, M., Pintacuda, G., Coker, H., Bowness, J.S., Ule, J., and Brockdorff, N. (2021). Acute depletion of METTL3 implicates N6-methyladenosine in alternative intron/exon inclusion in the nascent transcriptome. Genome research *31*, 1395-1408.

Weinberg, D.N., Papillon-Cavanagh, S., Chen, H., Yue, Y., Chen, X., Rajagopalan, K.N., Horth, C., McGuire, J.T., Xu, X., and Nikbakht, H. (2019). The histone mark H3K36me2 recruits DNMT3A and shapes the intergenic DNA methylation landscape. Nature *573*, 281-286.

Weinberg, D.N., Rosenbaum, P., Chen, X., Barrows, D., Horth, C., Marunde, M.R., Popova, I.K., Gillespie, Z.B., Keogh, M.-C., and Lu, C. (2021). Two competing mechanisms of DNMT3A recruitment regulate the dynamics of de novo DNA methylation at PRC1-targeted CpG islands. Nature genetics *53*, 794-800.

Weisenberger, D.J., Velicescu, M., Cheng, J.C., Gonzales, F.A., Liang, G., and Jones, P.A. (2004). Role of the DNA methyltransferase variant DNMT3b3 in DNA methylation. Molecular Cancer Research *2*, 62-72.

Wickham, H. (2016). Data analysis. In ggplot2 (Springer), pp. 189-201.

Wierzbicki, A.T., Haag, J.R., and Pikaard, C.S. (2008). Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. Cell *135*, 635-648.

Wilks, A.F., Cozens, P.J., Mattaj, I.W., and Jost, J.-P. (1982). Estrogen induces a demethylation at the 5'end region of the chicken vitellogenin gene. Proceedings of the National Academy of Sciences *79*, 4252-4255.

Wu, H., D'Alessio, A.C., Ito, S., Xia, K., Wang, Z., Cui, K., Zhao, K., Eve Sun, Y., and Zhang, Y. (2011). Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. Nature *473*, 389-393.

Wu, J.C., and Santi, D. (1987). Kinetic and catalytic mechanism of HhaI methyltransferase. Journal of Biological Chemistry *262*, 4778-4786.

Wutz, A., Rasmussen, T.P., and Jaenisch, R. (2002). Chromosomal silencing and localization are mediated by different domains of Xist RNA. Nature genetics *30*, 167-174.

Wysocka, J., Swigut, T., Xiao, H., Milne, T.A., Kwon, S.Y., Landry, J., Kauer, M., Tackett, A.J., Chait, B.T., and Badenhorst, P. (2006). A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. Nature *442*, 86-90.

Xie, L., Pelz, C., Wang, W., Bashar, A., Varlamova, O., Shadle, S., and Impey, S. (2011). KDM5B regulates embryonic stem cell self-renewal and represses cryptic intragenic transcription. The EMBO journal *30*, 1473-1484.

Xiong, Z., and Laird, P.W. (1997). COBRA: a sensitive and quantitative DNA methylation assay. Nucleic acids research *25*, 2532-2534.

Xu, C., Liu, K., Lei, M., Yang, A., Li, Y., Hughes, T.R., and Min, J. (2018). DNA sequence recognition of human CXXC domains and their structural determinants. Structure *26*, 85-95. e83.

Xu, Q., Xiang, Y., Wang, Q., Wang, L., Brind'Amour, J., Bogutz, A.B., Zhang, Y., Zhang, B., Yu, G., and Xia, W. (2019). SETD2 regulates the maternal epigenome, genomic imprinting and embryonic development. Nature genetics *51*, 844-856.

Xu, T.-H., Liu, M., Zhou, X.E., Liang, G., Zhao, G., Xu, H.E., Melcher, K., and Jones, P.A. (2020). Structure of nucleosome-bound DNA methyltransferases DNMT3A and DNMT3B. Nature *586*, 151-155.

Yagi, M., Kabata, M., Tanaka, A., Ukai, T., Ohta, S., Nakabayashi, K., Shimizu, M., Hata, K., Meissner, A., and Yamamoto, T. (2020). Identification of distinct loci for de novo DNA methylation by DNMT3A and DNMT3B during mammalian development. Nature communications *11*, 1-14.

Yamada, N., Hasegawa, Y., Yue, M., Hamada, T., Nakagawa, S., and Ogawa, Y. (2015). Xist exon 7 contributes to the stable localization of Xist RNA on the inactive X-chromosome. PLoS genetics *11*, e1005430.

Yamaguchi, S., Hong, K., Liu, R., Inoue, A., Shen, L., Zhang, K., and Zhang, Y. (2013a). Dynamics of 5-methylcytosine and 5-hydroxymethylcytosine during germ cell reprogramming. Cell research *23*, 329-339.

Yamaguchi, S., Hong, K., Liu, R., Shen, L., Inoue, A., Diep, D., Zhang, K., and Zhang, Y. (2012). Tet1 controls meiosis by regulating meiotic gene expression. Nature *492*, 443-447.

Yamaguchi, S., Shen, L., Liu, Y., Sendler, D., and Zhang, Y. (2013b). Role of Tet1 in erasure of genomic imprinting. Nature *504*, 460-464.

Yang, H., Bai, D., Li, Y., Yu, Z., Wang, C., Sheng, Y., Liu, W., Gao, S., and Zhang, Y. (2022). Allele-specific H3K9me3 and DNA methylation co-marked CpG-rich regions serve as potential imprinting control regions in pre-implantation embryo. Nature Cell Biology *24*, 783-792.

Yarychkivska, O., Shahabuddin, Z., Comfort, N., Boulard, M., and Bestor, T.H. (2018). BAH domains and a histone-like motif in DNA methyltransferase 1 (DNMT1) regulate de novo and maintenance methylation in vivo. Journal of Biological Chemistry *293*, 19466-19475.

Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P.K., Kivioja, T., Dave, K., and Zhong, F. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science *356*, eaaj2239.

Yoon, H.-G., Chan, D.W., Reynolds, A.B., Qin, J., and Wong, J. (2003). N-CoR mediates DNA methylation-dependent repression through a methyl CpG binding protein Kaiso. Molecular cell *12*, 723-734.

Young, A.N., Perlas, E., Ruiz-Blanes, N., Hierholzer, A., Pomella, N., Martin-Martin, B., Liverziani, A., Jachowicz, J.W., Giannakouros, T., and Cerase, A. (2021). Deletion of LBR N-terminal domains recapitulates Pelger-Huet anomaly phenotypes in mouse without disrupting X chromosome inactivation. Communications biology *4*, 1-8.

Zamudio, N., Barau, J., Teissandier, A., Walter, M., Borsos, M., Servant, N., and Bourc'his, D. (2015). DNA methylation restrains transposons from adopting a chromatin signature permissive for meiotic recombination. Genes & development *29*, 1256-1270.

Zemach, A., McDaniel, I.E., Silva, P., and Zilberman, D. (2010). Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science *328*, 916-919.

Zeng, H., Horie, K., Madisen, L., Pavlova, M.N., Gragerova, G., Rohde, A.D., Schimpf, B.A., Liang, Y., Ojala, E., and Kramer, F. (2008). An inducible and reversible mouse genetic rescue system. PLoS genetics *4*, e1000069.

Zeng, Y., Ren, R., Kaur, G., Hardikar, S., Ying, Z., Babcock, L., Gupta, E., Zhang, X., Chen, T., and Cheng, X. (2020). The inactive Dnmt3b3 isoform preferentially enhances Dnmt3b-mediated DNA methylation. Genes & development *34*, 1546-1558.

Zhang, H., Lang, Z., and Zhu, J.-K. (2018a). Dynamics and function of DNA methylation in plants. Nature reviews Molecular cell biology *19*, 489-506.

Zhang, T., Termanis, A., Özkan, B., Bao, X.X., Culley, J., de Lima Alves, F., Rappsilber, J., Ramsahoye, B., and Stancheva, I. (2016). G9a/GLP complex maintains imprinted DNA methylation in embryonic stem cells. Cell reports *15*, 77-85.

Zhang, X., Smits, A.H., van Tilburg, G., Ovaa, H., Huber, W., and Vermeulen, M. (2018b). Proteome-wide identification of ubiquitin interactions using UbIA-MS. Nature protocols *13*, 530-550.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S.W.-L., Chen, H., Henderson, I.R., Shinn, P., Pellegrini, M., and Jacobsen, S.E. (2006). Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. Cell *126*, 1189-1201.

Zhang, Y., Jurkowska, R., Soeroes, S., Rajavelu, A., Dhayalan, A., Bock, I., Rathert, P., Brandt, O., Reinhardt, R., and Fischle, W. (2010). Chromatin methylation activity of Dnmt3a and Dnmt3a/3L is

guided by interaction of the ADD domain with the histone H3 tail. Nucleic acids research *38*, 4246-4253.

Zhang, Y., Liu, T., and Meyer, C. a, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). Genome Biol *9*, R137.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., and Li, W. (2008). Model-based analysis of ChIP-Seq (MACS). Genome biology *9*, 1-9.

Zhang, Y., Xiang, Y., Yin, Q., Du, Z., Peng, X., Wang, Q., Fidalgo, M., Xia, W., Li, Y., and Zhao, Z.-a. (2018c). Dynamic epigenomic landscapes during early lineage specification in mouse embryos. Nature genetics *50*, 96-105.

Zhao, J., Ohsumi, T.K., Kung, J.T., Ogawa, Y., Grau, D.J., Sarma, K., Song, J.J., Kingston, R.E., Borowsky, M., and Lee, J.T. (2010). Genome-wide identification of polycomb-associated RNAs by RIP-seq. Molecular cell *40*, 939-953.

Zhu, H., Geiman, T.M., Xi, S., Jiang, Q., Schmidtmann, A., Chen, T., Li, E., and Muegge, K. (2006). Lsh is involved in de novo methylation of DNA. The EMBO journal *25*, 335-345.

Żylicz, J.J., Bousard, A., Žumer, K., Dossin, F., Mohammad, E., da Rocha, S.T., Schwalb, B., Syx, L., Dingli, F., and Loew, D. (2019). The implication of early chromatin changes in X chromosome inactivation. Cell *176*, 182-197. e123.

Appendix 1

XX Karyotyping of cell lines

The loss of one X chromosome among the XX ES cells when grown in culture, resulting in XO karyotype is a well reported issue in the XCI field and has direct implication on the conclusions drawn from the experiments performed in this work. I have therefore made use of several methods to ensure the XX karyotype among all the cell lines. In section 2.1.24 I have described the allele specific PCR and IF used for this purpose. In addition to the methods described in section 2.1.24, I also calculated allelic ratio of input samples for all the experiments after NGS to ensure the equal retention of both X chromosomes in all the cell lines. Figures A.1- A.4 shows the allelic ratio of the corresponding input DNA for all the cells lines used for ChIP and MBD-seq . Allelic ratio of 0.5 indicates the XX karyotype.

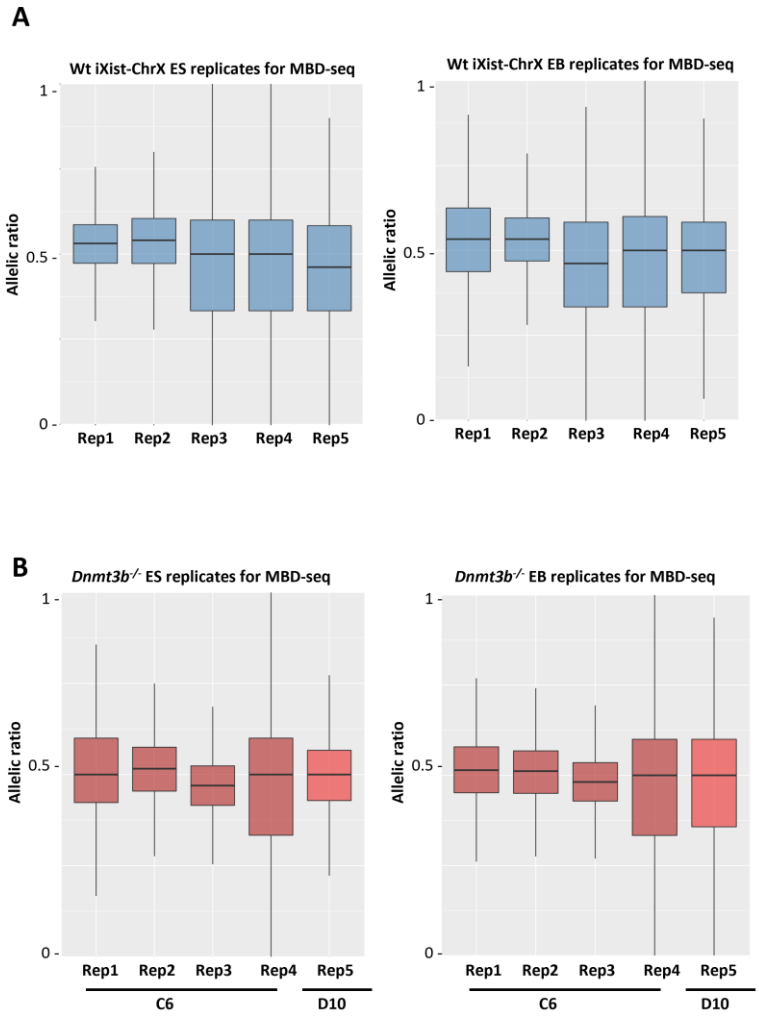


Figure A.1: Allelic ratio calculation for wt iXist-ChrX (A) and *Dnmt3b*$^{-/-}$ (B) ES and EBs used for MBD-seq.

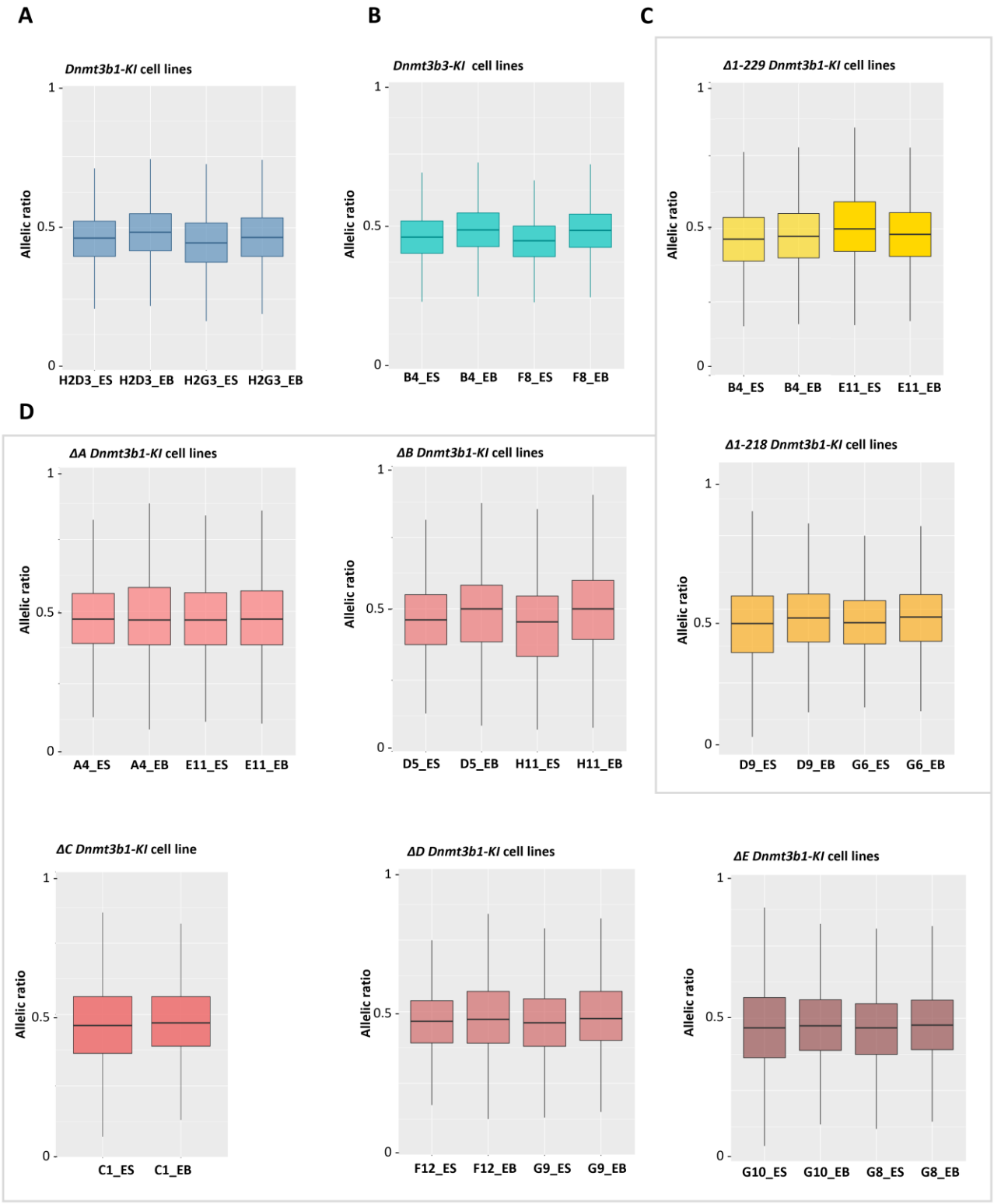


Figure A.2: XX karyotyping for KI lines. (A) Allelic ratio calculation for two clones each of *Dnmt3b1-KI*, *Dnmt3b3-KI* (B), *Δ1-229 Dnmt3b1 KI* (C, top), *Δ1-218 Dnmt3b1 KI* lines (C, bottom) and complementary cell lines expressing *Dnmt3b1* with smaller deletions in the N-terminal domain (D) in ES and EBs. Only one clone of *ΔC Dnmt3b1-KI* was used.

