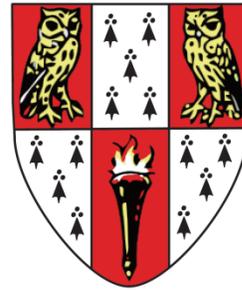
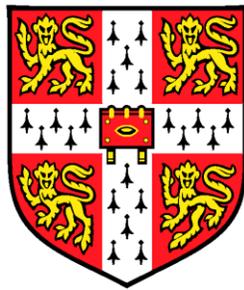


This dissertation is submitted to the University of Cambridge for the  
degree of Doctor of Philosophy by

Nelly Nikolaeva Olova

**Exploring functional significance of asymmetric DNA  
methylation in early mammalian development**



University of Cambridge ~ Hughes Hall

The Babraham Institute

September 2014



## **Declaration**

The research presented in this dissertation was performed in the Epigenetics ISP at the Babraham Institute, Cambridge, under the supervision of Prof. Wolf Reik FRS. This dissertation, or any part of it, has not been previously submitted for a degree, diploma or other qualification at the University of Cambridge nor any other university. It is the result of my own work except where specifically stated in the text.

Nelly Olova

September 2014

## **Statement of length**

This thesis does not exceed the 60,000-word limit stipulated by the Degree Committee.



## Acknowledgements

Working on this project has been a long and lonely path through the years, and it would not have been possible without the support and help of many friends and colleagues.

I am grateful to Wendy Dean and Fátima Santos, who joined me in the last year of this work, and have given much of their time to help, support, advise, encourage and share the wealth of their knowledge. It has been a pleasure to work with you and have your support!

I also want to especially thank Felix Krueger for all his patience and amazing efficiency while working on my countless bioinformatics requests through the years. Simon Andrews has always provided help in critical times, and has always been able to resolve even the most entangled tasks!

I am thankful to Miguel Branco who mentored me during the first three years, for his critical advices and different perspectives, and all the help in planning complex experiments and approaching difficult questions.

I am thankful to all Reik lab members, past and present, and especially Rebecca Berrens and Ferdinand von Meyenn, who, alongside with Fátima Santos, helped with experiments in the critical last few months before my submission. I have also received help and advice from Tamir Chandra and Tim Hore, Gabriella Ficz, and the rest of past and current lab members.

I am thankful for all the work that the staff at BI Sequencing and Mass spectrometry facilities have put into this project – Kristina Tabbada, David Oxley and Judith Webster.

Anne Segonds-Pichon has always been very helpful in sorting out my statistical puzzles!

I have been lucky to have a bunch of precious PhD students to commiserate with, like Rebecca Berrens, Heather Burgess, Julian Peat, Emilia Dimitrova, Alex Murray, Harry Armstrong and the rest of the PhD student gang, who were there to receive the usual PhD rant and remind me that I was not alone in this boat.

I am also very thankful to Gabi and Solenn for the continuous logistical support of late night rides home from work, which helped greatly reduce the bus-chasing stress levels, plus the included package of nice conversations and moral support along the way.

I inevitably owe a big ‘thank you’ to my lovely college friends - Roberto, Konstantin, Michel, Bastiaan, Nik, Ollie, Günel, Marek, Stefan and Steve, who had their ways to always bring a wide smile on my face and served as our Hughes-only ‘Academics Anonymous’ club. You were

my college family! I can't thank enough Leana for all the facebook chat hours and the true friendship! Also the Bulgarian circle of friends, who were a source of great joy and challenges, and taught me very precious life lessons! I thank the Hughes Hall Boat Club for having me for those two intense years, and letting me be part of their competitive crews, despite my numerous obligations and time-restrictions! Rowing has been my most successful de-stress medication!

I cannot thank enough to my family who have had to put up with largely losing me from their lives, and have nevertheless always been supportive and understanding to my ambitions and desire to travel and achieve new far away heights! I know I can never make up for the time lost and your life moments, which I have missed, and I can only hope for your forgiveness!

Beyond all, I would like to thank my supervisor Wolf Reik for granting me the honour to be a PhD student in his lab! I am grateful for all the freedom I have been given to explore my ideas, for keeping me on track, for challenging my knowledge and views, and for the full support I have received for my collaborations and grant applications! Thank you for always being on the other side of the email, no matter where in the world you have been!

It has been an absolute privilege to be a part of the Reik team, where I have learned so many mind-set changing lessons for science, myself, people, work and life in general!

## Summary

DNA methylation is an epigenetic modification important for the regulation of transcriptional activity in processes like genomic imprinting, retrotransposon silencing and centromeric stabilisation. It is also crucial for correct embryonic development and the differentiation of embryonic stem cells (ESCs) into diverse cell types. Although in mammals DNA methylation occurs predominantly in the symmetric CG context, it has been shown that certain cell types and tissues (ESCs, oocytes, primordial germ cells) have substantial amounts of methylation outside of the CG dinucleotide, which is asymmetric. Its presence in early developmental stages related to toti- or pluri-potency raises the intriguing possibility that non-CG context methylation may have a role in differential gene expression during those stages, contributing to their transcriptional plasticity.

I investigated the distribution, dynamics and functional significance of non-CG methylation in early mouse development. Most methods for DNA methylation analysis are either targeted for the analysis of CG methylation or, in the case of bisulfite sequencing, suffer from potentially confounding issues. I have compared existing approaches to detect methylation outside of CG context and developed novel tools, namely the use of antibodies against non-CG methylation, to either analyse global levels of non-CG methylation, or validate its genomic distribution. With these, I have evaluated the role of components of the DNA methylation machinery and have followed the dynamic changes of non-CG context methylation throughout development. My analysis reveals that the highest levels of non-CG methylation in the mouse are present in the mature oocyte and the zygote. The enzymes responsible for establishing and maintaining its levels are the *de novo* Dnmts (3a and 3b), among which the activity of Dnmt3a2 towards CH seems regulated, suggesting a specific rather than an unspecific role. In ES cells, CH methylation correlates with active histone marks e.g. H3K4me3, and inversely with H3K27me3. The distribution of mCH, both in ESC naïve and primed pluripotency states, is very heterogeneous, while its nuclear distribution is very homogenous. mCH is physically recognised by a number of pluripotency factors such as Oct4 and Sox2, as well as by other DNA modification-sensitive proteins like MeCP2, Foxk1 and Foxk2. Moreover, mCH cannot be hydroxylated by the Tet family of enzymes, and repels proteins involved in the initiation of base-excision repair (AID and RPA), thus potentially escaping active demethylation in the zygote. In summary, my results show that mCH is a valid methylation mark, with a functional significance in early mouse development.



## Abbreviations

5mC - 5-methylcytosine

5hmC - 5-hydroxymethylcytosine

5fC – 5-formylcytosine

5caC – 5-carboxycytosine

ES cells, ESCs – Embryonic Stem Cells; mESCs - from mouse, hESCs – from human

pMEFs – primary mouse embryonic fibroblasts

WT – wild type

CGIs (CG islands) – areas in the genome (200-500bp) with higher levels of CG dinucleotides than the average for the genome

BS-seq(uencing) – next generation sequencing of bisulphite-converted DNA

(h)MeDIP – (hydroxy)methylated DNA ImmunoPrecipitation

ICM – Inner Cell Mass, the cells in the preimplantation embryo that give rise to the foetus

DMR – differentially methylated region (usually associated with parental imprints)

IAP - Intracisternal A Particle, a family of retrotransposons normally heavily silenced

BLAST - Basic Local Alignment Search Tool

IMR90 – human embryonic lung fibroblast cell line

iPS cells – Induced Pluripotent Stem cells

KD – knock-down, a targeted and specific reduction in the expression of a particular gene

KO – knock-out, i.e a genotype having a null mutation or missing a gene

NNA - nearest neighbour analysis, a method for quantifying the nucleotides next to a cytosine

RdDM – RNA-directed DNA methylation, a *de novo* methylation mechanism in plants

TKO – Triple Knock-Out – ES cell line missing functional Dnmt1, Dnmt3a and Dnmt3b.

B6 (C57BL/6 J) – a common inbred strain of laboratory mouse, commonly called as ‘black 6’.

LC-MS – Liquid Chromatography coupled with Mass Spectrometry



## Table of contents

Acknowledgements .....	3
<b>1 Introduction .....</b>	<b>1</b>
1.1 Epigenetic DNA modifications and their role in development .....	1
1.1.1 Occurrence and distribution of DNA modifications .....	1
1.1.2 Biological significance and role of DNA methylation.....	2
1.2 Occurrence of non-CG DNA modifications in different organisms and cell types .....	3
1.3 Establishment and maintenance of non-CG methylation .....	7
1.3.1 Dnmt1 – the maintenance methyltransferase .....	8
1.3.2 Dnmt3 – the de novo methyltransferase family .....	10
1.3.3 Dnmt2 – the enigmatic Dnmt .....	13
1.3.4 Other protein and chromatin factors in DNA methylation.....	15
1.3.5 Non-coding RNA pathways in the regulation of DNA methylation .....	16
1.4 Removal of non-CG methylation .....	18
1.5 Biological role of CHH and CHG methylation .....	20
1.6 Techniques for the analysis of non-CG context methylation .....	24
1.6.1 Methods for DNA methylation analysis – an overview.....	24
1.6.2 BS-seq and the analyses of non-CG methylation.....	27
1.6.2.1 Genomic coverage and depth .....	27
1.6.2.2 Computing power .....	28
1.6.2.3 Bisulphite conversion efficiency .....	28
1.6.2.4 Amplification bias .....	29
1.6.2.5 DNA degradation.....	30
1.6.3 Additional tools facilitating DNA methylation analysis .....	30
1.6.4 Technological perspectives .....	32
1.7 Aims and hypothesis.....	32
<b>2 Materials and methods.....</b>	<b>35</b>
2.1 Cell lines and culture .....	35
2.2 Mouse tissues.....	36
2.3 Molecular Biology.....	36
2.3.1 Isolation of DNA and RNA from cells and tissues .....	36
2.3.2 Bisulphite conversion of genomic DNA .....	36
2.3.3 Amplification (PCR) of major satellite .....	37
2.3.4 T-A cloning .....	37
2.3.5 Sanger sequencing.....	38

2.3.6	Synthesis of M13 fragments .....	38
2.3.7	Next generation whole genome bisulphite sequencing (BS-seq, WGBS).....	38
2.3.8	Reduced representation bisulphite sequencing with MspJI (meRRBS) .....	39
2.3.9	Mass spectrometry .....	40
2.3.10	In vitro DNA methylation assays.....	40
2.3.11	Nearest Neighbour Analysis (NNA).....	40
2.3.12	Luminometric Methylation Assay (LUMA).....	41
2.3.13	Gene synthesis .....	41
2.3.14	Gateway cloning .....	41
2.3.15	Antibody concentration.....	42
2.3.16	DNA ELISA assays .....	43
2.3.17	Immunofluorescence and cell imaging .....	44
2.3.18	Preparation of metaphase spreads.....	44
2.3.19	cDNA synthesis from total RNA .....	45
2.3.20	Quantitative PCR (qPCR) for gene expression.....	45
2.3.21	Dot blot for measurement of total 5mC or 5hmC .....	46
2.3.22	Methylated DNA Immunoprecipitation assay (MeDIP).....	46
2.3.23	Protein binders pull down from nuclear extract.....	47
2.3.24	Tet1 oxidation assay .....	47
2.3.25	FACS-sorting of ES cells.....	48
2.3.26	Whole Genome Amplification (WGA).....	48
2.4	Bioinformatics.....	48
2.4.1	Data mapping.....	48
2.4.2	FastQC analysis of raw and mapped reads data .....	49
2.4.3	Repeat consensus analysis .....	49
2.4.4	Analysis of MeDIP and WGBS datasets .....	49
2.4.5	Analysis of ChIP datasets .....	50
2.4.6	In silico digestion.....	51
2.4.7	MspJI-RRBS base-calling.....	51
2.4.8	WRC sequence analysis.....	51
<b>3</b>	<b>Genomic distribution of non-CG methylation: a classical approach .....</b>	<b>53</b>
3.1	Introduction .....	53
3.2	Aims .....	54
3.3	Results .....	54
3.3.1	Non-CG methylation in MeDIP-seq.....	54
3.3.2	Low resolution BS-seq datasets: non-CG BS-conversion artefacts.....	62

3.3.3	Distribution of non-CG methylation in low resolution BS-seq datasets .....	66
3.3.4	Bisulphite treatment and DNA degradation .....	73
3.3.5	Measuring global levels of non-CG methylation in the mouse genome .....	78
3.3.5.1	NNA .....	79
3.3.5.2	LUMA .....	80
3.4	Discussion.....	84
<b>4</b>	<b>Novel tools and techniques for the analysis of non-CG methylation .....</b>	<b>89</b>
4.1	Introduction .....	89
4.2	Aims .....	90
4.3	Results .....	91
4.3.1	Enzymes specific for non-CG context methylation .....	91
4.3.1.1	Restriction enzymes sensitive to non-CG context methylation.....	91
4.3.1.2	Synthesis, cloning and validation of RlaI.....	93
4.3.1.3	Context specific bacterial methyl-transferases .....	94
4.3.2	MspJI-based methylation enrichment RRBS (meRRBS) .....	94
4.3.3	Development of context-specific antibodies for 5-methylcytosine.....	97
4.3.3.1	Antigen synthesis and monoclonal antibody production.....	98
4.3.3.2	Validation strategy.....	99
4.3.3.3	Avidin-biotin DNA ELISA .....	100
4.3.3.4	Results: ELISA screen.....	101
4.3.3.5	Results: immunofluorescence (IF) .....	103
4.3.3.6	Antibody concentration and purification.....	105
4.3.4	DNA ELISA.....	106
4.3.4.1	DNA adsorption to the 96-well polystyrene plate.....	107
4.3.4.2	Optimising assay conditions for genomic DNA.....	109
4.3.4.3	Quantitative capacity of the DNA ELISA assay .....	112
4.3.4.4	Enhancing sensitivity and dynamic range of the mCA ELISA assay ..	113
4.4	Discussion.....	115
<b>5</b>	<b>Genomic distribution of non-CG methylation: results from novel techniques.....</b>	<b>117</b>
5.1	Introduction .....	117
5.2	Aims .....	117
5.3	Results .....	118
5.3.1	CH methylation levels in ES and differentiated cells .....	118
5.3.2	Genomic distribution of non-CG methylation .....	119
5.3.3	Dynamics of non-CG methylation throughout the stages of cell cycle.....	125
5.3.4	Dynamics of non-CG methylation throughout mammalian development .....	127

5.3.5	Discussion.....	131
<b>6</b>	<b>Establishment and maintenance of non-CG methylation.....</b>	<b>133</b>
6.1	Introduction.....	133
6.2	Aims.....	134
6.3	Results.....	135
6.3.1	Investigating a role for Dnmt2.....	135
6.3.2	mCH contribution of the canonical Dnmts.....	139
6.3.3	Active demethylation of non-CG context.....	145
6.4	Discussion.....	149
<b>7</b>	<b>Functional significance of non-CG methylation.....</b>	<b>155</b>
7.1	Introduction.....	155
7.2	Aims.....	155
7.3	Results.....	156
7.3.1	Correlation of mCH with histone marks and pluripotency factors.....	156
7.3.2	Identifying protein binders for non-CG context methylation.....	162
7.3.3	Effect of global mCH decrease on pluripotency in mESC.....	168
7.4	Discussion.....	173
<b>8</b>	<b>General discussion and future perspectives.....</b>	<b>179</b>
8.1	A role in transcriptional regulation.....	180
8.2	Global methylation buffer.....	182
8.3	Resistance to active demethylation.....	183
<b>9</b>	<b>References.....</b>	<b>187</b>
<b>10</b>	<b>Appendix.....</b>	<b>215</b>
10.1	Additional tables.....	215
10.2	Additional figures.....	232

## Table of figures

Figure 1. Two phases of global DNA methylation erasure during early mammalian development. ....	3
Figure 2. Percentage of CN methylation in the mouse. ....	4
Figure 3. Percentage of methylation in human ES cells (H1) and differentiated fibroblasts (IMR90). ....	5
Figure 4. Domain structure and similarities of the three classes of mouse Dnmts. ....	8
Figure 5. A schematic of the evolutionary conservation of Dnmt2-like proteins in protists, plants, fungi and animals. ....	14
Figure 6. A structural comparison of the three members of the Tet family of oxygenases. ....	20
Figure 7. Analysis of MeDIP strand asymmetry. ....	56
Figure 8. Bisulphite conversion and cloning of major satellite single repeat. ....	58
Figure 9. MeDIP-seq symmetric and asymmetric peak analysis. ....	59
Figure 10. Feature enrichment of asymmetric and symmetric peaks, normalized to unbound genomic fraction. ....	61
Figure 11. Methylation levels in CG, CHG and CHH contexts in low coverage BS-seq datasets and conversion artefacts. ....	63
Figure 12. C>T transitions in all four cytosine contexts. ....	65
Figure 13. Non-CG methylation in mouse repeat sequences. ....	67
Figure 14. Non-CG methylation in mouse LINE1 sequences. ....	69
Figure 15. Methylation in genomic features in J1 ES and pMEFs in CG (upper), CHG (middle) and CHH (lower) contexts. ....	71
Figure 16. Methylation in mouse replication origins for chromosome 11. ....	72
Figure 17. Bisulphite-induced degradation of DNA. ....	73
Figure 18. Biased degradation of C-rich DNA after bisulphite treatment. ....	75
Figure 19. Effect of DNA modifications on the extent of DNA degradation by bisulphite. ....	77
Figure 20. A schematic representation of the NNA method. ....	79
Figure 21. Results from the nearest neighbor analysis. ....	80
Figure 22. A schematic principle of the LUMA assay. ....	81
Figure 23. LUMA assay results. ....	82
Figure 24. CCWGG LUMA trial. ....	83
Figure 25. MspJI family of restriction enzymes specifically cutting 5mC-containing DNA. ....	92
Figure 26. In vitro digestion of C <sup>m</sup> CWGG-containing oligonucleotides with RlaI. ....	93
Figure 27. Characterisation of MspJI enzyme's restriction specificity. ....	95
Figure 28. Strategy for calling 5mC positives from the MspJI-meRRBS reads. ....	96

Figure 29. MspJI-RRBS positive methylation calls.....	97
Figure 30. The specificity of the currently available antibody and the difference with the context-specific antibodies. ....	98
Figure 31. Optimisation of the avidin-biotin ELISA assay: .....	101
Figure 32. Specificity of mCA (2C8) and mCG (10E2) supernatants in comparison to the commercially available 5mC antibody.....	102
Figure 33. Immunofluorescence of mouse zygotes.....	104
Figure 34. Concentration of the supernatant of 2C8 mCA and 10E2 mCG clones and subsequent Protein G resin purification.....	105
Figure 35. Optimisation of DNA adsorption.....	108
Figure 36. Validation of the ELISA direct DNA binding assay for the different DNA modifications.....	110
Figure 37. Mass spec validation of the quantitative potential of direct DNA binding ELISA assay. ....	112
Figure 38. Enhancing the sensitivity of DNA ELISA.....	114
Figure 39. DNA ELISA for global mCA and mCG levels in pluripotent and differentiating mammalian cells.....	119
Figure 40. Feature enrichment of CG (upper panel), CHG (middle panel) and CHH (lower panel) methylation in mES and pMEFs .....	120
Figure 41. Immunofluorescence imaging of mouse ES cells (upper panel) and embryonic fibroblasts (lower panel). ....	122
Figure 42. Immunofluorescence imaging of human ES cells (upper panel) and embryonic fibroblasts (lower panel). ....	124
Figure 43. Immunofluorescence imaging of mouse ES metaphase spreads. ....	125
Figure 44. Methylation dynamics throughout the mES cell cycle. ....	126
Figure 45. Global methylation in a panel of B6 mouse adult tissues.....	128
Figure 46. Global levels of methylation during mouse early development .....	129
Figure 47. Immunofluorescence of growing oocyte (GVO).....	130
Figure 48. Estimation of global 5mC and 5hmC levels in Dnmt2-KO mESCs.....	136
Figure 49. Expression of canonical Dnmts in the Dnmt2-KO ES cell lines. ....	137
Figure 50. Methylation in the Dnmt-TKO mES cell line and other Dnmt2-only species. ....	138
Figure 51. ELISA for global mCA and mCG levels in a panel of Dnmt KO mESCs and Dicer-KO. ....	140
Figure 52. Direct correlation between Dnmts' expression levels and global mCA methylation. ....	142
Figure 53. Direct correlation between Dnmts' expression levels and global mCG methylation. ....	

.....	143
Figure 54. Expression of Dnmt3a and Dnmt3b in the different phases of the cell cycle of WT mESCs. ....	144
Figure 55. Tet1 activity on 5mC in different context. ....	146
Figure 56. CA and CG methylation in AID-KO cells. ....	147
Figure 57. Methylation in the AID targeting motif WRC. ....	149
Figure 58. Correlation of mCA (A-D) and mCG (E-H) methylation with active histone modifications. ....	157
Figure 59. Correlation of mCA (A-D) and mCG (E-H) methylation with repressive histone modifications. ....	158
Figure 60. Immunofluorescence of WT J1 mES cells. ....	159
Figure 61. Correlation of mCA and mCG methylation with transcription. ....	160
Figure 62. Correlation of mCA (A-D) and mCG (E-H) methylation with pluripotency factor binding sites. ....	161
Figure 63. Correlation of mCA and mCG methylation with early development (Tet3) and differentiation factors (Tcf3). ....	162
Figure 64. Validation of a second bait for a protein binders experiment. ....	163
Figure 65. Protein pull-down result. ....	165
Figure 66. A schematic representation of biological function associations for the mCA binders and repellers. ....	167
Figure 67. <i>In vitro</i> binding assay of the recombinant MeCP2 protein and DNA fragments with varying methylation context. ....	167
Figure 68. Dnmt3a and Dnmt3b double knockout (DKO) inducible mES cell line – time course post KO-induction. ....	169
Figure 69. Global mCG (A) and mCA (B) levels in E14 WT cells' time course towards ground state reprogramming. ....	171
Figure 70. Expression of pluripotency factors and developmentally related genes. ....	172
Figure 71. A schematic representation of the conservation of strand information in paired end NGS datasets. ....	232
Figure 72. Log2 expression of Dnmts and Np95 in a panel of Dnmt-KO and Np95-KO mES cell lines used for this project. ....	233
Figure 73. Log2 expression of Dnmts and Np95 in a panel of tissues. ....	233
Figure 74. A comparison between ELISA and LC-MS measurements of absolute gained levels of 5hmC after Tet1 oxidation in three different methylation contexts. ....	234
Figure 75. A schematic representation of the SILAC-based technique workflow for identification of nuclear protein binders. ....	235



## Table of tables

Table 1. A summary of methods for analysing DNA methylation.....	25
Table 2. qPCR programmes outline. ....	45
Table 3. Restriction enzymes sensitive to 5mC in non-CG context (in red) and their non-sensitive isoschizomers, where available (in black).....	92
Table 4. Mapping efficiency of the MspjI meRRBS datasets .....	95
Table 5. Oligonucleotide sequences used for the validation of mCA and mCG sera. ....	99
Table 6. Antibodies tested for their feasibility for the ELISA assay and the results of their performance .....	110
Table 7. Mass comparison of antigen between genomic DNA and the DNA fragment controls for ELISA.....	111
Table 8. Comparison of global differences of mCA and mCG in mESCs revealed by IF.....	132
Table 9. Histone modifications analysed and their corresponding genomic signatures.....	157
Table 10. Summary of main characteristics of non-CG methylation according to cell type. ....	180
Table 11. Base composition of the mouse genome and contribution of each cytosine context ..	215
Table 12. Table of used oligonucleotides.....	216
Table 13. M13-derived PCR fragments.....	217
Table 14. Table of used antibodies.....	218
Table 15. Illumina adapters and primers .....	218
Table 16. Table of consensus repeat sequences .....	219
Table 17. Unspecific cutters (target CN).....	220
Table 18. Specific cutters (target a specific dinucleotide – CA, CT or CG).....	220
Table 19. Full screen of mCA monoclonal supernatants.....	221
Table 20. Full screen of mCG monoclonal supernatants.....	224
Table 21. Datasets used for the calculation of global mCG and mCH levels in mouse early development. ....	226
Table 22. Datasets used for the ChIP analysis of histone marks and protein binders .....	227
Table 23. Statistically significant and marginally significant mCA protein readers and their major functions .....	228
Table 24. Statistically significant and marginally significant mCA protein repellors and their major functions.....	229
Table 25. GO and functional terms for nuclear pull-down identified mCA binders.....	230
Table 26. GO and functional terms for nuclear pull-down identified mCA repellors.....	231



## ***1 Introduction***

### **1.1 Epigenetic DNA modifications and their role in development**

Embryonic stem cells (ES cells) have the same genome as lineage-committed cells, yet possess the unique properties of self-renewal and pluripotency. It is a fundamental question how the identical genome sequence in a multicellular organism gives rise to the huge diversity of cell types, each possessing different gene expression profiles and cellular functions. It is now widely accepted that epigenetic mechanisms play a critical role in the process of cellular differentiation and the maintenance of a differentiated state (Hawkins et al. 2010; Hemberger et al. 2009).

Epigenetic modifications include chemical modifications of DNA and histones – the main components of chromatin, and constitute an additional layer of information which influences the activity of the underlying genetic information (Law & Jacobsen 2010). The addition of a methyl group to a cytosine base (resulting in 5-methyl-cytosine) is an epigenetic mark present in the DNA of all vertebrates and flowering plants, many fungi, invertebrates and protists, as well as many bacterial species (Goll & Bestor 2005). Another modification – 5-hydroxymethyl-cytosine has been known for nearly 40 years in animal DNA but it has been just recently ‘rediscovered’ and is still poorly characterised (Penn et al. 1972; Kriaucionis & Heintz 2010; Tahiliani et al. 2010).

#### ***1.1.1 Occurrence and distribution of DNA modifications***

In mammals, it has been long believed that DNA methylation occurs almost exclusively in symmetrical CG context and it is estimated that ~70-80% of all CG dinucleotides throughout the genome are methylated (Ehrlich et al. 1982). In plants DNA methylation occurs in three sequence contexts – the symmetric CG and CHG, and the asymmetric CHH context (where H is T, A or C), although regions enriched in CHH methylation have pretty similar methylation densities on both strands and show propensity for symmetrical methylation (Cokus et al. 2008). Genome-wide, DNA methylation in plants occurs predominantly in transposons and other repeat sequences, with marked enrichment in pericentromeric regions, although some is also found in the gene bodies of highly expressed genes (Cokus et al. 2008; Feng et al. 2010; Zhang et al. 2006; Lister et al. 2008a). CG islands in plants are mostly methylated, unlike CG islands in mammals which are to

a large extent unmethylated (Zhang et al. 2006; Meissner et al. 2008; Strausman et al. 2009; Laurent et al. 2010; Feng et al. 2010). In mammalian genomes, methylation density is more uniform throughout the length of the chromosomes and the different genomic regions and slightly higher only in sub-telomeric regions. Moreover, methylation in CHG and CHH contexts has recently also been described for mammalian ES cells, with the distinction from plants that they are both highly asymmetrical (Lister et al. 2009; Laurent et al. 2010).

### ***1.1.2 Biological significance and role of DNA methylation***

Thirty-six years ago it was first proposed that cytosine DNA methylation in eukaryotes could act as an inherited epigenetic mark that activates or represses genes during cellular differentiation (Holliday & Pugh 1975; Riggs 1975). Since then, key roles of DNA methylation in mammals have been described in cellular processes such as regulation of gene activity, silencing of transposable elements, X-inactivation, genomic imprinting and also different diseases, including cancer. Altogether, DNA methylation has been shown to have a crucial role in the embryonic development and cellular reprogramming in mammals (Strausman et al. 2009; Reik 2007; Okano et al. 1999; Li et al. 1992; Feng et al. 2011).

Very early in development, briefly after fertilisation, a genome-wide erasure of DNA methylation takes place, and only gametic imprints and a limited number of other sequences are left methylated (Figure 1) (Reik et al. 2001; Rideout III et al. 2011). This event of global methylation erasure is necessary for the acquisition of totipotency and then pluripotency in the very first stages of embryonic development (Morgan et al. 2005). Around the time of implantation, a round of *de novo* methylation takes place, which determines the methylation patterns of the first two embryonic lineages. At this point the epigenetic profile of the inner cell mass (ICM) cells is established from which ES cells are derived (Hemberger et al. 2009; Reik et al. 2001). The ES cell state has currently been shown to have the highest level of DNA methylation throughout development which drops gradually in the course of differentiation (Lister et al. 2009; Laurent et al. 2010). Since most of the differentially methylated regions (DMRs), which are by definition rich in CG-content, show an increase in methylation with differentiation (Laurent et al. 2010), the decrease in methylation occurring with the loss of pluripotency is mainly due to the loss of non-CG methylation (Lister et al. 2009).