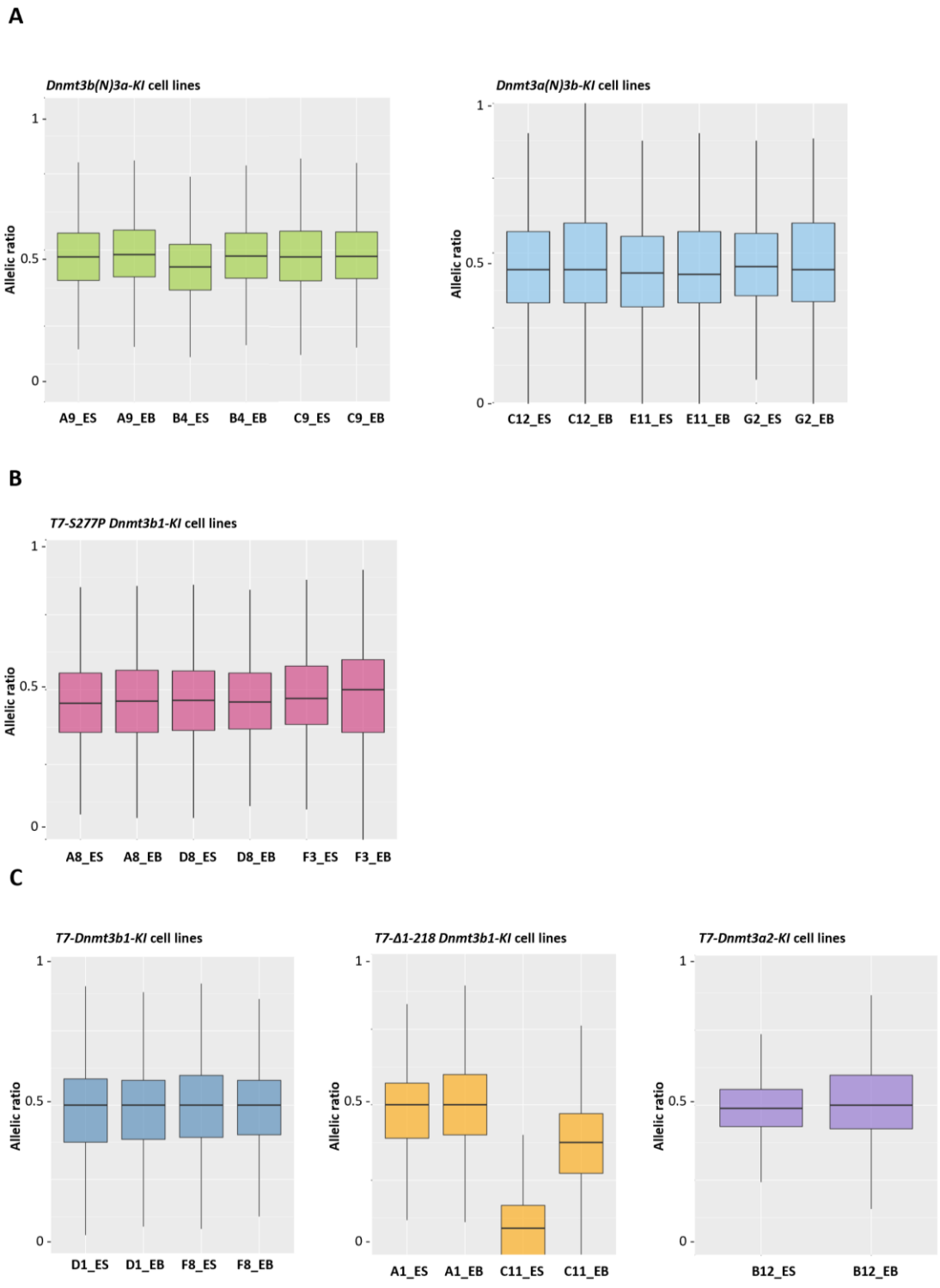


Figure A.3: Allelic ratio calculation for estimating XX karyotype of three clones each of complementary lines (A) targeted with chimeric *Dnmt3b* constructs, *Dnmt3b(N)3a-* and *Dnmt3a(N)3b-KI* lines in ES cells and EBs. (B) Same as (A) for three clones of *T7-S277P Dnmt3b1-KI*. (C) Same as (A and B) for two clones of *T7-Dnmt3b1-, T7-Δ1-218 Dnmt3b1-* and one clone of *T7-Dnmt3a2-KI line. T7-Δ1-218 Dnmt3b1* clone C11 lost X chromosome from *M.musculus Castaneous* carrying Tet-On Xist and was excluded from the analysis.

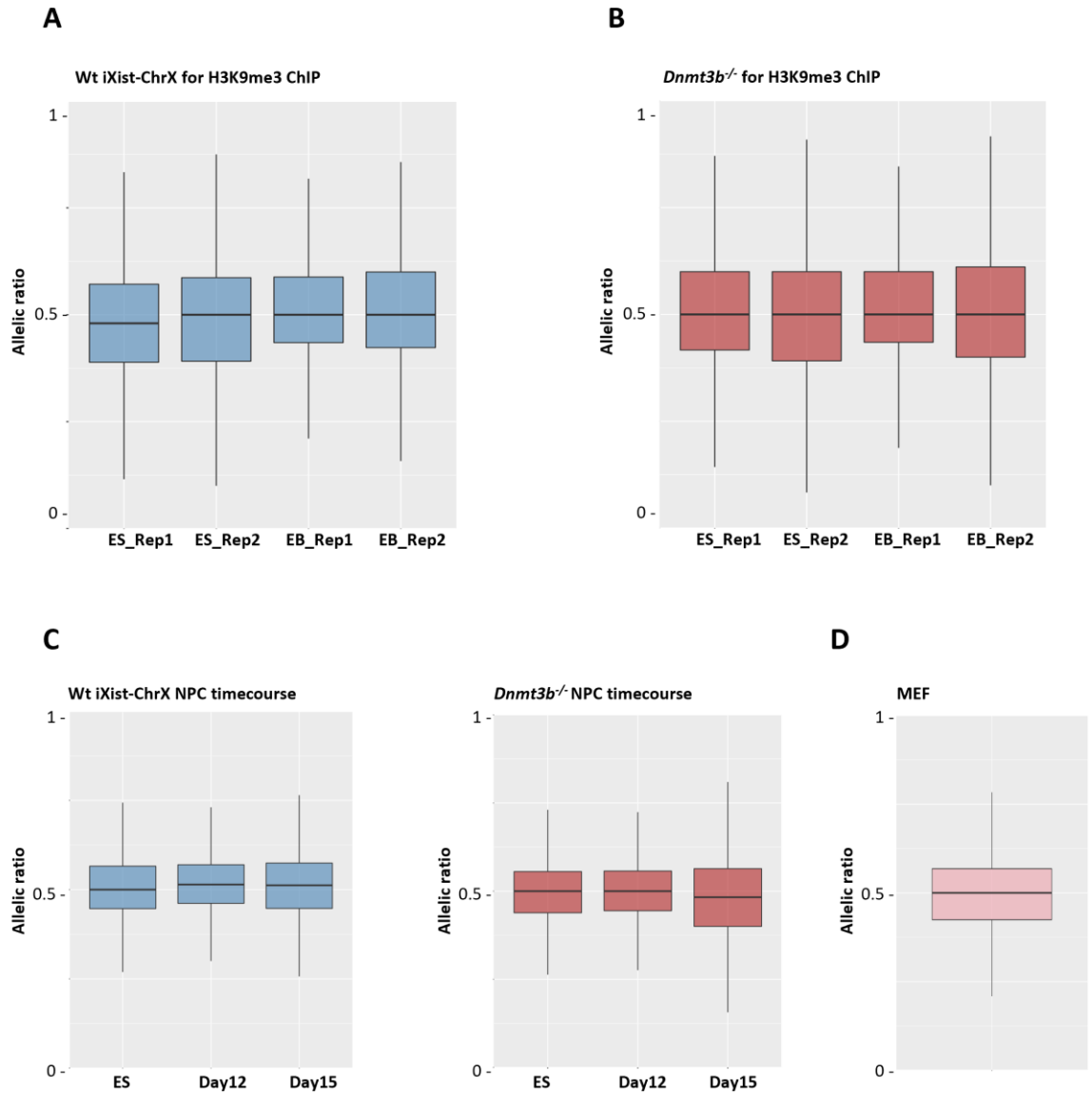


Figure A.4: Estimation of XX-karyotype of wt iXist-ChrX and *Dnmt3b*$^{-/-}$ lines (A and B) using allelic ratio of input samples form ES cells and EBs for H3K9me3 ChIP-seq. (C) Allelic ratio calculation for XX karyotype estimation of wt iXist-ChrX (left) and *Dnmt3b*$^{-/-}$ line (right) during NPC timecourse used for MBD-seq. (D) Allelic ratio calculation for MEF line used for comparing sensitivity of for BioCAP- and MBD-seq.

Appendix 2

Correlation between BioCAP-, MBD-, and ChIP-seq replicates

Metaplots, heatmaps and genome browser tracks in chapter 3 and 4 show one representative dataset out of three replicates, which were obtained from atleast two independent clones for each cell line. Figure A.5 shows high reproducibility among the BioCAP-, MBD- and ChIP-seq replicates.

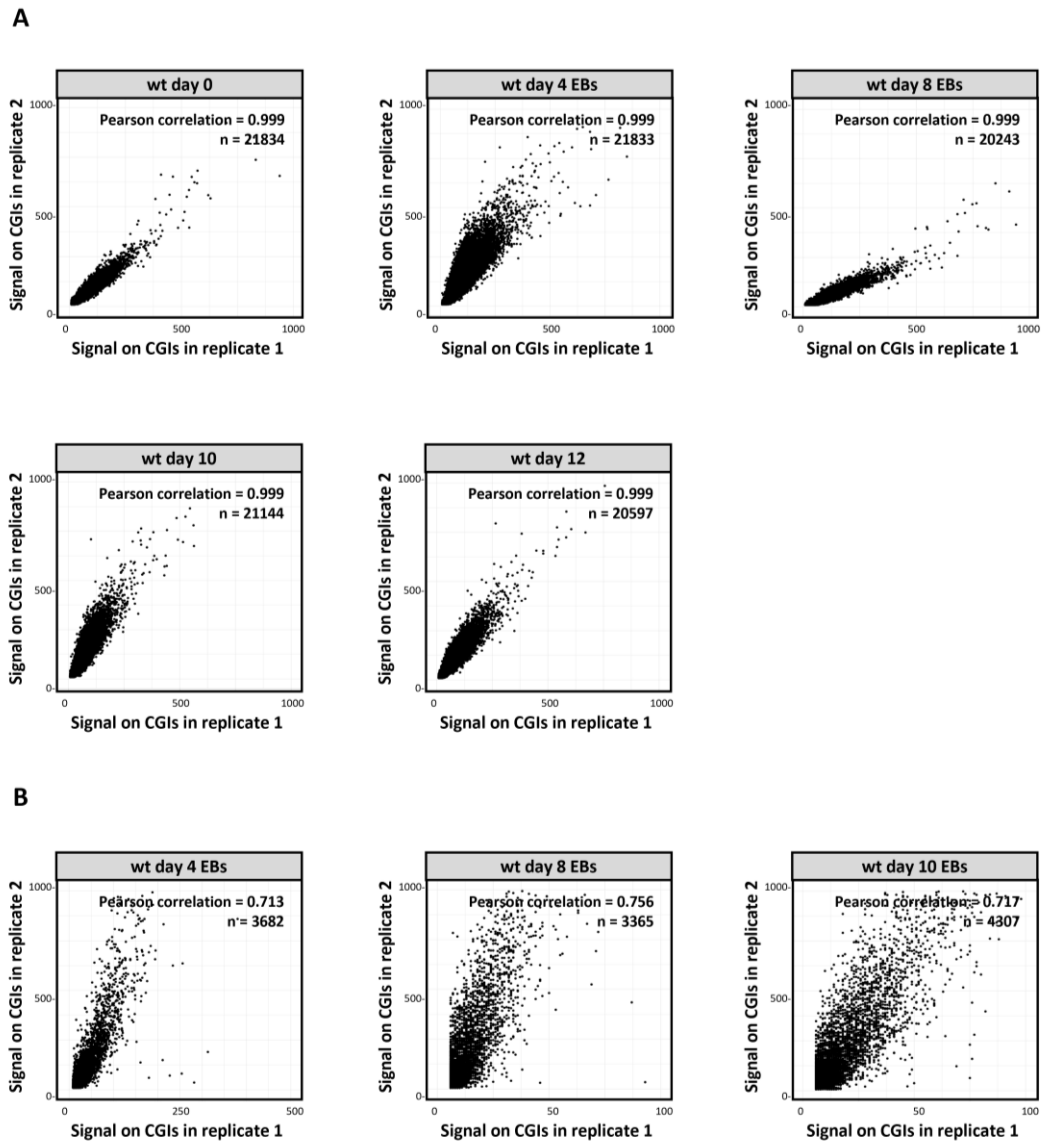
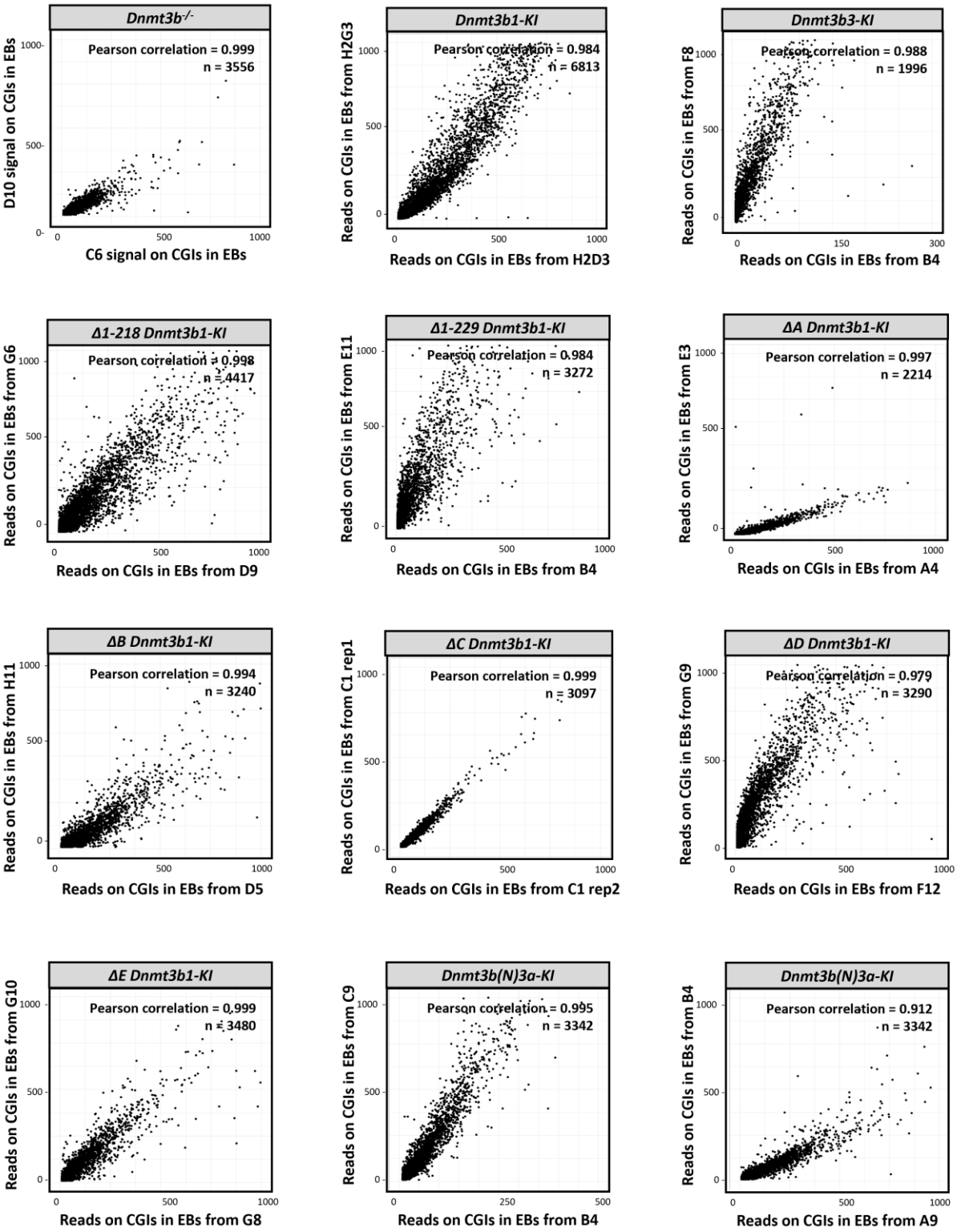
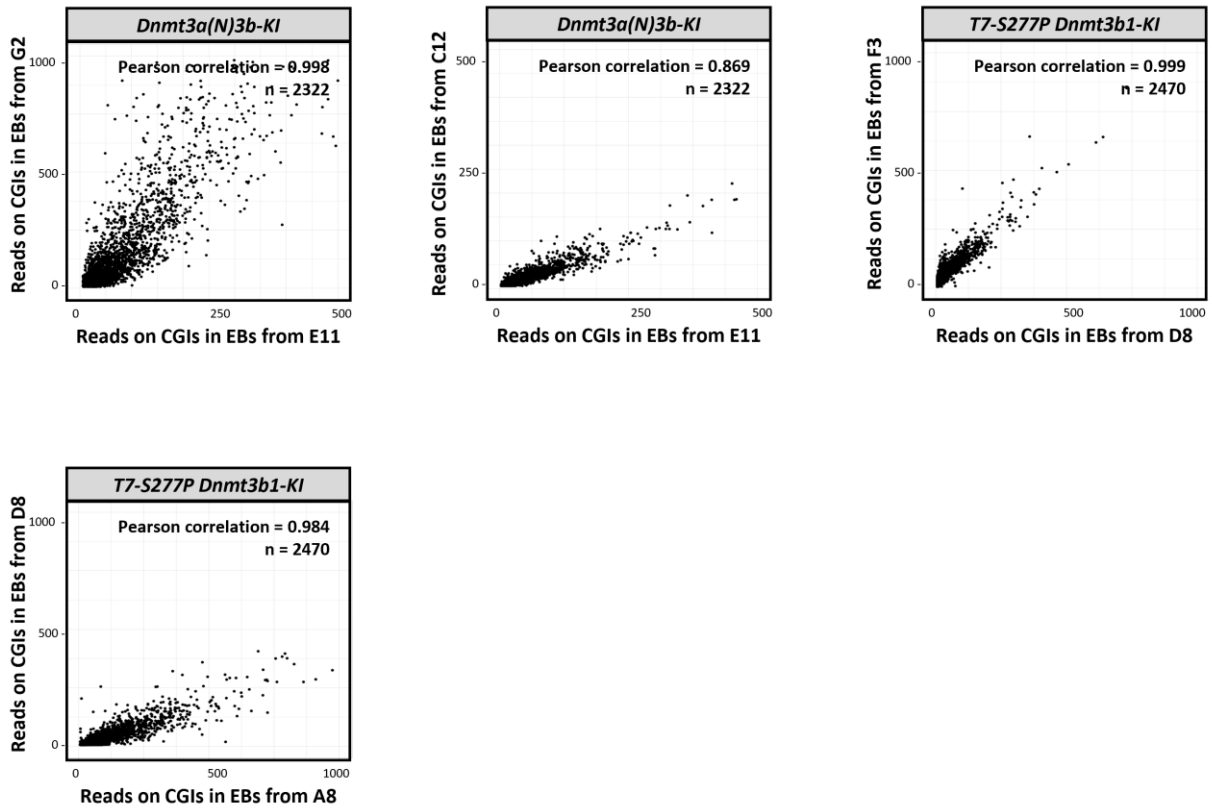


Figure A.5: Correlation plots showing reproducibility among replicates. (A) Scatter plots showing the correlation among BioCAP-seq replicates for wt iXist-ChrX line at five mentioned timepoints during EB differentiation. (B) Correlation plots showing reproducibility among MBD-seq replicates for wt iXist-ChrX line at day 4, 8 and 10 EB differentiation. Each dot represents the number of reads for a CGI at a given timepoint. The number of CGIs confidently captured (cut off =10 reads) in BioCAP- or MBD-seq for a given timepoint and the Pearson correlation among the replicates are show on top right corner.

C


Dnmt3b-/-
Pearson correlation = 0.999
n = 3556
D10 signal on CGIs in EBs
C6 signal on CGIs in EBs
Dnmt3b1-KI
Pearson correlation = 0.984
n = 6813
Reads on CGIs in EBs from H2G3
Reads on CGIs in EBs from H2D3
Dnmt3b3-KI
Pearson correlation = 0.988
n = 1996
Reads on CGIs in EBs from F8
Reads on CGIs in EBs from B4
Δ1-218 Dnmt3b1-KI
Pearson correlation = 0.998
n = 4417
Reads on CGIs in EBs from G6
Reads on CGIs in EBs from D9
Δ1-229 Dnmt3b1-KI
Pearson correlation = 0.984
n = 3272
Reads on CGIs in EBs from E11
Reads on CGIs in EBs from B4
ΔA Dnmt3b1-KI
Pearson correlation = 0.997
n = 2214
Reads on CGIs in EBs from E3
Reads on CGIs in EBs from A4
ΔB Dnmt3b1-KI
Pearson correlation = 0.994
n = 3240
Reads on CGIs in EBs from H11
Reads on CGIs in EBs from D5
ΔC Dnmt3b1-KI
Pearson correlation = 0.999
n = 3097
Reads on CGIs in EBs from C1 rep1
Reads on CGIs in EBs from C1 rep2
ΔD Dnmt3b1-KI
Pearson correlation = 0.979
n = 3290
Reads on CGIs in EBs from G9
Reads on CGIs in EBs from F12
ΔE Dnmt3b1-KI
Pearson correlation = 0.999
n = 3480
Reads on CGIs in EBs from G10
Reads on CGIs in EBs from G8
Dnmt3b(N)3a-KI
Pearson correlation = 0.995
n = 3342
Reads on CGIs in EBs from C9
Reads on CGIs in EBs from B4
Dnmt3b(N)3a-KI
Pearson correlation = 0.912
n = 3342
Reads on CGIs in EBs from B4
Reads on CGIs in EBs from A9

C (continued)



D

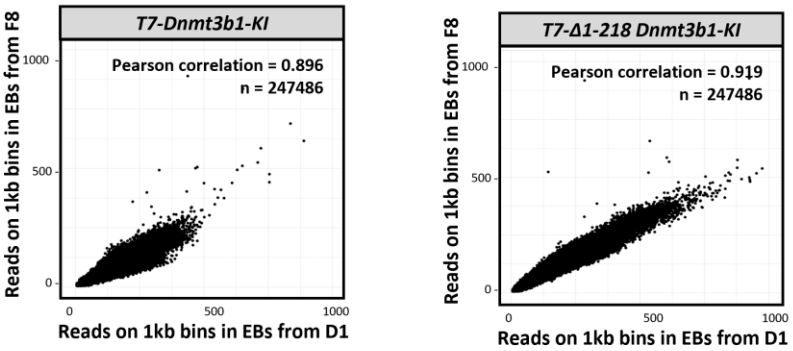


Figure A.5: Correlation plots showing reproducibility among replicates. (C) Plots showing correlation between MBD-seq datasets generated form individual clones for *Dnmt3b*$^{-/-}$ or complementation lines after 12 days of EB differentiation. (D) Correlation plots showing reproducibility among ChIP-seq replicates for *T7-Dnmt3b1-KI* and *T7-Δ1-218 Dnmt3b1-KI* lines. Each dot represents the number of reads for a CGI (C) or 1kb intervals (D) at day 12 EBs. The number of CGIs (or 1kb intervals in (D)) confidently captured (cut off =10 reads) and Pearson correlation is shown on top right.

Appendix 3

Catalytic assays for DNMT3B rescue constructs

The chimeric constructs *Dnmt3b(N)3a* and *Dnmt3a(N)3b* used in the complementation experiments were carefully designed to theoretically avoid disruption of protein structure. However, the possibility of conformational changes that render the protein catalytically inactive cannot be excluded. To address this caveat, I tried to test the catalytic activity of the complementation constructs. I made use of a *Dnmt1, Dnmt3a, Dnmt3b* triple knockout *TKO* J1 mouse ES cell line used earlier by (Okano et al., 1999) which are expected to be devoid of DNA methylation, therefore serving as a baseline/control. The enzymes to be tested (DNMT3B rescue constructs used to develop complementation lines) were introduced in this *TKO* cell line; therefore, any rise in cytosine methylation would be indicative of the catalytic activity of the test enzyme. I used two methods for estimating catalytic activity (1) LC/MS detection of methylated cytosine (mC) in the *TKO* complementation lines and (2) colorimetric ELISA-based assay (schematic in Figure A.5 and A.6). In agreement with the previously published reports, even the wildtype J1 expressing all DNMT proteins showed low mC/C ratio (mC/C= 4%). Although the complementation of *TKO* lines is widely used for the analysis of the catalytic activity of DNMTs, the levels of mC detected by LC/MS even upon the complementation with DNMT3B1 were very close to baseline (< 0.2% in comparison to 0.03% in *TKO* line) giving less confidence in the conclusions drawn from the experiment (Figure A.6). These low levels of mC/C ratio has also been reported earlier by (Baubec et al., 2015).

The second approach used a kit where an *in vitro* DNA methyltransferase reaction was carried out using nuclear extract of *TKO* lines expressing DNMT3B rescue constructs. The conversion of C to 5mC on substrate DNA is detected in a colorimetric ELISA-based method (see schematic in Figure A.6). In this method, purified protein control provided with the kit as control showed much greater catalytic activity in comparison to nuclear extracts of *TKO* cells expressing DNMT3B1. The exact values for OD/h/mg were inconsistent between experiments conducted at different times. Confounding factors such as insoluble contaminants and DNA fragments in the nuclear extracts could interfere with this colorimetric assay. Furthermore, this assay used naked DNA as substrate which does not simulate physiological substrate of DNMTs. Thus, although the chimeric DNMT3B constructs showed catalytic activity in these assays, I am not confident to draw solid conclusions from the catalytic assays (Figure A.7)

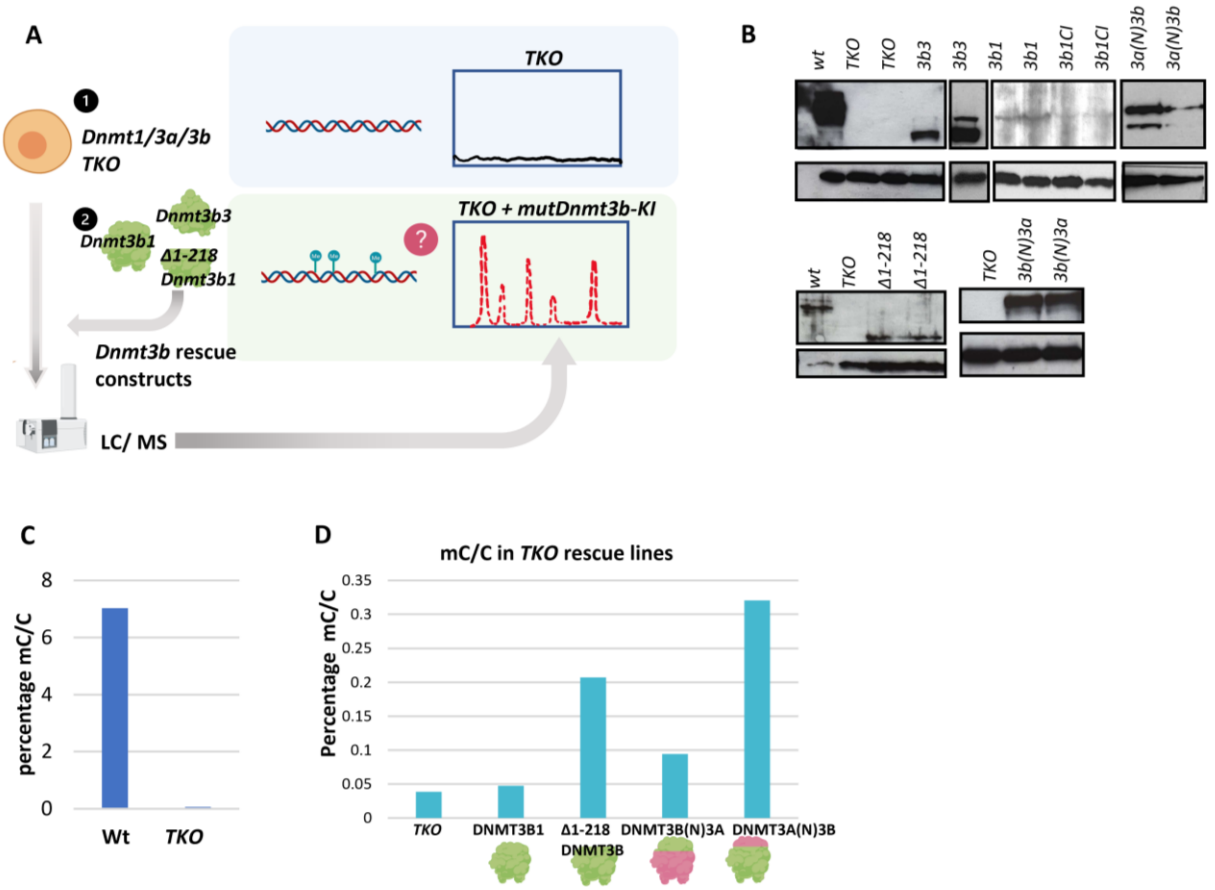


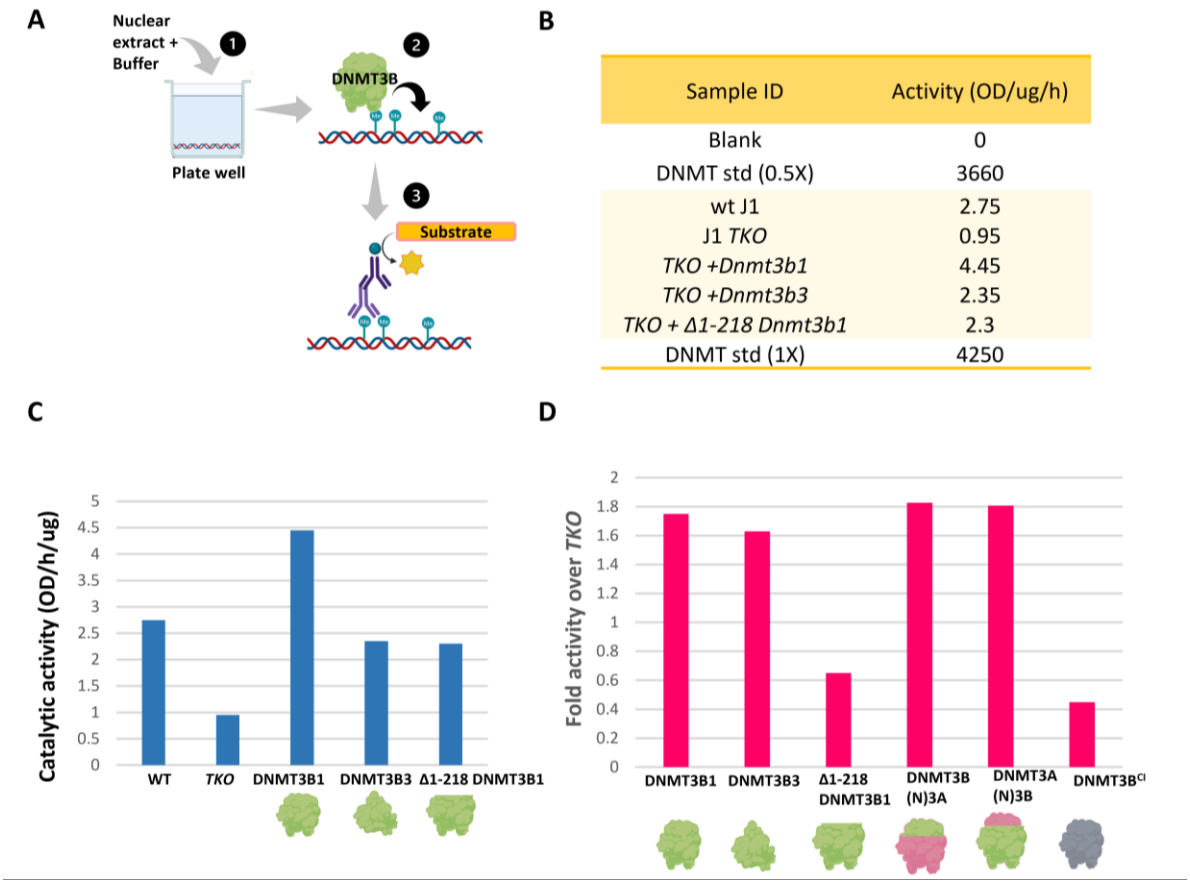
Figure A.6: LC/MS based assays to test the catalytic activity of DNMT3B rescue constructs.

A) Schematic explaining the LC/MS-based assay. The *TKO* line is expected to have basal levels of DNA methylation, therefore, serving as a good control. The *TKO* line could be transfected with *Dnmt3b* test constructs. If catalytically active, these DNMT3B rescue constructs would lead to the methylation of DNA. The DNA from these cells can be extracted and analysed by LC/MS which would calculate the proportion of 5mC/C. Methyl cytosine levels higher that *TKO* would indicate catalytic activity of the introduced rescue construct.

B) Western blot showing the successful integration and expression on complementation constructs in *TKO* line. Here wt represents wt J1 line, *TKO* is triple knock out line and the TKO complementation cell lines have been labelled by rescue construct targeted, 3b3 = *Dnmt3b3,* 3b1 = *Dnmt3b1*, 3b1CI = catalytically inactive *Dnmt3b1*^{CI} obtained by and used in the publication Nowialis et al., 2019. 3a(N)3b = *Dnmt3a(N)3b*, 3b(N)3a = *Dnmt3b(N)3a*, Δ1-218 = *Δ1-218 Dnmt3b1.*

C) Methyl cytosine to unmethylated cytosine ratio (mC/C) in wt and *TKO* J1 cells.

D) mC/C values for the *TKO* cell line expressing DNMT3B rescue constructs as obtained from LC/MS. All the rescue *TKO* lines show mC/C levels higher than *TKO* line. Contrary to expectation, DNMT3B1 show the lowest value for mC/C among the *TKO* complementation lines.



Sample ID	Activity (OD/ug/h)
Blank	0
DNMT std (0.5X)	3660
wt J1	2.75
J1 *TKO*	0.95
TKO +Dnmt3b1	4.45
TKO +Dnmt3b3	2.35
TKO + Δ1-218 Dnmt3b1	2.3
DNMT std (1X)	4250

Figure A.7: ELISA based assays to test the catalytic activity of DNMT3B rescue constructs.

A) Schematic showing the working of ELISA-based assay (provided by Abcam) for testing the catalytic activity of *Dnmt3b* rescue constructs. The assay uses a 96-well plate which is coded by substrate DNA. For every reaction, one well is used. The reaction buffer which is supplemented with SAM, is loaded on the well (1) along with 5 mg of the nuclear extract from the cell lines expressing the proteins of interest (*TKO* line expressing DNMT3B rescue construct) (2). The reaction is allowed to be carried out for 2 hours. After the incubation is complete the reaction buffer is removed and washed, and the primary antibody is added which recognises 5mC. The primary antibody in turn is recognised by secondary antibody, which is conjugated with horseradish peroxidase (HRP). HRP leads to colour change in presence of its substrate which can be detected using a platereader (3). The catalysis is expressed as a change in fluorescence per hour (of reaction) per mg of protein in the nuclear extract (OD/h/mg).

B) Table showing the catalytic activity of wt, *TKO* and DNMT3B rescue constructs (DNMT3B1, DNMT3B3, Δ1-218 DNMT3B1). The catalysis shown by purified proteins provided with the kit (shown as DNMT std) is magnitudes higher than the tested constructs.

C) Graph showing the data displayed in table in (B).

D) Graph showing the catalytic activity of DNMT3B rescue constructs. The graph shows average values from two replicates. Catalytically inactive DNMT3B3 shows unexpectedly high value of catalytic activity. Inconsistent with LC/MS assay, Δ1-218 DNMT3B1 construct showed lower catalytic activity.

Appendix 4

Differential methylation analysis of complementation lines expressing DNMT3B1 with smaller deletions in the N-terminus

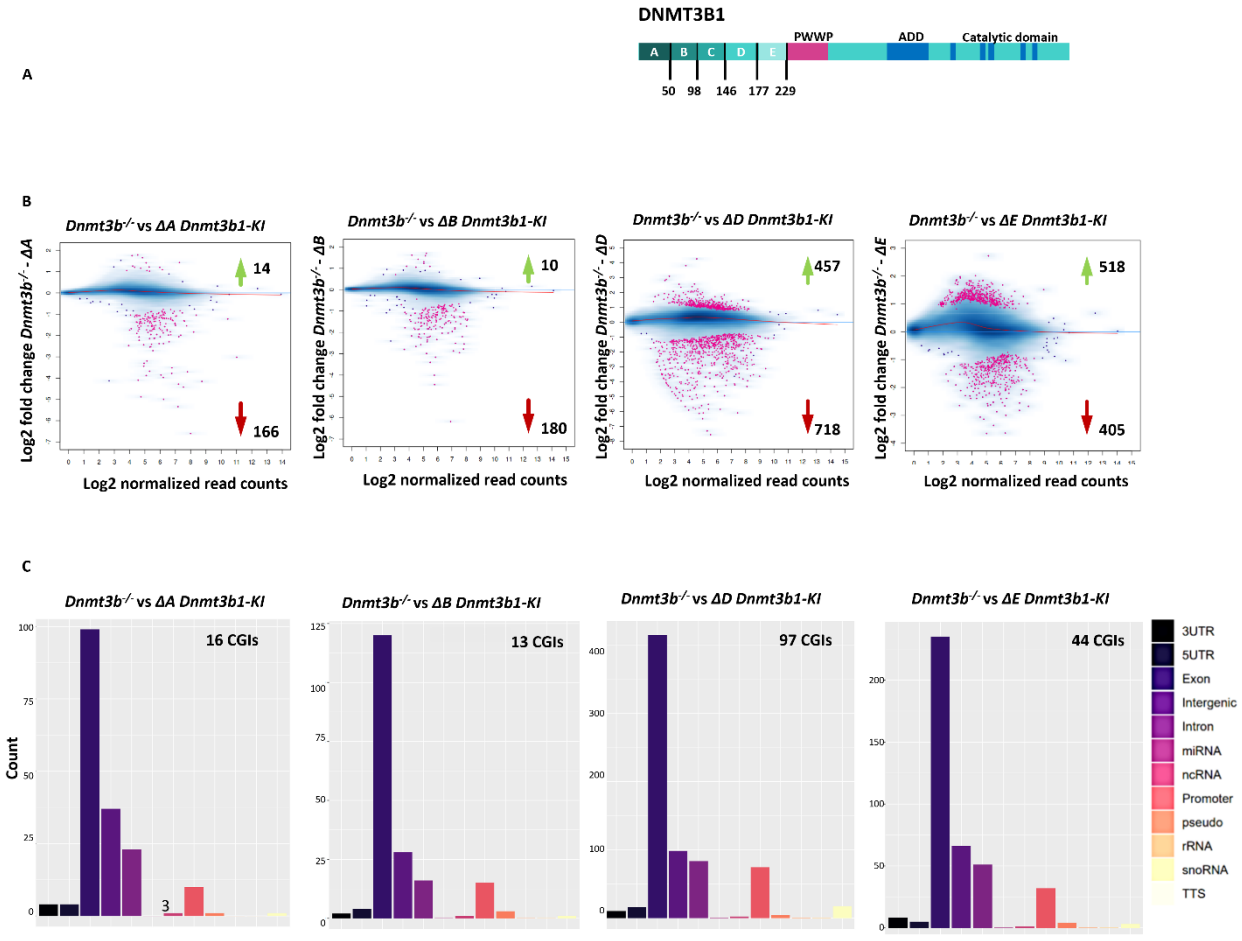


Figure A.8: Differentially methylated regions found in complementation lines expressing *Dnmt3b* with smaller deletions in the N-terminal domain at day 12 EBs.