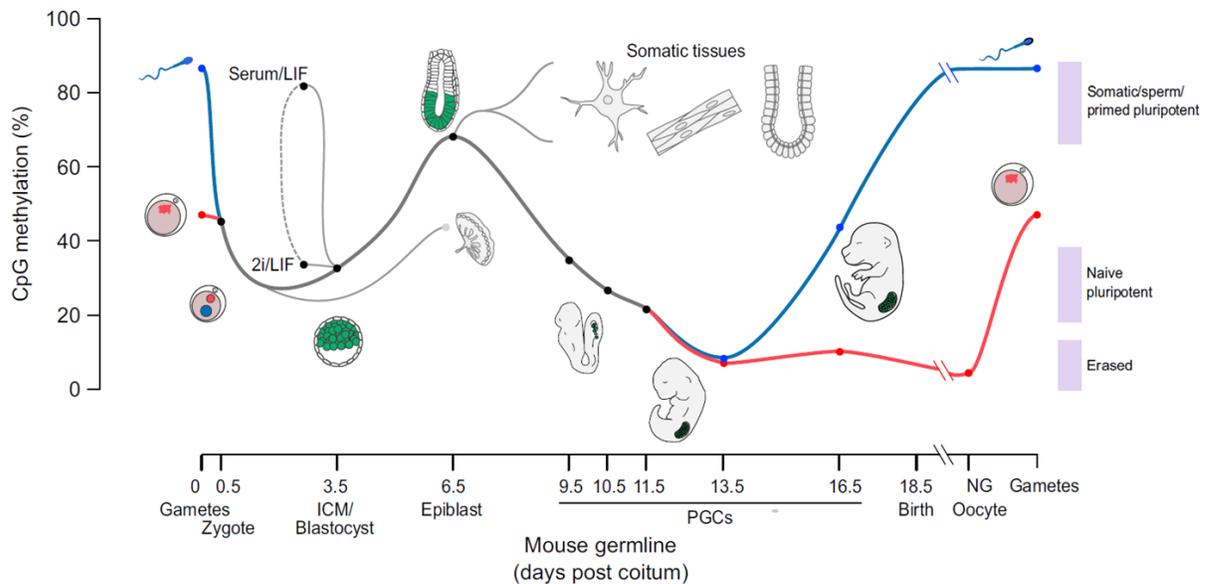


Figure 1. Two phases of global DNA methylation erasure during early mammalian development. Adapted from (Lee et al. 2014).

## 1.2 Occurrence of non-CG DNA modifications in different organisms and cell types

Methylation of cytosine residues outside the CG dinucleotide in mammals has been reported as far back as the 1970s and 1980s (Salomon & Kaye 1970; Harbers et al. 1975; Sneider 1980; Grafstrom et al. 1985). The first quantitative investigation of the ratio of all four methylated CN dinucleotides, however, was only published in 1987, in DNA from human spleen (Woodcock et al. 1987). The reported methylation percentage per context was much higher for CG dinucleotides (37.9% mCG/all CG), with merely 2% each for CA and CT, and 1.1% for CC context. However, the absolute numbers of methylated cytosines were 45.5% for mCG, and the remaining 54.5% of methylated cytosine was in CH context (21% CA, 22% CT and 11% CC, respectively), meaning it was actually higher in overall than mCG. These values, however, have not been subsequently confirmed. Since then, non-CG methylation in human cells has been reported in different instances, either in exogenous DNA integrants, or in different endogenous regions of the genome such as repeat sequences, origins of replication or gene coding sequences (Toth et al. 1990; Clark et al. 1995; Woodcock et al. 1997; Tasheva & Roufa 1994a; Tasheva & Roufa 1994b; Franchina & Kay 2000; Malone et al. 2001).

Further in 2000, the first investigation on ES cells established that non-CG methylation accounts for a quarter of the total methylation in the mouse ES cell genome, decreasing to less than 6% in differentiated tissues, while the overall amount of CG methylation increases (Figure 2) (Ramsahoye et al. 2000).

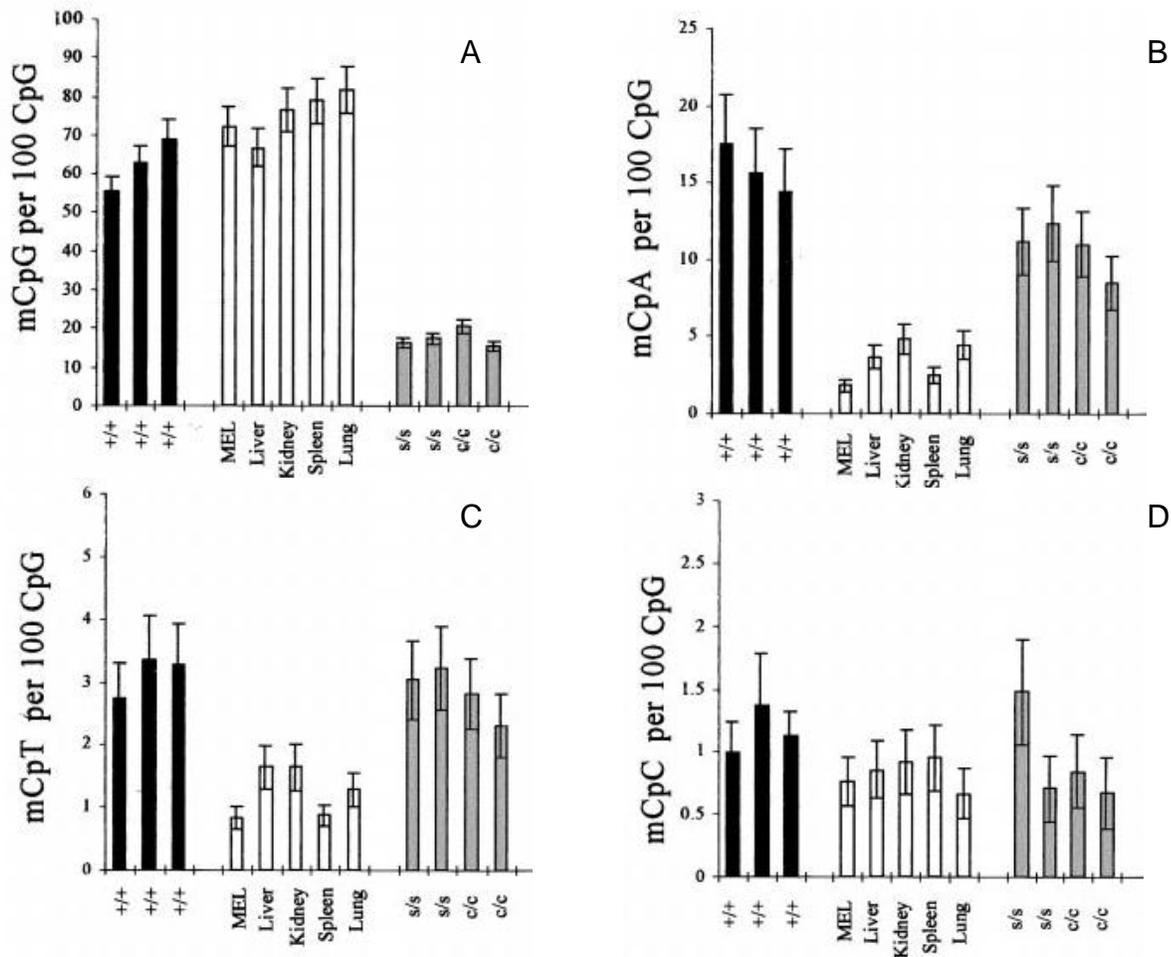


Figure 2. Percentage of CN methylation in the mouse. ES cells (black bars), tissues (white bars) and Dnmt1 null ES cell lines (grey bars). **A.** mCG; **B.** mCA; **C.** mCT, and **D.** mCC. Since all values are normalised for CG methylation, the amount of non-CG methylation can be estimated roughly as a quarter of the total. Adapted from (Ramsahoye et al. 2000).

These results have been confirmed by the reports of Lister et al. and Laurent et al. ten years later, who describe similar figures and arrive at similar conclusions for non-CG methylation in human ES cells (Lister et al. 2009; Laurent et al. 2010). With high resolution bisulphite sequencing and coverage of 94% of all cytosines in the genome, Lister et al. demonstrated that a

quarter (24.5%) of all methylated cytosines in ES cells are in a non-CG context. This amount is reduced to 0.02 % in foetal lung fibroblasts (Figure 3) and also decreases upon differentiation of the same ES cell line (Lister et al. 2009; Laurent et al. 2010).

In addition, analysis of the same primary foetal lung fibroblasts (IMR90) induced to pluripotency (iPS cells) reveals restored non-CG methylation at the same loci as in control ES cells, suggesting that the presence of asymmetric non-CG methylation is characteristic of an embryonic stem-cell state. While CG methylation shows high consistency of methylated loci throughout the cell population, and most of the CG loci are 80-100% methylated, CHG and CHH loci are methylated only in 10-40% of the ES cell population (Lister et al. 2009).

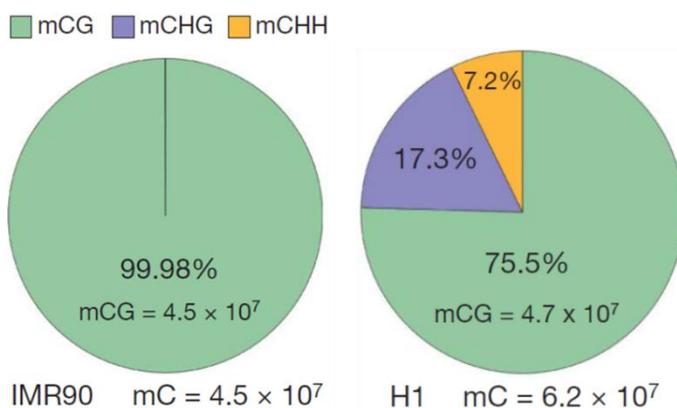


Figure 3. Percentage of methylation in human ES cells (H1) and differentiated fibroblasts (IMR90). The overall amount of non-CG methylation in ES cells (purple and orange) is 24.5% and it disappears in fibroblasts while CG methylation stays roughly the same as shown by the total number of analysed cytosines ( $4.7 \times 10^7$  in hESCs and  $4.5 \times 10^7$  in IMR90). Adapted from (Lister et al. 2009).

The more recently described DNA modification in mammals, hydroxymethylation of cytosine (5hmC), also shows a significant fraction in an asymmetric and non-CG context in mouse ES cells (Ficz et al. 2011).

Non-CG methylation has also been observed in *Drosophila melanogaster*, where it is the predominant type of DNA modification, with 0.7% methylation for CT dinucleotides, 0.3% for CA, 0.2% for CC and only 0.1% for CG dinucleotides. DNA methylation in *Drosophila* is characteristic at low levels only for the very early stages of embryonic development (1-3 hour larvae), and quickly decreases to barely detectable levels with the progression of development (Lyko, Whittaker, et al. 2000; Gowher et al. 2000; Kunert et al. 2003; Phalke et al. 2009). Only traces of non-CG methylation, however, are found present in the honeybee and silkworm

methylomes (Lyko et al. 2010; Xiang et al. 2010).

Asymmetric non-CG methylation was also described in the parasitic protozoan *Entamoeba histolytica* (Fisher et al. 2004). This is also the main type of methylation found in protists from the genus *Dictyostelium*. Although their overall levels of DNA methylation are again very low, with an estimated value of around 0.2% for the whole genome, methylation has been observed exclusively in CA, CT and CC dinucleotides (Kuhlmann et al. 2005).

Non-CG methylation occurs commonly in plant genomes (Law & Jacobsen 2010) even amongst protist algae (Feng et al. 2010). In *Arabidopsis thaliana*, genome-wide DNA methylation mapping reveals levels of around 23% methylation for CG, 6% for CHG and 1.6% for CHH contexts (Cokus et al. 2008; Feng et al. 2010) or an overall of 45% non-CG methylation (Lister et al. 2008a) and even higher methylation levels were described for poplar and rice (Feng et al. 2010). In plants, non-CG methylation is most abundant in repeat sequences and less enriched in gene bodies, while gene promoters are normally not methylated in plants (Henderson & Jacobsen 2008). Unlike CG context, which is either unmethylated or 80-100% methylated in the cell population, the level of methylation between cells at individual CHH loci is lower, up to 50%, while for CHG loci the methylation levels show high variability – between 20-100%. This is similar to the observations in mammalian ES cells and may indicate either that the non-CG methylation pattern is characteristic only for a particular cell type from the floral tissue (or ES cell population respectively) or that this type of methylation is more variable between the same type of cells (Cokus et al. 2008; Lister et al. 2008a; Lister et al. 2009).

Non-CG context methylation has also been reported in mouse oocytes and remained detectable on the maternal pronucleus after fertilization, although it did not persist further than the 2-cell stage (Imamura et al. 2005; Haines et al. 2001). It was not found after prolonged oocyte culturing, in the blastocyst or cultured somatic cells. In both reports, its presence has been related to the amount and dynamics of CG methylation. In addition, non-CG methylation has also been reported in male spermatogonia (Grandjean et al. 2007), and a number of other studies have recently reported varying but significant levels in mammalian tissues (Lee, Jin, et al. 2010; Fuso et al. 2010; Yan et al. 2011). It therefore remains an open question whether mCH is a mark characteristic for stem cells and related to pluripotency, or it is common for a wider variety of tissues.

### 1.3 Establishment and maintenance of non-CG methylation

Methylation on DNA is mediated by a conserved group of enzymes called DNA (cytosine-5) methyltransferases (DNA MTs, or Dnmts for mammals). Such enzymes exist in bacteria, as well as higher organisms like plants and vertebrates (Goll & Bestor 2005; Jurkowski & Jeltsch 2011). Bacterial DNA MTs have very selective sequence specificity, defined by a region termed ‘target recognition domain’. Such a domain does not exist in eukaryotic MTs, and their target selection is not a function of innate sequence specificity (Goll & Bestor 2005). Less is known about the sequence targeting and specificity for *de novo* methylation in eukaryotic cells and in mammalian cells in particular, for both the symmetric and asymmetric sites. In plants, cytosine methylation in all sequence contexts is controlled by the function of several DNA-methyltransferases – MET1, DRM1 and DRM2, CMT3 and the more recently characterised CMT2 (Zemach et al. 2013; Goll & Bestor 2005). MET1 is the equivalent of the mammalian Dnmt1 maintenance methyltransferase, which copies the symmetric CG methylation upon cell division. DRM1 and DRM2 are structurally related to the mammalian Dnmt3 family and establish *de novo* CG, CHG and CHH methylation at each round of DNA synthesis (Grandjean et al. 2007). The chromomethylase family of enzymes does not have an equivalent in mammals and they methylate CH-context only: CMT3 is responsible for the maintenance solely of the symmetric CHG methylation, while CMT2 methylates predominantly CHH context (Goll & Bestor 2005; Stroud et al. 2014). Thus, CHG context methylation in plants is established by the activities of CMT3, DRM2 and CMT2, while CHH methylation is in the competency of CMT2 and DRM2 (Stroud et al. 2014). In mammals, experimental results regarding the enzyme specificity and activity for *de novo* methylation in non-CG sites have been controversial, maybe because of the high redundancy in function of the mammalian DNA-methyltransferases (Okano & Li 2002). Some authors even discuss the existence of an as yet unidentified *de novo* methyltransferase activity (Lorincz et al. 2002).

In 2009 Cedar and colleagues suggested two principally distinct mechanisms for *de novo* DNA methylation of CG context sequences (Straussman et al. 2009). The first stage of global genomic methylation during early development they define as ‘untargeted’ methylation, which happens ‘by default’ and requires active protection of CG islands and other sequences meant to stay unmethylated. The second stage is a follow-on methylation of specific genes and individual sequences and leads to cellular differentiation; it is a ‘targeted’ event, facilitated by chromatin

marks and possibly other factors, like non-coding RNAs (ncRNAs) (Straussman et al. 2009). If this model would be true, then the *de novo* non-CG methylation should possibly be the type of ‘targeted’ event, since a very small number of non-CG context cytosines is methylated, and at very specific loci (Lister et al. 2009). Recently, a stochastic model for DNA methylation has been proposed, whereby the methylation of each cytosine position is dynamic and determined by the local rates of methylation and demethylation. This model favours the targeting of *de novo* and maintenance of non-CG methylation by processes not involving a second DNA strand (Jeltsch & Jurkowska 2014).

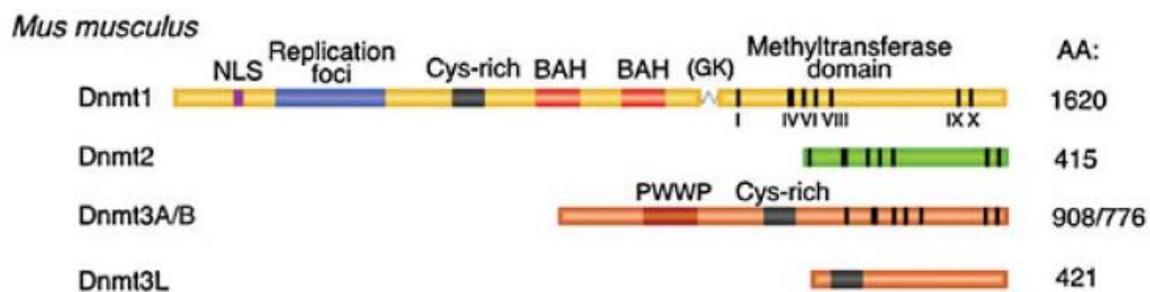


Figure 4. Domain structure and similarities of the three classes of mouse Dnmts. Adapted from (Goll & Bestor 2005). Details about the domains can be found in the main text.

Three DNA methyltransferases are operational in mammals, namely Dnmt1, Dnmt2 and Dnmt3. A non-catalytically functional member of this family is Dnmt3L (Figure 4).

### 1.3.1 *Dnmt1 – the maintenance methyltransferase*

Dnmt1 was the first biochemically characterized mammalian DNA methyltransferase and it is known as the maintenance methyltransferase (R. Z. Jurkowska, Jurkowski, et al. 2011; Goll & Bestor 2005; Cheng & Blumenthal 2008). It is a large protein, built of 1620 amino acids in mouse and 1616 amino acids in human. Its N-terminal part has multiple domains with diverse functions, such as protein-binding (to Proliferating cell nuclear antigen PCNA, DNA methyltransferase associated protein 1 Dmap1), stability, nuclear localization, centromere targeting, DNA binding domains (CXXC, or cysteine-rich, and KG linker), including some domains with yet unclear function in Dnmt1 (BAH1 and BAH2). The C-terminal domain is the catalytic domain, a feature shared with all Dnmts, although it is not catalytically active in an isolated form (R. Z. Jurkowska, Jurkowski, et al. 2011; Goll & Bestor 2005).

Dnmt1 is ubiquitously expressed throughout mammalian development, and is the major methyltransferase in somatic tissues (R. Z. Jurkowska, Jurkowski, et al. 2011; Goll & Bestor 2005). Spatially and functionally, Dnmt1 is closely associated with the replication machinery and has a strong processive activity on hemimethylated DNA where it transfers methyl groups to the newly synthesised non-methylated strand (Vilkaitis et al. 2005; Okano & Li 2002; Leonhardt & Page 1992). A mechanism has been proposed that the enzyme copies the methylation patterns from the parental methylated strand to the daughter strand and thereby maintains the mitotic inheritance of DNA methylation (Riggs 1975; Holliday & Pugh 1975). Dnmt1 has an extraordinarily strong affinity towards hemi-methylated CG sites (10-40 fold higher than for unmethylated CGs) (Song et al. 2011; Song et al. 2012), which greatly facilitates its role in maintaining pre-existing symmetric CG modifications. However, this specificity would not maintain asymmetric non-CG context methylation, as indeed has been suggested in some reports (Ramsahoye et al. 2000; Dodge et al. 2002). On the other hand, it has also been shown, that Dnmt1 has a self-inhibiting property, which stops it from *de novo* methylating entirely unmethylated CG sites, thus giving preference to the hemi-methylated CGs. This would mean that, under specific circumstances like chemical modifications or allosteric protein interactions, this conformationally-restricted property abrogated and *de novo* activity becomes once again possible (R. Z. Jurkowska, Jurkowski, et al. 2011). Indeed, there have been observations that Dnmt1-null mice and ES cells are deprived of non-CG methylation (Woodcock et al. 1998; Grandjean et al. 2007) or have decreased level of *de novo* methylation in general (Lorincz et al. 2002). This could also be an effect of the cooperative activity of Dnmt1 with the *de novo* Dnmts (Okano & Li 2002; Lorincz et al. 2002; Fatemi et al. 2002; Kim et al. 2002; Liang et al. 2002) which has also been proposed for plants (Singh et al. 2008). A recent report has confirmed a previously detected *in vitro* activity of mouse Dnmt1 towards CT dinucleotides (Ross et al. 2010; Suetake et al. 2003). A non-processive *de novo* methylation activity for Dnmt1 has also been identified in a number of *in vitro* studies (Fatemi et al. 2001; Vilkaitis et al. 2005; Goyal et al. 2006), including in non-CG (CHG) context (Lee, Jang, et al. 2010). Moreover, overexpression of Dnmt1, leads to the methylation of previously unmethylated sequences (Takagi et al. 1995; Vertino et al. 1996; Biniszkiwicz et al. 2002). Indeed a firm role in setting up *de novo* methylation marks in cooperation with the Dnmt3s has recently been fully recognized for Dnmt1 (Jeltsch & Jurkowska 2014). However, to what extent it possesses a Dnmt3-independent *de novo* activity *in vivo*, remains an open question.

Functionally, Dnmt1 is indispensable for mammalian development. Dnmt1-KO mice have variably decreased methylation levels, and embryonic lethality shortly at E10 (Li et al. 1992). Mutations in Dnmt1 cause loss of imprints (Li et al. 1993) and misregulation of X chromosome inactivation (Beard et al. 1995; Sado et al. 2000), apoptosis in affected cells including PGCs (Takashima et al. 2009; Jackson-Grusby et al. 2001) and general genomic instability leading to chromosomal aberrations (Chen et al. 1998). Interestingly, embryonic stem cells lacking Dnmt1 are viable and show no growth or morphological abnormalities, despite their strongly reduced (by 60%) level of DNA methylation (Li et al. 1992). These cells, however, die when induced to differentiate and have high rates of mitotic recombination (Chen et al. 1998). Dnmt1 is also implicated in multiple types of cancers, where its loss, on one hand, contributes to the genomic instability and higher mutation rates (Gaudet et al. 2003; Eden et al. 2003), while on the other its overexpression is crucial to maintain the high division rate and survival characteristic of cancer cells (Gravina et al. 2013; Chen et al. 2007; Robertson et al. 1999).

Dnmt1 is known to have a wide and varying number of interaction partnerships. In addition to the previously listed, relevant partners are the *de novo* Dnmts 3a and 3b, the methylated DNA binder MeCP2 and its obligatory maintenance methylation partner Np95 (Uhr1) (R. Z. Jurkowska, Jurkowski, et al. 2011).

Dnmt1 has several isoforms – the main ones being the ubiquitous somatic isoform (Dnmt1s) and the oocyte isoform (Dnmt1o), which is shorter at the N-terminus and is much more stable and long lived, accumulated in very high levels in the oocyte (Goll & Bestor 2005; R. Z. Jurkowska, Jurkowski, et al. 2011).

### ***1.3.2 Dnmt3 – the de novo methyltransferase family***

The two catalytically active members of the Dnmt3 family are two closely related MTs which share very high homology – Dnmt3a and Dnmt3b (Masaki Okano et al. 1998). They are considered the mammalian *de novo* Dnmts (Aoki et al. 2001), which establish the DNA methylation patterns during early development (Reik 2007). They are smaller proteins than Dnmt1, again with an N-terminal regulatory domain and a C-terminal catalytic domain. Although the catalytic domain is highly conserved, their N-terminal domains are unrelated, which speaks of potentially non-redundant functions. All of the Dnmt3 members possess an ADD domain (ATRX-DNMT3A-DNMT3L), and Dnmt3a and 3b also have a PWWP domain, rich in prolines

(P) and tryptophanes (W). The ADD domain is the main protein interaction platform, binding with transcriptional factors and multiple classes of chromatin modifying activities and binding proteins (Goll & Bestor 2005; R. Z. Jurkowska, Jurkowski, et al. 2011; Freitag & Selker 2005). In addition, the ADD domain prevents binding to an H3 tail with di- or trimethylated in lysine 4 (K4me<sub>2/3</sub>), and facilitates the binding of unmethylated H3K4 tails (Ooi et al. 2007; Otani et al. 2009; Zhang et al. 2010). H3K4me<sub>2</sub> and me<sub>3</sub> are marks of actively transcribed chromatin, particularly associated with promoters and transcription start sites (TSS), and the expression of tissue specific genes (Lauberth et al. 2013; Pekowska et al. 2010; Barski et al. 2007). The PWWP domain is important for the localisation of the Dnmt3s to heterochromatin by binding to the H3K36me<sub>3</sub> mark, and thus facilitating methylation of nucleosomal DNA (Dhayalan et al. 2010). The Dnmt3s also interact with Np95 (Uhrf1) which connects them to the H3K9me<sub>3</sub> and heterochromatin (Arita et al. 2012).

Unlike Dnmt1, both Dnmt3s have a strong affinity towards unmethylated DNA (Gowher & Jeltsch 2001). *In vitro* as well as *in vivo* both enzymes show a strong affinity towards CG dinucleotides (Masaki Okano et al. 1998), although substantial activity has been reported towards non-CG cytosine residues as well (Dodge et al. 2002; Suetake et al. 2003). While Dnmt3b shows a higher non-CG activity *in vitro* (Gowher & Jeltsch 2001; Aoki et al. 2001), *in vivo* results implicate Dnmt3a as more important for establishing the patterns of non-CG methylation (Lee, Jang, et al. 2010; Mund et al. 2004; Lyko et al. 1999; White et al. 2002). It has been reported that unlike Dnmt1, the Dnmt3s exhibit strong preferences for the flanking sequence around the CG site, which limits their methylation activity (Handa & Jeltsch 2005; R. Z. Jurkowska, Siddique, et al. 2011; Lin et al. 2002). Both Dnmt3s oligomerise in either hetero- or homodimers or together with Dnmt3L (R. Jurkowska et al. 2011). It is interesting to note, that their overall *in vitro* activity is not very strong, and is weaker than the *de novo* activity of Dnmt1. Nevertheless, they are the main *de novo* methyltransferases, and their activity cannot be compensated for by Dnmt1.

Both Dnmt3a and Dnmt3b have isoforms, some of which are active, and others are shorter inactive forms (Aoki et al. 2001; R. Z. Jurkowska, Jurkowski, et al. 2011; Goll & Bestor 2005). A specific shorter isoform of Dnmt3a, called Dnmt3a2 is functional in germ cells and early embryogenesis, at the times of the two major phases of *de novo* establishment of methylation. It is expressed from an intronic promoter and lacks a part of its N-terminus, but is as active as Dnmt3a, which is the ubiquitous form in low levels in somatic tissues. Dnmt3a2 is referred to as

the ‘euchromatic’ Dnmt3a due to its cellular localization (Chen et al. 2002). Dnmt3b also has a number of isoforms, potentially regulated in a developmental fashion. Interestingly, all the isoforms expressed in ES cells (Dnmt3b1, -3b6, -3b7 and -3b8) contain conserved catalytic domains, which differ from the isoforms expressed in somatic tissues (Dnmt3b2, -3b3, -3b4 and -3b5). Dnmt3b1 and 3b2 are the main active forms, although 3b3 and 3b6 have recently been shown to have an enzymatic activity as well. Notably, none of the Dnmt3s is expressed very highly in somatic tissues, although their expression is detectable in most, but their peaks in expression are found in early development (R. Z. Jurkowska, Jurkowski, et al. 2011).

The non-catalytic member of the family is the Dnmt3L (Dnmt3-like), which lacks the PWWP domain, and key parts of the catalytic domain (Aapola et al. 2000; Hata et al. 2002; Che´din et al. 2002). It is expressed only in the germ line and early development (Lucifero et al. 2004; Salle et al. 2004) and binds to both Dnmt3a and Dnmt3b, but not directly to the DNA (Suetake et al. 2004). The CG methylation activity of Dnmt3a is 20-fold higher in partnership with Dnmt3L (Kareta et al. 2006). It has proven essential for the *de novo* methylation in germ cells (Hata et al. 2002; Bourc’his et al. 2001; Webster et al. 2005; Bourc’his & Bestor 2004; Hata et al. 2006) and it has been proposed, that it recruits the Dnmt3a2 isoform to specific loci unmethylated at H3K4 via the ADD domain (Ooi et al. 2007; Otani et al. 2009; Jia et al. 2007). It is not known whether Dnmt3L also recruits the *de novo* Dnmt3s at non-CG sites, but both early male PGCs (E16.5) and maturing oocytes (GVOs) have elevated levels of non-CG methylation (Seisenberger et al. 2012; Shirane et al. 2013). In ESCs, where there is no need to establish imprinted regions, Dnmt3L releases the catalytic Dnmt3s from heterochromatin and targets them to gene bodies of housekeeping genes. It also reduces the extent of DNA methylation on bivalent promoters, by competing for the formation of functional hetero-duplexes with the catalytic Dnmt3s (Neri et al. 2013).