A) Schematic showing the sub-regions of DNMT3B N-terminus deleted for the experiment.

B) MA plot showing the differentially methylated regions in different N-terminal domain deletions. The numbers of hypo and hypermethylated regions in the complementation lines in comparison to *Dnmt3b*$^{-/-}$ cells are indicated by up and downwards facing arrows. The number of differentially methylated regions are more of *ΔD* and *ΔE-Dnmt3b1-KI* lines as compared to *ΔA* and *ΔB-Dnmt3b1-KI* lines. *ΔC Dnmt3b1-KI* line is not shown here due to absence of biological replicates for this line.

C) The enrichment of hypermethylated peaks found in complementation lines shown in (B) over genomic features. Most of the regions methylated by these DNMT3B1 constructs reside within exons.

Appendix 5

Key information regarding CGI methylation on Xi

CGI_Start	CGI_End	Day 0 (RPKM)	Day 12 (RPKM)	Max methylation	Associated Gene	Gene silencing speed	Expression of associated gene	Day with Allelic ratio= 1	Methylation dynamics	DNMT3B dependence	SMCHD1 dependence
10485674	10486150	0.0	0.0	Day0	Tspan7	medium	intermediate	Day0	Unmeth	NA	NA
11664228	11665429	763.1	495.5	Day0	H2al1j	NA	NA	Day8	slow	Dep	NA
5467794	5468844	293.5	346.9	Day12	Mycs	NA	NA	Day10	Very Slow	Dep	NA
5558052	5558407	0.0	5.5	Day10	Mycs	NA	NA	Day10	Very Slow	Dep	NA
6081998	6082801	4519.7	8407.7	Day12	AU022751	NA	intermediate	Day12	Very Slow	Dep	NA
6082863	6083532	1232.6	3391.7	Day12	AU022751	NA	intermediate	Day4	Fast	Dep	NA
6092012	6092468	209.6	5.5	Day0	AU022751	NA	intermediate	Day8	slow	Dep	NA
6399570	6401029	0.0	352.4	Day12	Shroom4	medium	intermediate	Day8	slow	Dep	NA
6780132	6780498	0.0	0.0	Day8	Dgkk	medium	low	Day0	Unmeth	NA	NA
7318470	7319604	0.0	280.8	Day12	Clcn5	medium	high	Day8	slow	Indep	Dep
7374853	7376449	0.0	1046.2	Day12	Usp27x	NA	intermediate	Day4	Fast	Dep	NA
7422552	7422980	0.0	0.0	Day8	2010204K13Rik	medium	high	Day0	Unmeth	NA	NA
7573136	7574828	142.6	1222.3	Day12	Ppp1r3f	slow	low	Day10	Very Slow	Dep	NA
7610581	7611111	184.5	115.6	Day0	Cacna1f	NA	NA	Day4	Fast	Dep	NA
7638429	7639050	184.5	110.1	Day0	Syp	NA	intermediate	Day8	slow	Dep	Dep
7648605	7648980	0.0	11.0	Day10	Prickle3	medium	low	Day4	Fast	Dep	Dep
7657075	7657603	150.9	302.8	Day12	Prickle3	medium	low	Day8	slow	Dep	Dep
7664998	7665352	125.8	594.7	Day12	Plp2	medium	high	Day4	Fast	Dep	NA
7679787	7680264	0.0	16.5	Day10	Magix	NA	low	Day8	slow	Dep	NA
7697066	7697442	0.0	71.6	Day10	Gpkow	slow	intermediate	Day8	slow	Dep	NA
7714167	7714557	0.0	77.1	Day10	Wdr45	slow	high	Day8	slow	Dep	NA
7728108	7728907	0.0	99.1	Day10	Praf2	slow	high	Day4	Fast	Dep	Dep
7750378	7751133	0.0	66.1	Day12	Mir684-1	NA	NA	Day8	slow	Dep	NA
7762401	7763173	0.0	231.3	Day12	Tfe3	slow	high	Day8	slow	Dep	NA
7810167	7810528	369.0	1459.1	Day12	Mir1198	NA	NA	Day4	Fast	Dep	NA
7841006	7843131	16.8	1112.2	Day12	Otud5	slow	high	Day8	slow	Dep	Dep
7877930	7878976	436.0	198.2	Day0	Pim2	slow	intermediate	Day8	slow	Dep	NA
7883957	7884727	0.0	93.6	Day12	Slc35a2	slow	high	Day8	slow	Dep	NA
7898927	7899907	0.0	291.8	Day12	Timm17b	slow	high	Day8	slow	Dep	NA
7919748	7920143	192.9	154.2	Day0	Pcsk1n	NA	intermediate	Day12	Very Slow	Dep	Dep
7921905	7922849	478.0	462.5	Day0	Pcsk1n	NA	intermediate	Day8	slow	Dep	Dep
7922779	7923156	125.8	16.5	Day0	Pcsk1n	NA	intermediate	Day10	Very Slow	Dep	Dep
7947492	7948524	0.0	88.1	Day12	Hdac6	slow	high	Day8	slow	Dep	Dep
8073770	8074947	16.8	539.6	Day12	Suv39h1	slow	high	Day8	slow	Dep	NA
8096484	8096958	0.0	181.7	Day12	Gm6787	NA	low	Day8	slow	Dep	NA
8132138	8132685	0.0	0.0	Day0	Wdr13	slow	intermediate	Day0	Unmeth	NA	NA
8145139	8145954	0.0	820.4	Day12	Rbm3os	NA	low	Day4	Fast	Indep	Dep
8175313	8176213	0.0	363.4	Day12	Tbc1d25	slow	intermediate	Day4	Fast	Dep	Dep
8192517	8193090	0.0	0.0	Day10	Ebp	slow	high	Day0	Unmeth	NA	NA
8205989	8206841	0.0	33.0	Day10	Porcn	slow	high	Day0	Unmeth	NA	NA
8252172	8252570	0.0	16.5	Day12	Ftsj1	slow	intermediate	Day8	slow	Dep	NA
8892279	8893957	0.0	1398.5	Day12	B630019K06Rik	NA	low	Day8	slow	Dep	Dep
9199639	9201046	33.5	1277.4	Day12	Lancl3	medium	low	Day4	Fast	Dep	Dep
9272316	9273498	595.4	446.0	Day0	Xk	slow	intermediate	Day8	slow	Dep	NA
10216361	10217067	0.0	126.6	Day12	Rpgr	slow	intermediate	Day8	slow	Indep	Dep
10484754	10485522	0.0	66.1	Day12	Tspan7	medium	intermediate	Day8	slow	Dep	NA
10650451	10650826	50.3	253.3	Day10	Mid1ip1	NA	intermediate	Day10	Very Slow	Dep	Dep
10717095	10717762	0.0	49.6	Day12	Mid1ip1	NA	intermediate	Day0	Unmeth	NA	Dep
10717884	10718799	0.0	209.2	Day12	Mid1ip1	NA	intermediate	Day0	Unmeth	NA	Dep
11806827	11808096	628.9	280.8	Day0	Bcor	slow	high	Day4	Fast	Dep	Dep
11998778	11999578	268.3	93.6	Day0	Bcor	slow	high	Day8	slow	Dep	Dep
12005648	12007732	0.0	159.7	Day12	Bcor	slow	high	Day8	slow	Dep	Dep
12076654	12077922	737.9	93.6	Day0	Bcor	slow	high	Day8	slow	Dep	Dep
12078347	12081760	1542.9	1701.4	Day12	Bcor	slow	high	Day8	slow	Dep	Dep
12086820	12089286	955.9	688.3	Day0	Bcor	slow	high	Day4	Fast	Dep	Dep
12089573	12090176	511.5	181.7	Day0	Bcor	slow	high	Day8	slow	Dep	Dep
12090913	12092179	41.9	297.3	Day12	Bcor	slow	high	Day8	slow	Dep	Dep
12127508	12129218	67.1	137.7	Day10	Bcor	slow	high	Day10	Very Slow	Dep	Dep
12129178	12129625	0.0	22.0	Day12	Bcor	slow	high	Day0	Unmeth	NA	Dep
12133291	12133974	75.5	60.6	Day0	Bcor	slow	high	Day8	slow	Dep	Dep
12134427	12136115	192.9	121.1	Day0	Bcor	slow	high	Day10	Very Slow	Dep	Dep
12136273	12137793	528.3	88.1	Day0	Bcor	slow	high	Day8	slow	Dep	Dep
12139834	12140741	503.1	71.6	Day0	Bcor	slow	high	Day8	slow	Indep	Dep
12151714	12152114	192.9	22.0	Day0	Bcor	slow	high	Day8	slow	Dep	Dep
12155786	12156258	83.9	0.0	Day0	Bcor	slow	high	Day10	Very Slow	Dep	Dep
12158332	12159235	360.6	247.8	Day0	Bcor	slow	high	Day8	slow	Dep	Dep
12159546	12160696	352.2	291.8	Day0	Bcor	slow	high	Day8	slow	Dep	Dep
12252364	12252819	0.0	0.0	Day8	Bcor	slow	high	Day8	slow	Dep	Dep
12587347	12588189	0.0	143.2	Day10	Atp6ap2	slow	high	Day8	slow	Dep	NA
12652168	12652681	276.7	1255.4	Day12	Gm4984	NA	intermediate	Day4	Fast	Dep	NA
12673115	12673760	0.0	5.5	Day10	1810030O07Rik	slow	high	Day10	Very Slow	Dep	Dep
12761265	12762748	0.0	501.1	Day12	Med14	slow	intermediate	Day8	slow	Dep	Dep
12936771	12937356	461.2	985.6	Day12	AA414768	NA	intermediate	Day4	Fast	Dep	NA
13041576	13042178	0.0	16.5	Day12	5730405O15Rik	slow	low	Day0	Unmeth	NA	NA
13070816	13073187	0.0	847.9	Day12	Usp9x	slow	high	Day8	slow	Dep	Dep
13280700	13282349	0.0	214.7	Day12	Ddx3x	slow	high	Day8	slow	Indep	NA
13461519	13462204	352.2	77.1	Day0	Nyx	NA	low	Day10	Very Slow	Dep	Dep
13485801	13487142	981.1	11122.2	Day12	Nyx	NA	low	Day4	Fast	Dep	Dep
13845492	13847241	0.0	798.4	Day12	Cask	medium	intermediate	Day4	Fast	Dep	Dep
16618172	16619159	0.0	38.5	Day12	Maoa	slow	intermediate	Day10	Very Slow	Indep	Dep
17319281	17319683	117.4	104.6	Day0	Efhc2	slow	intermediate	Day4	Fast	Dep	NA
17513429	17513806	92.2	148.7	Day12	Fundc1	slow	intermediate	Day4	Fast	Dep	NA
18161881	18163999	0.0	5.5	Day12	Kdm6a	slow	high	Day4	Fast	Dep	NA
19237267	19238235	0.0	5.5	Day10	Mir222	NA	NA	Day8	slow	Dep	NA
20059332	20061348	1023.0	1508.7	Day12	Chst7	medium	intermediate	Day8	slow	Dep	Dep
20291187	20292284	385.7	341.4	Day0	Slc9a7	medium	low	Day8	slow	Dep	NA
20425060	20426571	0.0	578.1	Day12	Jade3	medium	intermediate	Day8	slow	Dep	Dep
20616904	20617535	0.0	49.6	Day10	Rbm10	slow	high	Day8	slow	Dep	Dep
20617696	20618401	0.0	49.6	Day12	Ndufb11	slow	high	Day0	Unmeth	NA	NA
20688088	20689301	0.0	401.9	Day12	Cdk16	slow	high	Day8	slow	Dep	Dep
20703814	20704678	159.3	159.7	Day12	Usp11	slow	high	Day0	Unmeth	NA	NA
20848482	20848960	0.0	11.0	Day10	Araf	slow	high	Day8	slow	Dep	NA
20920232	20921323	905.6	572.6	Day0	Syn1	medium	low	Day10	Very Slow	Dep	NA
20950265	20950660	0.0	16.5	Day12	Elk1	slow	intermediate	Day4	Fast	Dep	NA
20961480	20962317	8.4	44.0	Day10	Uxt	slow	high	Day12	Very Slow	Dep	NA
21061877	21062268	0.0	77.1	Day10	Zfp182	slow	intermediate	Day0	Unmeth	NA	NA
21360693	21361191	268.3	1156.3	Day12	Gm5124	NA	low	Day4	Fast	Indep	NA
23364513	23365261	134.2	11.0	Day0	Klhl13	slow	high	Day0	Unmeth	NA	NA
23692624	23693592	0.0	27.5	Day12	Wdr44	slow	intermediate	Day0	Unmeth	NA	NA
25752570	25752961	8.4	99.1	Day12	Gm1993	NA	NA	Day0	Unmeth	NA	NA
35888448	35889627	134.2	253.3	Day10	Dock11	medium	intermediate	Day0	Unmeth	NA	NA
36111746	36112669	0.0	407.4	Day12	Il13ra1	medium	intermediate	Day8	slow	Dep	NA
36328260	36329803	721.1	1051.7	Day12	Lonrf3	NA	intermediate	Day0	Unmeth	NA	NA
36329720	36330557	0.0	126.6	Day12	Lonrf3	NA	intermediate	Day0	Unmeth	NA	NA
36597772	36598881	0.0	1327.0	Day12	Pgrmc1	medium	high	Day4	Fast	Dep	NA
36644678	36645594	150.9	346.9	Day12	Akap17b	medium	intermediate	Day0	Unmeth	NA	NA
36743570	36743924	259.9	539.6	Day12	Slc25a43	NA	intermediate	Day0	Unmeth	NA	NA
36795518	36796876	0.0	892.0	Day12	Slc25a5	escapee	high	Day8	slow	Dep	NA
36807697	36808299	0.0	148.7	Day12	Mdrl	NA	NA	Day0	Unmeth	NA	NA
36863876	36864431	0.0	11.0	Day10	C330007P06Rik	NA	high	Day8	slow	Dep	NA
36874045	36875235	100.6	825.9	Day12	Ube2a	NA	high	Day12	Very Slow	Dep	NA
36902327	36903673	0.0	77.1	Day12	Nkrf	NA	intermediate	Day0	Unmeth	NA	NA