Much like Dnmt1, both Dnmt3a and 3b are critical for the early phases of mammalian development. Dnmt3b knockout embryos die around E9.5, despite that their global methylation levels are higher than in the Dnmt1 knockout, while Dnmt3a knockout embryos develop to term and die shortly after birth (Okano et al. 1999). In contrast, Dnmt3L knockout mice are perfectly viable and without any discernable morphological abnormalities (Hata et al. 2002). Male mice, however, are sterile, due to impaired spermatogenesis (Bourc’his & Bestor 2004; Webster et al. 2005), and female mice cannot deliver viable pups as they die before midgestation (Bourc’his et al. 2001). Dnmt3L has an essential role in the establishment of maternal imprints, and in

silencing of repetitive elements during spermatogenesis. In these processes, it partners with Dnmt3a which delivers the catalytic activity, and during this developmental timeframe Dnmt3b is dispensable. On the contrary, later in development Dnmt3b seems to cooperate with Dnmt1 for the maintenance of methylation, and conditional knockout MEFs for Dnmt3b and not 3a, have decreased global levels of methylation, accompanied by genomic instability and chromosomal aberrations (Dodge et al. 2005). Mouse ESCs mutant for 3a or 3b show mild global demethylation, showing that both Dnmt3s play a role in the maintenance during that stage (Chen et al. 2003; Okano et al. 1999).

### ***1.3.3 Dnmt2 – the enigmatic Dnmt***

The function and biological roles of Dnmt2 have been less well characterized and are both controversial and enigmatic in comparison with the rest of the mammalian Dnmts. Dnmt2 is the most conserved and widely distributed member of the Dnmt family, with orthologues found in dozens of organisms ranging within all evolutionary kingdoms (Figure 5) (Goll & Bestor 2005; Ponger & Li 2005; Schaefer & Lyko 2010b). Moreover, in most protists, fungi, nematodes and insects, Dnmt2 is the only Dnmt family protein encoded in their genomes and, in some of them, the only DNA methyltransferase at all. Dnmt2 is also the only mammalian DNA methyltransferase with orthologues found in bacteria – the genus *Geobacter* (Goll & Bestor 2005; Ponger & Li 2005).

When Dnmt2 was first described in 1998 by Okano and colleagues, it did not show any CG *de novo* methylation activity *in vitro*. Dnmt2 null ES cells displayed normal *de novo* CG-methylation activity towards endogenous and exogenous virus DNA and thus it was concluded that Dnmt2 was a non-essential DNA-methyltransferase (M Okano et al. 1998).

Further research has shown that a Dnmt2 orthologue is responsible for the methylation of isolated cytosine residues in the early stages of embryo development in *Drosophila melanogaster*. The expression of the enzyme is developmentally regulated and strictly corresponds to the time frame of methylation occurrence in the fly genome (Kunert et al. 2003; Lyko 2001; Lyko, Whittaker, et al. 2000). Moreover, the preferred dinucleotide targets are CA and CT (0.4% methylation on average), without a consensus sequence specificity, while less preference has been shown towards CC and CG (0.1% on average) (Lyko, Ramsahoye, et al. 2000; Kunert et al. 2003). Given the fact that non-CG methylation is present in the early stages of

mammalian embryogenesis (Ramsahoye et al. 2000), when Dnmt2 is also expressed, this raises the possibility that Dnmt2 may be a non-CG methyltransferase – a CT/A-specific methyltransferase, and the earlier attempts to demonstrate its activity on DNA (M Okano et al. 1998) may have been misled by their focus on CG (Kunert et al. 2003; Lyko 2001).

It has been subsequently shown that both human and mouse Dnmt2 orthologues form covalent complexes with DNA, suggesting an active catalytic function (Liu et al. 2003; Dong et al. 2001). Further research has proved *in vitro* the DNA methylation activity of the human Dnmt2, albeit very low, and again demonstrated its high affinity towards non-CG dinucleotides in comparison to the other *de novo* Dnmts (Hermann et al. 2004). Overexpression of mouse Dnmt2 in *Drosophila* shows a weak but significant DNA methylation activity and a strict specificity towards non-CG (Mund et al. 2004).

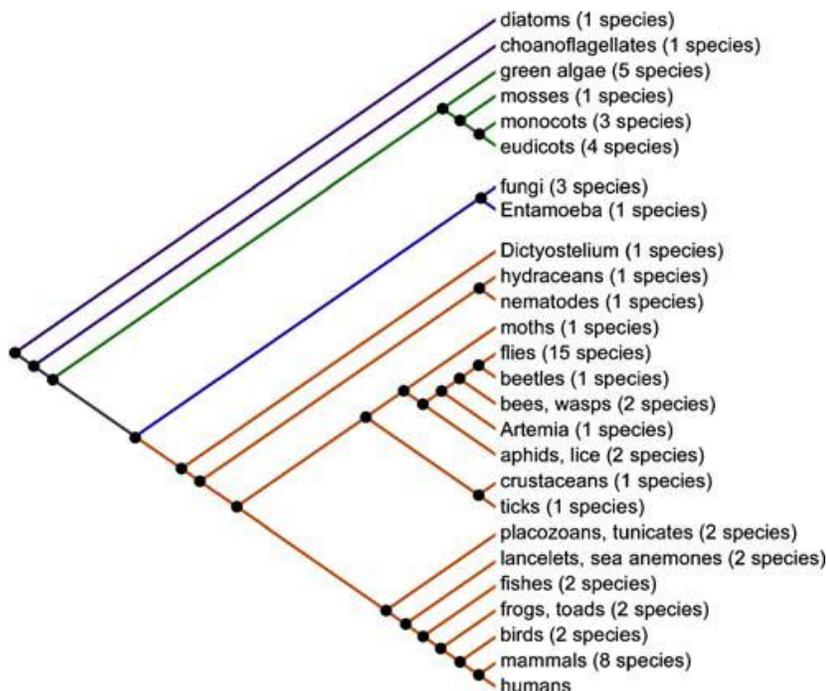


Figure 5. A schematic of the evolutionary conservation of Dnmt2-like proteins in protists, plants, fungi and animals. Phylogenetically related orthologues have been retrieved for 65 species and total 108 protein sequences. Adapted from (Schaefer & Lyko 2010b).

A Dnmt2 orthologue was later shown to be responsible for the asymmetric methylation observed in the genome of *Dictyostelium*. In this group of protists, the Dnmt2 orthologue DnmA is the only DNA methyltransferase encoded in their genome and has a high similarity to its mammalian counterpart (Kato et al. 2006). No obvious sequence context for the enzyme specificity has been identified but all the methylated cytosines are within non-CG dinucleotides (Kuhlmann et al. 2005).

After continued lack of evidence for an *in vivo* DNA methylation activity of Dnmt2 in mammals, it was clearly and reproducibly demonstrated that Dnmt2 enzymes originating from human, mouse, *Drosophila* and *Arabidopsis* can successfully and very specifically methylate an aspartic acid transport RNA – tRNA<sup>Asp</sup> (Goll et al. 2006; Jurkowski et al. 2008; Hengesbach et al. 2008; Schaefer et al. 2009). It appears that the anticodon loops of different Dnmt2-encoding species have conserved sequences, while the tRNA<sup>Asp</sup> in *Caenorhabditis elegans* and *Saccharomyces cerevisiae*, who do not have Dnmt2 homologues, have diverged in their anticodon sequence (Goll et al. 2006). The *in vitro* experiments with purified Dnmt2 enzyme, however, do not exclude the possibility that *in vivo* other factors may facilitate an activity towards DNA, since it has already been shown that Dnmt2 can bind and form stable covalent bonds with DNA (Dong et al. 2001). Moreover, it has been demonstrated that Dnmt2 utilizes a DNA methyltransferase mechanism to methylate the tRNA<sup>Asp</sup> and is not a classical RNA-methyltransferase (Jurkowski et al. 2008). Recently it was also shown by comparative structural analysis that DNA was the original substrate of the ancestral Dnmt2 enzyme, and the adaptation to an RNA substrate occurred subsequently (Jurkowski & Jeltsch 2011). These facts leave the question open for a possibility that *in vivo* Dnmt2 has some remaining low activity on DNA, or that this activity could be stimulated under particular circumstances.

#### ***1.3.4 Other protein and chromatin factors in DNA methylation***

Studies in *Neurospora*, *Arabidopsis* and *Drosophila* have shown that mutations in homologues of the Su(var)3-9 histone-3-lysine-9 methyltransferase (H3K9 MT) result in reduced levels of DNA methylation (Kunert et al. 2003; Tamaru et al. 2003; Jackson et al. 2002; Weissmann et al. 2003), and this is especially the case with non-CG methylation (Law & Jacobsen 2010). This has led to the notion that histone methylation may promote DNA methylation as an evolutionary conserved pathway existing in mammals too (Freitag & Selker 2005). Indeed, *de novo* DNA methylation in mammals has been shown to decrease in Suv39h null mutants, mainly at repeat sequences (Peters et al. 2003; Lehnertz et al. 2003; Martens et al. 2005; Fuks et al. 2003). Similar observations have been made for another mammalian H3K9 MT, G9a, which is necessary to recruit Dnmt3a and Dnmt3b for triggering *de novo* methylation at gene promoters during cellular differentiation (Feldman et al. 2006; Dong et al. 2008; Epsztejn-Litman et al. 2008). Direct interactions with Dnmt3s have been reported also for two more

members of the mammalian H3K9 MT family – EZH1 and SETDB1 (Li et al. 2006; Vire et al. 2006). It has been shown that the interactions of the Dnmts with the H3K9 MT G9a involve domains independent of their histone methyltransferase function, and thus mutations in their histone methyltransferase function (and lack of H3K9me3 respectively) do not affect the methylation on DNA. This suggests that DNA methylation is not dependent on the histone methylation *per se*, but rather on the recruitment function of the histone methyltransferases. (Cedar & Bergman 2009).

Interestingly, an orthologue of SETDB1 in *Drosophila melanogaster* (dSETDB1) has been shown recently to bind methylated CA sequences and recruit Dnmt2 to target loci for DNA methylation and silencing (Gou et al. 2010).

The most studied protein, which does not belong to the Dnmt family but is closely related to DNA methylation, is Np95 (or Uhrf1). It is an E3 ubiquitin ligase which has a role in chromatin modification, replication-linked DNA methylation maintenance and the DNA damage response (Mistry et al. 2008; Nishiyama et al. 2013). It has been described to work tightly with Dnmt1 to facilitate its function in the maintenance of pre-existing CG methylation, similarly to the way in which Dnmt3L works together with the Dnmt3s. Np95 binds specifically to hemimethylated DNA and loss of function mutations lead to severe loss of DNA methylation (Sharif et al. 2007; Bostick et al. 2007). Np95 has also been shown to interact with Dnmt3a and Dnmt3b, which may suggest a role for Np95 in *de novo* methylation as well (Meilinger et al. 2009).

LSH1 (lymphoid-specific helicase 1 or HELLS) is a chromatin-remodeling factor also shown to have a role in the maintenance of DNA methylation. Although Lsh-KO mice develop normally, their genomes are heavily demethylated, both at repeats and at single copy genes (Dennis et al. 2001). Lsh-deficient cells are found to have reactivation of repeat expression, at pericentric satellites and LTR elements, but not at single copy genes (Huang et al. 2004; Dunican et al. 2013; Yu et al. 2014). This suggests that chromatin structure and DNA methylation are directly related and interdependent.

### ***1.3.5 Non-coding RNA pathways in the regulation of DNA methylation***

It is well established that all *de novo* methylation in plants as well as the maintenance of asymmetric CHH methylation are induced through an siRNA-facilitated mechanism known as

RNA-directed DNA methylation (RdDM), in cooperation with histone modifying enzymes (Jackson et al. 2002; Chan et al. 2004; Chan et al. 2006). Non-CG methylation in plant *drm1 drm2 cmt3* triple-KO *Arabidopsis thaliana* mutants can be recovered by sequence-specific signals, dependent on the siRNA pathway and H3K9 methylation (Chan et al. 2006). Thus, plant non-CG methylation shows a different recruitment behaviour from CG methylation, which cannot be recovered in a *met1* mutant after backcrossing to wild type, demonstrating different potential of utilising non-CG methylation for gene regulation (Henderson & Jacobsen 2008). siRNAs are also involved in chromatin remodelling in *Saccharomyces pombe* (Verdel 2004; Noma et al. 2004), which however, do not have DNA methylation. Knockouts of RNAi pathway genes in *Arabidopsis* and *S. pombe* lead to a loss in asymmetric DNA methylation or H3K9 methylation at repetitive loci, respectively (Cam et al. 2005; Volpe 2005; Zilberman et al. 2004). The current knowledge about the mechanism of RdDM in plants, however, involves a number of plant-specific proteins (Matzke et al. 2009), making it unclear whether a similar mechanism exists in mammals.

In the protist *Dictyostelium* dependence of asymmetric DNA methylation on RNAi pathway has also been suggested. DNA methylation was reduced and, as a result, gene expression of particular genes elevated in KOs of a number of proteins involved in the yeast RNAi pathway (Kuhlmann et al. 2005). This observation is particularly noteworthy, since the enzyme responsible for the DNA methylation in protists is a high-homology orthologue of the mammalian Dnmt2, while DRM2 in plants is homologous to the Dnmt3 family in mammals (Kato et al. 2006).

Interestingly, in mammals, it has been suggested that Suv39h-dependent DNA methylation of satellites could be initiated by a Dicer-mediated mechanism in which RNA-duplexes from the satellite sequences, after processing, are targeted back to the pericentromeric regions in the nucleus to initiate heterochromatinisation (Cedar & Bergman 2009; Fukagawa et al. 2004; Sugiyama et al. 2005; Kanellopoulou et al. 2005). Knocking out Dicer in mouse ES cells leads to a loss of DNA methylation and H3K9 methylation at centromeres (Sugiyama et al. 2005; Kanellopoulou et al. 2005) and an aberrant accumulation of satellite transcripts (Fukagawa et al. 2004). It has also been shown that siRNAs bind Dnmt3a (Weinberg et al. 2006) and endogenous siRNAs complementary to retrotransposons are abundantly expressed in mouse oocytes (Tam et al. 2008; Watanabe et al. 2008). Moreover, synthetic RNA can target complementary DNA sequences for DNA and histone methylation, further implicating the RNAi pathway in the

process of targeting chromatin modifications to the correct sequences (Weinberg et al. 2006; Kim et al. 2006; Morris et al. 2004). Indeed non-coding RNAs seem involved in recruiting DNA and histone methylases at imprinted loci and during X-chromosome inactivation (Law & Jacobsen 2010; Nagano 2010; Zhao et al. 2008)) but also at non-imprinted autosomal loci (Tufarelli et al. 2003). There is recent evidence that piRNAs are involved in *de novo* methylation in primordial germ cells (PGCs), acting upstream of the Dnmts (Aravin et al. 2008; Kuramochi-Miyagawa et al. 2008).

Quite the opposite, Imamura et al. have reported that overexpression of an antisense RNA in rat induced CG-island demethylation and non-CG methylation on the sense strand of the sphingosine kinase 1 gene (*Sphk1*) (Imamura et al. 2004). In line with this, it has recently been reported that Dnmt1 interacts with non-coding RNAs from the *CEBPA* locus and this blocks its activity, leading to demethylation of *CEBPA*. Moreover, it has been suggested, that this mechanism is functional at numerous loci genome-wide (Di Ruscio et al. 2013).

## 1.4 Removal of non-CG methylation

The removal of 5mC has been a heavily investigated topic in the last fifteen years and it is currently known that a few mechanisms are involved in this process.

The removal of the methylation mark can either be through passive or active mechanisms. A few reports have claimed that the levels of non-G context methylation are dynamic, and imply that it can therefore be subject to active or passive removal (Imamura et al. 2005; Fuso et al. 2010; Haines et al. 2001). The passive demethylation is replication-dependent, and it occurs in highly proliferating cells (Franchini et al. 2012). It is based on an impeded maintenance of methylation, usually associated with down-regulation of Dnmt1 or Np95 (Kagiwada et al. 2013), or their physical exclusion from the nucleus and the replication foci (Lee et al. 2014; Seisenberger et al. 2012). This results in a passive ‘dilution’ of the 5mC marks through the cycles of replication. This type of demethylation would particularly affect non-CG context methylation, because it does not have a known mechanism of maintenance, and would either need to be actively established *de novo* after each course of replication, or is bound to be passively quantitatively reduced.

There is more than one mechanism known for active demethylation (Bhutani et al. 2011; Wu & Zhang 2010). It is termed ‘active’ because it involves the active removal or modification

of 5mC in a replication independent manner (or at least cell division independent). Therefore this type of demethylation can also occur in non-replicating cells (Kohli & Zhang 2013). Active and regulated demethylation has already been suggested for non-CG context in the *myogenin* gene during muscle differentiation, although the mechanism has not been investigated (Fuso et al. 2010). One group of mechanisms involve the DNA repair machinery of the BER pathway (base excision repair) and proteins like AID (Activation-induced (cytidine) deaminase), UNG2 (or UDG2) (Uracil-DNA glycosylase) and potentially TDG (Thymidine-DNA glycosylase) (Santos et al. 2013; Hajkova 2010; Popp et al. 2010). This mechanism relies on cytosine deamination by AID, followed by a short- or long-patch repair by the glycosylases – UNG2 in case an unmethylated cytosine is targeted by AID (to form uracil), and TDG in case a methylated cytosine is targeted by AID (to form thymidine) (Franchini et al. 2012). The *in vivo* significance of the first pathway (deaminating unmethylated cytosines to uracil) is not yet fully validated, however it seems a very plausible mechanism, since fertilized oocytes maternally deleted for TDG do not show any impairment in the active demethylation of the paternal pronucleus (Santos et al. 2013). In addition, the substrate affinity of AID is around 10-fold weaker towards methylated or hydroxymethylated cytosines in comparison to unmethylated cytosine (Nabel et al. 2012). The conserved recognition target sequence for AID is WRC (where W is A or T, and R is A or G), and since this sequence does not have any specificity for cytosine context, the expectation is that it should equally affect both types of DNA methylation – CG and non-CG. Interestingly, the highest sequence preference of UNG2 is towards uracils followed by an A or T, and the least preference is for U followed by a G (Doseth et al. 2012).

Another pathway for active demethylation is through the chemical modification (oxidation) of the methyl group on 5mC. A recently identified family of three enzymes, the Ten-eleven translocation methylcytosine dioxygenases or Tets (Tet1, Tet2 and Tet3) are responsible for oxidising 5mC to 5hmC, 5fC and 5caC (Figure 6) (Tahiliani et al. 2010; Ito et al. 2011). Although this is not a physical removal of the methylation, the effect of oxidising the methyl group chemically reverses the effect of methylation by neutralising its strong hydrophobic and strand duplex stabilising effect (Thalhammer et al. 2011). A mechanism to maintain the 5hmC mark has not yet been reported (Lee et al. 2014; Shen & Zhang 2013), and therefore the oxidation derivatives of 5mC are either passively diluted in the course of cell division, or removed by the repair mechanisms, most likely involving BER (Kohli & Zhang 2013; Shen & Zhang 2013). It

has been shown that TDG, in addition to recognising T:G mismatches, has an *in vitro* activity on 5fC and 5caC (not 5hmC) and has therefore been proposed to act downstream for their active removal (He et al. 2011; Maiti & Drohat 2011). Another possibility is a recently reported *in vitro* property of the Dnmt enzymes to remove oxidised methylation derivatives and thus contribute to restoring unmethylated cytosine (Chen et al. 2013; Liutkeviciute et al. 2009).

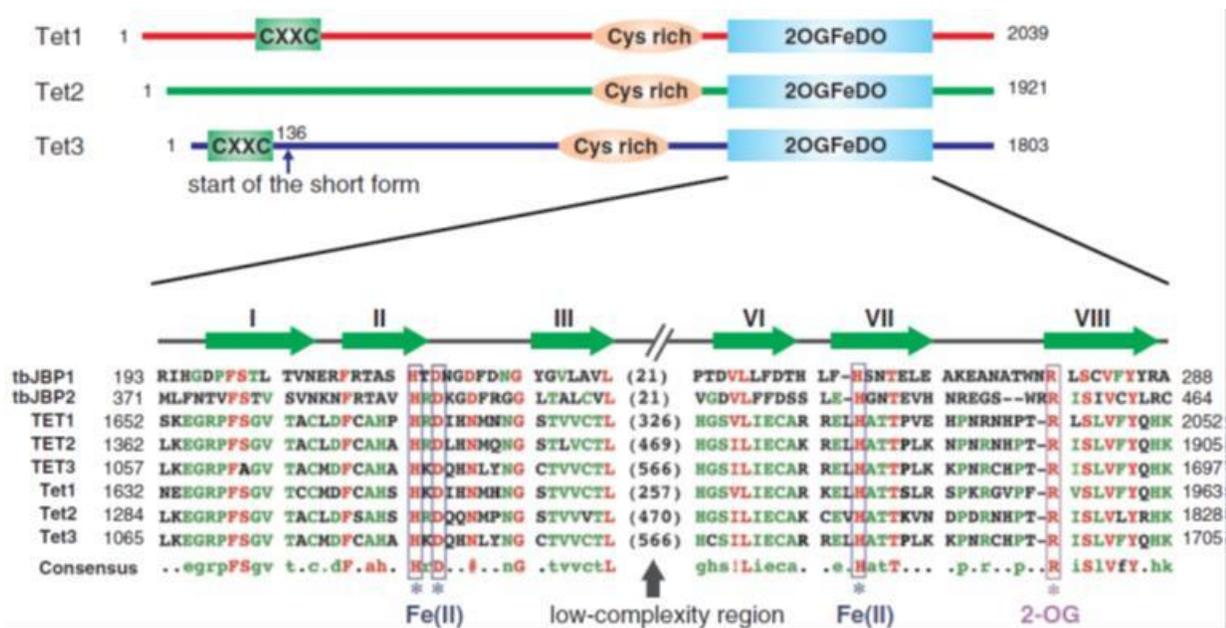


Figure 6. A structural comparison of the three members of the Tet family of oxygenases. A domain outline (upper panel) and a multiple alignment of the conserved catalytic domain (lower panel) between human, mouse, and the bacterial oxygenases, which served for their original identification in mammals (adapted from Shen and Zhang, 2013).

It is currently not known if the Tet enzymes have any context limitations for their oxidation activity on 5mC, although it is very clear that they strongly affect CG context - it is the predominant context of sperm and hence the male pronucleus, which is affected by global genome-wide hydroxylation shortly after fertilisation (Santos et al. 2013). The possibility that Tet enzymes might be able to discriminate between mCG and mCA would be worth investigating.

## 1.5 Biological role of CHH and CHG methylation

DNA methylation in the symmetrical dinucleotide has long been appreciated as a requisite requirement for development (Lee et al. 2014). CG methylation has been implicated in genome

defence and transcriptional regulation. Silencing of transcription is achieved by methylation of CGIs and repeat sequences (Reik 2007), while gene body methylation is often associated with higher gene expression (Seisenberger et al. 2012; Ball et al. 2009; Lister et al. 2009; Laurent et al. 2010). In contrast, the biological significance of non-CG methylation, has so far remained elusive.

The role of CH methylation has been best characterised in plants, where cytosine methylation in non-CG sequence contexts is tightly linked to inactive chromatin and transcriptional gene silencing (Volpe et al. 2010; Stroud et al. 2014). This includes pericentromeric repeats (macro- and minisatellites), DNA type and retrotransposons, telomere repeats and some single copy genes, generally conferring genome stability (Cokus et al. 2008; Lister et al. 2008a; Vrbsky et al. 2010). Early studies on the *drm1 drm2 cmt3* triple-KO *Arabidopsis thaliana*, which lacks CH context methylation, have shown distinct phenotypes and developmental defects in mutant plants, suggesting that non-CG methylation is implicated in the regulation of developmentally important genes (Chan et al. 2006). The presence of non-CG methylation on repeat sequences within the gene promoter of the *SDC* gene (*SUPPRESSOR OF drm1 drm2 cmt3*) has later been shown to silence gene expression and its absence was enough to cause the *drm1 drm2 cmt3* developmental phenotypes (Henderson & Jacobsen 2008). Recently, the varying decrease in levels of non-CG methylation in *drm1*, *drm2*, *cmt3* and *cmt2* mutants has been implicated in the transcriptional derepression of transposable elements (TEs) in a quantitative fashion (Stroud et al. 2014). Interestingly, a fraction of protein coding genes were also found silenced by non-CG methylation, although under the control of only CMT3 and DRM2, and not CMT2. These observations validate that non-CG methylation in plants acts primarily as a repressor of transcription. Moreover, the 24-nt siRNAs, which target the *de novo* methylation, are strongly dependent on non-CG methylation through a possible self-reinforcing loop mechanism that involves H3K9 methylation, which in itself depends largely on non-CG methylation and controls the biogenesis of the siRNAs. At the same time, global loss of non-CG methylation in plants induces histone acetylation, not only in transcriptionally derepressed TEs and genes, but in a genome-wide fashion, leading to relaxation of chromocentres and overall decompactisation of heterochromatin. Thus, non-CG methylation is of crucial importance for plant genomes to maintain a condensed heterochromatic state (Stroud et al. 2014).

In *Dictyostelium*, similarly to plants, non-CG methylation has been found in two retrotransposons – Skipper and DIRS-1, where it was shown to contribute to the silenced

unexpressed state (Kuhlmann et al. 2005). After depletion of DNA methylation using Dnmt2-null cells their expression levels rose and mobility of Skipper within the genome was restored. Interestingly the non-CG methylation of Skipper in the wild type genome was localized only within the coding sequences and only on the antisense strand. In DIRS-1 the methylation was found in the control LTR region but also on the antisense strand only (Kuhlmann et al. 2005).

In *Drosophila*, the Dnmt2-mediated asymmetric methylation has been suggested to play a role in the maintenance of retrotransposon silencing and telomere integrity during early embryonic development. Depletion of DNA methylation in Dnmt2 null strains causes enhanced expression of retrotransposon sequences and loss of subtelomeric repeat clusters (Phalke et al. 2009; Schaefer & Lyko 2010a). Nevertheless, complete depletion of asymmetric methylation in the fly genome by RNA interference does not have any detectable consequences for embryonic development (Kunert et al. 2003). This is consistent with the lack of phenotypic effects in *Su(var)3-9* mutant fly strains (Tschiersch et al. 1994), which have been shown to lack DNA methylation almost completely (Kunert et al. 2003). However, overexpression of mouse Dnmt3a in *Drosophila* leads to severe developmental defects or lethality, attributed to disruption in cell cycle progression and abnormal chromosome condensation (Lyko et al. 1999; Weissmann et al. 2003), which indicates that, nevertheless, DNA methylation or the lack of it may have a functional role in *Drosophila* (Lyko 2001). This is also suggested by the strong conservation between Dnmt-2 mediated DNA methylation in Drosophilids and *Anopheles* mosquitoes over more than 250 million years of evolution (Marhold et al. 2004).

Recently, a role of non-CG hydroxymethylation has been proposed in the regulation of alternative splicing in honey bees, since both 5mC and 5hmC have been observed in gene bodies, but only 5hmC in introns (Cingolani et al. 2013).

In mammals, no definitive roles have been ascribed for non-CG methylation at present, which is not surprising given the limited reports, which are available.

Gowher and Jeltsch (2001) hypothesised at least three possible roles for non-CG methylation. First, it could have a role in epigenetic gene regulation for specific targets. Second, it could contribute to global genome silencing in the early stages of development, in advance of the establishment of heterochromatic marks. And third, it could function to activate Dnmt1 for *de novo* methylation of specific sequences (Gowher & Jeltsch 2001). In light of later discoveries more roles could be hypothesised or envisioned but these three certainly still remain valid.

In mammalian ES cells, as opposed to plant cells, analysis of high resolution genome-wide

methylation maps shows twice as high non-CG methylation density within gene bodies, compared to other genomic sequences, with the highest density in exon sequences. Surprisingly, significant enrichment of asymmetric non-CG methylation and of hydroxy-methylation was found on the antisense strand of the non-CG enriched sets of genes (Lister et al. 2009). Moreover, strand-specific RNA-seq analysis in those sets of genes showed that highly expressed genes contain threefold higher non-CG methylation density than non-expressed genes (Lister et al. 2009). Another interesting observation by Lister et al. (2009) is the mosaicism of methylation in non-CG sites within the ES cell population, where 85% of the sites were methylated only in a fraction of the genomes (10-40%), unlike CG methylation where 77% of mCG sites were methylated in 80-100% of the genomes. This striking heterogeneity of asymmetric methylation in the ES cell population may point to differences in the epigenomes of individual ES cells, which may be linked to the unique set of epigenetic marks each cell gains before the process of differentiation sets off (Lister et al. 2009). The difference in the amounts of non-CG methylation in human ES cells and fibroblasts suggests it may have a role in the origin and maintenance of the pluripotent lineage (Lister et al. 2009).

A possible role for non-CG methylation in telomere silencing has been suggested by Gonzalo et al. (2006), who showed that in ES cells lacking either Dnmt1 or Dnmt3a and 3b together, telomeric sequences have an increased frequency of recombination between sister chromatids. This suggests that DNA methylation may also be required to maintain the ends of chromosomes in a heterochromatic state, in order to prevent recombination at telomeres, which could lead to translocations, genomic instability and aneuploidies (Gonzalo et al. 2006). Telomere sequences lack the canonical CG dinucleotides, so this effect would be completely due to non-CG methylation.

Non-CG methylation has also been observed in some origins of replication and has been suggested to play a role in the mechanism of origin licensing (Tasheva & Roufa 1994a; Tasheva & Roufa 1994b). Interestingly, the CA, CC and CT methylation was not observed in replication arrested  $G_0$  cells and seemed dynamically regulated during the cell cycle. The cytosines rapidly demethylated when cellular growth and DNA replication were arrested in an exponentially growing replicating cell culture, and they quickly remethylated when the arrest was reversed (Tasheva & Roufa 1994a; Tasheva & Roufa 1994b).

The occurrence of non-CG methylation has been observed in human cancer cells compared to adult healthy tissues, potentially leading to higher incidence of C→T transition mutations and

hence genomic instability (Kouidou et al. 2005; Kouidou et al. 2006), although, generally, non-CG methylation is believed not to be characteristic for cancer cells (Dyachenko et al. 2010). This raises the possibility that, being abundant in ES cells, non-CG methylation may be a also feature of cancer stem cells (Dyachenko et al. 2010).

To summarise, non-CG methylation does not yet have a confirmed role in mammalian cells. Having in mind that its peak in appearance is very early in development and it correlates with gene bodies of highly expressed genes, it is not very likely that it has a common origin or function with the plant non-CG methylation (Dyachenko et al. 2010).

## **1.6 Techniques for the analysis of non-CG context methylation**

### ***1.6.1 Methods for DNA methylation analysis – an overview***

Epigenetics is a relatively young scientific field, and many of the methods it employs are also young and novel, a number of them having been developed within the last 10 years (Table 1). The current methods used for the analysis of DNA methylation have been broadly divided into three groups, based on their leading principle of methylation detection: methods involving bisulphite treatment of DNA, methods based on affinity or 5mC binding properties, and methods based on enzyme digestion (Table 1) (Zilberman & Henikoff 2007; Laird 2010). There is a fourth group of techniques, used for the direct detection of global amounts of 5mC in a digested DNA sample – such as high pressure liquid chromatography (HPLC), mass spectrometry (MS), and capillary electrophoresis, as well as the thin layer chromatography (TLC) based methods (direct digestion or in nearest neighbour analysis – NNA) (Woodcock et al. 1987; Nyce et al. 1986; Grafstrom et al. 1985; Mund et al. 2004). This group has regained its significance in recent years, especially in regard to the identification and analysis of novel DNA modifications like hydroxy- and formyl-cytosine (Kriaucionis and Heintz 2010; Tahiliani et al. 2010; Booth et al. 2012), but also for global analysis of 5mC levels (Ficz et al. 2013; Oda et al. 2013; Senner et al. 2012; J. Wang et al. 2014; Barciszewska et al. 2014; Zhang et al. 2014). More recently, several nano-techniques have been developed which are not widely used yet, discussed in detail elsewhere (Shanmuganathan et al. 2013).

Table 1. A summary of methods for analysing DNA methylation. MS-REA = Methyl-sensitive restriction endonuclease assay (Bird & Southern 1978; Cedar et al. 1979), MSP = Methylation-specific PCR (Herman et al. 1996), MCAM = methylated CG island amplification and Differential methylation hybridization methods (DMH): HELP = HpaII tiny fragment enrichment by ligation mediated PCR, Mmass = Microarray-based methylation assessment of single samples (reviewed in (Laird 2010)), RLGS = Restriction landmark genomic scanning (Kawai et al. 1993), LUMA = Luminometric methylation assay (Karimi et al. 2006), COBRA = Combined bisulphite restriction analysis (Xiong & Laird 1997), MethyLight = Methylation sensitive qPCR (Eads et al. 2000), Ms-SNuPE = Methylation-sensitive single nucleotide primer extension (Gonzalvo & Jones 2002; Wu et al. 2008), RRBS = Reduced representation bisulphite sequencing (Meissner et al. 2005), BS-seq/WGBS = Whole genome bisulphite sequencing (Lister et al. 2008b), PBAT = Post-bisulphite adaptor tagging (Miura et al. 2012), MIRA = Methylated CG island recovery assay (Rauch & Pfeifer 2005), IF = Immunofluorescence, MeDIP = Methylated DNA immunoprecipitation (Weber et al. 2005; Down et al. 2008), HPLC = High pressure liquid chromatography (Grafstrom et al. 1985), TLC = Thin layer chromatography (Harbers et al. 1975), MS = mass-spectrometry, NNA – Nearest neighbour analysis (Grafstrom et al. 1985), MS-SSCA = Methyl-sensitive single strand conformation analysis

Colour coding: red = CG-context biased, orange = no context information, green = information on all contexts, i.e. useful for non-CG context methylation; \* Discussed in text

Methods	Established	Context	non-CG useful
<b>Restriction enzyme-based methods – MspI/HpaII (CCGG), BstUI (CGCG), MaeII (ACGT)</b>			
MS-REA	1978	CG-context only	No
MSP	1996		No
RLGS	1993		No
Array-based (MCAM, HELP, Mmass, etc)	2002 - 2008		No
LUMA	2006		Maybe*
<b>Bisulphite (BS) conversion-based methods</b>			
BS-conversion, cloning and Sanger sequencing	1992	All contexts	Yes
COBRA	1997	CG-context	No
MethyLight	2000	No context	No
Bisulphite pyrosequencing	2003	CG- context	No
Ms-SNuPE (array-based: EpiTYPER)	2002 (2008)	CG- context	No
Illumina bead arrays (Infinium, GoldenGate)	2006 - 2009	CG-bias	Maybe*
RRBS (MspI-digest)	2005	CG-bias	Maybe*
MethylC-seq (BS-seq,WGBS)	2008	All contexts	Yes
PBAT	2012		Yes
<b>Affinity-based (Antibody or 5mC binding proteins)</b>			
Dot blot (for 5mC)	2002	Global 5mC, no context	No
5mC IF	2002		No
MIRA	2005		
MeDIP (+seq)	2005 (2008)		No (Maybe*)
MeCAP-seq	2010	CG-bias	No
<b>Other methods</b>			
HPLC, TLC, capillary electrophoresis, MS	1900s	No context	No
Nearest Neighbour Analysis (NNA)	1961	All contexts	Yes
MS-SSCA	1999	No context	No

As emphasised in Table 1, the classical analysis of DNA methylation has been primarily focused on the context of CG methylation. As such, many of the methods developed to date enrich for this type of methylation either by sequence-specific restriction enzymes (MspI/HpaII, Xma/SmaI, MaeII, BstUI) or with 5mC recognition proteins (like MBDs, MeCP2), and hence fail to provide information for other cytosine contexts (Table 1) (Laurent et al. 2010; Laird 2010; Zilberman & Henikoff 2007). Most of the remaining techniques including the chromatography-, mass spectrometry- and antibody-based methods, detect total 5mC per DNA sample, and therefore do not provide any context information. The same is the situation with the methyl-sensitive McrBC enzyme sometimes used for the degradation of highly methylated DNA— its recognition sequence  $R^{me}C...R^{me}C$  (R being A or G) does not account for the context of methylation 3' of the cytosine base (Zilberman & Henikoff 2007; Laird 2010).

For this reason, none of these techniques is suitable for the analysis of non-CG context methylation. The most suitable techniques, used historically for the identification of non-CG methylation, have been the locus specific bisulphite conversion and sequencing and the NNA analysis (Araujo et al. 1998; Clark et al. 1995; Grafstrom et al. 1985; Ramsahoye et al. 2000). More recently, bisulphite conversion has been coupled with high depth next generation sequencing (>10-fold genome coverage) to yield whole genome data on the distribution of non-CG methylation (Lister et al. 2009; Shirane et al. 2013; Lister et al. 2013). In addition, some other methods can potentially be used to indirectly derive data for non-CG context methylation, for example the asymmetric peaks detected in MeDIP datasets may correspond to non-CG context methylation (Ficz et al. 2011). CG-biased methods like the RRBS or the Illumina bead arrays (Table 1) can also contain information on non-CG context methylation, although they are both designed around enrichment for CGIs (Ziller et al. 2011). Lastly, another restriction enzyme-based method – LUMA – originally described for the MspI/HpaII enzyme pair (Karimi et al. 2006), has recently been modified to work with the enzymes AjiI/Psp6I for the estimation of global  $CCWGG$  (W = A/T) methylation (Barrès et al. 2009; Yan et al. 2011).

It therefore seems that, in addition to the overall research focus on CG methylation, the current scarcity in understanding about a potential biological function of the non-CG context methylation, can also be contributed to technological limitations.

## ***1.6.2 BS-seq and the analyses of non-CG methylation***

Bisulphite conversion followed by DNA sequencing is the only technique to date which provides base pair resolution data for DNA methylation and has established as the ‘gold standard’ to identify presence and distribution of both CG and non-CG methylation. It has, however, been suggested that its validity for non-CG context should be confirmed in parallel via another independent technique (Laird 2010). The data output from this analysis is the absolute level of methylation for each individual cytosine for a sequence, therefore false positives will contribute to the individual methylation value. Conversion artefacts have called into question the validity of a non-CG positive signal (Rein et al. 1997; Harrison et al. 1998), and in the classical bisulphite Sanger sequencing approach have been dealt with via strict rules for the design of PCR primers (Henderson et al. 2010; Warnecke et al. 2002). In NGS, these rules cannot be applied. I will focus briefly on a few factors related to the NGS analysis of non-CG context methylation, which point to the limitations of this technique in a few different ways, and are important to take into consideration.

### ***1.6.2.1 Genomic coverage and depth***

The degree of CG methylation in mammals is on average 70-80% for the CG dinucleotide (Ehrlich et al. 1982; Goll & Bestor 2005). Although this value is quite high, in order to estimate the correct percentage of methylation in a given cytosine position, between 10 and 20 sequences (clones) of the same position are standardly used. For regions with lower methylation status, more than 10 clones would be needed in order to correctly estimate the methylation percentage. When it comes to next generation sequencing (NGS), the usual depth for low coverage datasets is between 3- and 5-fold – they are regarded as useful for low resolution comparative analysis of multiple organisms or cell lines in CG context, and for global genomic elements only, but are not adequate for the evaluation of individual cytosine positions (Feng et al. 2010). Deeper ‘high resolution’ analysis, which provides correct information on individual CG calls, requires above 15-fold average depth per position, reaching even more than 30-fold in present day (Lister et al. 2009; Li et al. 2010; Stadler et al. 2011). Non-CG context methylation is on average below 25% per cytosine position, meaning that even higher depths would be needed in order to achieve similar level of accuracy as for the CG-context analysis. This situation is almost analogous to the high depth requirements for the bisulphite-based techniques for base resolution analysis of

hydroxyl-methylcytosine (Huang et al. 2012). Therefore, so far only the high resolution datasets have proven informative for the analysis of non-CG context methylation (Lister et al. 2011; Stadler et al. 2011; Shirane et al. 2013), making the NGS approach financially intractable at present for a routine genome-wide analysis of non-CG context DNA methylation.

### *1.6.2.2 Computing power*

The depth of coverage required for data reliability directly leads to the next specificity of non-CG context analysis in NGS datasets and this is the capacity of the computing facilities. The higher the depth, the longer it takes to calculate methylation in individual positions or even in genomic features, and the more server memory and processing power need to be utilised. In addition, the distribution of CG context in mouse genome Build 37 used in this study, is only 1/25<sup>th</sup> of the genomic cytosine content or only 4% from the total cytosine, making it only a fraction of the individual frequencies of the other three cytosine contexts – 36% for CA, 35% for CT and 25% for CC (see Table 0). This means that all calculations performed for non-CG context will require roughly 20-fold more processor power and time, than the same analysis performed for the CG context. These requirements create restrictions owing to the extended capacity for computing power and accompanying cost per sample.

### *1.6.2.3 Bisulphite conversion efficiency*

For long, the amount of non-CG positive calls served to determine conversion efficiency in a treated DNA sample, and thus served as an internal control and not a methylation signal in itself, preventing the need to use actual unmethylated controls (Genereux et al. 2008; Chhibber & Schroeder 2008). Bisulphite conversion of cytosine has a high chemical efficiency; however it never reaches 100% (Harrison et al. 1998). Good efficiency values lie above 95%, usually aiming for above 98-99% of the total unmethylated cytosine content. Due to the repressed representation of CG in mammalian genomes (4%, as discussed above), conversion artefacts are mostly seen in non-CG context (the remaining 96% of cytosine), and they are often dispersed as individual false positives throughout the whole genome (Warnecke et al. 2002). Therefore a single false positive value in a high depth dataset is not very likely to affect the evaluation of CG-context methylation with an average of 70% per position. This is not the case for non-CG context methylation, where individual conversion artefacts are very likely to affect the estimation of ‘real’ non-CG 5mC

calls, amounting for 5-20% on average per position. For this reason, authors often apply different ‘filters’ to exclude possible false positives from the final analysis – the most common being to exclude all reads containing more than 3 positive calls in non-CG context from the final data analysis (Lister et al. 2009; Lister et al. 2011), or defining a cutoff value below which the methylation signal is considered as conversion ‘noise’ (Raddatz et al. 2013). However it is not known how much of what is real signal is filtered out and how much is actual artefact, and what proportion of the artefacts actually remain in the validated data. This is because no alternative method currently exists which can identify DNA modifications at a single base resolution without bisulphite conversion. With the development of nanopore sequencing for the analysis of DNA modifications (such as the technology offered by Pacific Biosciences) (Flusberg et al. 2010; Coupland et al. 2012) this might be possible to do in the near future.

#### *1.6.2.4 Amplification bias*

It is well documented that due to DNA polymerase sequence preferences, there is a detectable bias towards CG context in the non-bisulphite converted whole genome NGS datasets accompanied with a much lower representation of AT-rich DNA regions (Quail et al. 2012; Oyola et al. 2012). AT-rich genomes have therefore proven difficult to amplify, a problem that can be extrapolated to the bisulphite-converted whole genome (WGBS) datasets in particular, which become highly AT-rich after conversion, irrespective of the original base composition of the sequenced genome. Commercial suppliers compete to engineer polymerases that show less context preference, but the issue still hasn’t been fully resolved (Oyola et al. 2012) (see also KAPA HiFi Uracil + leaflet at <http://www.bio360.net/z/KAPABiosystems/>). Therefore, highly methylated and unconverted CG-rich regions are unlikely to suffer from this problem but will be amplified preferentially, contributing to CG-bias and potential overestimation of the total number of positive methylation signals. The cytosine-rich but hypomethylated sequences including CH context, on the other hand, will certainly be greatly affected by this bias, with reduced representation and sequencing depth, necessary to complete an accurate analysis. However, whether those biases really affect bisulphite-converted datasets in this way, however, has not been evaluated in depth to date.

### 1.6.2.5 DNA degradation

It is well documented that the treatment of DNA with sodium bisulphite is a chemically aggressive process and leads to high levels of DNA degradation, i.e. loss of input material (Grunau et al. 2001; Shiraishi & Hayatsu 2004). This has for a long time dictated the amount of starting material used for bisulphite analysis, making it very difficult to produce methylation data from biological sources with very low cell numbers (Miura et al. 2012; Smallwood & Kelsey 2012). To date there is no comprehensive research into contexts of base sequence degradation during bisulphite mutagenesis. A recent study has shown that the degradation is triggered at the cytosine, and is a direct consequence of the bisulphite attack, contradicting earlier suggestions that it was due to an acidic attack on the purines (Tanaka & Okamoto 2007; Piperi & Papavassiliou 2011). It is therefore an open question whether the overall cytosine content in DNA is affected by the chemical attack to a larger extent than the other three bases, and whether the surrounding sequence context or methylation status would have an influence.

### 1.6.3 Additional tools facilitating DNA methylation analysis

Molecular tools such as methylation-sensitive restriction enzymes, anti-methylcytosine antibodies and methyl-binding proteins have been widely used for the development of a wealth of techniques as discussed in 1.6.1., although the broadly used and commercially available restriction enzymes target specifically CG context. A limited number of studies have used enzymes like the BstNI (BstOI) and AjiI together with their methylation-sensitive counterparts EcoRII or Psp6I to characterize methylation in CCWGG context (Woodcock et al. 1987; Franchina & Kay 2000; Barrès et al. 2009). Some studies have used the MspI enzyme and its sensitivity to the first cytosine in the CCGG sequence (Crowther et al. 1989), but considering the fact that CC methylation is the lowest observed non-CG context methylation in mammals, this tool is not very useful. Another group of restriction enzymes are the methylation-conditional endonucleases, like the McrBC (Sutherland et al. 1992), which digest only methylated DNA. Because of its peculiar recognition sequence  $R^{me}C$  ( $N_{40-3000}\dots$ )  $R^{me}C$  (in which R is A or G) the enzyme is naturally targeted to highly methylated regions (such as DMRs). It has been used for the degradation of highly methylated DNA (enrichment of unmethylated DNA) (Lippman et al. 2004; Rollins et al. 2006) or for the detection of 5mC (Gowher et al. 2000). Although the recognition context of the enzyme is not CG-specific, due to the low levels of non-CG

methylation, the lack of identified non-CG DMRs and the inability to distinguish between CG and non-CG, the properties of this enzyme do not seem relevant for the detection of non-CG methylation.

In 2011 another group of methylation-specific restriction enzymes has been characterized and named after its main component - the MspJI family of restriction endonucleases (Cohen-Karni et al. 2011). All of the group members cleave 9 bp (on the forward strand) and 13 bp (on the reverse strand) downstream of their recognition motif, and while half are enriching for CG-context, the other half are enriching for all methylation contexts or even mCA or mCT (RlaI). Therefore, this group of enzymes might be useful for the analysis of non-CG methylation and is worth investigating.

Other methylcytosine enrichment tools, like the well-studied and widely used methyl-binding proteins (MBDs, MeCP2) bias towards CG context and especially highly methylated CGIs (Bock et al. 2010; Zilberman & Henikoff 2007; Robinson et al. 2010). The antibodies generated against methylcytosine on the other hand, do not differentiate the sequence context, and will not be useful for non-CG analysis, unless a way to ‘subtract’ the CG information is applied. There are currently no known proteins which specifically bind to methylated non-CG context.

Additional tools utilised for studying DNA methylation include the enzymes responsible for generating it - the DNA methyltransferases. Apart from the widely used eukaryotic methyltransferases, bacterial enzymes have also become an extremely useful tool – the best known of which is the M.SssI methylase (Renbaum et al. 1990). This enzyme methylates cytosines exclusively in the CG dinucleotide, and has therefore proven very useful to study CG methylation. It is usually used for *in vitro* DNA methylation – vectors, short fragments or genomic DNA – in order to analyse the properties of CG-binding proteins, or restriction enzymes, or often as a positive control (a calibrator) in different techniques (Goll et al. 2006; Laurent et al. 2010; Guo et al. 2011). It is also a common tool to study nucleosome positioning and genome organization (Fatemi et al. 2005; Gal-Yam et al. 2006; Miranda et al. 2010) and has been used for targeted CG-methylation and gene repression (van der Gun et al. 2010). Bacterial methyltransferases specific for other DNA sequence contexts are not used or commercially available.

### **1.6.4 Technological perspectives**

As is the situation with the recently discovered novel types of DNA modifications (Kriaucionis & Heintz 2010; Ito et al. 2011; He et al. 2011; Tahiliani et al. 2010), new techniques may need to be developed or old techniques adapted through the use of new molecular tools, in order to achieve comprehensive characterization of non-CG methylation, beyond the mere ‘sequencing after bisulphite treatment’ analysis. Recent advances regarding hydroxy-methylcytosine (hmC) and formylcytosine (fC) have shown that additional understanding about the localisation, dynamics and biological functions of each novel modification must be paralleled by the development of unique techniques for its analysis and detection (Ficz et al. 2011; Booth et al. 2012; Raiber et al. 2012; Yu et al. 2012). In this regard, it becomes clear (Table 1) that the development or use of a few types of molecular tools can provide access to a variety of methodological possibilities for such in-depth analysis; and these are:

- Context-specific anti-5mC antibodies → several quantitative and descriptive techniques are based on the use of antibodies
- Non-CG context 5mC binding proteins or restriction enzymes → nearly half of the existing current techniques employ the use of such proteins for CG context
- Methyltransferases methylating exclusively non-CG contexts would also help teach us more about the nature of these modifications and their functional significance

Therefore, in order to achieve the aims of this work, alongside the use of the classical DNA methylation techniques, it is important to explore novel opportunities - modify existing methods or develop molecular tools for the accurate characterisation of non-CG context methylation.

## **1.7 Aims and hypothesis**

In mammalian development, each cell lineage acquires a unique set of epigenetic marks during differentiation. The epigenetic profile of each lineage represents a complex combination of histone marks and DNA modifications, as well as chromatin interacting factors, which work together to define the gene expression profile of the lineage. Although there is some plasticity of epigenetic marks, in general differentiation is associated with gain of modifications involved in

lineage specification and restoration to pluripotency requires removal of these modifications in the germline and early embryo (Hemberger et al. 2009). While this pattern is well established for some histone modifications and DNA methylation in the symmetric CG context, the opposite has been shown for the overall distribution and abundance of asymmetric non-CG methylation in ES compared to differentiated cells (Lister et al. 2009; Laurent et al. 2010; Ramsahoye et al. 2000).

As a summary, non-CG methylation is a poorly characterized phenomenon in mammals, typical for their very early stage of development, and proposed to have a role in pluripotency (Lister et al. 2009; Ziller et al. 2011). Since pluripotent cells are a model of the ICM component of the preimplantation stage mammalian blastocyst (Nichols et al. 2009), then its accumulation is a brief event in the course of development and its peak coincides with the wave of global genomic methylation occurring in the blastocyst post fertilization. In the context of the currently proposed two stages of CG methylation (Straussman et al. 2009), its appearance falls within the first phase of global untargeted CG genome methylation, which occurs around the ES cell stage – the phase with the highest density of methylation in the genome. Its distribution is very heterogeneous in the ES cell population and its enrichment is highest in actively expressed genes, and particularly in the antisense strand of the exons (Lister et al. 2009).

In addition, non-CG methylation is evolutionarily more conserved than CG methylation and in organisms like Drosophilids it is also characteristic only for the very early stages of development (Lyko, Ramsahoye, et al. 2000; Lyko 2001). The conservation of this pattern of distribution up to mammals indicates a possibility for an important conserved role in those very early stages of development. I therefore hypothesised a role of non-CG methylation in the establishment and maintenance of a pluripotent cell state, possibly through a role in regulation of gene expression. It is important to note, that DNA methylation *per se* has been claimed non-essential for the self-renewal capability of pluripotent cells (Tsumura et al. 2006), while it is absolutely crucial for their ability to differentiate (Schmidt et al. 2012). Therefore, the role of non-CG methylation might be expected in the latter, where creating intra-population methylation heterogeneity could pre-determine or influence differentiation pathways before the exit from pluripotency.

I propose to investigate this hypothesis by addressing the following questions:

- 1) What is the genomic distribution of non-CG methylation in ES cells, and its wider dynamics during key developmental stages in the mouse?
- 2) How is the asymmetric methylation established *de novo* and maintained; are there dedicated Dnmts involved in this process (like Dnmt2), what are the individual contributions of Dnmt3a and Dnmt3b?
- 3) Does non-CG methylation have a role in transcriptional regulation in ES cells and does that influence their pluripotency state?
- 4) If necessary, can the development of novel tools and techniques facilitate in depth analysis of non-CG methylation?

## 2 Materials and methods

### 2.1 Cell lines and culture

The cell lines and tissues used in these studies were derived or obtained by the host laboratory for use in epigenetic studies. Mouse lines derived de novo from strategically relevant genetically modified loci were generated under licensing in keeping with the Animals Scientific Procedure ACT 1986 and under Home Office permission to Prof Wolf Reik. Human cell lines, embryonic stem cells (hESC) were obtained as a gift and used under the relevant MTAs.

The J1 ES cell line (129S4/SvJae) was purchased from ATCC (Cat. SCRC-1010); the E14 ES cell line has been derived from the E14 cell line strain 129P2/OlaHsd and is a long standing reagent in the host laboratory. The Dnmt2 KO, Dnmt3a and 3b single and double KO ES cells, the Dnmt1/3a/3b-KO and Np95-KO (129/Ola derived) are J1 or E14 derived and a gift from Masaki Okano (M Okano et al. 1998; Okano et al. 1999; Tsumura et al. 2006), the Dnmt1<sup>s/s</sup>-KO line is a gift from En Li (Lei et al. 1996), and the Dicer-KO lines are a gift from Neil Brockdorff, University of Oxford (Nesterova et al. 2008). The tamoxifen inducible Dnmt3a/4b-DKO ES cells are a kind gift from H. Koseki, RIKEN Center for Integrative Medical Sciences Laboratory for Developmental Genetics.

All wild type (WT) and Dicer-KO cell lines were grown on a  $\gamma$ -irradiated pMEF feeder layer while all Dnmt-pathway KO lines were grown on gelatine. All ES cell lines were cultured at 37°C and 5% CO<sub>2</sub> in complete ES medium (DMEM 4500 mg/L glucose, 4 mM L-glutamine and 110 mg/L sodium pyruvate, 15% foetal bovine serum, 100 U of penicillin/100  $\mu$ g of streptomycin in 100 mL medium, 0.1mM non-essential amino acids, 50  $\mu$ M  $\beta$ -mercaptoethanol, 10<sup>3</sup>U/ml LIF ESGRO<sup>®</sup>) (Ficz et al. 2011). Naïve mES cells were grown in serum-free N2B27 medium (Cat. DMEM/F12: GIBCO 21331; Neurobasal: GIBCO 21103; N2: Stem Cells SF-NS-01-005; B27: GIBCO 17504-044), supplemented with 10<sup>3</sup> U/ml LIF, 1 mM Mek inhibitor PD0325901 and 3 mM Gsk3b inhibitor CHIR99021 (Ficz et al. 2013).

Primary mouse embryonic fibroblasts were derived from E13.5 embryos (C57BL/6J x CBA/Ca F1 x F1, called further for short 'F1') and cultured for 1 to 3 passages in DMEM 4500 mg/L glucose, 4 mM L-glutamine and 110 mg/L sodium pyruvate, 10% foetal bovine serum, 100 U of penicillin/100  $\mu$ g of streptomycin in 100 mL medium, 50  $\mu$ M  $\beta$ -mercaptoethanol.

H9 hES cells were cultured and provided by Peter Rugg-Gunn (Babraham Institute) and IMR90 human lung fibroblasts (ATCC, Cat. CCL-186) - by Tamir Chandra (Reik lab).

DNA from C57BL/6 J (B6) ES cells TMBD10, B6 pMEFs and AID-KO pMEFs (Popp et al. 2010) as well as from iPS cells derived from those MEF genotypes was provided by Inês Milagre (Reik lab).

## **2.2 Mouse tissues**

DNA from B6 adult mouse tissues was kindly provided by Tim Hore (Reik lab). DNA from C57BL/6 J (B6) E11.5 and adult forebrain and hindbrain, as well as E11.5 placenta, were provided by Heather Burgess (Reik lab). B6 E19.5 placenta was provided by Myriam Hemberger (Babraham Institute).

Mouse embryos and oocytes were collected by Fátima Santos according to standard procedures (Hogan et al., 1994) and were derived from C57BL/6J animals.

## **2.3 Molecular Biology**

### ***2.3.1 Isolation of DNA and RNA from cells and tissues***

Genomic DNA from ES cell lines was extracted using Qiagen kits: All Prep DNA/RNA kit, Qiagen DNeasy Blood and Tissue kit and QIAmp DNA micro kit and stored in EB buffer (Qiagen) or RNase free water (Qiagen). All genomic DNAs were quantitated by Quant-iT™ PicoGreen® dsDNA Assay Kit on a CytoFluor II microplate reader (PerSeptive Biosystems) and stored at -20°C.

RNA was extracted either with the Qiagen All Prep DNA/RNA kit or using QIAzol® (Qiagen). Total RNAs were quantitated with NanoDrop® ND-1000 Spectrophotometer and stored at -20°C or -80°C.

### ***2.3.2 Bisulphite conversion of genomic DNA***

Genomic DNA was treated with sodium bisulphite using any of the following kits: Qiagen Epiect Bilulfite Kit (FFTP protocol), Imprint DNA Modification kit from Sigma-Aldrich (1-step or 2-step) and EZ DNA Methylation™ Kit (Zymo Research) according to manufacturer's

instructions. The following minor modification was applied: the temperature incubation programme in the Qiagen kit was doubled (i.e. 10 hours instead of 5 hours in total). Conversion with 9 M ammonium bisulphite was performed at 70°C for 30 minutes as described (Hayatsu et al. 2004).

### ***2.3.3 Amplification (PCR) of major satellite***

One µg of bisulphite converted genomic DNA was diluted 1:100 and 5 µl of the dilution subjected to amplification with HotStart Taq (Qiagen) in 50 µl volume, 200 nM primer, 200 µM dNTPs, 2 mM MgCl<sub>2</sub>, 1.0 unit of enzyme. Amplification was done with an initial step at 94°C for 15 minutes followed by 35 cycles of 20 seconds at 94°C, 20 seconds at 55°C, and 20 seconds at 72°C, with a final step at 72°C for 3 minutes. The amplified DNA was loaded on a 2% agarose gel, the 370bp band excised and purified with a Qiagen Gel Extraction kit according to kit instructions. The fragments were cloned and sequenced as described in 2.3.4.