CGI_Start	CGI_End	Day 0 (RPKM)	Day 12 (RPKM)	Max methylation	Associated Gene	Gene silencing speed	Expression of associated gene in ES	Day with Allelic ratio= 1	Methylation dynamics	DNMT3B dependence	SMCHD1 dependence
36989271	36989867	0.0	49.6	Day10	Sept6	NA	NA	Day0	Unmeth	NA	NA
37048296	37049990	687.6	396.4	Day0	Sowahd	NA	NA	Day0	Unmeth	NA	NA
37084904	37085377	0.0	104.6	Day12	Rpl39	NA	high	Day0	Unmeth	NA	NA
37109851	37110482	0.0	66.1	Day12	Upf3b	slow	intermediate	Day0	Unmeth	NA	NA
37126842	37127556	0.0	192.7	Day12	Nkap	slow	intermediate	Day8	slow	Dep	NA
37190817	37192202	293.5	534.1	Day12	Ndufa1	NA	high	Day0	Unmeth	NA	NA
38120987	38122001	1081.7	4476.4	Day12	Rhox13	NA	intermediate	Day4	Fast	Dep	NA
38180931	38181407	159.3	192.7	Day12	Zbtb33	slow	intermediate	Day8	slow	Dep	Dep
38189407	38190530	0.0	732.3	Day12	Zbtb33	slow	intermediate	Day4	Fast	Dep	Dep
38252307	38252664	0.0	0.0	Day8	Tmem255a	NA	low	Day0	Unmeth	NA	NA
38456144	38456572	0.0	22.0	Day10	Lamp2	slow	high	Day8	slow	Dep	NA
38564289	38565188	0.0	110.1	Day12	Cul4b	medium	high	Day0	Unmeth	NA	NA
38634835	38635562	0.0	126.6	Day12	C1galt1c1	NA	intermediate	Day0	Unmeth	NA	NA
41910947	41912166	0.0	258.8	Day12	Thoc2	medium	high	Day0	Unmeth	NA	NA
42148662	42150694	0.0	853.4	Day12	Stag2	medium	high	Day8	slow	Dep	NA
42489649	42490070	545.0	1404.0	Day12	Sh2d1a	NA	low	Day8	slow	Dep	NA
44367418	44368562	176.1	506.6	Day12	Dcaf12l2	NA	low	Day0	Unmeth	NA	NA
44789061	44790387	360.6	633.2	Day12	Dcaf12l1	NA	intermediate	Day0	Unmeth	NA	NA
47891979	47892619	109.0	38.5	Day0	Smarca1	fast	intermediate	Day0	Unmeth	NA	NA
48033685	48034936	0.0	429.5	Day12	Apln	NA	low	Day4	Fast	Dep	Dep
48207950	48208923	8.4	132.1	Day12	Zdhhc9	fast	intermediate	Day8	slow	Dep	NA
48284730	48285561	0.0	27.5	Day10	9530027J09Rik	NA	low	Day8	slow	Dep	Dep
48285470	48286172	167.7	231.3	Day12	9530027J09Rik	NA	low	Day8	slow	Dep	Dep
48314518	48315220	67.1	11.0	Day0	Bcorl1	fast	intermediate	Day8	slow	Dep	Dep
48340699	48342623	0.0	649.7	Day12	Bcorl1	fast	intermediate	Day8	slow	Dep	Dep
48343396	48344372	0.0	27.5	Day12	Bcorl1	fast	intermediate	Day10	Very Slow	Dep	Dep
48344627	48346085	8.4	60.6	Day12	Bcorl1	fast	intermediate	Day8	slow	Dep	Dep
48409780	48410366	0.0	0.0	Day10	Elf4	fast	low	Day0	Unmeth	NA	Dep
48452790	48454770	0.0	1139.8	Day12	Elf4	fast	low	Day4	Fast	Dep	Dep
48462613	48463288	0.0	5.5	Day12	Elf4	fast	low	Day10	Very Slow	Dep	Dep
48513108	48513884	0.0	192.7	Day12	Aifm1	fast	high	Day8	slow	Indep	Dep
48519128	48519851	159.3	104.6	Day0	Rab33a	NA	low	Day8	slow	Dep	NA
48593996	48594747	0.0	60.6	Day12	Zfp280c	fast	intermediate	Day8	slow	Dep	NA
48623486	48624019	0.0	44.0	Day10	Slc25a14	medium	intermediate	Day10	Very Slow	Dep	NA
48694857	48695509	0.0	77.1	Day12	Rbmx2	medium	intermediate	Day8	slow	Dep	Dep
49470401	49471004	117.4	126.6	Day12	Arhgap36	NA	low	Day10	Very Slow	Dep	Dep
49480197	49480690	50.3	38.5	Day0	Arhgap36	NA	low	Day4	Fast	Dep	Dep
49493556	49494435	218.0	11.0	Day0	Arhgap36	NA	low	Day10	Very Slow	Dep	Dep
49495171	49495569	92.2	192.7	Day12	Arhgap36	NA	low	Day0	Unmeth	NA	Dep
50373707	50374125	997.9	1861.0	Day12	Olfr1324	NA	NA	Day10	Very Slow	Indep	NA
50841098	50842100	218.0	77.1	Day0	Stk26	NA	NA	Day8	slow	Dep	NA
51017784	51019177	0.0	501.1	Day12	Rap2c	NA	intermediate	Day0	Unmeth	NA	NA
51204599	51206357	654.1	1145.3	Day12	Mbnl3	slow	intermediate	Day0	Unmeth	NA	NA
51679789	51681668	8.4	936.0	Day12	Hs6st2	fast	intermediate	Day12	Very Slow	Dep	Dep
52164164	52165665	0.0	479.0	Day12	Gpc4	fast	high	Day8	slow	Dep	Dep
52613273	52614535	201.2	330.4	Day12	Gpc3	fast	high	Day10	Very Slow	Dep	Dep
52744148	52744760	100.6	181.7	Day12	Kis2	fast	intermediate	Day0	Unmeth	NA	NA
52911750	52912757	0.0	814.9	Day12	Phf6	fast	intermediate	Day4	Fast	Dep	Dep
52987876	52988900	8.4	396.4	Day12	Hprt	medium	low	Day8	slow	Dep	Dep
53056639	53057603	0.0	286.3	Day12	C430049B03Rik	fast	intermediate	Day8	slow	Dep	NA
53269352	53269777	0.0	11.0	Day12	Fam122b	fast	intermediate	Day8	slow	Dep	NA
53716244	53716956	259.9	1905.1	Day12	AW822252	NA	low	Day4	Fast	Dep	NA
53756633	53757220	779.8	958.1	Day12	1700013H16Rik	NA	intermediate	Day8	slow	Dep	NA
53775623	53776076	184.5	930.5	Day12	Zfp36l3	NA	intermediate	Day8	slow	Dep	NA
53974513	53975087	226.4	440.5	Day12	Xlr	NA	low	Day12	Very Slow	Dep	Dep
53975754	53976156	50.3	22.0	Day0	Xlr	NA	low	Day10	Very Slow	Dep	Dep
56346162	56346700	0.0	181.7	Day12	Zfp449	medium	intermediate	Day0	Unmeth	NA	NA
56374299	56375156	427.7	170.7	Day0	Smim10l2a	NA	NA	Day8	slow	Indep	NA
56455130	56456274	0.0	192.7	Day12	Ints6l	NA	NA	Day8	slow	Dep	NA
56597604	56598414	0.0	88.1	Day12	Mmgt1	slow	high	Day0	Unmeth	NA	NA
56609638	56610458	0.0	192.7	Day12	Slc9a6	medium	intermediate	Day8	slow	Dep	Dep
56731006	56732303	92.2	473.5	Day12	Fhl1	fast	high	Day8	slow	Dep	Dep
57053041	57054230	0.0	655.2	Day12	Htatsf1	slow	intermediate	Day4	Fast	Dep	Dep
57392407	57393061	0.0	176.2	Day12	Rbmx	fast	intermediate	Day4	Fast	Dep	Dep
57503265	57504044	234.8	335.9	Day10	Gpr101	NA	NA	Day4	Fast	Indep	NA
58030153	58030889	184.5	71.6	Day0	Zic3	medium	high	Day8	slow	Dep	Dep
58030906	58032364	327.0	578.1	Day12	Zic3	medium	high	Day8	slow	Indep	Dep
58038054	58039353	419.3	346.9	Day0	Zic3	medium	high	Day8	slow	Dep	Dep
59133751	59134938	251.6	512.1	Day12	Fgf13	fast	low	Day8	slow	Dep	NA
59566759	59568379	536.7	534.1	Day0	Fgf13	fast	low	Day8	slow	Dep	NA
60178465	60178966	100.6	11.0	Day0	Mcf2	medium	high	Day0	Unmeth	NA	NA
60403173	60405107	0.0	941.5	Day12	Atp11c	medium	intermediate	Day8	slow	Dep	Dep
60483285	60483708	385.7	688.3	Day12	Gm7073	NA	NA	Day0	Unmeth	NA	Dep
60546999	60548656	419.3	770.8	Day12	Gm715	NA	low	Day8	slow	Dep	NA
60891703	60894853	1106.9	2351.1	Day12	Sox3	NA	low	Day4	Fast	Dep	NA
61709373	61710012	125.8	655.2	Day12	Ldoc1	NA	intermediate	Day4	Fast	Dep	NA
64275554	64276198	109.0	38.5	Day0	Slitrk4	NA	low	Day8	slow	Dep	NA
64276419	64277210	75.5	38.5	Day0	Slitrk4	NA	low	Day0	Unmeth	NA	NA
68678107	68679525	0.0	754.3	Day12	Fmr1	slow	high	Day8	slow	Indep	Dep
68821083	68821506	0.0	1222.3	Day12	Fmr1nb	fast	high	Day0	Unmeth	NA	NA
69359731	69361341	0.0	616.7	Day12	Aff2	fast	low	Day8	slow	Dep	NA
70364769	70365313	0.0	132.1	Day12	Ids	medium	intermediate	Day8	slow	Dep	NA
70476633	70477329	0.0	181.7	Day12	Tmem185a	NA	NA	Day8	slow	Dep	NA
71049649	71050487	33.5	440.5	Day12	Mamld1	fast	low	Day0	Unmeth	NA	NA
71214882	71215302	0.0	49.6	Day12	Mtm1	medium	intermediate	Day4	Fast	Dep	NA
71364433	71365407	0.0	644.2	Day12	Mtmr1	medium	intermediate	Day8	slow	Dep	NA
71492457	71492906	511.5	335.9	Day0	Cd99l2	fast	intermediate	Day4	Fast	Dep	NA
71555510	71557431	0.0	2367.6	Day12	Hmgb3	medium	intermediate	Day4	Fast	Dep	Dep
71663725	71665020	402.5	847.9	Day12	Gpr50	NA	low	Day8	slow	Dep	Dep
71816289	71817263	8.4	594.7	Day12	Vma21	slow	intermediate	Day4	Fast	Indep	NA
71962049	71963233	385.7	291.8	Day0	Prrg3	fast	intermediate	Day8	slow	Dep	NA
72274722	72275076	0.0	269.8	Day12	Gabre	fast	low	Day4	Fast	Dep	NA
72824933	72825627	276.7	517.6	Day12	Gabrq	NA	low	Day8	slow	Dep	NA
72917989	72918635	0.0	11.0	Day10	Cetn2	NA	high	Day12	Very Slow	Dep	NA
72987088	72987811	184.5	473.5	Day12	Zfp185	fast	intermediate	Day4	Fast	Dep	NA
73228041	73229385	0.0	280.8	Day12	F8a	NA	intermediate	Day8	slow	Dep	Dep
73342399	73343216	0.0	297.3	Day12	Zfp275	medium	intermediate	Day0	Unmeth	NA	NA
73400925	73401325	109.0	864.5	Day12	Zfp92	NA	low	Day4	Fast	Dep	Dep
73410850	73411308	75.5	286.3	Day10	Zfp92	NA	low	Day4	Fast	Indep	Dep
73421819	73422641	1090.1	2868.7	Day12	Zfp92	NA	low	Day4	Fast	Indep	Dep
73458780	73459236	0.0	22.0	Day10	Haus7	slow	high	Day8	slow	Dep	Dep
73502593	73504042	251.6	1040.6	Day12	Atp2b3	NA	low	Day4	Fast	Dep	NA
73637105	73637666	67.1	286.3	Day12	Dusp9	fast	high	Day4	Fast	Dep	Dep
73639133	73639848	184.5	88.1	Day0	Dusp9	fast	high	Day0	Unmeth	NA	Dep
73639873	73640704	620.5	2323.6	Day12	Dusp9	fast	high	Day0	Unmeth	NA	Dep
73659569	73660409	352.2	1051.7	Day12	Pnck	NA	intermediate	Day8	slow	Dep	NA
73672670	73674477	0.0	881.0	Day12	Slc6a8	slow	high	Day8	slow	Dep	NA
73715982	73716718	0.0	335.9	Day12	Bcap31	slow	high	Day8	slow	Dep	NA
73716960	73717515	41.9	49.6	Day10	Bcap31	slow	high	Day0	Unmeth	NA	NA
73786525	73787627	0.0	319.4	Day12	Ssr4	slow	high	Day8	slow	Dep	Dep
73794632	73795050	369.0	616.7	Day12	Ssr4	slow	high	Day10	Very Slow	Dep	Dep
73823310	73823954	167.7	154.2	Day0	Pdzd4	medium	high	Day4	Fast	Dep	NA
73824041	73825357	0.0	440.5	Day12	Pdzd4	medium	high	Day8	slow	Dep	NA
73880382	73881057	150.9	253.3	Day12	L1cam	fast	intermediate	Day4	Fast	Dep	Dep
73911003	73911668	134.2	726.8	Day12	Arhgap4	slow	intermediate	Day4	Fast	Indep	NA
73921306	73922364	0.0	418.5	Day12	Naa10	slow	high	Day4	Fast	Dep	Dep
73965728	73966234	0.0	44.0	Day12	Hcfc1	slow	high	Day12	Very Slow	Dep	Dep
73966361	73967974	8.4	302.8	Day12	Hcfc1	slow	high	Day8	slow	Dep	Dep
74023202	74024469	0.0	1123.2	Day12	Irak1	medium	intermediate	Day4	Fast	Indep	NA
74084604	74085966	0.0	396.4	Day12	Mecp2	slow	intermediate	Day8	slow	Dep	Dep
74245704	74246694	25.2	600.2	Day12	Flna	medium	high	Day4	Fast	Dep	Dep
74249542	74250255	0.0	236.8	Day12	Flna	medium	high	Day8	slow	Dep	Dep
74254397	74255375	0.0	655.2	Day12	Emd	slow	high	Day4	Fast	Dep	Dep

CGI_Start	CGI_End	Day 0 (RPKM)	Day 12 (RPKM)	Max methylation	Associated Gene	Gene silencing speed	Expression of associated gene in ES	Day with Allelic ratio= 1	Methylation dynamics	DNMT3B dependence	SMCHD1 dependence
74282415	74282943	0.0	115.6	Day12	Taz	slow	high	Day4	Fast	Dep	Dep
74296806	74297588	0.0	11.0	Day10	Atp6ap1	escapee	high	Day8	slow	Dep	Dep
74304822	74305684	16.8	60.6	Day12	Gdi1	medium	high	Day8	slow	Dep	NA
74312728	74313676	0.0	413.0	Day12	Fam50a	slow	high	Day4	Fast	Dep	NA
74328984	74329392	0.0	143.2	Day12	Plxna3	medium	intermediate	Day4	Fast	Dep	NA
74367981	74368915	0.0	726.8	Day12	Ubl4a	NA	NA	Day4	Fast	Dep	NA
74372970	74373389	0.0	341.4	Day10	Gm44504	NA	NA	Day4	Fast	Indep	NA
74392859	74393365	0.0	11.0	Day10	Fam3a	medium	intermediate	Day8	slow	Dep	NA
74428625	74429384	8.4	33.0	Day12	G6pdx	medium	high	Day0	Unmeth	NA	NA
75095726	75096410	0.0	38.5	Day10	Dkc1	slow	high	Day4	Fast	Indep	Dep
75130564	75131234	0.0	49.6	Day12	Mpp1	medium	intermediate	Day0	Unmeth	NA	NA
75143111	75143583	310.3	1260.9	Day12	Mpp1	medium	intermediate	Day4	Fast	Indep	NA
75155553	75155978	109.0	258.8	Day12	Smim9	NA	NA	Day4	Fast	Dep	NA
75416063	75416879	0.0	66.1	Day10	Cmc4	NA	NA	Day8	slow	Dep	NA
75475304	75475761	134.2	1640.8	Day12	Vbp1	slow	high	Day0	Unmeth	NA	NA
75577798	75578281	0.0	0.0	Day8	Rab39b	NA	intermediate	Day0	Unmeth	NA	NA
75966192	75966939	930.8	4790.3	Day12	Pls3	fast	high	Day4	Fast	Dep	NA
77510585	77512121	0.0	567.1	Day12	Tbl1x	fast	high	Day8	slow	Dep	NA
77795183	77796389	8.4	192.7	Day10	Prkx	medium	intermediate	Day8	slow	Dep	Dep
78583537	78583939	0.0	22.0	Day12	Prrg1	fast	low	Day8	slow	Dep	NA
78856765	78857124	1844.8	3132.9	Day12	Fam47c	NA	NA	Day4	Fast	Dep	NA
79517213	79517607	8.4	126.6	Day12	Cfap47	NA	NA	Day4	Fast	Indep	NA
81070367	81071426	50.3	11.0	Day0	Tmem47	slow	high	Day10	Very Slow	Dep	NA
85573560	85574663	0.0	1035.1	Day12	Tab3	medium	intermediate	Day8	slow	Dep	Dep
85776231	85777354	0.0	55.1	Day12	Gk	NA	NA	Day10	Very Slow	Dep	Dep
86191587	86192860	335.4	220.2	Day0	Nr0b1	NA	high	Day4	Fast	Indep	NA
93286368	93286723	184.5	242.3	Day12	Arx	NA	NA	Day8	slow	Dep	Dep
93288326	93289658	754.7	682.8	Day0	Arx	NA	NA	Day8	slow	Dep	Dep
93296970	93297696	209.6	192.7	Day0	Arx	NA	NA	Day8	slow	Dep	Dep
93298956	93299513	159.3	77.1	Day0	Arx	NA	NA	Day8	slow	Dep	Dep
93631878	93632496	0.0	352.4	Day12	Pola1	slow	high	Day4	Fast	Dep	Dep
93831526	93832538	0.0	622.2	Day12	Pdk3	fast	high	Day8	slow	Dep	Dep
93922026	93922573	377.3	5203.2	Day12	AU015836	NA	intermediate	Day4	Fast	Dep	NA
93922693	93923048	234.8	2268.5	Day12	AU015836	NA	intermediate	Day4	Fast	Dep	NA
93980462	93981079	293.5	2527.3	Day12	AU015836	NA	intermediate	Day4	Fast	Dep	NA
93988622	93989005	167.7	1547.2	Day12	AU015836	NA	intermediate	Day4	Fast	Dep	NA
94122125	94123811	0.0	1756.4	Day12	Zfx	slow	intermediate	Day12	Very Slow	Dep	NA
94177525	94178399	0.0	66.1	Day10	Eif2s3x	slow	high	Day8	slow	Dep	NA
94212391	94213395	0.0	11.0	Day12	Eif2s3x	slow	high	Day0	Unmeth	NA	NA
94234174	94235633	0.0	616.7	Day12	Klhl15	fast	intermediate	Day4	Fast	Dep	Dep
94366552	94367330	0.0	115.6	Day12	Fam90a1b	NA	NA	Day8	slow	Dep	NA
94541170	94542444	0.0	148.7	Day12	Maged1	fast	high	Day8	slow	Dep	Dep
94635870	94636726	100.6	401.9	Day12	Gspt2	medium	intermediate	Day8	slow	Dep	NA
94773656	94774334	1031.4	4476.4	Day12	Zxda	medium	low	Day4	Fast	Dep	NA
94990478	94990920	125.8	385.4	Day12	Spin4	fast	low	Day4	Fast	Indep	NA
95025896	95026915	0.0	77.1	Day12	Spin4	fast	low	Day8	slow	Dep	NA
95166350	95166711	0.0	5.5	Day12	Arhgef9	fast	low	Day0	Unmeth	NA	NA
95196302	95196717	16.8	11.0	Day0	Arhgef9	fast	low	Day0	Unmeth	NA	NA
95444307	95445120	0.0	88.1	Day12	Amer1	fast	intermediate	Day4	Fast	Dep	NA
95711452	95712158	0.0	132.1	Day12	Zc3h12b	fast	low	Day4	Fast	Dep	Dep
95956486	95957207	0.0	55.1	Day12	Las1l	slow	high	Day4	Fast	Dep	Dep
96095741	96096433	0.0	269.8	Day12	Msn	fast	high	Day0	Unmeth	NA	NA
98149788	98150234	8.4	5.5	Day0	Ar	fast	low	Day0	Unmeth	NA	NA
98150442	98151164	58.7	66.1	Day12	Ar	fast	low	Day4	Fast	Dep	NA
98889960	98891295	8.4	16.5	Day12	Ophn1	fast	low	Day0	Unmeth	NA	NA
98936726	98938594	0.0	1156.3	Day12	Yipf6	slow	intermediate	Day10	Very Slow	Dep	Dep
99042501	99043234	58.7	44.0	Day0	Stard8	fast	intermediate	Day4	Fast	Dep	NA
99135591	99137344	16.8	1448.1	Day12	Efnb1	NA	intermediate	Day0	Unmeth	NA	Dep
99216379	99217445	310.3	1007.6	Day12	Efnb1	NA	intermediate	Day4	Fast	Dep	Dep
99469580	99470188	0.0	22.0	Day10	Pja1	medium	high	Day4	Fast	Dep	NA
99470244	99470971	0.0	225.7	Day10	Pja1	medium	high	Day4	Fast	Dep	NA
99819602	99821925	1115.2	2615.4	Day12	Tmem28	fast	intermediate	Day0	Unmeth	NA	NA
99853209	99853641	150.9	16.5	Day0	Tmem28	fast	intermediate	Day8	slow	Dep	NA
99975380	99976192	259.9	275.3	Day12	Eda	fast	intermediate	Day4	Fast	Dep	NA
99976228	99976697	100.6	22.0	Day0	Eda	fast	intermediate	Day8	slow	Dep	NA
1E+08	1E+08	846.9	3650.5	Day12	Otud6a	NA	intermediate	Day4	Fast	Dep	NA
1.01E+08	1.01E+08	0.0	170.7	Day12	Kif4	medium	high	Day8	slow	Dep	NA
1.01E+08	1.01E+08	0.0	203.7	Day12	Dlg3	fast	intermediate	Day8	slow	Dep	NA
1.01E+08	1.01E+08	8.4	176.2	Day12	Dlg3	fast	intermediate	Day0	Unmeth	NA	NA
1.01E+08	1.01E+08	486.3	3733.1	Day12	Tex11	fast	intermediate	Day4	Fast	Indep	NA
1.01E+08	1.01E+08	0.0	236.8	Day12	Snx12	medium	high	Day4	Fast	Indep	Dep
1.01E+08	1.01E+08	8.4	77.1	Day12	Foxo4	fast	high	Day8	slow	Dep	Dep
1.01E+08	1.01E+08	0.0	242.3	Day10	Med12	medium	intermediate	Day0	Unmeth	NA	NA
1.01E+08	1.01E+08	0.0	66.1	Day12	Nlgn3	NA	low	Day8	slow	Dep	Dep
1.01E+08	1.01E+08	159.3	501.1	Day12	Gjb1	NA	intermediate	Day12	Very Slow	Dep	NA
1.01E+08	1.01E+08	0.0	0.0	Day10	Zmym3	medium	high	Day0	Unmeth	NA	NA
1.01E+08	1.01E+08	0.0	60.6	Day12	Zmym3	medium	high	Day0	Unmeth	NA	NA
1.01E+08	1.01E+08	0.0	363.4	Day12	Nono	medium	high	Day8	slow	Dep	NA
1.02E+08	1.02E+08	0.0	33.0	Day12	Taf1	slow	high	Day10	Very Slow	Dep	Dep
1.02E+08	1.02E+08	0.0	0.0	Day0	Ogt	slow	high	Day0	Unmeth	NA	NA
1.02E+08	1.02E+08	679.2	5098.6	Day12	Cxcr3	NA	NA	Day4	Fast	Dep	NA
1.02E+08	1.02E+08	1249.4	4806.8	Day12	8030474K03Rik	NA	NA	Day4	Fast	Dep	NA
1.02E+08	1.02E+08	679.2	1013.1	Day12	Nhsl2	fast	low	Day12	Very Slow	Dep	NA
1.02E+08	1.02E+08	0.0	88.1	Day12	Nhsl2	fast	low	Day4	Fast	Indep	NA
1.02E+08	1.02E+08	360.6	743.3	Day12	Rtl5	NA	NA	Day4	Fast	Dep	NA
1.02E+08	1.02E+08	352.2	2152.9	Day12	Rps4x	medium	high	Day4	Fast	Indep	NA
1.02E+08	1.02E+08	0.0	38.5	Day12	Rps4x	medium	high	Day12	Very Slow	Dep	NA
1.02E+08	1.02E+08	972.7	1833.5	Day12	Cited1	NA	intermediate	Day8	slow	Dep	NA
1.03E+08	1.03E+08	8.4	666.2	Day12	Phka1	fast	intermediate	Day8	slow	Indep	Dep
1.03E+08	1.03E+08	369.0	787.4	Day12	Gm5126	NA	low	Day12	Very Slow	Dep	NA
1.03E+08	1.03E+08	343.8	11.0	Day0	Nap1l2	NA	intermediate	Day0	Unmeth	NA	NA
1.03E+08	1.03E+08	150.9	275.3	Day12	Cdx4	NA	low	Day4	Fast	Dep	NA
1.03E+08	1.03E+08	0.0	99.1	Day12	Chic1	medium	intermediate	Day10	Very Slow	Dep	NA
1.03E+08	1.03E+08	125.8	379.9	Day12	Tsix	escapee	intermediate	Day0	Unmeth	NA	NA
1.03E+08	1.03E+08	536.7	1806.0	Day12	Xist	escapee	low	Day0	Unmeth	NA	NA
1.03E+08	1.03E+08	0.0	60.6	Day10	Jpx	slow	low	Day4	Fast	Dep	NA
1.04E+08	1.04E+08	0.0	187.2	Day12	Ftx	medium	intermediate	Day8	slow	Dep	Dep
1.04E+08	1.04E+08	67.1	143.2	Day12	Ftx	medium	intermediate	Day4	Fast	Indep	Dep
1.04E+08	1.04E+08	0.0	187.2	Day12	Ftx	medium	intermediate	Day4	Fast	Dep	Dep
1.04E+08	1.04E+08	201.2	121.1	Day0	Zcchc13	NA	low	Day8	slow	Dep	NA
1.04E+08	1.04E+08	41.9	280.8	Day12	Slc16a2	fast	low	Day4	Fast	Dep	NA
1.04E+08	1.04E+08	0.0	754.3	Day12	Rlim	fast	high	Day4	Fast	Dep	NA
1.04E+08	1.04E+08	830.1	2912.7	Day12	Nexmif	NA	NA	Day4	Fast	Dep	NA
1.04E+08	1.04E+08	0.0	115.6	Day12	Abcb7	medium	high	Day8	slow	Dep	Dep
1.04E+08	1.04E+08	0.0	33.0	Day10	Uprt	fast	intermediate	Day4	Fast	Dep	NA
1.05E+08	1.05E+08	184.5	126.6	Day0	Zdhhc15	fast	intermediate	Day8	slow	Dep	Dep
1.05E+08	1.05E+08	92.2	170.7	Day12	Magee2	NA	low	Day4	Fast	Dep	NA
1.05E+08	1.05E+08	0.0	38.5	Day10	Pbdc1	slow	high	Day10	Very Slow	Dep	NA
1.05E+08	1.05E+08	0.0	501.1	Day12	Magee1	NA	intermediate	Day4	Fast	Dep	Dep
1.05E+08	1.05E+08	83.9	418.5	Day12	E530001F21Rik	NA	low	Day0	Unmeth	NA	NA
1.06E+08	1.06E+08	33.5	5.5	Day0	Fgf16	NA	NA	Day12	Very Slow	Dep	NA
1.06E+08	1.06E+08	545.0	638.7	Day12	Fgf16	NA	NA	Day0	Unmeth	NA	NA
1.06E+08	1.06E+08	0.0	600.2	Day12	Atrx	medium	high	Day4	Fast	Dep	NA
1.06E+08	1.06E+08	0.0	121.1	Day12	Magt1	fast	high	Day8	slow	Dep	Dep
1.06E+08	1.06E+08	8.4	60.6	Day12	Atp7a	medium	intermediate	Day8	slow	Dep	NA
1.06E+08	1.06E+08	16.8	126.6	Day12	Pgk1	fast	high	Day8	slow	Dep	Dep
1.06E+08	1.06E+08	360.6	677.2	Day12	Taf9b	fast	intermediate	Day8	slow	Indep	NA
1.06E+08	1.06E+08	92.2	88.1	Day0	Fndc3c1	fast	intermediate	Day4	Fast	Dep	NA
1.07E+08	1.07E+08	8.4	49.6	Day12	Itm2a	fast	high	Day0	Unmeth	NA	NA
1.08E+08	1.08E+08	930.8	2488.7	Day12	4933401B06Rik	NA	NA	Day0	Unmeth	NA	NA
1.08E+08	1.08E+08	0.0	66.1	Day12	2610002M06Rik	medium	intermediate	Day10	Very Slow	Dep	Dep
1.09E+08	1.09E+08	0.0	33.0	Day12	Hmgn5	fast	intermediate	Day0	Unmeth	NA	NA
1.11E+08	1.11E+08	218.0	203.7	Day10	Pou3f4	NA	low	Day4	Fast	Indep	NA
1.11E+08	1.11E+08	25.2	16.5	Day0	Pou3f4	NA	low	Day0	Unmeth	NA	NA
1.11E+08	1.11E+08	209.6	137.7	Day0	Rps6ka6	fast	high	Day4	Fast	Dep	NA
1.12E+08	1.12E+08	494.7	2037.2	Day12	Hdx	NA	intermediate	Day0	Unmeth	NA	NA
1.13E+08	1.13E+08	561.8	1343.5	Day12	Zfp711	NA	low	Day0	Unmeth	NA	NA