### ***2.3.4 T-A cloning***

For sequencing of bisulphite converted DNA regions of interest, PCR fragments were cloned into pGEM-T using the pGEM-T Easy Vector Kit from Promega. Ligations were performed in 10 µl volume at 4°C overnight. Invitrogen's Subcloning Efficiency DH5α Competent Cells or One Shot Top 10 Chemically Competent E.coli were used for transformation and all steps were performed according to the manufacturer's instructions using 5µl of the pGEM-T ligation reaction. For each transformation, 100 µl and 900 µl of the transformation mixture were plated out onto LB plates containing 100 µg/ml ampicillin, for selection, and 40 µl of a 40 mg/ml X-gal solution, spread over the plate, for blue/white selection. Plates were incubated at 37°C overnight (o/n).

Colony screen was performed via PCR with the M13 primer pair (Appendix Table 12). White colonies were picked from the LB-Ampicillin plates with a 10µl pipette tip and transferred into a PCR tube containing the PCR reaction mix. The PCR reaction mix was prepared using Roche's Taq DNA Polymerase (25 µl volume, 300 nM primer, 200 µM dNTPs, 1.25 units enzyme). Amplification was carried out with an initial step at 94°C for 10 minutes followed by 35 cycles of 30 seconds at 94°C, 30 seconds at 55°C, and 30 seconds at 72°C, with a final step at

72°C for 10 minutes. Ten µl of the PCR reaction mix were loaded on a 2% agarose gel to verify the expected band size, 5 µl of the remaining PCR mix was sent for sequencing.

### **2.3.5 Sanger sequencing**

Sanger sequencing was carried out by the company Beckman Coulter Genomics. For bisulphite converted DNA, the sequencing results were analysed with QUMA (Kumaki et al. 2008) and visualised by a custom-made R script (from Miguel Branco).

### **2.3.6 Synthesis of M13 fragments**

PCR of M13 fragments – either unmethylated or enriched for methylcytosine or hydroxymethylcytosine, was performed using a standard dNTP mix (Bioline), or modified dm5CTPs (10 mM, NEB) or d5hmCTPs (100 mM, Bioline) instead of dCTPs. The PCR reaction mix was prepared using Dream Taq DNA Polymerase from Fermentas/Thermo (50 µl volume, 200 nM primer, 200 µM dNTPs, 1.25 units enzyme). Amplification was carried out with an initial step at 95°C for 2 minutes followed by 35 cycles of 30 seconds at 95°C, 20 seconds at 57°C, and 30 seconds (C, 5hmC) or 5 minutes (5mC) at 72°C, with a final step at 72°C for 7 minutes.

All M13-derived PCR fragments have been further purified with Qiagen PCR Purification kit or Thermo GeneJet PCR Purification kit, quantitated by Quant-iT™ PicoGreen® dsDNA Assay Kit on a CytoFluor II microplate reader (PerSeptive Biosystems) or on an Agilent 2100 Bioanalyzer system, and stored at -20°C. One µl of each PCR product was checked on a standard DNA resolving 2% agarose gel to ensure amplification was successful. A full list of the oligos and the different fragments are presented in Appendix Table 12 and Appendix Table 13.

### **2.3.7 Next generation whole genome bisulphite sequencing (BS-seq, WGBS)**

The amount of input material for WGBS was approximately 300 ng genomic DNA spiked with a 2 kb PCR fragment from M13mp18, 1:10 000. The DNA was fragmented via sonication with a Covaris E220 instrument with the 300 bp programme, in a total volume of 50-85 µl. Early Access Methylation Adapter Oligos (Illumina) were ligated to the fragmented DNA with the NEB Next DNA Library Prep Master Mix Set for Illumina, according to the manufacturer's instructions and purified after each step with Agencourt® AMPure® XP beads. Half of the

eluted adapter-ligated DNA was kept as a backup, the other half was subsequently bisulphite converted as described in 2.3.2. The bisulfite-treated DNA was eluted in 20 µl RNase free water (Qiagen). Again, half of the eluted DNA was kept, the other half was amplified using PfuTurbo Cx Hotstart DNA Polymerase (Agilent Technologies): volume 50 µl, 300 µM dNTPs, 400 nM primer, 2.5 units enzyme, with an initial step at 98°C for 30 seconds followed by 15 cycles of 98°C for 10 seconds, 65°C for 30 seconds, and 72°C for 30 seconds, followed by a final elongation step at 72°C for 5 minutes. In some instances indexed adapter-specific primers for Illumina were used from the iPCRtagT system (Quail et al. 2012). Each library was checked on an Agilent 2100 Bioanalyzer system, for average fragment size and concentration, and subsequently quantitated via KAPA Library Quantification Kit (KAPA Biosystems).

Paired-end 100 bp next generation sequencing (NGS) was performed on an Illumina HiSeq system at the facility at the Wellcome Trust Sanger Institute (WTSI).

### ***2.3.8 Reduced representation bisulphite sequencing with MspJI (meRRBS)***

The amount of input material was 150 ng intact genomic DNA. The DNA was digested with MspJI enzyme for 4 hours at 37°C adapting manufacturer's instructions (NEB) for 18 µl reaction volume. NEB Next kit Klenow exo- (1 µl), dNTP mix (0.2 mM) and dATP (2mM) were added directly to this mix for end repair and A-tailing and incubated for 40 min at 37°C, and then inactivated at 75°C for 15 min. Methylated adapters (Illumina) were added to the mixture (100 nM), together with 2 µl T4 Ligase (NEB Next kit), 1 mM ATP and incubated overnight at 4°C. Successful adapter ligation was tested via a test PCR: volume 50 µl, 0,5 µl DNA, 300 µM dNTPs, 400 nM primer P1 and P2 (Illumina), 2.5 units Phusion DNA Polymerase and Phusion High-Fidelity PCR Buffer 5x (NEB Next kit), with programme conditions as described in 2.3.7. The amplified libraries were run on a 2% agarose gel.

The adapter-ligated DNA was bisulphite converted with the Sigma Imprint DNA Modification kit using the 2-step protocol according to manufacturer's instructions and eluted in 20 µl RNase free water (Qiagen). Half of the eluted DNA was kept as back up, the other half was amplified with KAPA HiFi HotStart Uracil+ 2x Ready Mix (KAPA Biosystems): volume 50 µl, 10 µl DNA, 200 nM PCR primers PE 1.0 and 2.0, with an initial step at 98°C for 45 seconds followed by 15 cycles of 98°C for 15 seconds, 65°C for 30 seconds, and 72°C for 30 seconds, and a final elongation step at 72°C for 1 minute. Each library was barcoded with

indexed adapter-specific primers from the iPCRtagT system (Quail et al. 2012). The libraries were purified with Agencourt® AMPure® XP beads, processed as described for BS-seq in 2.3.7 and sequenced at the WTSI.

### **2.3.9 Mass spectrometry**

Mass spectrometry was performed by David Oxley, Head of the BI mass spectrometry facility. Genomic DNA was first quantitated by Quant-iT™ PicoGreen® dsDNA Assay Kit and digested into single nucleosides overnight at 37°C with a DNA Degradase Plus™ (Zymo Research) according to manufacturer's instructions. Variable amounts between 300-1000 ng were used for the digestion depending on sample availability. They were then submitted to the BI mass spectrometry facility, where 50 pg per sample was analysed by LC-MS/MS on a Thermo Q-Exactive mass spectrometer coupled to a Proxeon nanoLC. Three replicates of each sample were analysed and the amounts of C, 5mC and 5hmC were quantified in fmoles.

### **2.3.10 In vitro DNA methylation assays**

The *in vitro* DNA methylation was performed with 0.5 – 1.0 µg of genomic DNA (TKO ES) or M13-derived PCR fragments (see Appendix Table 13) with either M.SssI (NEB) or M.CviPI (NEB), according to the manufacturer's instructions. The reactions were incubated for 2 hours at 37°C, the DNA purified with GeneJet PCR Purification kit (Thermo) and quantitated by Quant-iT™ PicoGreen® dsDNA Assay Kit on a CytoFluor II microplate reader (PerSeptive Biosystems).

### **2.3.11 Nearest Neighbour Analysis (NNA)**

The nearest neighbour analysis was performed following the published protocol (Ramsahoye 2000) with the following modifications. One µg of DNA was digested o/n at 37°C with 10 units of FokI (gDNAs) or DpnII (PCR fragments) and 10 units of RNase A, purified with Qiagen QiaQuick PCR purification kit, and eluted in 30 µl dH<sub>2</sub>O. The DNA was quantified with PicoGreen® dsDNA Assay as in 2.3.1 and labelled with 1 U Klenow (Invitrogen) and 10 µCi <sup>32</sup>P-dNTP (PerkinElmer NEG502A) as in the original protocol. The labelled DNA was precipitated with 100% ethanol/ 3 M sodium acetate at -20°C for minimum 1 hour, and the pellet

resuspended in micrococcal nuclease digestion mix with 0.2 units MNase (Pharmacia Biotech) and 2 µg PDE II (Sigma-Aldrich) for 4 hours at 37°C. After digestion 0.5 µl was spotted onto a glass backed cellulose plate (Merck) and TLC performed as in original protocol.

### ***2.3.12 Luminometric Methylation Assay (LUMA)***

The originally published protocol was followed closely for this assay (Karimi et al. 2011). Two hundred (200) ng of DNA were digested o/n with either of the following: MspI, HpaII (Fermentas), AjiI, Psp6I, Bsp19I (SibEnzyme) together with one control enzyme: EcoRI, CviII or Tsp509I (NEB) using temperatures and buffer conditions as advised by the manufacturer. The digestion was simultaneous when temperature requirements were the same, or in two steps if they were different. The digested DNA was analysed on the Pyrosequencer at the Department of Physiology, Development and Neuroscience (PDN) by Mitsutero Ito.

### ***2.3.13 Gene synthesis***

Gene synthesis was performed using the GeneArt® Gene Synthesis platform by Life Technologies (available on <http://www.lifetechnologies.com>). The DNA sequence was reverse translated from the protein sequence and *AttB1* and *AttB2* sequences flanking the gene were added in order to facilitate downstream Gateway cloning. All genes were received in a Gateway entry vector.

### ***2.3.14 Gateway cloning***

The cloning was performed according to the instructions in the Gateway Technology Rev 1.0 user manual. In brief, the gene of interest was cloned into a pDONR vector via a BP reaction containing 10 µl of 50 ng vector with gene of interest, 150 ng pDONR vector, and 2 µl BP Clonase™ II enzyme mix in TE buffer pH 8.0, incubated 1 hour at 25°C, terminated with Proteinase K digest and transformed into One Shot ® Top10™ Chemically Competent E. coli Cells (Invitrogen).

Transformation was done according to the manufacturer's instructions, i.e. incubation on ice for 30 minutes, heat-shock by incubating at 42°C for 30 seconds, incubation with 250 µl of S.O.C. Medium at 37°C for 1 hour with shaking. 10 µl and 100 µl of each transformation were

plated out onto kanamycin selective (100 µg/ml) pre-warmed agar plates.

After selecting colonies, mini-prepping plasmid with Qiagen Mini Prep kit and digesting diagnostically with restriction enzymes to verify the correct insert size, the second Gateway sub-cloning reaction was performed – LR reaction. The 10 µl mix contained 180 ng pDONR with gene of interest, 100 ng pQLinkHD vector, and 2 µl LR Clonase™ II enzyme mix in TE buffer pH 8.0. It was incubated 1 hour at 25C, terminated with Proteinase K digest and transformed into One Shot® Top10™ Chemically Competent E. coli Cells (Invitrogen) as described above and plated onto selective pre-warmed plates containing 100 µg/ml ampicillin.

Plasmid was mini-prepped from selected colonies with the Qiagen Mini Prep kit according to manufacturer's instructions, and digested diagnostically with restriction enzymes to verify the correct insert size. The plasmid preps with correct insert size were sent for Sanger sequencing to Beckman Coulter Genomics using PQE-REV and PQE-FOR primers (Appendix Table 12) to verify correct sequence before downstream applications.

### ***2.3.15 Antibody concentration***

I have used both Amicon centrifugal filters (Millipore) and Protein G agarose (Innova Biosciences) for concentrating the hybridoma supernatants. A 100K cut-off Amicon filter was used which will allow the filtration of proteins less than 100 kDa, but will retain IgG which has around 160 kDa. First, 10 ml of supernatant were concentrated 20 x with the Amicon Ultra-15 Centrifugal Filter at 4,000 x g for 21 minutes at 4°C to a final volume of 500 µl. The column filter was 'washed' with PBS to collect the proteins immobilized on the membrane and the final volume made up to 700 µl. Thus the final concentration of the supernatant was ~14-fold.

The Protein G resin was pre-washed twice with PBS, 40 µl were added to each concentrate and incubated at 4 C overnight. The supernatant was kept as a concentrate of the original hybridoma supernatant, and the resin with the bound IgG was washed twice with PBS. The bound antibody was eluted with 20 µl 50mM Glycine, pH 1.9, and after mixing, was immediately added to a vial with 2 µl of 1M Tris-HCl, pH 8.0; 1.5M NaCl; 1mM EDTA; 0.5% sodium azide which brings the pH up to ~7.5. The elution step was repeated two more times with the same volumes, each fraction kept separately.

### 2.3.16 DNA ELISA assays

For the direct DNA coating assay, varying amounts of DNA were diluted in 50 mM acetate buffer pH 5.0, denatured at 99°C for 5 min, incubated immediately on ice for 10 min and loaded on the plate in decreasing dilutions, 50 µl/well. For quantitative comparison between samples OliGreen<sup>TM</sup> was added at a 1:800 dilution to the buffer and the amount of DNA imaged on a PHERAstar FS instrument (BMG Labtech) at 485nm excitation and 525nm emission.

For the avidin-biotin ELISA assay, the following conditions were used: 100 µl of 250 ng/well NeutrAvidin (Pierce/Thermo Scientific) were incubated in 50 mM Carbonate buffer pH 9.6 for 1 hour at 37°C or overnight at 4°C, followed by 1 hour at 37°C incubation of 50 µl of 10 ng/well biotin-labelled oligonucleotides in PBS.

The subsequent steps of both assays are the same: blocking for 1 hour at 37°C or overnight at 4°C with 2% BSA in PBS, incubation with a mix of primary and secondary antibodies for 1-2 hours at 37°C – the dilutions of the primary antibody are specified for each experiment in the results section, usually 1:2000, while the secondaries were used at 1:5000-1:10 000. The signal was developed with a 1 x TMB substrate solution (eBioScience) at RT, for 1 – 15 min and stopped with 1N sulphuric acid. For quantitative comparison between samples, the assay was developed with the chemiluminescent SuperSignal ELISA Femto Maximum Sensitivity Substrate (Pierce/Thermo Scientific). In both cases the signal was measured with a PHERAstar FS instrument. All assays were performed with flat bottom medium binding 96-well microplates (Greiner).

For the DNA binding experiment with recombinant WT and R111G MeCP2 ds oligonucleotides were used. Equimolar amounts of F and R oligos (F is 5'-biotinylated) were incubated at 99°C for 5 min and then allowed to re-anneal at RT for 1 hour. They were attached to the avidin plate as described above, using 40 ng/well. After the blocking, an additional binding step was performed with MeCP2 for 60-90 min at RT in the following binding buffer: 0.1% BSA, 0.5 mM DTT, 1 mM EDTA in 1x PBST. The rest of the protocol is the same as described above, but the antibody binding was performed in the binding buffer at RT. The result was developed with 1 x TMB substrate solution (eBioScience). The recombinant MeCP2 WT and KO proteins were kindly provided by Prof Sir Adrian Bird (University of Edinburgh). All antibodies used are described in Appendix Table 14.

### ***2.3.17 Immunofluorescence and cell imaging***

Cells were either grown directly on sterile cover slips or cytopspun onto microscope slides with a Shandon Cytospin 2 Centrifuge for 3 minutes at 300 rpm. They were fixed in 2% PFA for 30-60 minutes at room temperature, washed in PBS and permeabilised in PBS 0.5% Triton X-100 for 1h at room temperature (RT). After washing, they were blocked in PBS 0.05% Tween-20, 1% BSA (BS) for 1h at RT or overnight at 4°C and subsequently incubated 1 hour at RT or overnight at 4°C with a primary antibody in BS (see Appendix Table 14). After 30-60 minutes of washing, the cells were incubated with 1:500 BS dilution of a secondary fluorescently labelled antibody (Life Technologies Molecular Probes) for 30 minutes at RT in the dark. Cells stained for DNA modifications had an extra step of 2N HCl treatment for 30 minutes after permeabilisation, as described (Ficz et al. 2011), and an extra permeabilisation step with 2% PFA for 10 min if this was performed together with staining for a protein. After 30 minutes washing the cells were stained with DNA stain (DAPI or YOYO) and mounted with Vectashield (Vector) or SlowFade Gold (Molecular Probes).

The slides were imaged with either: 1) Zeiss 510 META Point scanning confocal equipped with Zeiss510 META Digital microscope camera AxioCam, 2) Olympus FV 1000 System confocal and 3) Nikon A1R MP Multiphoton Confocal Microscope, at 40x and 63x magnifications. Super resolution images were taken by Simon Walker on a Nikon SIM/STORM microscope.

Staining and imaging of oocytes and embryos was performed by Fátima Santos as published (Santos et al. 2013).

### ***2.3.18 Preparation of metaphase spreads***

Metaphase chromosome spreads were prepared as described (Novo et al. 2013) with slight modifications. WT ES cells were treated with 0.05 µg/ml colcemid for 30 minutes at 37°C. After harvesting they were incubated with hypotonic solution (8g/l NaCitrate, 75 mM KCl) for 20 min at 37°C and fixed with freshly prepared ethanol/acetic acid 3:1 at -20°C overnight. Metaphase chromosomes were spread onto glass slides and air-dried overnight. Air-dried slides were rehydrated and fixed in 2% PFA, and processed further as in 2.3.17.

### 2.3.19 cDNA synthesis from total RNA

All total RNA preparations before cDNA synthesis were treated two times with DNA-free™ DNase Treatment and Removal kit (Ambion) following the manufacturer's instructions (2-step treatment). cDNA synthesis was performed from 100-500 ng RNA with the RevertAid First Strand cDNA Synthesis kit (Fermentas) or SuperScript III First Strand Synthesis System (Invitrogen, Life Technologies) following manufacturer's protocol; random hexamers were used for gene expression or satellite analysis, and strand specific satellite oligos for the strand-specific major and minor satellite analysis (Appendix Table 12).

### 2.3.20 Quantitative PCR (qPCR) for gene expression

qPCR was set up manually or with Bravo Automated Liquid Handling Platform (Agilent Technologies). cDNA preparations for manual preparation were diluted 1:50 - 1:100 times in RNase-free water (provided with kit) and 5 µl per sample were used in triplicate for each quantitative real-time PCR reaction. For automated set up, cDNA was diluted 1:7 – 1:30 fold and 2.5 µl per sample were used for each qPCR reaction. The reactions were set up using either Brilliant II or Brilliant III Ultra-Fast SYBR Green QPCR Master Mix (Agilent Technologies), or Platinum SYBR Green qPCR SuperMix-UDG (Invitrogen, LT). qPCR was performed on Mx3005P™ (Stratagene) and C1000 Touch Thermal cycler 384 Real Time System (BioRad), the cycling conditions are outlined in Table 2. Primer pair efficiencies were calculated with the BioRad CFX Manager from a standard dilution curve for each master mix, and sample expression values were normalised for each pair.

Table 2. qPCR programmes outline.

	Brilliant II	Brilliant III	Platinum
Annealing	-	-	50°C for 2 min
Hot start	95°C for 10 minutes	95°C for 3 minutes	95°C for 2 minutes
Denaturation	95°C for 30 seconds	95°C for 8 seconds	95°C for 3 seconds
Annealing	58°C for 30 seconds	-	-
Polymerisation	72°C for 30 seconds	60°C for 15 seconds	58°C for 30 seconds
Number of cycles	40	40	40
Melting curve	Yes	Yes	Yes

For primer pairs with similar efficiencies, the ratios between Ct values of target genes were normalized relative to the Ct of a reference gene (Schmittgen & Livak 2008). All samples were normalised to at least one reference gene (Hspcb and Atp5b for gene expression), and subsequently to the value of the relevant study control. All primer sequences are listed in Appendix Table 12.

### ***2.3.21 Dot blot for measurement of total 5mC or 5hmC***

DNA was denatured at 99°C for 5 minutes and spotted onto Hybond<sup>TM</sup>-N+ nitrocellulose membrane (GE Healthcare). The membrane was UV cross-linked and incubated overnight with 10% non-fat milk and 1% BSA in PBT (PBS + 0.1% Tween20) at 4°C followed by >1 hour incubation with either 1:250 anti-5mC or 1:500 anti-5hmC antibody (see Appendix Table 14). Membranes were washed 4x with PBT, incubated for 30 minutes with HRP-conjugated anti-mouse or anti-rabbit antibodies (GE Healthcare; 1:10,000 in blocking solution), washed with PBT, and developed using the ECL Plus detection system (GE Healthcare). The membranes were exposed to a high performance chemiluminescence film (Hyperfilm<sup>TM</sup> ECL, Amersham).

### ***2.3.22 Methylated DNA Immunoprecipitation assay (MeDIP)***

For the 5-methylcytosine context-specific quantitative MeDIP-Seq experiments, 1.0 µg of genomic DNA was used as input material. Fragmentation was performed on a Covaris E220 instrument with the 150 bp programme, in a total of 85µl. For end repair, A-tailing and adapter ligation of the input material, the NEB Next DNA Sample Prep Master Mix Set for Illumina was used, as described in 2.3.7. Each input sample was barcoded with Illumina TruSeq adapters as indicated in Appendix Table 15. All reactions were cleaned up using AMPure XP beads (Agencourt) and eluted in a final volume of 30-50µl with buffer EB (Qiagen). All adapter-attached samples were quantitated on an Agilent 2100 Bioanalyzer system and subjected to a test PCR reaction for verification of the efficiency of adapter ligation (as in 2.3.8, but with NEB Next 2x Phusion mix).

Equal amounts of the measured input material from the DNAs of interest were then mixed, denatured at 99°C for 10 minutes in a heating block and immediately put on ice to prevent re-annealing. After 10 minutes, 1:10 µl of 10x IP buffer (100 mM Na-Phosphate pH 7.0, 1.4 M NaCl, 0.5% Triton X-100) and 1:100 µl of a mouse monoclonal antibody against 5meCA or

5meCG (Reik lab, Babraham Institute) were added. Samples were incubated with the antibody at 4°C for two hours on a rotating wheel. To capture the DNA-immuno-complex, 40 µl of Dynabeads Protein G (Novex/Life Sciences) were washed and blocked for two hours at 4°C with PBS-BSA 0.1%, and added to the DNA-antibody IP mix. After a further incubation at 4°C for two hours, the beads were collected with a DynaMag<sup>TM</sup>-2 Magnet (Life Technologies) and the supernatant taken out. After three washes at RT with 1x IP buffer, the beads were resuspended in 200 µl proteinase K digestion buffer (50 mM Tris pH 8.0, 10 mM EDTA, 0.5% SDS). After this, 35 µg of proteinase K (Roche) was added and the samples were incubated at 55°C for 30 minutes using a shaking heating block at 800 rpm. The beads were collected with the magnet, the bound fraction was taken out and cleaned up with Agencourt® AMPure® XP beads, eluted in 20 µl RNase free water (Qiagen). For enrichment of adapter-ligated fragments, PCR was performed with NEB Next 2x Phusion mix: volume 50 µl, 5 µl DNA (pulled down or input), 200 nM TruSeq PCR primers 1.0 and 2.0, programme conditions as in 2.3.7 for 12 cycles for mCA and mCG pull-downs and 6 cycles for the input material.

The amplified libraries were purified with Agencourt® AMPure® XP beads, processed as described for BS-seq in 2.3.7 and sequenced at the WTSI.

### ***2.3.23 Protein binders pull down from nuclear extract***

The protein pull down from ES cell nuclear extracts was performed by our collaborator Michiel Vermeulen as described (Spruijt et al. 2013).

### ***2.3.24 Tet1 oxidation assay***

Recombinant active Tet1 protein was purchased from Active Motif. Equimolar amounts of F and R oligos (F is 5'-biotinylated) were incubated at 99°C for 5 min and then allowed to re-anneal at RT for 1 hour. The oxidation reaction was performed as instructed by the manufacturer, in 10 µl reaction volume: 200 ng DNA, 2.5 µg Tet1 in 50 mM HEPES pH 6.8, 50 µM Fe(NH<sub>4</sub>)<sub>2</sub>(SO<sub>4</sub>)<sub>2</sub>, 2 mM ascorbate and 1 mM α-ketoglutarate, and incubated for 3 hours at 37°C.

A quarter of the reaction volume (2.5 µl, 50ng oligos) were diluted in 200 µl of EB buffer (Qiagen) and denatured at 99°C for 5 min, incubated immediately on ice for 10 min and loaded on a pre-coated with NeutrAvidin polystyrene plate in 6x dropping dilutions with EB buffer,

50 µl/well. The rest of the ELISA was performed as in 2.3.15, incubated with either the anti-mC or anti-hmC antibodies (Appendix Table 14), and developed with the chemiluminescent SuperSignal ELISA Femto Maximum Sensitivity Substrate from Pierce/Thermo Scientific.

### **2.3.25 FACS-sorting of ES cells**

J1 ES cells were grown on feeder cells in standard serum + LIF conditions as described in 2.1, in 6 x 15 cm dishes to 60-70% confluence. Prior to FAC-sorting, they were trypsinised and incubated in dense concentration on 2 x non-gelatinised 15 cm dishes for 35 min for the feeders to attach. They were then spun at 300g x 3 min and washed in 1 x PBS, resuspended in 5 ml cold 70% ethanol and incubated for 3 hours at 4°C. They were then washed in 5 ml PBS and incubated for 10 min at RT, spun and resuspended in 1 ml PBS and sieved through a 40 nm filter. After 10 min at RT 80ul of 0.5 mg/ml PI + 0.6% Igepal-630 were added to the cells and left for 2 hours at RT in the dark or o/n at 4°C.

Where indicated, EdU pulse labelling was performed for 15 minutes at 37°C at 20 µM final EdU concentration, followed by harvesting and cell fixation in 70 % ethanol.

The cells were sorted with a FACS Aria system and the sorts kept at 4°C prior to further manipulation.

### **2.3.26 Whole Genome Amplification (WGA)**

Whole genome amplification was carried out with Qiagen REPLI-g Mini kit, following manufacturer's instructions. For mouse gDNA, 380ng of starting material were used, yielding ~12ug of WGA DNA, while for human gDNA - 90ng starting material was used, yielding ~7.4ug WGA DNA. The WGA DNA was used without further purification, from a 1:10 working stock, and further dilutions towards its final applications.

## **2.4 Bioinformatics**

### **2.4.1 Data mapping**

In all the various analyses of genome-wide datasets, the mouse NCBIM37 genome build was used as the reference genome. Data mapping was carried out by Felix Krueger, Simon

Andrews or Phil Ewels (Bioinformatics Facility, Babraham Institute) using Bowtie (Langmead et al. 2009) or Bismark (Krueger & Andrews 2011). Mapped reads were further analysed with the SeqMonk software developed by Simon Andrews (<http://www.bioinformatics.babraham.ac.uk/projects/seqmonk/>) and custom-made R and Perl scripts.

#### ***2.4.2 FastQC analysis of raw and mapped reads data***

Each NGS dataset generated in this study was subjected to a raw data quality analysis with the BI-developed tool FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>) by the staff at the BI Bioinformatics facility. Additional FastQC analysis of the mapped reads of selected published datasets was performed by Felix Krueger and the further analysis for the bisulphite degradation study was carried out by the student.

#### ***2.4.3 Repeat consensus analysis***

The repeat analysis for mouse satellites and other repeats was carried out by Felix Krueger, including the telomere degradation analysis. A similar pipeline was performed for unconverted MeDIP data, as well as for the WGBS-seq and MspJI meRRBS datasets. Depending on the type of library, either Bowtie or Bismark were used to align the trimmed raw reads against a consensus repeat sequence for each repeat class. The alignments were performed with high stringency allowing for only one base mismatch ( $n=1$ ). For MeDIP data, the percentage of reads aligning to a repeat sequence was scored, separated by strand, and for BS-Seq data the methylation results were assessed for each cytosine from the consensus sequence. The methylation analysis was done by the student in SeqMonk: a custom-made genome was used for each repeat consensus, probes were generated for each base as read position, and methylation was calculated as percentage of methylated calls versus total cytosine calls for each probe. The data was exported and analysed further in Excel and visualised with GraphPad Prism 6.0.

#### ***2.4.4 Analysis of MeDIP and WGBS datasets***

The feature and global methylation analysis was carried out with SeqMonk by the student. Phil Ewels helped with calculating cytosine content and context distribution in selected features

used for the analysis of MeDIP and MspJI data.

For SeqMonk analysis of MeDIP-seq data, tiling probes were created for every 500bp without overlap. Feature probes coordinates were extracted from the NCBI37 annotation, except for promoters which were determined as spanning -1000 to +200 around the transcription start site and CG islands were annotated based on published coordinates (Illingworth et al. 2010). Feature read numbers were normalized to the total read number per dataset. Extreme outlier probes with a log<sub>2</sub> value above 10 were excluded from further analysis as they likely represent mapping artefacts (Ficz et al. 2011). Highly asymmetric peaks were selected ('Forward-biased' and 'Reverse-biased') on regions covered by at least 10 reads, and strand to strand ratio higher than log<sub>2</sub> value of 0.5, meaning one strand has to be at least three times more represented than the other. Peak enrichment was normalized to the unbound and represented directly or as log<sub>2</sub> enrichment.

For BS-seq analysis, probes were generated over selected genomic features in SeqMonk and methylation was calculated as a percentage of positive (methylated) calls versus all calls for each individual probe, and then averaged for the feature. Initial filtering of outliers was performed as for MeDIP-seq, and those regions were excluded from further analysis.

The analyses over high depth WGBS datasets (Lister et al. 2009; Stadler et al. 2011) were carried out by Felix Krueger or Simon Andrews (BI Bioinformatics) with custom-made scripts.

#### ***2.4.5 Analysis of ChIP datasets***

ChIP (Chromatin Immuno-Precipitation) data peak calling was performed by the student with the MACS peak caller (Zhang et al. 2008) integrated in SeqMonk, and peak coordinates exported for further analysis. Whole cell extract (WCE) or Input datasets (provided with published ChIP datasets) were used for normalisation for the MACS peak calling. Outlier peaks with high read coverage potentially due to sequencing artefacts were removed. For mCA or mCG enrichment, the ChIP peaks were overlaid with mCA or mCG peak coordinates and enrichment was calculated by the 'Quantitation trend plot' feature over either mCA or mCG in SeqMonk.

### **2.4.6 *In silico* digestion**

*In silico* digestion of the mouse genome by restriction enzymes was performed by Felix Krueger.

### **2.4.7 *MspJI-RRBS* base-calling**

The base-calling was carried out by Felix Krueger. The justification for strand and read usage and final ‘true positive’ call selection, was carried out by both the student and Felix Krueger, based on the nature of *MspJI* digestion and library read generation. As a result, after adapter trimming, only position #16 from read 2’s were selected as ‘true positive’ methylation calls for further analysis (read 1’s contained an error of the unspecifically digested 5’ ends). The rest of the analysis was performed as described in 2.4.4 for WGBS.

### **2.4.8 *WRC* sequence analysis**

The AID consensus sequence (*WRC*) analysis was carried out by Simon Andrews (Head, BI Bioinformatics) with custom R and Perl scripts.



### **3 Genomic distribution of non-CG methylation: a classical approach**

#### **3.1 Introduction**

At the time of starting this project, only one study had attempted a detailed investigation of the whole genome distribution of this type of methylation, and it was done in human cell lines (Lister et al. 2009). In order to approach the question for its biological relevance in the mouse as a model animal, a good starting point was to analyse the non-CG 5mC genomic distribution in the mouse, and Lister's data could provide a good point of comparison, since it is done on embryonic stem cells and fibroblasts. Another important factor for my analysis, though, have been the existing scepticism and doubts within the scientific community, related to a possible artefactual nature of the non-CG context methylation, and the fact that it had traditionally been used for estimation of conversion efficiency, and not as an actual signal in itself (Harrison et al. 1998; Araujo et al. 1998; Laird 2010; Genereux et al. 2008). In order to move this ongoing discussion further, it therefore seemed necessary to not only analyse existing non-CG methylation signals in the genome, but to address the issues around BS-conversion artefacts and the limitations of the existing techniques, finding ways to validate the real non-CG 5mC signals.

As discussed in 1.6, the techniques to study non-CG methylation are quite limited; however, a few have been successfully used to obtain qualitative and quantitative information of the cytosine neighbouring context and they are: bisulphite (BS) conversion and sequencing, nearest neighbour analysis (NNA) and a CCWGG modification of the LUMA technique. They seemed good candidates to provide both quantitative data on the global levels of genomic non-CG methylation (with LUMA and NNA), and also to obtain details on the genomic distribution of non-CG methylation in ES and differentiated cells (with BS sequencing). It is important to remind, however, that the genome wide BS-seq technique has to date been used to analyse non-CG genomic distribution only on high depth sequencing datasets (min 10-fold coverage) (Lister et al. 2009; Laurent et al. 2010; Shirane et al. 2013; Lister et al. 2013), which were not an option for this study. Because of the different concerns raised in 1.6 regarding the use of bisulphite conversion and the feasibility of low depth BS-seq datasets to deliver valid results for non-CG context, it was necessary to also address these issues and evaluate the suitability of the approach.

In addition, it was appropriate to employ and analyse other genome-wide data like datasets

of mouse ES cell lines obtained with MeDIP-seq, which have been available in the lab, to indirectly assess the distribution of asymmetric methylation, as has been suggested previously (Ficz et al. 2011).

## 3.2 Aims

1. To analyse indirectly genome-wide distribution of asymmetric methylation in mouse ES cells from the available MeDIP-seq datasets and validate the feasibility of using this approach
2. Perform a direct analysis of the genome-wide distribution of asymmetric methylation in mouse ES cells and differentiating primary embryonic fibroblasts (pMEFs) with whole genome low coverage BS-seq data and use that to validate the MeDIP-seq approach
3. Investigate the feasibility of using the low resolution BS-seq approach by addressing 1) conversion artefacts and 2) DNA degradation and amplification sequence biases
4. Assess global genomic levels of non-CG methylation in WT mouse ES cells, pMEFs and tissues using NNA and CCWGG-LUMA

For the purpose of the genome-wide analysis carried out in this and the following chapters cytosine methylation will be presented in three sequence contexts – CG, CHG and CHH, where CHG and CHH are both qualified as ‘non-CG’ (H being A, T or C according to universal nomenclature).

## 3.3 Results

### 3.3.1 *Non-CG methylation in MeDIP-seq*

One way to assess the patterns of genome-wide DNA methylation is to first enrich for the methylated fraction of the genome and subject only this portion of the genome to high throughput sequencing. This strategy might have considerable advantages specifically for low levels of methylation like in the non-CG context, and does not have the disadvantage of a chemical treatment and incomplete conversion. As discussed in 1.6 (Introduction), the two most commonly

used techniques rely on pulling down methylated DNA using either a methyl-DNA binding protein (MBD – MBDCap, MeCP2 – MeCAP) (Serre et al. 2010; Brinkman et al. 2010) or an antibody against 5-methylcytosine (MeDIP) (Weber et al. 2005; Down et al. 2008). The MBDCap/MeCAP recognise CG context only, while the MeDIP recognises methylated cytosine in any sequence context (Zilberman & Henikoff 2007; Robinson et al. 2010; Bock et al. 2010), making it a more appropriate enrichment technique for this study. There are clear advantages of this method over the BS-seq: only a small fraction of the genome is sequenced which leads to high coverage over all methylated regions, highly important for the non-CG context; and this is achieved with a relatively small sequencing effort, which could enable sample multiplexing. The downside of including an enrichment step, however, is that any direct quantitative information is lost and precise levels of methylation cannot be assessed as for BS-seq.

Since the MeDIP technique does not provide information on individual bases, on their neighbouring context and methylation status (the pulled down sonicated DNA fragments have an average length of 250bp), in order to be able to interpret this data, I have to look at asymmetrically represented genomic regions, which could be a sign of asymmetric methylation. In order to fulfil this requirement one of the DNA strands has to be pulled down more than the other, because of the asymmetric nature of non-CG methylation in mammals as discussed in 1.2 (Guo et al. 2013; Ziller et al. 2011). To validate this approach, I looked at the ratio of pulled down reads from several repeat classes using published repeat consensus sequences (listed in Appendix Table 16). The raw reads were aligned to the consensus sequence and the proportion of forward and reverse strand reads estimated in a number of MeDIP datasets (a schematic of the paired end technology which preserves original strand information is presented in Appendix Figure 71). I compared a WT mES cell line (J1) with a panel of differentiated samples: embryonic bodies (EB) derived from the J1 ES line, primary mouse fibroblasts (pMEFs) and sperm, together with an unmethylated control of a Dnmt3a/3b/1-KO ES cell line (TKO), which were already available in the lab and subsequently published (Ficz et al. 2011). The TKO line does not have any 5mC (Kaneda et al. 2004; Raddatz et al. 2013) and has been widely used as a negative control in methylation studies. I tested four repeat classes; for three of them (LINE1 5'UTR, SINE1B and IAP LTR1a) the strand representation was very symmetric, while the major satellite repeat showed a very strong strand asymmetry - the reverse strand was pulled down several-fold higher than the forward strand (Figure 7A). The unexpected observation was that the strand asymmetry

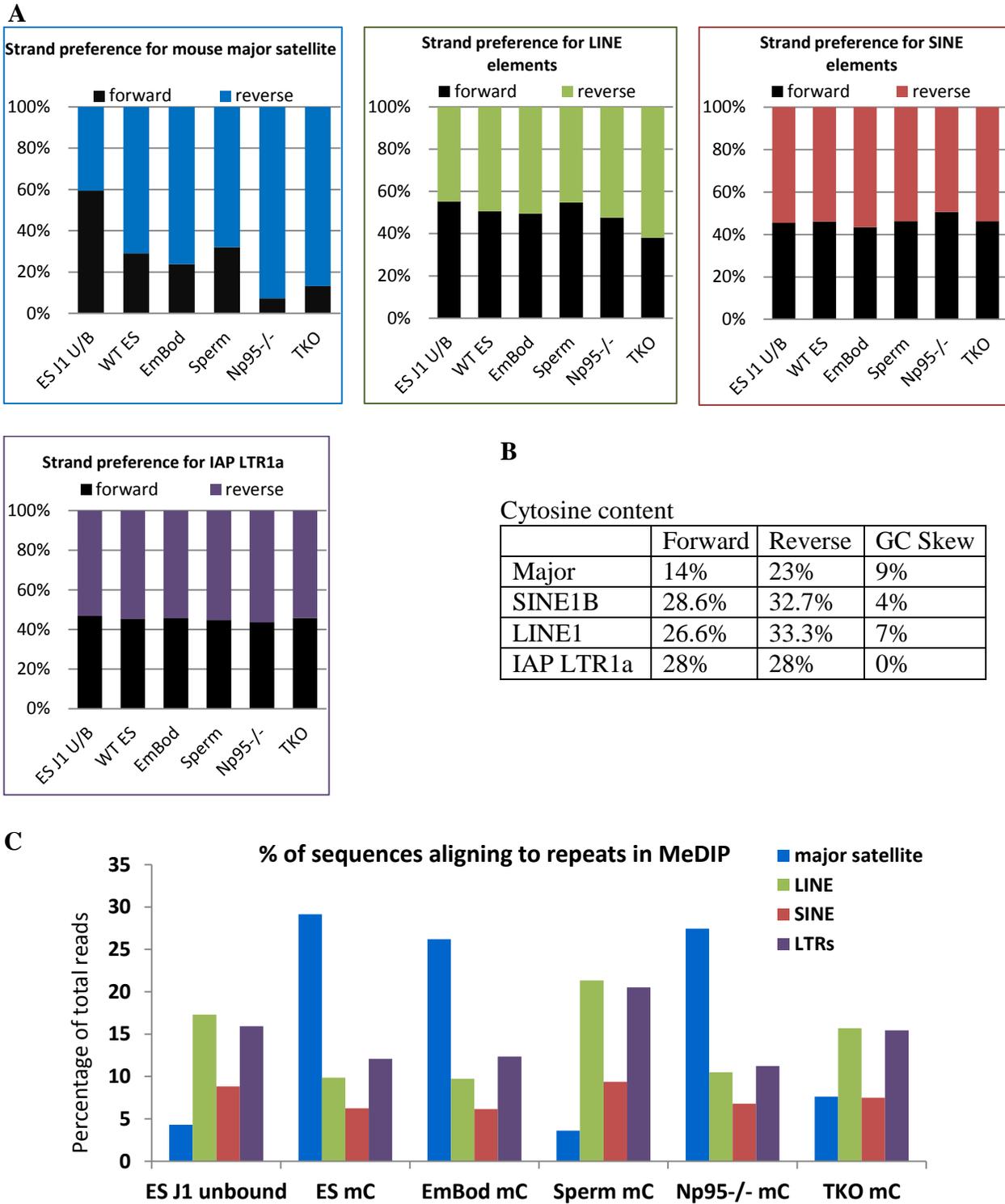


Figure 7. Analysis of MeDIP strand asymmetry. **A**. Asymmetric strand representation of repeats in the MeDIP pulled down DNA fraction in a panel of ES cell lines and tissues. **B**. Percentage of cytosine distribution between the two strands of each repeat consensus. **C**. Percentage of reads in the MeDIP datasets showing the fraction of each repeat in the pull down.

was strongest in the unmethylated TKO sample, suggesting that this can be merely an artefact and not a consequence of underlying asymmetric methylation. I calculated the cytosine content and it showed that the major satellite also has a cytosine strand asymmetry and the pulled down reverse strand is the cytosine-rich strand, composed of 65% more cytosine than the C-poor forward strand (Figure 7B). The remaining three analysed repeats have overall higher C-content or a relatively uniform cytosine distribution on both strands. Another important point is that the satellite repeat itself constituted a large fraction of the total reads in most MeDIP datasets, pointing out to a potential amplification artefact (Figure 7C), especially in that the pull down ratios do not reflect directly the samples' methylation levels – the sperm sample (very methylated) has a low satellite proportion while the relatively unmethylated Np95-KO had pulled down as much as the WT samples.

I decided to validate this result with a classical targeted bisulphite sequencing. I designed individual primer pairs for the amplification of bisulphite converted forward and reverse strands, which become non-complementary after bisulphite conversion and require separate amplification (see Appendix Table 12). These oligos yielded several sizes of major satellite bands for J1 ES cell and TKO gDNA (as a conversion control) and I cloned the band which spans for one and a half satellite consensus repeat – ~370bp. The results are shown in Figure 8. Both strands showed presence of non-CG methylation (in the CA context), which was more pronounced on the reverse strand (2.1% reverse vs 0.88% forward, conversion rate 99.71% as determined by the TKO). These findings therefore confirmed that there is an existing link between asymmetrically pulled down DNA strands and asymmetric methylation, although from the findings in Figure 7A and B, we know that there also is a clear link, between cytosine content and asymmetric strand enrichment in the MeDIP datasets.

To further address those issues and test the MeDIP approach, I performed a genome-wide feature analysis on ES cell lines with variable methylation levels – J1, E14 (WT), Np95-KO (reduced maintenance methylation) and TKO. Plotting the percentage of symmetric and asymmetric peaks for each dataset revealed that the less methylated genomes like the Np95-KO and the TKO had more asymmetric peaks, this value reaching almost 100% in the TKO (Figure 9A).

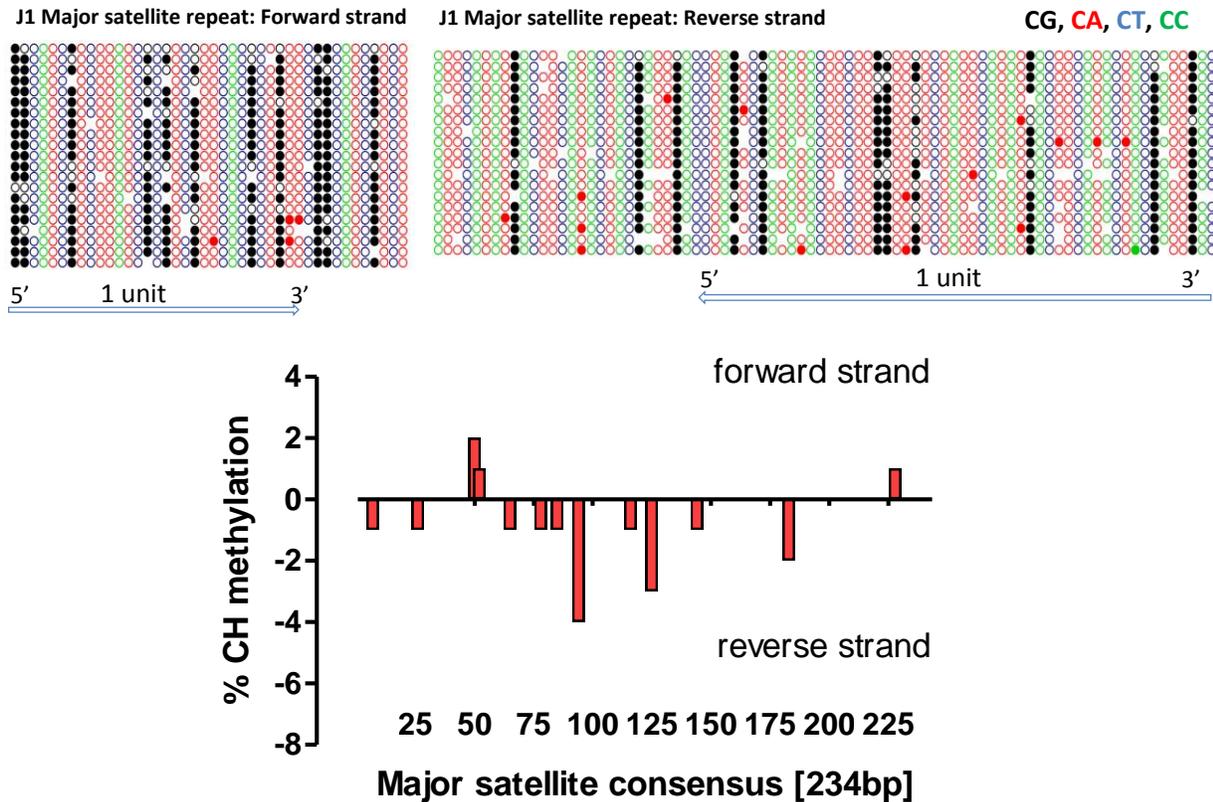


Figure 8. Bisulphite conversion and cloning of major satellite single repeat. J1 ES cell line forward strand (left upper) and reverse strand (right upper); the span of one unit is shown, with a coloured legend for each context. Lower panel – a quantitative representation of methylation positions and their methylation values for the satellite monomer on both strands; Note: the values on the reverse strand are not negative!

I then looked at the peak overlap with selected genomic features with coordinates obtained through SeqMonk directly from Ensembl Genome Browser. Some features, not available on Ensembl, such as ‘Introns’, ‘Intragenic regions’, ‘Promoters’ etc were custom made in SeqMonk as explained in 2.4.4. I calculated the number of peaks per feature, and normalised this to the unbound DNA fraction, to the cytosine content of each feature (which correlates to its length), and to the total number of reads per dataset. The results showed that the TKO unmethylated control had the highest relative enrichment for some of the features, but not for others (Figure 9B). Those features were mainly simple repeats like microsatellites and telomere repeat, which are very likely to be artefacts, especially given how C-rich they are – 50% for telomeres and some C-containing microsatellites. I therefore removed from the analysis the repetitive features where the negative control shows more than two-fold higher enrichment than the input (value of 1.0 is equal to input).

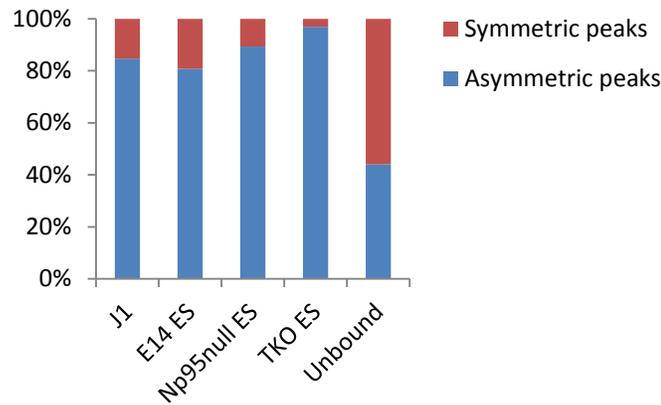
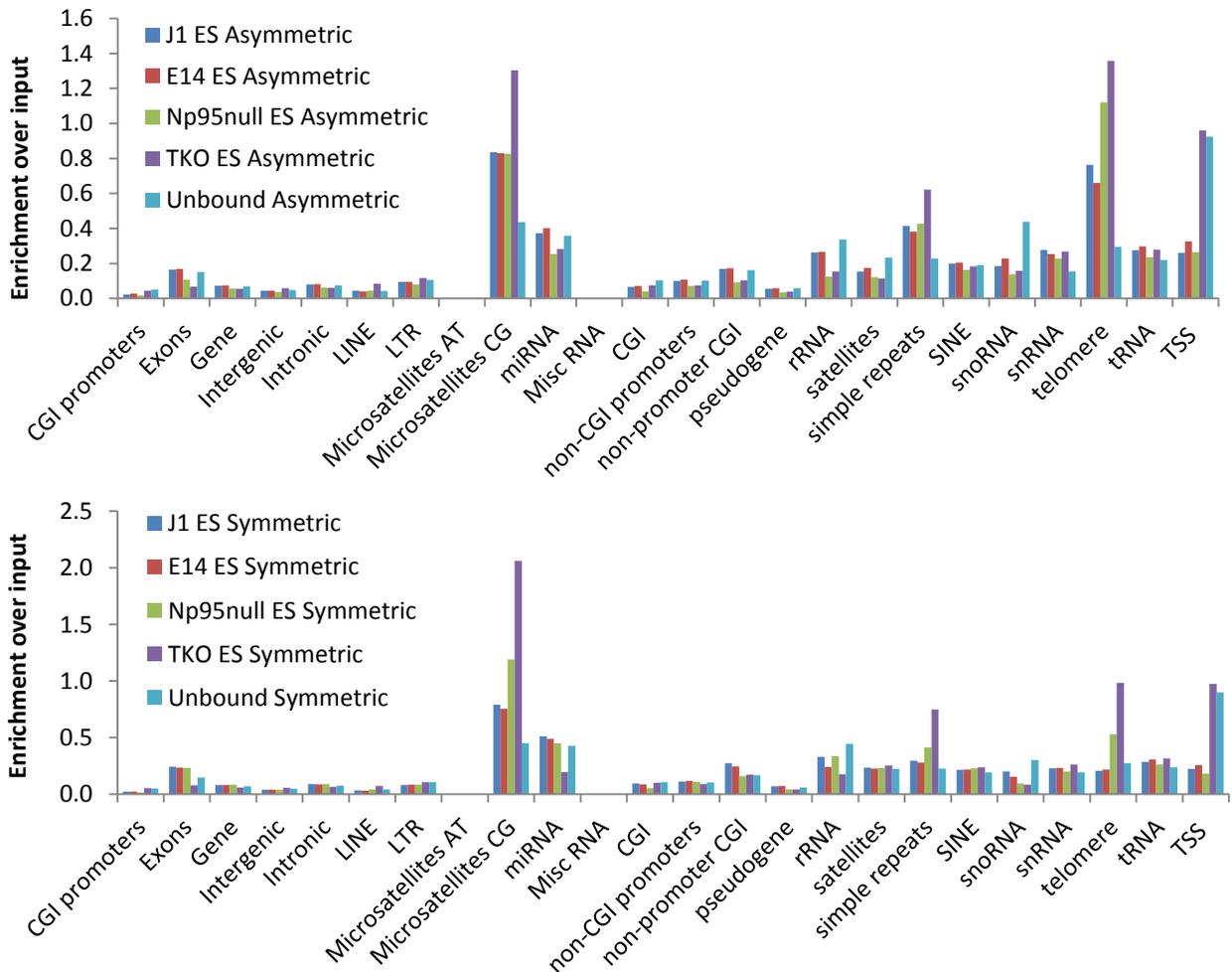
**A****B**

Figure 9. MeDIP-seq symmetric and asymmetric peak analysis. **A.** Percentage of symmetric and asymmetric peaks in the MeDIP-seq datasets included in this analysis. **B.** Feature analysis of peak enrichment, normalized to the unbound fraction, to peak counts per dataset and to the cytosine content of each feature (reflecting its genomic size). Each dataset represents pooled data from two biological replicates, published in (Ficz et al. 2011), and the TKO dataset was produced by the same authors.

I then analysed further the remaining features representing them as a  $\log_2$  of the values in Figure 9C, so that the enriched regions would appear as positive and the depleted as negative values along the Y axis (Figure 10). The results showed that the asymmetric feature enrichment pattern follows the pattern of the symmetric peaks, but with much lower values, showing that either both peak types reflect the CG methylation, or that the pattern of asymmetric methylation follows the pattern of CG methylation. Indeed (Ficz et al. 2011) have shown (fig. 3a and c) that a high proportion of the CG-enriched peaks are asymmetric (referred to in the paper as ‘biased’ from ‘strand biased’). The highly enriched features in WT ES in Figure 10A, such as exons, and introns, indeed have also been reported as enriched by the authors (Ficz et al. 2011, supplementary materials), while the CGIs, promoters, intergenic sequences, LTRs and LINEs are less enriched (Figure 10A). In addition, my results showed that some RNA loci were enriched over the background, the SINEs were also enriched in both peak fractions as well as the non-promoter CGIs were enriched in the symmetric methylation as expected. Interestingly, the mapped satellite units did not show enrichment, which is probably due to the initial data filtering. Only the snRNA and tRNA loci showed a higher enrichment in the asymmetric fraction, however the TKO showed similar enrichment, pointing to a potential artefact. Comparing the asymmetric peak distribution of WT and methylation compromised ES cell lines (Figure 10 B and C) revealed that the asymmetric fraction in general had very low enrichment in most features in comparison to the symmetric peaks. This means that those asymmetric peaks followed the dynamics of the asymmetric peaks of the unbound fraction to which they were normalised, and therefore most likely reflect amplification biases rather than real asymmetric methylation.

The fact that the completely unmethylated genome of the TKO yield many reads from the antibody pull down, which are exclusively asymmetric, shows that the pull down artefacts would manifest as asymmetric peaks. The unbound control on the other hand, which should represent an unbiased genomic sample, also shows around 45% of asymmetric peaks, while it is expected to have entirely symmetric unbiased strand distribution. This shows that alongside the antibody pull down artefacts, there are, in addition, polymerase amplification artefacts, potentially created by the known polymerase bias discussed in 1.6.2.4. As shown in (Ficz et al. 2011, fig. 3c), the composition of the asymmetric peaks is very rich in CH cytosines, which could explain both the antibody unspecific C-affinity pull down, in the absence of its endogenous 5mC target, and also

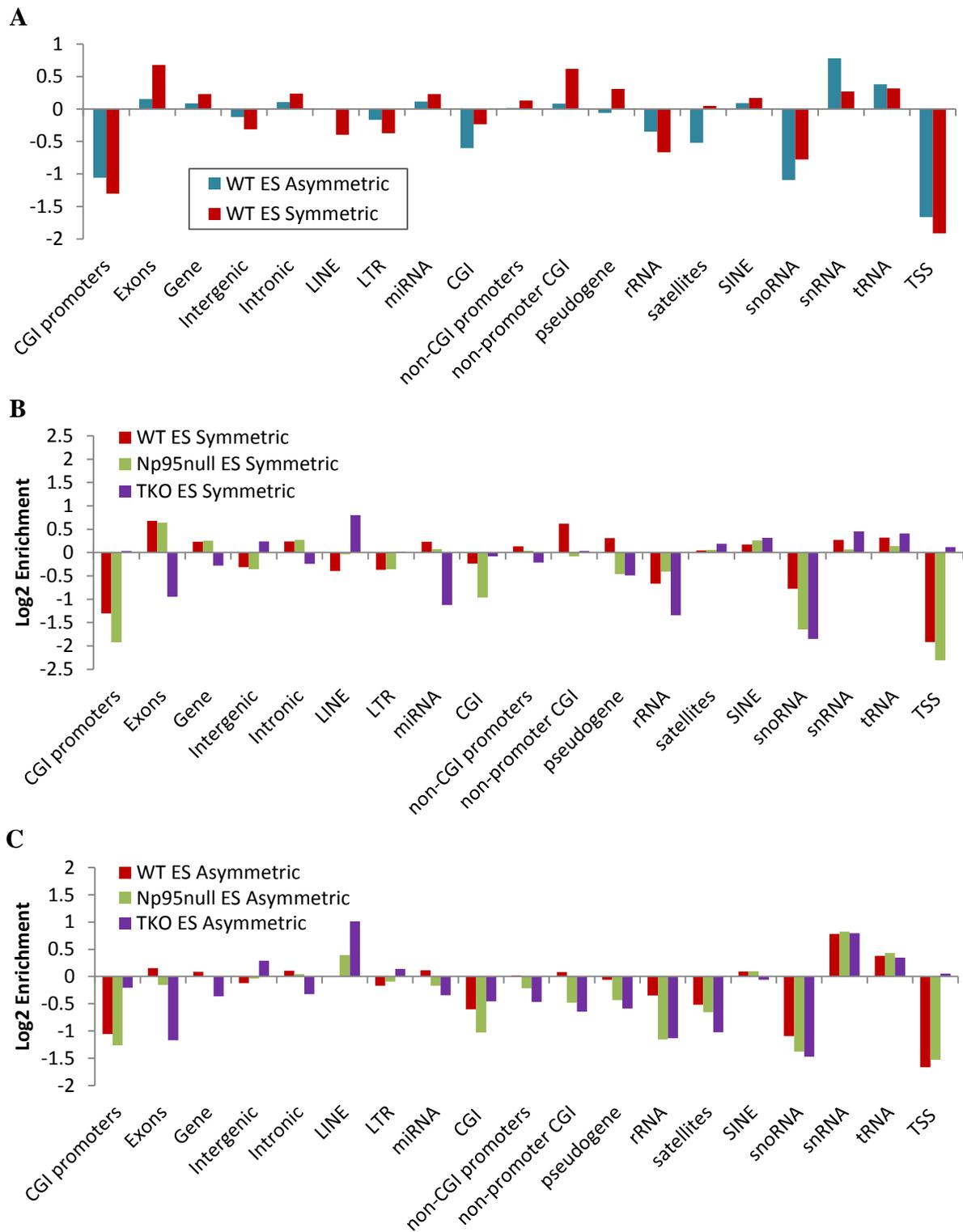


Figure 10. Feature enrichment of asymmetric and symmetric peaks, normalized to unbound genomic fraction. **A.** Comparison of the enrichment of symmetric and asymmetric peaks in WT ES cells. **B., C.** Enrichment of the symmetric (B) and asymmetric (C) peaks in WT and KO ES cell lines.

the polymerase preference to over-amplify C-rich regions. As shown in Figure 10 B and C, the asymmetric peaks which were not pull down artefacts overrepresented in the TKO samples, disappeared after normalisation to the unbound fraction and do not show any high specific feature enrichment.