CGI_Start	CGI_End	Day 0 (RPKM)	Day 12 (RPKM)	Max methylation	Associated Gene	Gene silencing speed	Expression of associated gene in ES	Day with Allelic ratio= 1	Methylation dynamics	DNMT3B dependence	SMCHD1 dependence
1.13E+08	1.13E+08	25.2	0.0	Day0	Dach2	medium	low	Day0	Unmeth	NA	NA
1.2E+08	1.2E+08	134.2	137.7	Day10	Pabpc5	NA	low	Day0	Unmeth	NA	NA
1.2E+08	1.2E+08	150.9	5.5	Day0	Pcdh11x	medium	NA	Day0	Unmeth	NA	NA
1.22E+08	1.22E+08	83.9	60.6	Day0	Nap1l3	NA	low	Day10	Very Slow	Dep	NA
1.23E+08	1.23E+08	486.3	2510.8	Day12	Cldn34c1	NA	NA	Day10	Very Slow	Dep	NA
1.24E+08	1.24E+08	0.0	0.0	Day0	Astx5	NA	NA	Day0	Unmeth	NA	NA
1.24E+08	1.24E+08	0.0	0.0	Day10	Vmn2r121	NA	NA	Day0	Unmeth	NA	NA
1.24E+08	1.24E+08	0.0	0.0	Day0	Vmn2r121	NA	NA	Day0	Unmeth	NA	NA
1.3E+08	1.3E+08	0.0	209.2	Day12	Diaph2	NA	NA	Day0	Unmeth	NA	NA
1.34E+08	1.34E+08	243.2	424.0	Day12	Gm26851	NA	NA	Day0	Unmeth	NA	NA
1.34E+08	1.34E+08	176.1	1145.3	Day12	Gm26851	NA	NA	Day10	Very Slow	Dep	NA
1.34E+08	1.34E+08	0.0	220.2	Day12	Pcdh19	NA	low	Day0	Unmeth	NA	NA
1.34E+08	1.34E+08	0.0	1051.7	Day12	Cstf2	slow	high	Day4	Fast	Dep	Dep
1.34E+08	1.34E+08	0.0	264.3	Day12	Cenpi	slow	intermediate	Day0	Unmeth	NA	Dep
1.34E+08	1.34E+08	2465.3	7813.1	Day12	Taf7l	NA	low	Day4	Fast	Dep	NA
1.34E+08	1.34E+08	184.5	1222.3	Day12	Taf7l	NA	low	Day4	Fast	Dep	NA
1.35E+08	1.35E+08	41.9	754.3	Day12	Timm8a1	NA	intermediate	Day8	slow	Dep	NA
1.35E+08	1.35E+08	0.0	60.6	Day12	Rpl36a	NA	high	Day0	Unmeth	NA	NA
1.35E+08	1.35E+08	0.0	88.1	Day12	Hnrnph2	NA	high	Day10	Very Slow	Dep	NA
1.35E+08	1.35E+08	218.0	1519.7	Day12	Hnrnph2	NA	high	Day0	Unmeth	NA	NA
1.35E+08	1.35E+08	0.0	176.2	Day12	Armcx4	fast	intermediate	Day8	slow	Dep	NA
1.35E+08	1.35E+08	125.8	534.1	Day12	Nxf2	NA	low	Day4	Fast	Dep	NA
1.35E+08	1.35E+08	0.0	181.7	Day12	Zmat1	fast	low	Day4	Fast	Dep	Dep
1.35E+08	1.35E+08	100.6	137.7	Day12	3632454L22Rik	NA	low	Day4	Fast	Dep	NA
1.36E+08	1.36E+08	0.0	104.6	Day12	Gprasp1	slow	high	Day8	slow	Dep	NA
1.36E+08	1.36E+08	0.0	198.2	Day12	Bhlhb9	medium	intermediate	Day8	slow	Dep	NA
1.36E+08	1.36E+08	109.0	264.3	Day12	Bex2	medium	high	Day4	Fast	Dep	NA
1.36E+08	1.36E+08	226.4	341.4	Day12	Bex4	escapee	high	Day10	Very Slow	Dep	NA
1.36E+08	1.36E+08	159.3	1547.2	Day12	Bex1	fast	high	Day12	Very Slow	Dep	NA
1.36E+08	1.36E+08	16.8	27.5	Day12	Bex3	NA	NA	Day0	Unmeth	NA	NA
1.37E+08	1.37E+08	0.0	82.6	Day12	Morf4l2	escapee	high	Day4	Fast	Indep	Dep
1.37E+08	1.37E+08	184.5	115.6	Day0	Rab9b	NA	low	Day4	Fast	Dep	NA
1.37E+08	1.37E+08	67.1	0.0	Day0	Slc25a53	slow	low	Day0	Unmeth	NA	Dep
1.37E+08	1.37E+08	0.0	33.0	Day12	Slc25a53	slow	low	Day4	Fast	Dep	Dep
1.37E+08	1.37E+08	0.0	401.9	Day12	Fam199x	fast	intermediate	Day4	Fast	Dep	NA
1.37E+08	1.37E+08	595.4	864.5	Day12	Esx1	NA	low	Day10	Very Slow	Dep	Dep
1.38E+08	1.38E+08	150.9	22.0	Day0	Il1rapl2	slow	NA	Day8	slow	Dep	NA
1.39E+08	1.39E+08	159.3	148.7	Day0	Nrk	fast	intermediate	Day0	Unmeth	NA	NA
1.39E+08	1.39E+08	0.0	396.4	Day12	Pwwp3b	NA	NA	Day4	Fast	Dep	NA
1.39E+08	1.39E+08	41.9	38.5	Day10	Trap1a	NA	high	Day0	Unmeth	NA	NA
1.39E+08	1.39E+08	0.0	5.5	Day8	Radx	NA	NA	Day0	Unmeth	NA	NA
1.4E+08	1.4E+08	8.4	594.7	Day12	Rnf128	fast	high	Day8	slow	Indep	NA
1.4E+08	1.4E+08	0.0	0.0	Day0	Tbc1d8b	fast	low	Day0	Unmeth	NA	NA
1.4E+08	1.4E+08	142.6	947.0	Day12	Morc4	fast	intermediate	Day8	slow	Dep	NA
1.4E+08	1.4E+08	402.5	952.5	Day12	Nup62cl	fast	high	Day4	Fast	Dep	NA
1.4E+08	1.4E+08	343.8	627.7	Day12	Nup62cl	fast	high	Day4	Fast	Dep	NA
1.4E+08	1.4E+08	444.4	627.7	Day12	Mir3475	NA	NA	Day12	Very Slow	Indep	Dep
1.4E+08	1.4E+08	0.0	44.0	Day10	Prps1	medium	high	Day8	slow	Dep	Dep
1.41E+08	1.41E+08	142.6	1503.2	Day12	Tsc22d3	fast	intermediate	Day8	slow	Dep	Dep
1.41E+08	1.41E+08	0.0	864.5	Day12	Mid2	fast	low	Day4	Fast	Dep	Dep
1.42E+08	1.42E+08	402.5	319.4	Day0	Irs4	NA	low	Day4	Fast	Indep	NA
1.42E+08	1.42E+08	360.6	1018.6	Day12	Irs4	NA	low	Day0	Unmeth	NA	NA
1.42E+08	1.42E+08	385.7	1717.9	Day12	Kcne1l	NA	low	Day4	Fast	Indep	NA
1.42E+08	1.42E+08	0.0	374.4	Day12	Acsl4	medium	high	Day8	slow	Dep	NA
1.43E+08	1.43E+08	0.0	253.3	Day12	Tmem164	fast	intermediate	Day8	slow	Dep	NA
1.43E+08	1.43E+08	0.0	291.8	Day12	Ammecr1	medium	intermediate	Day4	Fast	Dep	Dep
1.43E+08	1.43E+08	142.6	93.6	Day0	Chrdl1	fast	low	Day4	Fast	Dep	Dep
1.44E+08	1.44E+08	0.0	209.2	Day12	Alg13	slow	high	Day4	Fast	Indep	Dep
1.47E+08	1.47E+08	226.4	655.2	Day12	Htr2c	NA	NA	Day4	Fast	Dep	NA
1.47E+08	1.47E+08	0.0	0.0	Day0	Mir448	NA	NA	Day0	Unmeth	NA	NA
1.48E+08	1.48E+08	0.0	0.0	Day8	Lrch2	medium	intermediate	Day0	Unmeth	NA	NA
1.48E+08	1.48E+08	0.0	5.5	Day10	Lrch2	medium	intermediate	Day12	Very Slow	Dep	NA
1.5E+08	1.5E+08	0.0	0.0	Day0	Tmem29	slow	intermediate	Day0	Unmeth	NA	NA
1.51E+08	1.51E+08	0.0	77.1	Day12	Apex2	fast	intermediate	Day4	Fast	Dep	NA
1.51E+08	1.51E+08	0.0	187.2	Day10	Gnl3l	fast	high	Day4	Fast	Dep	NA
1.51E+08	1.51E+08	0.0	1227.9	Day12	Fgd1	fast	high	Day4	Fast	Dep	NA
1.51E+08	1.51E+08	0.0	38.5	Day10	Tsr2	medium	intermediate	Day4	Fast	Dep	Dep
1.51E+08	1.51E+08	251.6	104.6	Day0	Wnk3	medium	intermediate	Day8	slow	Dep	Dep
1.51E+08	1.51E+08	0.0	330.4	Day12	Fam120c	medium	low	Day4	Fast	Dep	Dep
1.52E+08	1.52E+08	0.0	451.5	Day12	Phf8	slow	intermediate	Day8	slow	Dep	Dep
1.52E+08	1.52E+08	0.0	5.5	Day12	Huwe1	slow	high	Day0	Unmeth	NA	NA
1.52E+08	1.52E+08	0.0	787.4	Day12	Huwe1	slow	high	Day8	slow	Dep	NA
1.52E+08	1.52E+08	0.0	27.5	Day10	Hsd17b10	fast	high	Day4	Fast	Dep	NA
1.52E+08	1.52E+08	0.0	82.6	Day12	Smc1a	slow	high	Day4	Fast	Dep	Dep
1.52E+08	1.52E+08	92.2	374.4	Day12	Iqsec2	fast	low	Day4	Fast	Indep	NA
1.52E+08	1.52E+08	0.0	0.0	Day4	Kdm5c	slow	high	Day0	Unmeth	NA	NA
1.52E+08	1.52E+08	16.8	484.5	Day12	Kdm5c	slow	high	Day0	Unmeth	NA	NA
1.52E+08	1.52E+08	8.4	22.0	Day12	Kantr	NA	NA	Day10	Very Slow	Dep	NA
1.52E+08	1.52E+08	0.0	126.6	Day12	Tspyl2	medium	high	Day8	slow	Dep	NA
1.52E+08	1.52E+08	16.8	11.0	Day10	Tspyl2	medium	high	Day0	Unmeth	NA	NA
1.53E+08	1.53E+08	469.6	440.5	Day0	Shroom2	fast	high	Day8	slow	Dep	NA
1.53E+08	1.53E+08	8.4	27.5	Day10	Gpr143	NA	low	Day10	Very Slow	Dep	NA
1.53E+08	1.53E+08	276.7	468.0	Day12	Usp51	NA	intermediate	Day0	Unmeth	NA	NA
1.53E+08	1.53E+08	184.5	181.7	Day10	Mageh1	fast	high	Day0	Unmeth	NA	NA
1.53E+08	1.53E+08	0.0	99.1	Day12	Rragb	fast	intermediate	Day0	Unmeth	NA	NA
1.53E+08	1.53E+08	0.0	11.0	Day10	Klf8	fast	intermediate	Day4	Fast	Dep	Dep
1.53E+08	1.53E+08	0.0	726.8	Day12	Ubqln2	slow	intermediate	Day8	slow	Dep	NA
1.55E+08	1.55E+08	0.0	11.0	Day12	Sat1	medium	high	Day10	Very Slow	Dep	Dep
1.55E+08	1.55E+08	0.0	104.6	Day12	Acot9	medium	high	Day8	slow	Indep	NA
1.55E+08	1.55E+08	8.4	374.4	Day12	Prdx4	medium	high	Day0	Unmeth	NA	NA
1.56E+08	1.56E+08	511.5	479.0	Day0	Ptchd1	fast	low	Day8	slow	Dep	NA
1.56E+08	1.56E+08	226.4	451.5	Day12	Ptchd1	fast	low	Day8	slow	Dep	NA
1.57E+08	1.57E+08	0.0	1255.4	Day12	Sms	slow	high	Day4	Fast	Dep	NA
1.58E+08	1.58E+08	427.7	3121.9	Day12	Yy2	NA	intermediate	Day4	Fast	Indep	NA
1.58E+08	1.58E+08	0.0	473.5	Day12	Mbtps2	slow	intermediate	Day4	Fast	Indep	NA
1.58E+08	1.58E+08	50.3	5.5	Day0	Klhl34	NA	low	Day8	slow	Dep	NA
1.58E+08	1.58E+08	687.6	1282.9	Day12	Cnksr2	fast	low	Day4	Fast	Dep	Dep
1.59E+08	1.59E+08	8.4	490.0	Day12	Rps6ka3	slow	intermediate	Day8	slow	Dep	NA
1.59E+08	1.59E+08	0.0	247.8	Day12	Eif1ax	slow	high	Day8	slow	Dep	Dep
1.59E+08	1.59E+08	712.8	831.4	Day12	Map7d2	medium	intermediate	Day10	Very Slow	Dep	Dep
1.6E+08	1.6E+08	0.0	715.8	Day12	Bclaf3	NA	NA	Day4	Fast	Dep	NA
1.6E+08	1.6E+08	0.0	1101.2	Day12	Sh3kbp1	medium	intermediate	Day8	slow	Dep	NA
1.6E+08	1.6E+08	545.0	2351.1	Day12	Map3k15	fast	intermediate	Day4	Fast	Dep	Dep
1.6E+08	1.6E+08	0.0	401.9	Day12	Pdha1	medium	high	Day8	slow	Dep	Dep
1.6E+08	1.6E+08	301.9	159.7	Day0	Adgrg2	NA	NA	Day8	slow	Dep	NA
1.61E+08	1.61E+08	0.0	242.3	Day12	Phka2	medium	intermediate	Day8	slow	Dep	Dep
1.61E+08	1.61E+08	0.0	330.4	Day12	Cdkl5	medium	intermediate	Day8	slow	Dep	Dep
1.61E+08	1.61E+08	0.0	0.0	Day8	Cdkl5	medium	intermediate	Day10	Very Slow	Dep	Dep
1.61E+08	1.61E+08	8.4	5.5	Day0	Scml2	slow	intermediate	Day10	Very Slow	Dep	NA
1.62E+08	1.62E+08	301.9	633.2	Day12	Rai2	medium	low	Day8	slow	Dep	Dep
1.62E+08	1.62E+08	494.7	534.1	Day12	Rai2	medium	low	Day4	Fast	Dep	Dep
1.62E+08	1.62E+08	0.0	1387.5	Day12	Nhs	fast	intermediate	Day4	Fast	Dep	Dep
1.63E+08	1.63E+08	352.2	732.3	Day12	Reps2	fast	low	Day4	Fast	Dep	Dep
1.63E+08	1.63E+08	0.0	1734.4	Day12	Rbbp7	medium	high	Day4	Fast	Dep	NA
1.63E+08	1.63E+08	0.0	93.6	Day12	Txlng	slow	high	Day8	slow	Dep	NA
1.63E+08	1.63E+08	159.3	429.5	Day12	Syap1	slow	intermediate	Day4	Fast	Dep	Dep
1.63E+08	1.63E+08	0.0	44.0	Day12	Ctps2	medium	intermediate	Day12	Very Slow	Dep	NA
1.64E+08	1.64E+08	134.2	451.5	Day12	Rnf138rt1	NA	low	Day4	Fast	Indep	Dep
1.64E+08	1.64E+08	0.0	1134.2	Day12	Ap1s2	medium	intermediate	Day4	Fast	Dep	Dep
1.64E+08	1.64E+08	0.0	534.1	Day12	Zrsr2	slow	intermediate	Day8	slow	Indep	Dep
1.64E+08	1.64E+08	0.0	429.5	Day12	Siah1b	slow	high	Day8	slow	Dep	Dep
1.64E+08	1.64E+08	109.0	11.0	Day0	Pir	medium	intermediate	Day10	Very Slow	Dep	NA
1.64E+08	1.64E+08	0.0	275.3	Day12	Piga	slow	intermediate	Day8	slow	Dep	Dep
1.66E+08	1.66E+08	0.0	11.0	Day12	Gemin8	slow	intermediate	Day8	slow	Dep	Dep
1.66E+08	1.66E+08	377.3	1139.8	Day12	Gpm6b	medium	low	Day8	slow	Dep	NA
1.66E+08	1.66E+08	0.0	291.8	Day12	Ofd1	slow	intermediate	Day8	slow	Dep	NA
1.66E+08	1.66E+08	0.0	115.6	Day12	Rab9	slow	high	Day8	slow	Dep	Dep