Taken together, the results from the MeDIP-seq data analysis show that it is very likely that a high proportion of the asymmetric peaks are pull-down or amplification artefacts, while another proportion indeed reflects existing underlying methylation, predominantly in CG context. A proportion of the asymmetric peaks undoubtedly reflect non-CG methylation, as confirmed by bisulphite analysis, which however would be impossible to track among the artefacts and the asymmetric CG methylation peaks. Therefore, the MeDIP-seq datasets generated with a general 5mC antibody, although containing both CG and non-CG methylation, as expected, are not suitable for the exclusive analysis of non-CG methylation.

### ***3.3.2 Low resolution BS-seq datasets: non-CG BS-conversion artefacts***

In order to characterise the distribution of non-CG methylation, an affordable technique should be available, which can be applied to a large number of samples and not have the limitations of cost and computing power of the high coverage BS-seq datasets. I therefore focused my next analysis on low coverage BS-seq datasets, generated previously in the lab (Seisenberger et al. 2012; Popp et al. 2010), together with datasets sequenced for this project (Raddatz et al. 2013). I first quantitated total methylation calls in the CG, CHG and CHH contexts. Filtering was applied for outliers – regions with very high read coverage, which could be either amplification artefacts, or reads from the major satellite units mapping to regions of interspersed repeat sequence, which do not actually belong to the particular locus. No filtering was applied for the exclusion of methylated cytosine calls as it was done in (Lister et al. 2009) due to the low coverage, and because the initial purpose of this study was to evaluate the contribution of conversion artefacts to the analysis of non-CG context methylation in low resolution datasets. The first results showed a higher amount of total CHG and CHH methylation in ES cells (36%) consistent with expectations. The primary mouse fibroblasts (pMEFs) contained less but nonetheless considerable amount of 5mC in non-CG context (22%) compared to the published human (0.02%) data (Lister et al. 2009) (Figure 11A).

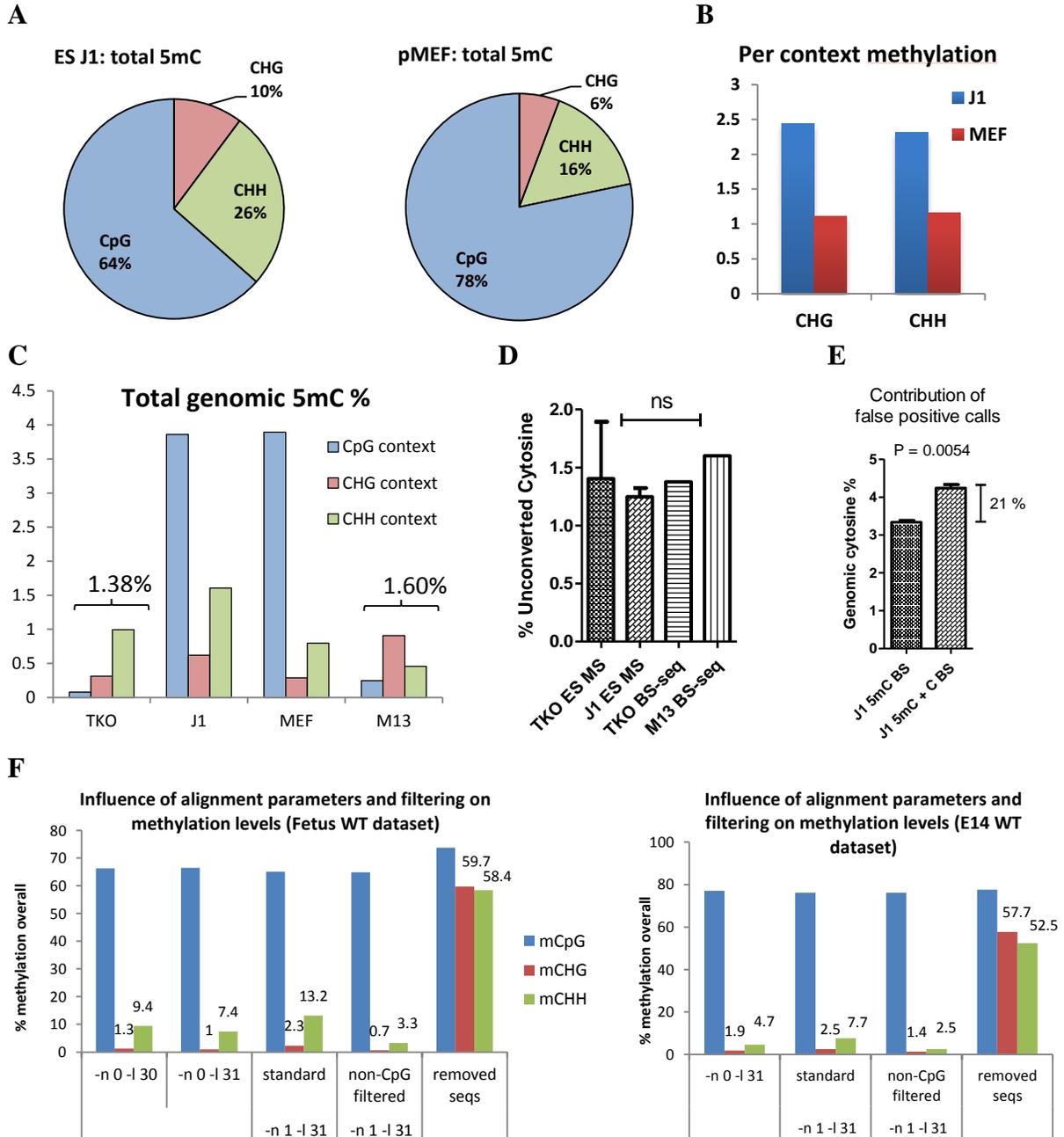


Figure 11. Methylation levels in CG, CHG and CHH contexts in low coverage BS-seq datasets and conversion artefacts. **A**. total proportion of positive calls in each context for the whole dataset normalised to total read count. **B**. CHG and CHH methylation as percentage of each context. **C**. Global levels of methylation per context presented as a percentage of total cytosine content for each dataset. **D**. Cytosine conversion efficiency after bisulphite treatment of the WT J1 and TKO ES samples as measured by a mass spectrometer (LC/MS) and BS-seq. **E**. Contribution of false positive calls (unconverted cytosine) to ‘global’ methylation after bisulphite treatment as measured by LC/MS. **F**. Applying various mapping stringencies and non-CG filtering for two datasets changes the estimate of non-CG but not of CG methylation; “n” is the number of allowed mismatches, and |30/31 is the number of consecutive bases for which the mismatch rule is applied. Standard conditions used in our facility’s pipeline are n = 1 mismatch per 31bp (length of 1 x Illumina GAllx sequencing read).

Similar results were observed when 5mC calls were calculated as a percentage per context - the methylation in pMEFs is still around half of that in ES cells (Figure 11B). To address the question whether this is a real estimate, or the direct result of a conversion artefact, I sequenced the methylome of the Dnmt3a/3b/1-KO ES cell line (TKO), which has no methylation (Raddatz et al. 2013). An M13 2kb PCR fragment was spiked in as an additional conversion control as advised in (Krueger et al. 2012). The total levels of CHG and CHH positive calls in both the TKO ES DNA (1.38%) and the M13 DNA (1.6%) were comparable and slightly higher than the pMEF DNA (1.25%), confirming that those levels of non-CG methylation can be attributed entirely to false positive calls (Figure 11C). I then measured by mass spectrometer the level of unconverted cytosines of bisulphite-treated J1 and TKO genomic DNA, and the amount of unconverted cytosine in both samples was between 1 and 2 % of the total, confirming the observed false positive levels in the TKO and M13 samples from BS-seq (Figure 11D). Thus the incomplete conversion contributes significantly to the false discovery rate of BS-dependent platforms where the unconverted cytosines will be interpreted as methylated cytosines and add up to more than 20% to the total 5mC value (Figure 11E). Lastly, we tried enhancing the alignment stringency of the low coverage BS-seq datasets and applied filtering of non-CG context as described (Lister et al. 2009). The results showed that any change in mapping conditions and filtering, affected dramatically the methylation values in both CHG and CHH context, but not the levels of mCG (Figure 11F). This points to an additional issue that the non-CG context positive calls, real or artefactual, can be manipulated by the data processing parameters. The real methylation values are also affected by this, as seen in (Figure 11E), where the E14 sample almost 'loses' its non-CG methylation with more stringent mapping parameters.

In search for possible validation approaches of the low level methylation signal in non-CG context in base resolution datasets, I looked at C>T transitions (or single nucleotide polymorphisms, SNPs) in non-CG context between the C57BL/6 J (B6) genomic annotation used for mapping and the sequenced J1 mES cells of the 129S4/SvJae mouse strain. It is known, that methylated cytosines cause higher rates of C>T transitions, which contribute to disease and in the context of evolution have led to the depletion of CG in mammalian genomes (Cooper et al. 2010). Therefore sites of C>T SNPs could be mapped and used to validate positions of real methylation as opposed to conversion artefacts. My analysis revealed that C>T transitions in the CG context

were indeed comparatively higher methylated (83% versus 74% average CG methylation) (Figure 12A). Among C>T SNPs in CH contexts, CA showed the highest 5mC value (3.6% out of 2.9% on average), which was a too low difference, however, to be reliably used as a validation marker. Curiously, post-mutation calculations revealed that methylated CG sites give rise to CA sites after deamination, which in its turn give rise to TA sites, providing a plausible explanation of the predominance of AT content (~60% for mammals, Appendix Table 11) in the genomes of organisms possessing CG DNA methylation (Figure 12B).

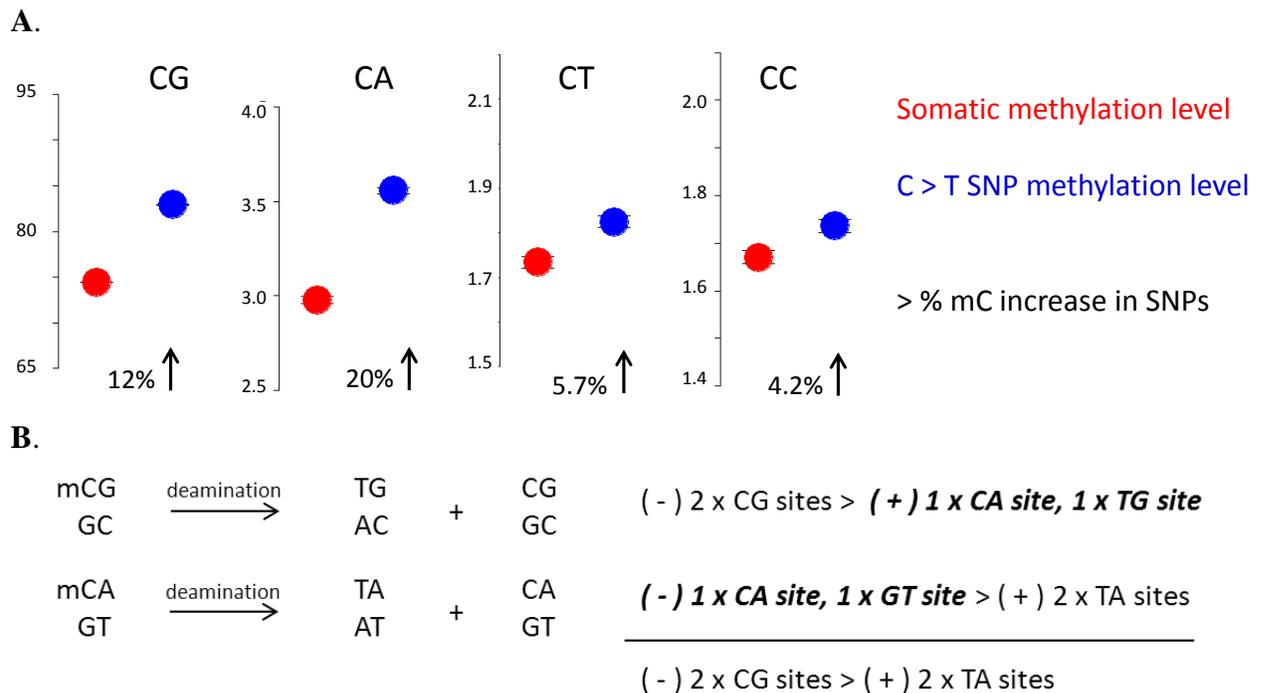


Figure 12. C>T transitions in all four cytosine contexts. **A.** Average percentage of methylation for each context in J1 mESC compared with polymorphic C>T transition sites between C57BL/6J (NCBIM37) and 129S4/SvJae mouse strains. **B.** A calculation, demonstrating why CA context is not proportionally reduced in the mouse genome as CG, despite its increased C>T transition rate.

I also explored the opposite possibility – if SNPs could contribute to artefacts in the methylation calls, undistinguishable from the conversion artefacts we observe in BS-seq. The effect of SNPs on methylation calling can contribute in three ways:

- 1) False negatives (when a T is actually a T rather than a converted cytosine);
- 2) Give wrong context in the case of G>H transitions, where methylated CGs could give false rise in CH methylation;

- 3) Mapping mismatches – if a large number of SNPs occur on one sequencing read, such reads will be excluded from analysis by the stringent mapping parameters.

I explored the first two possibilities. False C>T transition negatives would be exactly 1,699,651 sites out of 1,068,292,058 cytosines in the genome or 0.16% of all cytosines, which will not affect the result of final methylation. Regarding the false positives, (C)G>(C)H transitions are 528,608 sites out of a total of 42,685,558 CG sites, or the decrease in CG methylation due to transitions will be 1.24%. There are 1,025,760,420 cytosines in CH context, making CG>CH transitions 0,052% of CH context. Given that around 70% of CG context is methylated in mESC, this will mean that around 0.036% of CH calls will appear methylated due to wrongly mapped CG sites. This makes around 1/40<sup>th</sup> of the observed conversion artefacts.

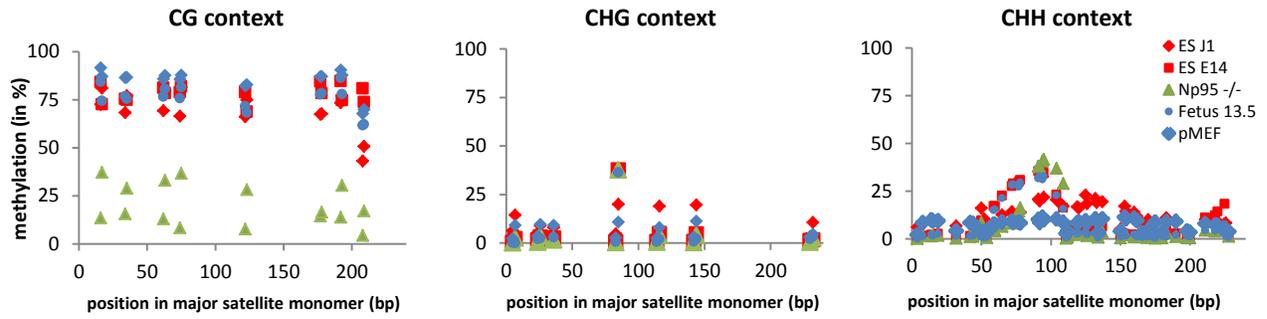
In summary, the conversion efficiency of bisulphite-based techniques has always been a main question for non-CG methylation because the ‘real’ signal looks like the conversion artefacts – sparsely distributed, individual values, with low % per position, and in non-CG context (Warnecke et al. 2002). A small amount of those artefacts is due to wrongly mapped methylated CG sites from the 129 genome, which correspond to CH context in the B6 genomic annotation. The current results suggest one should be extremely cautious when interpreting global non-CG context methylation data from low resolution BS-seq datasets. How this affects the actual analysis of non-CG methylation for genomic features has to be further assessed.

### ***3.3.3 Distribution of non-CG methylation in low resolution BS-seq datasets***

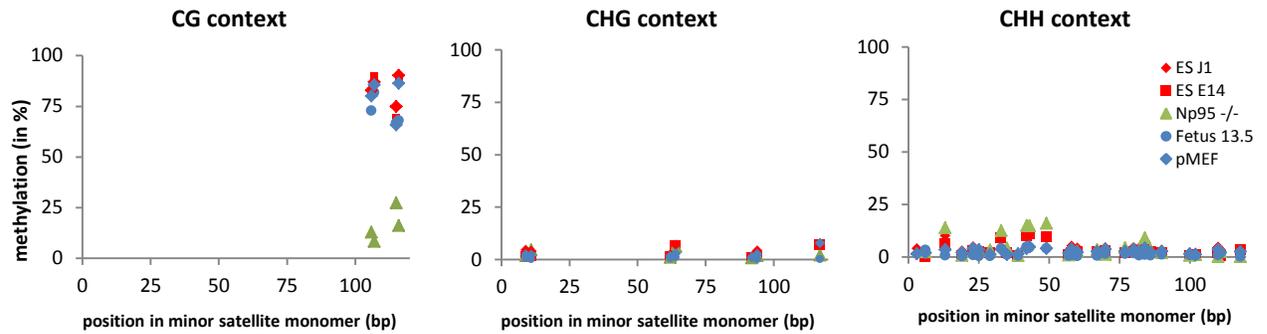
Despite the high levels of false discovery rate contributing to the global amounts of cytosine methylation, which makes the analysis of individual cytosines impossible, a global analysis of the annotated genomic features and unmappable repeat classes as for MeDIP-seq in 3.3.1, was nevertheless worth attempting.

I therefore first set to validate the repeat alignment results from the MeDIP analysis in 3.3.1 and repeated the same alignments for bisulphite converted DNA for two pairs of pluripotent and differentiated samples (J1 ES and pMEF (Seisenberger et al. 2012), and E14 ES, Np95<sup>-/-</sup> ES and E13.5 Fetus (Popp et al. 2010)). Since non-CG methylation should be more prevalent in pluripotent cells we would expect to see a higher value in the ESC than in lineage committed cells

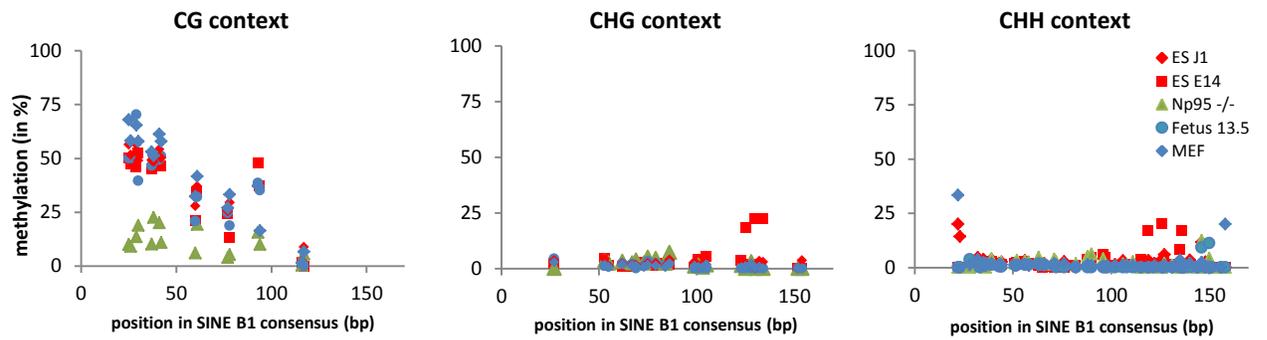
**A. Major satellite**



**B. Minor satellite**



**C. SINE1B**



**D. IAP LTR1a**

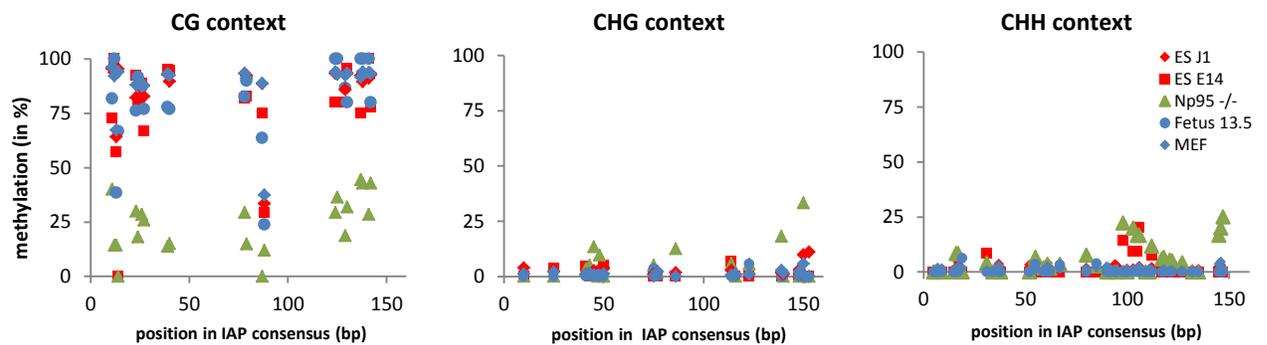


Figure 13. Non-CG methylation in mouse repeat sequences. The individual cell lines are given in figure legends – pluripotent cells are marked with ‘red’, differentiated with ‘blue’, DNA methylation-KO ES in ‘green’; GC context (left), CHG context (middle) and CHH context (right). A. Major satellite repeat; B. Minor satellite repeat; C. SINE1B consensus; D. IAP LTR1a.

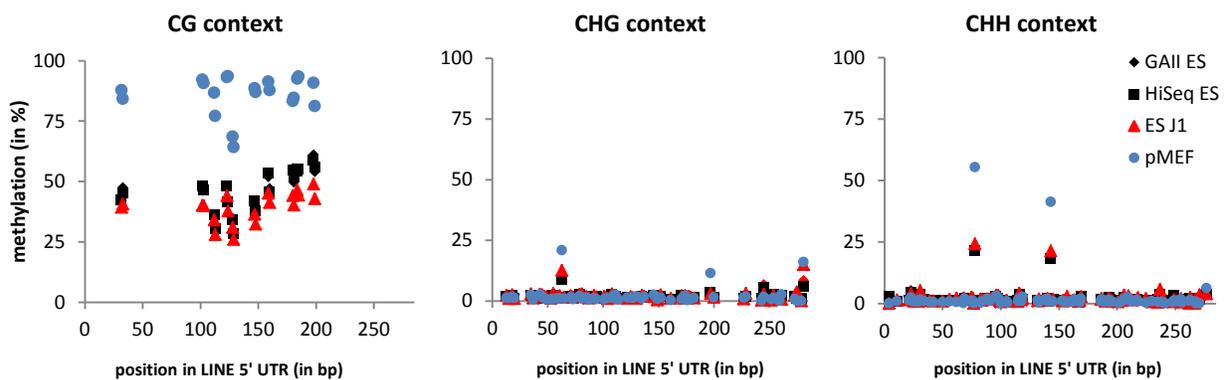
(Ramsahoye et al. 2000). I included in the analysis all repeats previously analysed in 3.3.1 together with the minor satellite and LINE 3' UTR consensus (see Appendix Table 16). The results confirmed the presence of non-CG methylation in the major satellite, again mostly on the reverse strand (strand information not given on figure), and surprisingly it was also present in the differentiated cells, although potentially to a lower extent (Figure 13 A, red vs blue colour). The minor satellite, SINE1B and IAP LTR1a showed lower amounts of non-CH methylation and, unlike the major satellite, it was present exclusively in the ES cells and not in the differentiated samples (Figure 13 B-D, red vs blue colour). There was no striking strand bias in those repeats and the figure omits strand-specific information, showing a summed up methylation from both strands along the length of each consensus.

Another interesting point from this analysis is the largely preserved non-CG methylation in Np95<sup>-/-</sup> ESCs, which maintains the pattern of the WT ES cells and not the lineage committed cells, despite that the CG methylation is heavily reduced. In the case of minor satellite and especially the IAP LTR1a we see slightly higher non-CG methylation levels than in the WT ESCs, which could lead to the speculation for a compensation mechanism over the lack of CG methylation. This could in fact be observed in all four repeats, because the highest levels of non-CG methylation are detected within the 'gaps' of CG methylation – as if the low levels of non-CG methylation were compensating for the lack of methylatable CG sequence in that region. However not all 'gaps' were compensated for, meaning some sequences potentially required a higher level of compensation than others, suggesting this might be biologically meaningful.

These results looked quite promising regarding the capacity of low coverage WGBS datasets for the assessment of repeat classes, for the LINE1 sequences, and I therefore I used the recently published mouse ES cell high depth dataset by (Stadler et al. 2011) as a comparison. The results are presented in Figure 14. The data for J1 and E14 WT ES cells from both datasets show a very coherent overlap ('red' vs 'black' on the figure). The interesting observation about LINE1 UTRs is that both of them have a few conserved individual CH positions with very high levels of methylation, which is a rarity for CH context, while they do not show larger stretches of low level non-CG methylation characteristic of other repeats. This result is strikingly reproducible and accurate in all datasets, the positions are conserved in all datasets and do not resemble a conversion artefact. Interestingly, in the 5'UTR those values are higher for the pMEF than the ES

cells, which also have only 30-50% CG methylation while the MEFs are fully methylated in CG. This observation highlights that the levels of non-CG methylation track the levels of CG methylation for the same region, again pointing to a potential ancillary role to the main CG methylation, and not necessarily only in ES cells. The levels of mCH therefore seem region specific, rather than cell specific, and it will be important to define where it occurs in the different cell types and what differences in its role the different locations might determine. The 3' end of the 3' UTR is excessively CH-rich and therefore those levels must be a result of bisulphite degradation and amplification bias over unconverted DNA.

### A. LINE1 5' UTR



### B. LINE1 3' UTR

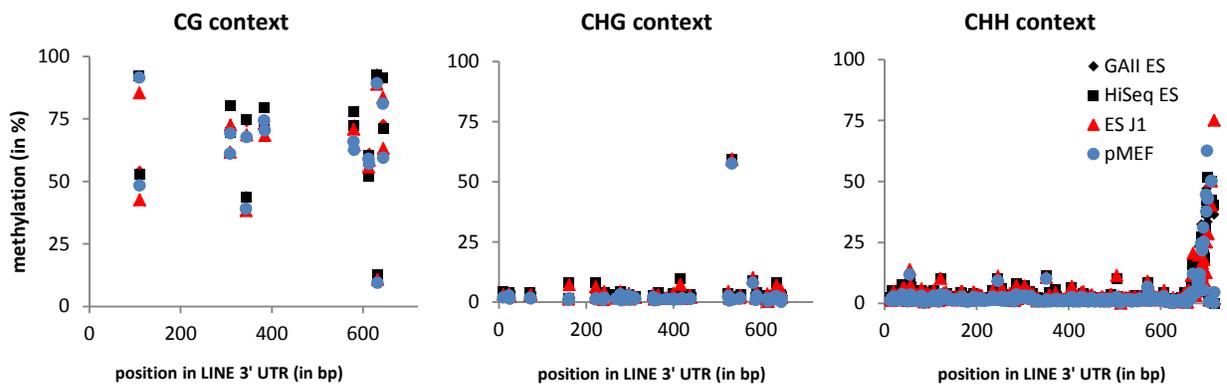


Figure 14. Non-CG methylation in mouse LINE1 sequences. Shown are low coverage J1 (red) and pMEF (blue) (Seisenberger et al. 2012), and high coverage mES datasets (Stadler et al. 2011) (black) – the latter in two groups as ‘GAll’ and ‘HiSeq’, depending on the Illumina machine used. **A.** LINE1 5' UTR. **B.** LINE1 3' UTR.