CGI_Start	CGI_End	Day 0 (RPKM)	Day 12 (RPKM)	Max methylation	Associated Gene	Gene silencing speed	Expression of associated gene in ES	Day with Allelic ratio= 1	Methylation dynamics	DNMT3B dependence	SMCHD1 dependence
1.67E+08	1.67E+08	0.0	506.6	Day12	Tceanc	slow	intermediate	Day8	slow	Dep	Dep
1.67E+08	1.67E+08	0.0	38.5	Day12	Tmsb4x	fast	high	Day8	slow	Dep	NA
1.67E+08	1.67E+08	0.0	77.1	Day12	Prps2	medium	high	Day8	slow	Dep	NA
1.68E+08	1.68E+08	109.0	27.5	Day0	Frmpd4	medium	NA	Day0	Unmeth	NA	NA
1.69E+08	1.69E+08	134.2	1029.6	Day12	Msl3	medium	intermediate	Day8	slow	Dep	Dep
1.69E+08	1.69E+08	201.2	38.5	Day0	Arhgap6	fast	low	Day8	slow	Indep	Dep
1.69E+08	1.69E+08	385.7	209.2	Day0	Arhgap6	fast	low	Day8	slow	Indep	Dep
1.69E+08	1.69E+08	25.2	5.5	Day0	Arhgap6	fast	low	Day8	slow	Dep	Dep
1.69E+08	1.69E+08	16.8	2053.8	Day12	Hccs	slow	intermediate	Day4	Fast	Dep	NA
1.7E+08	1.7E+08	0.0	5.5	Day10	Mid1	NA	high	Day4	Fast	Indep	NA
1.7E+08	1.7E+08	0.0	0.0	Day0	G530011O06Rik	NA	low	Day0	Unmeth	NA	NA
1.7E+08	1.7E+08	0.0	0.0	Day0	G530011O06Rik	NA	low	Day0	Unmeth	NA	NA
1.7E+08	1.7E+08	268.3	1266.4	Day12	G530011O06Rik	NA	low	Day4	Fast	Indep	NA
1.7E+08	1.7E+08	243.2	418.5	Day12	G530011O06Rik	NA	low	Day8	slow	Dep	NA
1.7E+08	1.7E+08	1006.2	2896.2	Day12	G530011O06Rik	NA	low	Day4	Fast	Indep	NA
1.7E+08	1.7E+08	1274.6	2345.6	Day12	G530011O06Rik	NA	low	Day4	Fast	Dep	NA
1.7E+08	1.7E+08	1392.0	4735.2	Day12	G530011O06Rik	NA	low	Day4	Fast	Indep	NA
1.7E+08	1.7E+08	310.3	88.1	Day0	G530011O06Rik	NA	low	Day4	Fast	Dep	NA
1.7E+08	1.7E+08	150.9	308.3	Day8	G530011O06Rik	NA	low	Day4	Fast	Dep	NA
1.7E+08	1.7E+08	8.4	0.0	Day0	G530011O06Rik	NA	low	Day0	Unmeth	NA	NA
1.7E+08	1.7E+08	1727.4	2108.8	Day8	G530011O06Rik	NA	low	Day8	slow	Dep	NA
1.7E+08	1.7E+08	0.0	0.0	Day10	G530011O06Rik	NA	low	Day10	Very Slow	Dep	NA
1.7E+08	1.7E+08	3152.9	2081.3	Day0	G530011O06Rik	NA	low	Day8	slow	Dep	NA
1.7E+08	1.7E+08	0.0	5.5	Day12	G530011O06Rik	NA	low	Day4	Fast	Dep	NA
1.7E+08	1.7E+08	226.4	275.3	Day10	G530011O06Rik	NA	low	Day12	Very Slow	Indep	NA

Table A1: Key information, regarding CGIs on Xi, their methylation dynamics, associated genes and dependence on DNMT3B and SMCHD1.

RPKM: Normalized number of reads /per million / kb.

Appendix 6

NGS datasets generated in this work

Dataset	Experiment	Replicate	Cell type	Raw NGS reads
Wt_D0_BioCAP	BioCAP-seq	1/2	ES day 0	53753855
Wt_D0_BioCAP	BioCAP-seq	2/2	ES day 0	43811627
Wt_D0_BioCAP_Input	BioCAP-seq Input	1/2	ES day 0	31030200
Wt_D0_BioCAP_Input	BioCAP-seq Input	2/2	ES day 0	40092812
Wt_D4_BioCAP	BioCAP-seq	1/2	EB day 4	48550131
Wt_D4_BioCAP	BioCAP-seq	2/2	EB day 4	56727517
Wt_D4_BioCAP_Input	BioCAP-seq Input	1/2	EB day 4	26425951
Wt_D4_BioCAP_Input	BioCAP-seq Input	2/2	EB day 4	63328065
Wt_D8_BioCAP	BioCAP-seq	1/2	EB day 8	37590126
Wt_D8_BioCAP	BioCAP-seq	2/2	EB day 8	50069518
Wt_D8_BioCAP_Input	BioCAP-seq Input	1/2	EB day 8	11129288
Wt_D8_BioCAP_Input	BioCAP-seq Input	2/2	EB day 8	36350068
Wt_D10_BioCAP	BioCAP-seq	1/2	EB day 10	58003714
Wt_D10_BioCAP	BioCAP-seq	2/2	EB day 10	50069518
Wt_D10_BioCAP_Input	BioCAP-seq Input	1/2	EB day 10	25696705
Wt_D10_BioCAP_Input	BioCAP-seq Input	2/2	EB day 10	55405275
Wt_D12_BioCAP	BioCAP-seq	1/2	EB day 12	59749476
Wt_D12_BioCAP	BioCAP-seq	2/2	EB day 12	38556307
Wt_D12_BioCAP_Input	BioCAP-seq Input	1/2	EB day 12	26585560
Wt_D12_BioCAP_Input	BioCAP-seq Input	2/2	EB day 12	62319289
MEF_BioCAP	BioCAP-seq	1/2	MEF	66680307
MEF_BioCAP	BioCAP-seq	2/2	MEF	56462798
MEF_MBD	MBD-seq	1/2	MEF	49909003
MEF_MBD	MBD-seq	2/2	MEF	43196903
MEF_MBD_Input	MBD-seq Input	1/1	MEF	36997644
Wt_D0_MBD	MBD-seq	1/3	ES day 0	8024178
Wt_D0_MBD	MBD-seq	2/3	ES day 0	9221058
Wt_D0_MBD	MBD-seq	3/3	ES day 0	6403526
Wt_D0_MBD Input	MBD-seq Input	1/4	ES day 0	52461288
Wt_D0_MBD Input	MBD-seq Input	2/4	ES day 0	52130394
Wt_D0_MBD Input	MBD-seq Input	3/4	ES day 0	11345730
Wt_D0_MBD Input	MBD-seq Input	4/4	ES day 0	38523690
Wt_D4_MBD	MBD-seq	1/2	EB day 4	8105824
Wt_D4_MBD	MBD-seq	2/2	EB day 4	24417970
Wt_D4_MBD Input	MBD-seq Input	1/2	EB day 4	40320765
Wt_D4_MBD Input	MBD-seq Input	2/2	EB day 4	10057808
Wt_D8_MBD	MBD-seq	1/2	EB day 8	4514446
Wt_D8_MBD	MBD-seq	2/2	EB day 8	21720094
Wt_D8_MBD Input	MBD-seq Input	1/2	EB day 8	38577742
Wt_D8_MBD Input	MBD-seq Input	2/2	EB day 8	8251978
Wt_D10_MBD	MBD-seq	1/3	EB day 10	23675375
Wt_D10_MBD	MBD-seq	2/3	EB day 10	26983616
Wt_D10_MBD	MBD-seq	3/3	EB day 10	27169731
Wt_D10_MBD Input	MBD-seq Input	1/2	EB day 10	29424282
Wt_D10_MBD Input	MBD-seq Input	2/2	EB day 10	9220734
Wt_D12_NPC_MBD	MBD-seq	1/1	NPC day 12	10291112
Wt_D12_NPC_MBD_Input	MBD-seq Input	1/1	NPC day 12	63497824
Wt_D15_NPC_MBD	MBD-seq	1/2	NPC day 15	11664764
Wt_D15_NPC_MBD	MBD-seq	2/2	NPC day 15	10830718
Wt_D15_NPC_MBD_Input	MBD-seq Input	1/1	NPC day 15	43886408
Wt_D12_EB_MBD	MBD-seq	1/5	EB day 12	8930514
Wt_D12_EB_MBD	MBD-seq	2/5	EB day 12	10079156
Wt_D12_EB_MBD	MBD-seq	3/5	EB day 12	17587802
Wt_D12_EB_MBD	MBD-seq	4/5	EB day 12	9460374
Wt_D12_EB_MBD	MBD-seq	5/5	EB day 12	20253024
Wt_D12_EB_MBD_Input	MBD-seq Input	1/3	EB day 12	42029734
Wt_D12_EB_MBD_Input	MBD-seq Input	2/3	EB day 12	52941142
Wt_D12_EB_MBD_Input	MBD-seq Input	3/3	EB day 12	12420452
C6_Dnmt3b$^{-/-}$_D0_MBD	MBD-seq	1/4	ES day 0	6251656

Dataset	Experiment	Replicate	Cell type	Raw NGS reads
C6_Dnmt3b$^{-/-}$_D0_MBD	MBD-seq	2/4	ES day 0	11438588
C6_Dnmt3b$^{-/-}$_D0_MBD	MBD-seq	3/4	ES day 0	6720924
C6_Dnmt3b$^{-/-}$_D0_MBD	MBD-seq	4/4	ES day 0	6646402
C6_Dnmt3b$^{-/-}$_D0_MBD_Input	MBD-seq Input	1/3	ES day 0	53826984
C6_Dnmt3b$^{-/-}$_D0_MBD_Input	MBD-seq Input	2/3	ES day 0	54452808
C6_Dnmt3b$^{-/-}$_D0_MBD_Input	MBD-seq Input	3/3	ES day 0	30193200
C6_Dnmt3b$^{-/-}$_D0_MBD_Input	MBD-seq Input	4/4	ES day 0	7955702
C6_Dnmt3b$^{-/-}$_D4_MBD	MBD-seq	1/2	EB day 4	43309174
C6_Dnmt3b$^{-/-}$_D4_MBD	MBD-seq	2/2	EB day 4	34781772
C6_Dnmt3b$^{-/-}$_D4_MBD_Input	MBD-seq Input	1/2	EB day 4	53054095
C6_Dnmt3b$^{-/-}$_D4_MBD_Input	MBD-seq Input	2/2	EB day 4	12465366
C6_Dnmt3b$^{-/-}$_D8_MBD	MBD-seq	1/2	EB day 8	23276090
C6_Dnmt3b$^{-/-}$_D8_MBD	MBD-seq	2/2	EB day 8	18614514
C6_Dnmt3b$^{-/-}$_D8_MBD_Input	MBD-seq Input	1/2	EB day 8	53603163
C6_Dnmt3b$^{-/-}$_D8_MBD_Input	MBD-seq Input	2/2	EB day 8	10949146
C6_Dnmt3b$^{-/-}$_D10_MBD	MBD-seq	1/2	EB day 10	15788430
C6_Dnmt3b$^{-/-}$_D10_MBD	MBD-seq	2/2	EB day 10	15788430
C6_Dnmt3b$^{-/-}$_D10_MBD_Input	MBD-seq Input	1/2	EB day 10	9916152
C6_Dnmt3b$^{-/-}$_D10_MBD_Input	MBD-seq Input	2/2	EB day 10	9916152
C6_Dnmt3b$^{-/-}$_D12_EB_MBD	MBD-seq	1/5	EB day 12	7573604
C6_Dnmt3b$^{-/-}$_D12_EB_MBD	MBD-seq	2/5	EB day 12	12444032
C6_Dnmt3b$^{-/-}$_D12_EB_MBD	MBD-seq	3/5	EB day 12	9177854
D10_Dnmt3b$^{-/-}$_D12_EB_MBD	MBD-seq	4/5	EB day 12	11728820
C6_Dnmt3b$^{-/-}$_D12_EB_MBD	MBD-seq	5/5	EB day 12	20159680
C6_Dnmt3b$^{-/-}$_D12_MBD_Input	MBD-seq Input	1/4	EB day 12	43680746
C6_Dnmt3b$^{-/-}$_D12_MBD_Input	MBD-seq Input	2/4	EB day 12	55238474
D10_Dnmt3b$^{-/-}$_D12_MBD_Input	MBD-seq Input	3/4	EB day 12	45792144
C6_Dnmt3b$^{-/-}$_D12_MBD_Input	MBD-seq Input	4/4	EB day 12	10173124
C6_Dnmt3b$^{-/-}$_D15_MBD	MBD-seq	1/2	NPC day 15	30081226
C6_Dnmt3b$^{-/-}$_D15_MBD	MBD-seq	2/2	NPC day 15	17463176
C6_Dnmt3b$^{-/-}$_D15_MBD_Input	MBD-seq Input	1/1	NPC day 15	5468722
B4_Dnmt3b3KI_D0_MBD	MBD-seq	1/2	ES day 0	27753048
F8_Dnmt3b3KI_D0_MBD	MBD-seq	2/2	ES day 0	22435802
B4_Dnmt3b3KI_D0_MBD_Input	MBD-seq Input	1/2	ES day 0	60139914
F8_Dnmt3b3KI_D0_MBD_Input	MBD-seq Input	2/2	ES day 0	66014882
B4_Dnmt3b3KI_D12_MBD	MBD-seq	1/3	EB day 12	13131846
B4_Dnmt3b3KI_D12_MBD	MBD-seq	2/3	EB day 12	16169232
F8_Dnmt3b3KI_D12_MBD	MBD-seq	3/3	EB day 12	4176108
B4_Dnmt3b3KI_D12_MBD_Input	MBD-seq Input	1/2	EB day 12	57183376
F8_Dnmt3b3KI_D12_MBD_Input	MBD-seq Input	2/2	EB day 12	57468248
B4_Δ1-229 Dnmt3b1KI_D0_MBD	MBD-seq	1/2	ES day 0	16717038
E11_Δ1-229 Dnmt3b1KI_D0_MBD	MBD-seq	2/2	ES day 0	30583612
B4_Δ1-229 Dnmt3b1KI_D0_MBD_Input	MBD-seq Input	1/2	ES day 0	53495214
E11_Δ1-229 Dnmt3b1KI_D0_MBD_Input	MBD-seq Input	2/2	ES day 0	27669162
B4_Δ1-229 Dnmt3b1KI_D12_MBD	MBD-seq	1/3	EB day 12	10445076
B4_Δ1-229 Dnmt3b1KI_D12_MBD	MBD-seq	2/3	EB day 12	14423392
E11_Δ1-229 Dnmt3b1KI_D12_MBD	MBD-seq	3/3	EB day 12	24964070
B4_Δ1-229 Dnmt3b1KI_D12_MBD_Input	MBD-seq Input	1/2	EB day 12	62466006
E11_Δ1-229 Dnmt3b1KI_D12_MBD_Input	MBD-seq Input	2/2	EB day 12	33216606
H2D3_Dnmt3b1 KI_D0_MBD	MBD-seq	1/2	ES day 0	19702058
H2G3_Dnmt3b1 KI_D0_MBD	MBD-seq	2/2	ES day 0	28340630
H2D3_Dnmt3b1 KI_D0_MBD_Input	MBD-seq Input	1/2	ES day 0	30511410
H2G3_Dnmt3b1 KI_D0_MBD_Input	MBD-seq Input	2/2	ES day 0	27850636
H2D3_Dnmt3b1 KI_EB_MBD	MBD-seq	1/3	EB day 12	27293988
H2G3_Dnmt3b1 KI_D12_MBD	MBD-seq	2/3	EB day 12	28270678
H2G3_Dnmt3b1 KI_D12_MBD	MBD-seq	3/3	EB day 12	38029278
H2D3_Dnmt3b1 KI_D12_MBD_Input	MBD-seq Input	1/2	EB day 12	29095328
H2G3_Dnmt3b1 KI_D12_MBD_Input	MBD-seq Input	2/2	EB day 12	25228392
A9_Dnmt3b(N)3a KI_D0_MBD	MBD-seq	1/3	ES day 0	19646090
B4_Dnmt3b(N)3a KI_D0_MBD	MBD-seq	2/3	ES day 0	25442290
C9_Dnmt3b(N)3a KI_D0_MBD	MBD-seq	3/3	ES day 0	14875102
A9_Dnmt3b(N)3a KI_D0_MBD_Input	MBD-seq Input	1/3	ES day 0	50203754
B4_Dnmt3b(N)3a KI_D0_MBD_Input	MBD-seq Input	2/3	ES day 0	28784144
C9_Dnmt3b(N)3a KI_D0_MBD_Input	MBD-seq Input	2/3	ES day 0	21157322

Dataset	Experiment	Replicate	Cell type	Raw NGS reads
A9_Dnmt3b(N)3a KI_D12_MBD	MBD-seq	1/3	EB day 12	9271138
B4_Dnmt3b(N)3a KI_D12_MBD	MBD-seq	2/3	EB day 12	10442394
C9_Dnmt3b(N)3a KI_D12_MBD	MBD-seq	3/3	EB day 12	19387172
A9_Dnmt3b(N)3a KI_D12_MBD_Input	MBD-seq Input	1/3	EB day 12	52639422
B4_Dnmt3b(N)3a KI_D12_MBD_Input	MBD-seq Input	2/3	EB day 12	29295082
C9_Dnmt3b(N)3a KI_D12_MBD_Input	MBD-seq Input	2/3	EB day 12	25556094
D9_Δ1-218 Dnmt3b1KI_D0_MBD	MBD-seq	1/2	ES day 0	33407708
G6_Δ1-218 Dnmt3b1KI_D0_MBD	MBD-seq	2/2	ES day 0	52397672
D9_Δ1-218 Dnmt3b1KI_D0_MBD_Input	MBD-seq Input	1/2	ES day 0	17787610
G6_Δ1-218 Dnmt3b1KI_D0_MBD_Input	MBD-seq Input	2/2	ES day 0	26580034
D9_Δ1-218 Dnmt3b1KI_D12_MBD	MBD-seq	1/3	EB day 12	30092304
D9_Δ1-218 Dnmt3b1KI_D12_MBD	MBD-seq	2/3	EB day 12	27704716
G6_Δ1-218 Dnmt3b1KI_D12_MBD	MBD-seq	3/3	EB day 12	50872226
G6_Δ1-218 Dnmt3b1KI_D12_MBD	MBD-seq	4/4	EB day 12	28927538
D9_Δ1-218 Dnmt3b1KI_D12_MBD_Input	MBD-seq Input	1/2	EB day 12	21558484
G6_Δ1-218 Dnmt3b1KI_D12_MBD_Input	MBD-seq Input	2/2	EB day 12	22132020
C12_Dnmt3a(N)3b KI_D0_MBD	MBD-seq	1/3	ES day 0	32770502
E11_Dnmt3a(N)3b KI_D0_MBD	MBD-seq	2/3	ES day 0	19511412
G2_Dnmt3a(N)3b KI_D0_MBD	MBD-seq	3/3	ES day 0	15215366
C12_Dnmt3a(N)3b KI_D0_MBD_Input	MBD-seq Input	1/3	ES day 0	7452690
E11_Dnmt3a(N)3b KI_D0_MBD_Input	MBD-seq Input	2/3	ES day 0	8561792
G2_Dnmt3a(N)3b KI_D0_MBD_Input	MBD-seq Input	2/3	ES day 0	7552784
C12_Dnmt3a(N)3b KI_D12_MBD	MBD-seq	1/3	EB day 12	2896797
E11_Dnmt3a(N)3b KI_D12_MBD	MBD-seq	2/3	EB day 12	10920572
G2_Dnmt3a(N)3b KI_D12_MBD	MBD-seq	3/3	EB day 12	14664850
C12_Dnmt3a(N)3b KI_D12_MBD_Input	MBD-seq Input	1/3	EB day 12	4164928
E11_Dnmt3a(N)3b KI_D12_MBD_Input	MBD-seq Input	2/3	EB day 12	4899570
G2_Dnmt3a(N)3b KI_D12_MBD_Input	MBD-seq Input	2/3	EB day 12	4367794
A4_ΔA Dnmt3b1KI_D0_MBD (Δ1-50)	MBD-seq	1/2	ES day 0	18353250
E3_ΔA Dnmt3b1KI_D0_MBD (Δ1-50)	MBD-seq	2/2	ES day 0	21845128
A4_ΔA Dnmt3b1KI_D0_MBD (Δ1-50)_Input	MBD-seq Input	1/2	ES day 0	16821500
E3_ΔA Dnmt3b1KI_D0_MBD (Δ1-50)_Input	MBD-seq Input	2/2	ES day 0	16083524
A4_ΔA Dnmt3b1KI_D12_MBD (Δ1-50)	MBD-seq	1/3	EB day 12	13311274
E3_ΔA Dnmt3b1KI_D12_MBD (Δ1-50)	MBD-seq	2/3	EB day 12	9954214
E3_ΔA Dnmt3b1KI_D12_MBD (Δ1-50)	MBD-seq	3/3	EB day 12	10201384
A4_ΔA Dnmt3b1KI_D12_MBD (Δ1-50)_Input	MBD-seq	3/3	EB day 12	11283798
E3_ΔA Dnmt3b1KI_D12_MBD (Δ1-50)_Input	MBD-seq Input	1/2	EB day 12	14132124
D5_ΔB Dnmt3b1KI_D0_MBD (Δ51-98)	MBD-seq	1/2	ES day 0	23131462
H11_ΔB Dnmt3b1KI_D0_MBD (Δ51-98)	MBD-seq	2/2	ES day 0	15726122
D5_ΔB Dnmt3b1KI_D0_MBD (Δ51-98)_Input	MBD-seq Input	1/2	ES day 0	17458842
H11_ΔB Dnmt3b1KI_D0_MBD (Δ51-98)_Input	MBD-seq Input	2/2	ES day 0	9161206
D5_ΔB Dnmt3b1KI_D12_MBD (Δ51-98)	MBD-seq	1/3	EB day 12	20393640
H11_ΔB Dnmt3b1KI_D12_MBD (Δ51-98)	MBD-seq	2/3	EB day 12	1441214
H11_ΔB Dnmt3b1KI_D12_MBD (Δ51-98)	MBD-seq	3/3	EB day 12	13508170
H11_ΔB Dnmt3b1KI_D12_MBD (Δ51-98)_Input	MBD-seq Input	1/2	EB day 12	11225440
D5_ΔB Dnmt3b1KI_D12_MBD (Δ51-98)_Input	MBD-seq Input	2/2	EB day 12	13241170
C1_ΔC Dnmt3b1KI_D0_MBD (Δ98-146)	MBD-seq	1/2	ES day 0	9316114
C1_ΔC Dnmt3b1KI_D0_MBD (Δ98-146)	MBD-seq	2/2	ES day 0	10030266
C1_ΔC Dnmt3b1KI_D0_MBD (Δ98-146)_Input	MBD-seq Input	1/2	ES day 0	14688048
C1_ΔC Dnmt3b1KI_D12_MBD (Δ98-146)	MBD-seq	1/2	EB day 12	12050160
C1_ΔC Dnmt3b1KI_D12_MBD (Δ98-146)	MBD-seq	2/2	EB day 12	11211256
C1_ΔC Dnmt3b1KI_D12_MBD (Δ98-146)_Input	MBD-seq Input	1/2	EB day 12	18637052
F12_ΔD Dnmt3b1KI_D0_MBD (Δ146-177)	MBD-seq	1/2	ES day 0	19015506
G9_ΔD Dnmt3b1KI_D0_MBD (Δ146-177)	MBD-seq	2/2	ES day 0	23216952
F12_ΔD Dnmt3b1KI_D0_MBD (Δ146-177)_Input	MBD-seq Input	1/2	ES day 0	24339532
G9_ΔD Dnmt3b1KI_D0_MBD (Δ146-177)_Input	MBD-seq Input	2/2	ES day 0	18338408
F12_ΔD Dnmt3b1KI_D12_MBD (Δ146-177)	MBD-seq	1/3	EB day 12	14128966
G9_ΔD Dnmt3b1KI_D12_MBD (Δ146-177)	MBD-seq	2/3	EB day 12	29474304
G9_ΔD Dnmt3b1KI_D12_MBD (Δ146-177)	MBD-seq	3/3	EB day 12	24855470
F12_ΔD Dnmt3b1KI_D12_MBD (Δ146-177)_Input	MBD-seq Input	1/2	EB day 12	16603766
G9_ΔD Dnmt3b1KI_D12_MBD (Δ146-177)_Input	MBD-seq Input	1/2	EB day 12	15937368
G8_ΔE Dnmt3b1KI_D0_MBD (Δ177-229)	MBD-seq	1/2	ES day 0	14316170
G10_ΔE Dnmt3b1KI_D0_MBD (Δ177-229)	MBD-seq	2/2	ES day 0	37114144
G8_ΔE Dnmt3b1KI_D0_MBD (Δ177-229)_Input	MBD-seq Input	1/2	ES day 0	16001734

Dataset	Experiment	Replicate	Cell type	Raw NGS reads
G10_ΔE Dnmt3b1KI_D0_MBD (Δ177-229)_Input	MBD-seq Input	2/2	ES day 0	15725396
G8_ΔE Dnmt3b1KI_D12_MBD (Δ177-229)	MBD-seq	1/3	EB day 12	13731114
G10_ΔE Dnmt3b1KI_D12_MBD (Δ177-229)	MBD-seq	2/3	EB day 12	13468194
G10_ΔE Dnmt3b1KI_D12_MBD (Δ177-229)	MBD-seq	3/3	EB day 12	12724414
G8_ΔE Dnmt3b1KI_D12_MBD (Δ177-229)_Input	MBD-seq Input	1/2	EB day 12	17973278
G10_ΔE Dnmt3b1KI_D12_MBD (Δ177-229)_Input	MBD-seq Input	1/2	EB day 12	17785576
A8_T7_S277P Dnmt3b KI_D0_MBD	MBD-seq	1/3	ES day 0	17653214
D8_ T7_S277P Dnmt3b KI_D0_MBD	MBD-seq	2/3	ES day 0	8219610
F3_ T7_S277P Dnmt3b KI_D0_MBD	MBD-seq	3/3	ES day 0	10226900
A8_ T7_S277P Dnmt3b KI_D0_MBD_Input	MBD-seq Input	1/3	ES day 0	10935510
D8_ T7_S277P Dnmt3b KI_D0_MBD_Input	MBD-seq Input	2/3	ES day 0	14379072
F3_ T7_S277P Dnmt3b KI_D0_MBD_Input	MBD-seq Input	2/3	ES day 0	14940218
A8_ T7_S277P Dnmt3b KI_D12_MBD	MBD-seq	1/3	EB day 12	14612312
D8_ T7_S277P Dnmt3b KI_D12_MBD	MBD-seq	2/3	EB day 12	7181686
F3_ T7_S277P Dnmt3b KI_D12_MBD	MBD-seq	3/3	EB day 12	12732058
A8_ T7_S277P Dnmt3b KI_D12_MBD_Input	MBD-seq Input	1/3	EB day 12	11725590
D8_ T7_S277P Dnmt3b KI_D12_MBD_Input	MBD-seq Input	2/3	EB day 12	14022810
F3_ T7_S277P Dnmt3b KI_D12_MBD_Input	MBD-seq Input	2/3	EB day 12	6604620
Wt_D0	chrRNA-seq	1	ES day 0	56490764
Wt_D12	chrRNA-seq	1	EB day 12	50527545
C6_Dnmt3b$^{-/-}$_D0	chrRNA-seq	1	ES day 0	54631993
C6_Dnmt3b$^{-/-}$_D12	chrRNA-seq	1	EB day 12	73704272
B4_Dnmt3b3KI_D0	chrRNA-seq	1/2	ES day 0	47713917
F8_Dnmt3b3KI_D0	chrRNA-seq	2/2	ES day 0	56026276
B4_Dnmt3b3KI_D12	chrRNA-seq	1/2	EB day 12	48686333
F8_Dnmt3b3KI_D12	chrRNA-seq	2/2	EB day 12	58661232
H2D3_Dnmt3b1 KI_D0	chrRNA-seq	1/2	ES day 0	47026597
H2G3_Dnmt3b1 KI_D0	chrRNA-seq	2/2	ES day 0	56303544
H2D3_Dnmt3b1 KI_D12	chrRNA-seq	1/2	EB day 12	54319239
H2G3_Dnmt3b1 KI_D12	chrRNA-seq	2/2	EB day 12	56730975
B4_Δ1-229 Dnmt3b1KI_D0	chrRNA-seq	1/2	ES day 0	54230062
E11_Δ1-229 Dnmt3b1KI_D0	chrRNA-seq	2/2	ES day 0	55665341
B4_Δ1-229 Dnmt3b1KI_D12	chrRNA-seq	1/2	EB day 12	56099522
E11_Δ1-229 Dnmt3b1KI_D12	chrRNA-seq	2/2	EB day 12	59351501
D1_T7_Dnmt3b1 KI_D0	ChIP-seq	1/2	ES day 0	127669390
F8_T7_Dnmt3b1 KI_D0	ChIP-seq	2/2	ES day 0	93890646
D1_T7_Dnmt3b1 KI_D0_Input	ChIP-seq Input	1/2	ES day 0	17775548
F8_T7_Dnmt3b1 KI_D0_Input	ChIP-seq Input	2/2	ES day 0	17915014
A1_T7_ Δ1-218_Dnmt3b1 KI_D0	ChIP-seq	1/2	ES day 0	61864108
C11_T7_ Δ1-218_Dnmt3b1 KI_D0	ChIP-seq	2/2	ES day 0	79317988
A1_T7_ Δ1-218_Dnmt3b1 KI_D0_Input	ChIP-seq Input	1/2	ES day 0	20255312
C11_T7_ Δ1-218_Dnmt3b1 KI_D0_Input	ChIP-seq Input	2/2	ES day 0	24851870
D1_T7_Dnmt3b1 KI_D12	ChIP-seq	1/2	EB day 12	76225134
F8_T7_Dnmt3b1 KI_D12	ChIP-seq	2/2	EB day 12	42298842
D1_T7_Dnmt3b1 KI_D12_Input	ChIP-seq Input	1/2	EB day 12	13492881
F8_T7_Dnmt3b1 KI_D12_Input	ChIP-seq Input	2/2	EB day 12	14559993
A1_T7_ Δ1-218_Dnmt3b1 KI_D12	ChIP-seq	1/2	EB day 12	100829393
C11_T7_ Δ1-218_Dnmt3b1 KI_D12	ChIP-seq	2/2	EB day 12	65508431
A1_T7_ Δ1-218_Dnmt3b1 KI_D12_Input	ChIP-seq Input	1/2	EB day 12	14999585
C11_T7_ Δ1-218_Dnmt3b1 KI_D12_Input	ChIP-seq Input	2/2	EB day 12	16020325
Wt_D0_H3K9me3_ChIP	ChIP-seq	1/2	ES day 0	33589164
Wt_D0_H3K9me3_ChIP	ChIP-seq	2/2	ES day 0	23710676
Wt_D0_H3K9me3_ChIP_Input	ChIP-seq Input	1/2	ES day 0	13386146
Wt_D0_H3K9me3_ChIP_Input	ChIP-seq Input	2/2	ES day 0	14757026
C6_Dnmt3b$^{-/-}$_D0_H3K9me3_ChIP	ChIP-seq	1/3	ES day 0	50863430
C6_Dnmt3b$^{-/-}$_D0_ H3K9me3_ChIP	ChIP-seq	2/3	ES day 0	95000498
C6_Dnmt3b$^{-/-}$_D0_ H3K9me3_ChIP	ChIP-seq	3/3	ES day 0	103700504
C6_Dnmt3b$^{-/-}$_D0_ H3K9me3_ChIP_Input	ChIP-seq Input	1/2	ES day 0	50863430
C6_Dnmt3b$^{-/-}$_D0_ H3K9me3_ChIP_Input	ChIP-seq Input	2/2	ES day 0	19473924
Wt_D12_H3K9me3_ChIP	ChIP-seq	1/2	EB day 12	13819275
Wt_D12_H3K9me3_ChIP	ChIP-seq	2/2	EB day 12	20129666
Wt_D12_H3K9me3_ChIP_Input	ChIP-seq Input	1/2	EB day 12	14825396
Wt_D12_H3K9me3_ChIP_Input	ChIP-seq Input	2/2	EB day 12	22536108
C6_Dnmt3b$^{-/-}$_D12_H3K9me3_ChIP	ChIP-seq	1/3	EB day 12	58213742

Dataset	Experiment	Replicate	Cell type	Raw NGS reads
C6_Dnmt3b$^{-/-}$_D12_ H3K9me3_ChIP	ChIP-seq	2/3	EB day 12	75956374
C6_Dnmt3b$^{-/-}$_D12_ H3K9me3_ChIP	ChIP-seq	3/3	EB day 12	67462411
C6_Dnmt3b$^{-/-}$_D12_ H3K9me3_ChIP_Input	ChIP-seq Input	1/2	EB day 12	58213742
C6_Dnmt3b$^{-/-}$_D12_ H3K9me3_ChIP_Input	ChIP-seq Input	2/2	EB day 12	14825396

Table A2: Summary for the NGS datasets generated in this work.

The number of reads described in this table present the values of raw aligned reads after QC thresholding (Phred Q score ≥ 30), deduplication, and prior to any normalization.