To summarise, the initial analysis of low depth BS-seq datasets on repeat consensus sequences yielded very good results. In both groups of comparative repeat analysis, the agreement between two different sets of datasets generated by different groups, show very good reproducibility, despite differences in coverage and depth. The results from the lower coverage datasets were entirely comparable with high depth datasets. This is not a surprise given that this analysis utilises alignment of multiple raw reads to a repeat consensus sequence, which gives a very high depth of alignment even in low-resolution datasets.

I proceeded with assessing the methylation of unique (non-repetitive) or non-unique but annotated genomic sequences (like the rRNA loci or pseudogenes). I chose a selection of features and calculated the relative methylation levels for each of the three contexts – CG, CHG and CHH. As seen in Figure 15, the levels of CG methylation (upper panel) are entirely as expected, with every feature highly methylated between 60-80%, apart from CGIs, transcription start sites (TSS) and CGI-containing promoters. In both CHG and CHH contexts however, the distinction between methylated and unmethylated features is not as dramatic and fluctuates around the genomic average and the TKO incomplete conversion value (Figure 15, middle and lower panel). This result shows that the positive calls in non-CG context are distributed more or less evenly among the features and do not enrich in any particular genomic locations. However, high depth analyses have revealed enrichment in gene bodies and some repeats in human and mouse (Lister et al. 2009; Ziller et al. 2011). In addition, unmethylated CGIs are not expected to have CH methylation, because they are protected from methylation. This means that the amount of information in the low coverage datasets is not enough to provide robust and adequate evaluation of CH methylation, and the majority of observed positive calls are actual false positives resulting from incomplete bisulphite conversion. Incomplete conversion has no reason to be enriched in one feature over another, and therefore is expected to be uniformly distributed along the whole genome, a result in keeping with our observation in Figure 15.

Thus, high depth repeat analysis from low coverage datasets is entirely valid for both CG and CH methylation, while feature analysis in CH context is dominated by artefacts and is not reliable.

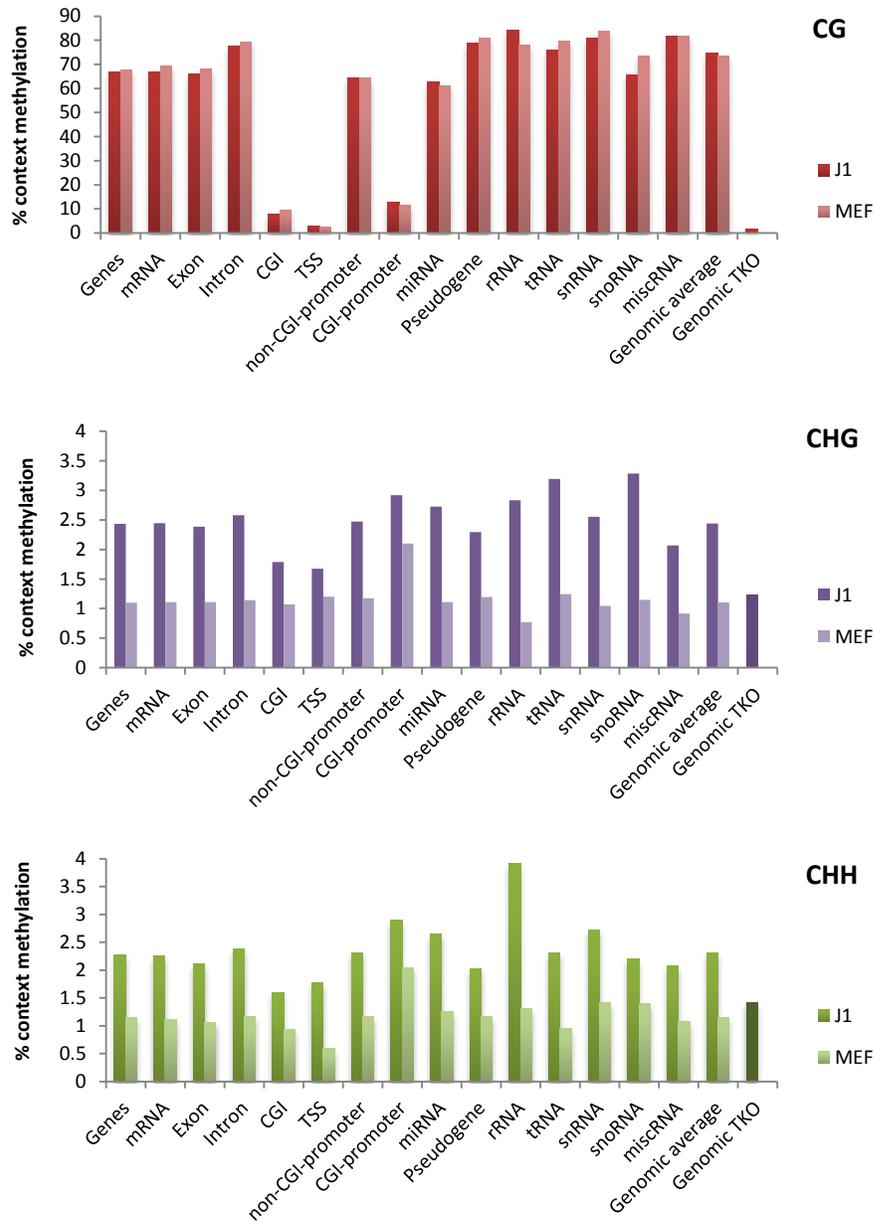


Figure 15. Methylation in genomic features in J1 ES and pMEFs in CG (upper), CHG (middle) and CHH (lower) contexts. The genomic false discovery rate per context of the TKO estimated in Figure 11 is added to each graph for clarity.

As a last test, I decided to verify a published observation, rather than use the low coverage datasets for identifying novel patterns of mCH localisation for which I cannot be 100% certain if the result is valid. It has been reported that non-CG methylation was observed at replication origins of actively replicating mammalian cells. It was proposed to play a role in the licensing of

origins, as it did not persist once the cells entered a replication arrest (Tasheva & Roufa 1994b). I have therefore decided to validate this observation, and check if I would detect more non-CG methylation in mouse ES cells than in pMEFs at origins in the low resolution BS-seq datasets. I used published annotation of replication origins for mouse for chromosome 11 to analyse methylation calls in cell type specific origins of ES cells and pMEFs (Cayrou et al. 2011).

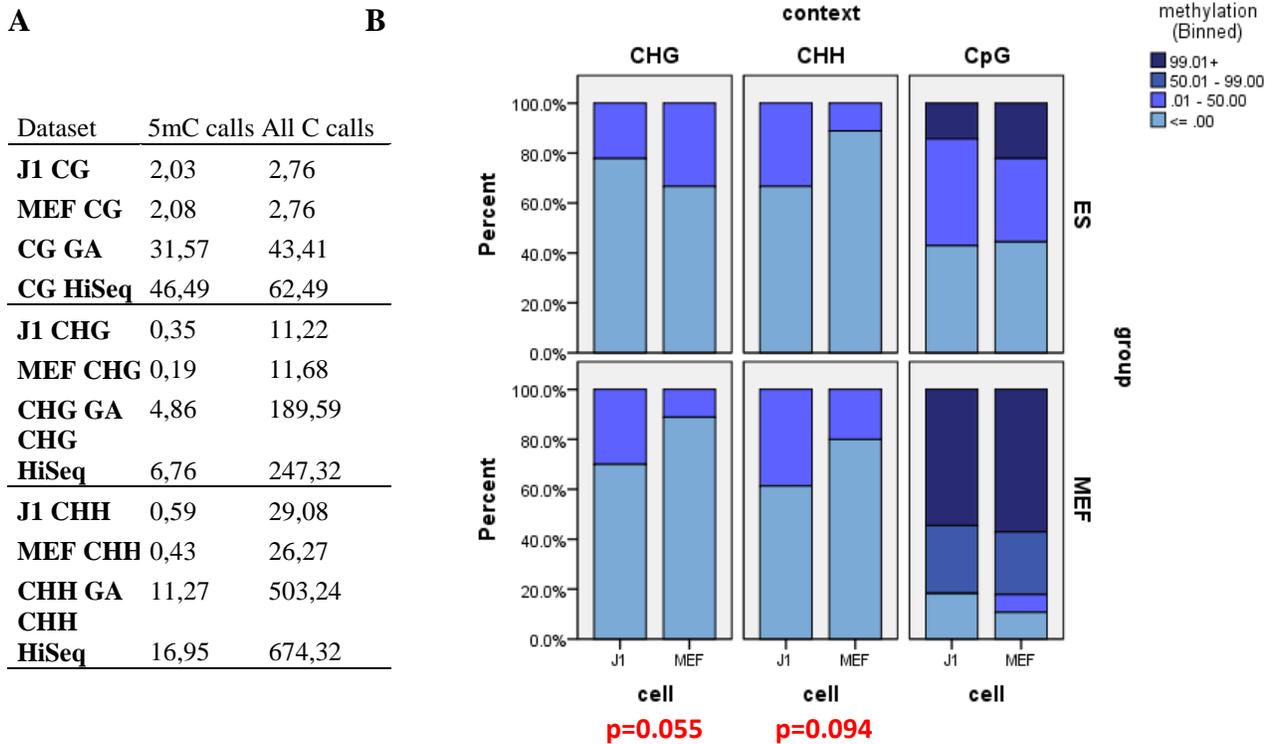


Figure 16. Methylation in mouse replication origins for chromosome 11. **A.** Average cytosine coverage per feature for low coverage J1 and pMEF and high coverage mES from (Stadler et al. 2011) – note that this data is split in two and appears as ‘GA’ and ‘HiSeq’, depending on the Illumina machine it was performed. **B.** Context methylation analysis in ES and pMEF specific replication origins (vertical axis groups) in J1 and MEFs (horizontal axis); the methylation ratios are binned in groups as shown in figure legend, and each group represented as a percentage of the total. Anne Segonds-Pichon, BI Bioinformatics helped for the statistical analysis.

However the cytosine coverage per origin of the low resolution BS-seq datasets was not enough for a statistical validity of the results (Figure 16A) and we had to perform the analysis on the high depth mES cell datasets (Stadler et al. 2011). The difference between CHG and CHH methylation in ES cells and pMEFs was not statistically significant to conclude that ES cells have more non-CG methylation in replication origins, for both cell type specific origin groups (Figure 16B).

### 3.3.4 Bisulphite treatment and DNA degradation

It has previously been reported that the unmethylated cytosine is the primary base targeted for degradation by bisulphite (BS) (Tanaka & Okamoto 2007), but so far no comprehensive study has been carried out to investigate whether this degradation bias affects the uniform genomic coverage of BS converted sequencing data. If indeed the unmethylated cytosines are the main source of DNA degradation, then we might expect a lower representation of C-rich regions in the datasets and a higher representation of highly methylated regions versus unmethylated regions (Figure 17A). This can, on one hand, lead to an overestimation of global genomic methylation levels (because the total cytosine pool is reduced), and on the other, can potentially have a negative effect on the coverage and depth of non-CG rich regions, which are two crucial parameters for its correct evaluation as discussed in 1.6.2 (Introduction) (Figure 11A).

I therefore set to answer the following questions: 1) Whether cytosine rich regions are indeed preferentially degraded by the BS treatment; 2) Whether cytosine modifications affect the degree of BS induced degradation; 3) Could this lead to any real biases in WGBS data.

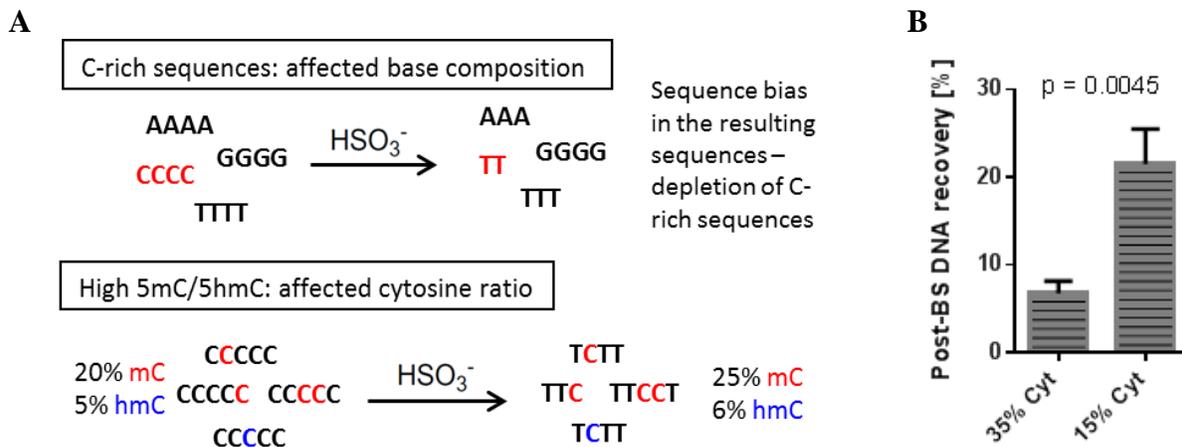


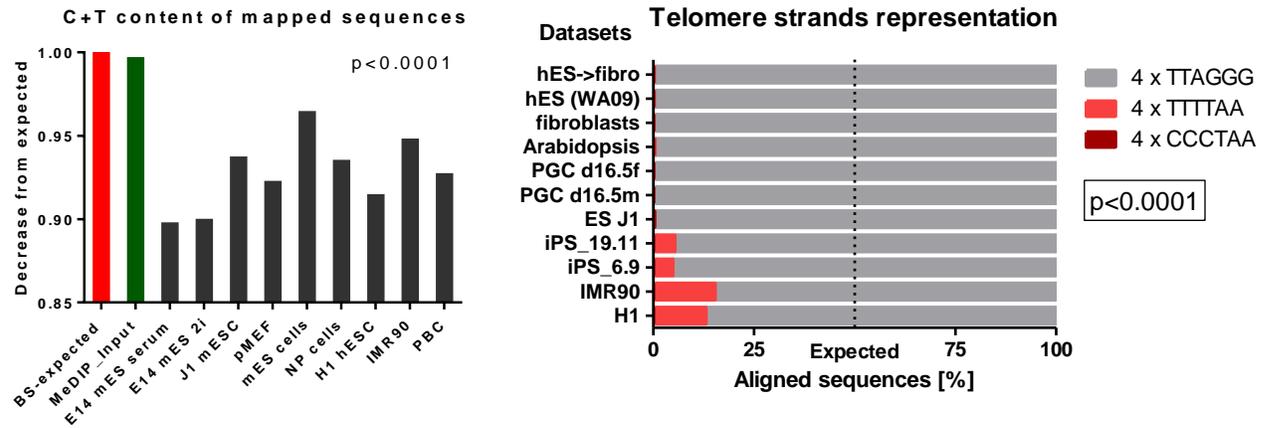
Figure 17. Bisulphite-induced degradation of DNA. **A.** A schematic representation of the expected consequences from the biased degradation in a cytosine-rich (upper) and highly methylated (lower) environments. **B.** DNA yield after bisulphite conversion of a C-rich and C-poor DNA PCR fragments shows higher degradation of the C-rich DNA.

For this purpose, I combined analysis of synthetic PCR fragments as well as of WGBS datasets. In order to address the first question I used two DNA fragments from the sequence of the M13 bacteriophage – one with 15% cytosine ('C-poor' fragment, R3) and the other with 35%

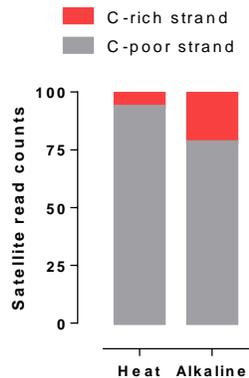
cytosine content ('C-rich' fragment, R5), the mapped mouse genomic average being 20.90% (see Appendix Table 13 for M13 sequences and Appendix Table 11 for genomic representation of cytosine contexts). I bisulphite treated both fragments with two widely used kits – Epiectect and Imprint (see 2.3.2), and measured the recovered DNA on an Agilent Bioanalyser high sensitivity DNA chip (Figure 17B). The results with both kits were very similar and the average recovery of the C-poor fragment was around 4-fold higher, meaning the cytosine content indeed affected the level of degradation.

I then looked at published WGBS datasets, where, due to the known polymerase AT-rich regions fail (Quail et al. 2012; Oyola et al. 2012) this problem might be further intensified. I first analysed the total cytosine content of a selection of published datasets, to see if the ratio of cytosine in the mapped sequences matches the genomic content, or is reduced. Due to the conversion of cytosine to uracil and then thymine after amplification, I summed the total cytosine and thymine content from each dataset's FastQC quality reports, and compared that to the genomic ratio. A non-converted 'Input' MeDIP-seq sample was used as a control dataset, which fully matched the genomic ratio (Figure 18A left). All of the analysed datasets, published by different labs and converted with a variety of BS conversion kits, showed a lower C + T content than the expected genomic value. I then looked into specific genomic C-rich sequences to investigate if this effect is seen at a regional level. I chose the telomeric repeat, as a unique and extreme case of continuous stretch of C-rich DNA, which is 100% asymmetric in its cytosine content – one strand represents the tandem repeat [CCCTAA]<sub>n</sub> (50% cytosine), while the complementary [TTAGGG]<sub>n</sub> strand is completely depleted of cytosine. Since the 6-mer tandem repeat can be quite degenerate (personal observations in the course of work), we looked for reads containing either one 12-mer stretch (2 telomeric units) or four of the single unit sequences, not necessarily in tandem. Both results showed a complete absence of the C-rich strand from the datasets, both in its converted and unconverted form (Figure 18A, right). Some datasets contained low levels of the converted TTTTAA -version of the [CCCTAA] consensus, which is very likely to be a native TTTTAA genomic sequence which cannot be filtered out but is not necessarily the converted telomere repeat. A very similar result was observed for a second tandem repeat - the major satellite consensus, which is also asymmetric in its cytosine content, although not as extreme as the telomere (14% cytosines on the forward strand and 23% on the reverse strand).

A



B



C

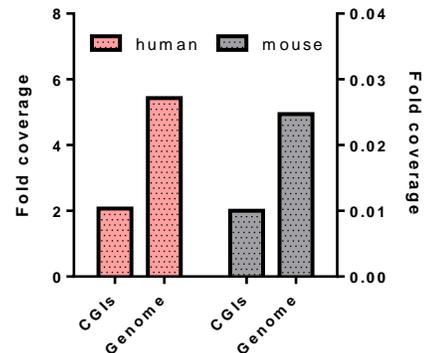


Figure 18. Biased degradation of C-rich DNA after bisulphite treatment. **A.** Cytosine depletion in published BS-seq datasets – left: decrease in total measured C + T (T is included because the majority of cytosines are converted to T's); right: telomere strands representation in a number of published datasets; Datasets used: MeDIP Input (Raiber et al. 2012), E14 serum and 2i (Ficz et al. 2013), J1, pMEF and PGC d16.5f, and d16.5m (Seisenberger et al. 2012), mES and NP (neuronal progenitor) cells (Stadler et al. 2011), H1 hESC, IMR90, iPS\_19.11 and iPS\_6.9 (Lister et al. 2009), PBC (peripheral blood cells) (Li et al. 2010), hES (WA09), hES > fibro and fibroblasts (Laurent et al. 2010), Arabidopsis (Lister et al. 2008b). **B.** Biased representation of C-rich and C-poor major satellite strands in a WGBS datasets after heat or alkaline denaturation in BS-conversion protocols. **C.** Affected coverage of CGIs due to the biased cytosine degradation in low (mouse) and high (human) coverage datasets (Lister et al. 2009; Popp et al. 2010).

The result showed that the reverse C-rich strand represented only 1% of the total reads mapping to satellite in the J1 dataset, the forward strand taking 99% of all mapped reads (Figure 18B). For both the telomere and satellite repeats however, it is clear that the C/G strand asymmetry will create an AT -asymmetry after BS-conversion, which will render the T-rich (ex-C-rich) strand in a much less favoured position in regard to DNA polymerases, in addition to the

degradation, creating this extreme bias where the C-strand is virtually missing. I therefore looked also at CG-islands (CGIs), which should have a more uniform CG content on both strands. Both for human and mouse ESC datasets, one of which high depth and the second - low depth dataset, the coverage of the CGIs was lower than the genomic average (Figure 18C).

To summarise, all results so far pointed towards a cumulative coverage bias in BS-seq datasets towards C-low sequences and reduced coverage and depth of unmethylated C-rich sequences, as a result of the interplay between bisulphite induced degradation of DNA and polymerase bias.

I next addressed the effect of methylation on BS-induced degradation. I used again the M13 bacteriophage fragments, this time with methylated or hydroxymethylated dNTPs against the unmethylated control fragments, and treated them with the same two BS-conversion kits (Figure 19A). Although the quantitation of DNA on the Agilent bioanalyser is based on of fluorescent DNA intercalating dyes, which show a marked preference for unmethylated versus methylated DNA (Ioannou et al. 2010), the summarised result from both kits showed that, both methylated and hydroxymethylated probes recovered better and suffered less degradation, as expected from (Tanaka & Okamoto 2007).

It was then important to see how much this affects WGBS datasets and the interpretation of our results. For this purpose, I sequenced the completely unmethylated gDNA from a *Dnmt1/3a/3b* triple-KO (TKO) ES cell line and an *in vitro* methylated DNA from the same sample. I used a bacterial GC-methylase *M.CviPI* (NEB), which methylates around 20% of cytosines in the mouse genome in all contexts, according to our *in silico* digestion calculations and the genomic occurrence of the GC dinucleotide presented in Appendix Table 11. After NGS of all samples the C + T content of the methylated TKO data (meTKO) was higher and closer to the genomic expected value, than the value in the unmethylated sample (Figure 19B). However, a comparison with WT mESC lines converted with the same kits, and containing only around 3.5% of total genomic 5mC, showed that the WT mouse genome was behaving much like the unmethylated sample and 3.5% of biological 5mC did not make a difference for the levels of global degradation it suffered. This meant that the susceptibility to degradation of native biologically methylated mammalian gDNA was as high as for a completely unmethylated DNA.

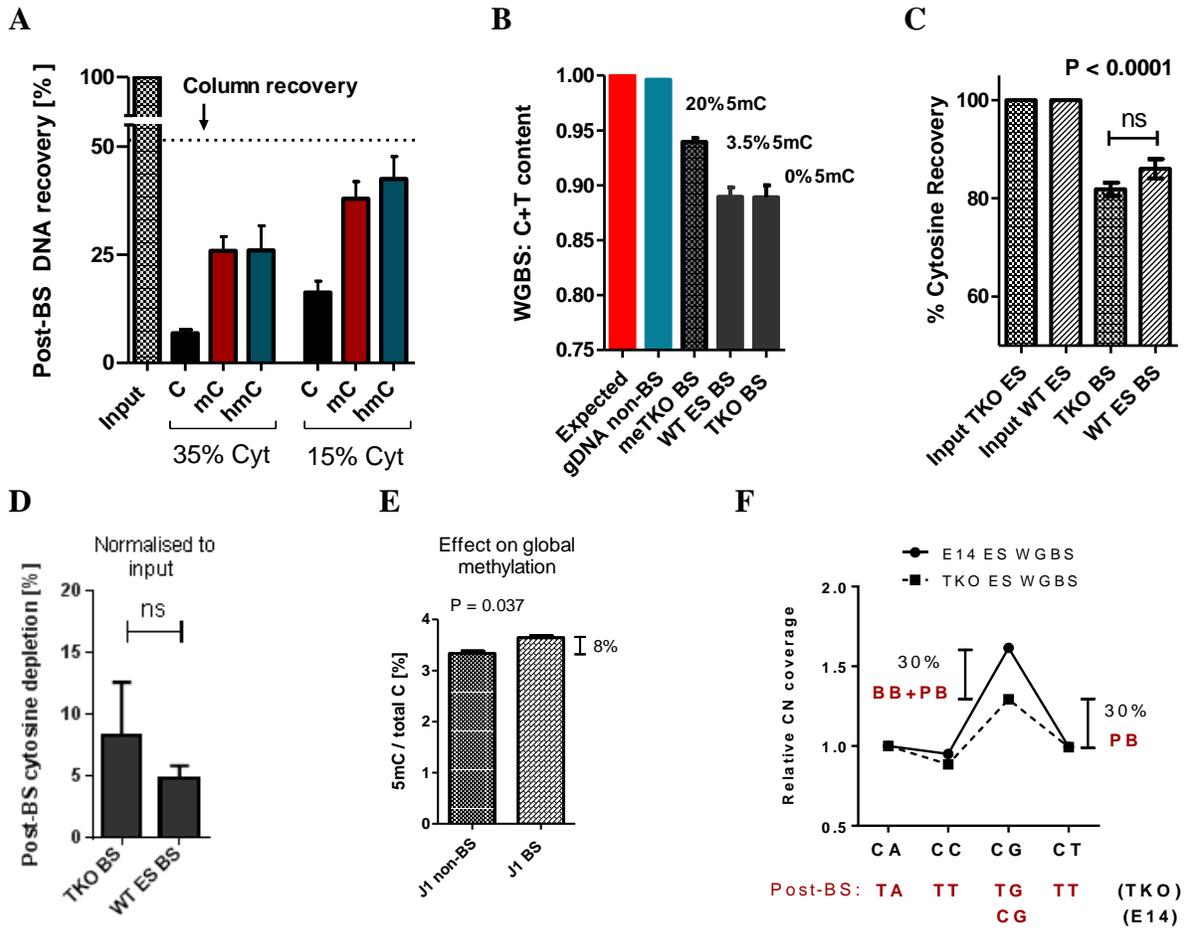


Figure 19. Effect of DNA modifications on the extent of DNA degradation by bisulphite. **A**. DNA yield after bisulphite conversion of unmodified and 5mC- or 5hmC-modified C-rich and C-poor DNA PCR fragments. **B**. Decrease in genomic cytosine content in selected WGBS datasets by total measured C + T (T is included because the majority of cytosines are converted to T's) - *in vitro* 20% methylated TKO (meTKO), unmethylated TKO and WT ES DNA (3.5% mC). **C**. Mass spectrometry measurement of total cytosine decrease in bisulphite treated TKO and WT ES DNA. **D**. Extent of cytosine decrease (in percent) from the WT and TKO gDNA measured by mass spectrometer and normalized to T content to account for the column losses from BS purification. **E**. Change in global methylation in WT ES cells with and without bisulphite conversion measured by mass spectrometer. **F**. Coverage of each dinucleotide pair in the WT E14 (mCG=80%) and TKO (mCG=0%) datasets (value normalized to CA = 1).

However, since the effect of cytosine degradation is surely paralleled by the shown polymerase bias against AT-rich sequences, in order to verify that the observed result was an effect of the bisulphite degradation bias and not merely the polymerase bias, I repeated the experiment with a quantitative measurement on a mass spectrometer (LC/MS) (Figure 19C). The initially observed decrease in cytosine content was around 15-20%, but after normalisation to thymidine to account for the purification losses, the total cytosine decrease amounted around 5%

loss from the input content (Figure 19D). This however contributed to an 8% of increase in the total methylation levels measured in the J1 gDNA before and after BS-conversion, showing that the degradation of unmethylated cytosine indeed leads to an overestimation of global genomic 5mC levels (Figure 19E). To see how affected further this will be in WGBS datasets, I assessed the extent of polymerase preference on GC-rich sequences by plotting the coverage of all four CN dinucleotides for two datasets – methylated and unmethylated. Three of the dinucleotides convert entirely into AT-sequences, and all three have 30% lower coverage than the unmethylated CG dinucleotide, which has an even increased coverage with further 30% in the WT mESC with 80% CG methylation level. Interestingly, we did not observe a higher coverage depth in the WT dataset in the CA context, which is the second highest methylated context in ES cells (Stadler et al. 2011). This shows that CA methylation levels are not high enough to cause a bias, and the CA context suffers as harsh degradation as the rest of unmethylated CT and CC contexts in both WT and unmethylated gDNA. The polymerase bias can however reach quite high levels in CG-rich regions.

### ***3.3.5 Measuring global levels of non-CG methylation in the mouse genome***

As outlined in 3.2 Aims it is important to be able to assess global genomic levels of non-CG methylation in my pluripotent and differentiated samples of interest. There has been conflicting evidence for the presence of non-CG methylation both in pluripotent, but also in various differentiated cell types or tissues, and it is not very clear if it is limited to the pluripotent state or a more general phenomenon.

The total genomic methylation in mouse ES cells is variable between 3-4% of total cytosine (Ficz et al. 2013), and this value is lower for human DNA (Ehrlich et al. 1982). mCH methylation constitutes a sizable proportion of the total methylation in ES cells, accounting for approximately one quarter of the total 5mC (Ramsahoye et al. 2000; Lister et al. 2009). To test this, I used two published techniques to estimate global non-CG methylation: luminometric methylation assay (LUMA) for CCWGG, and nearest neighbour analysis (NNA).

I have investigated which alternative enzymes would be useful to increase the sensitivity or change the specificity of both techniques, in addition to the published ones, and the results from an *in silico* digest and calculations are summarised in Appendix Table 17 and Appendix Table 18.

## 3.3.5.1 NNA

The principle of NNA is described in Figure 20.

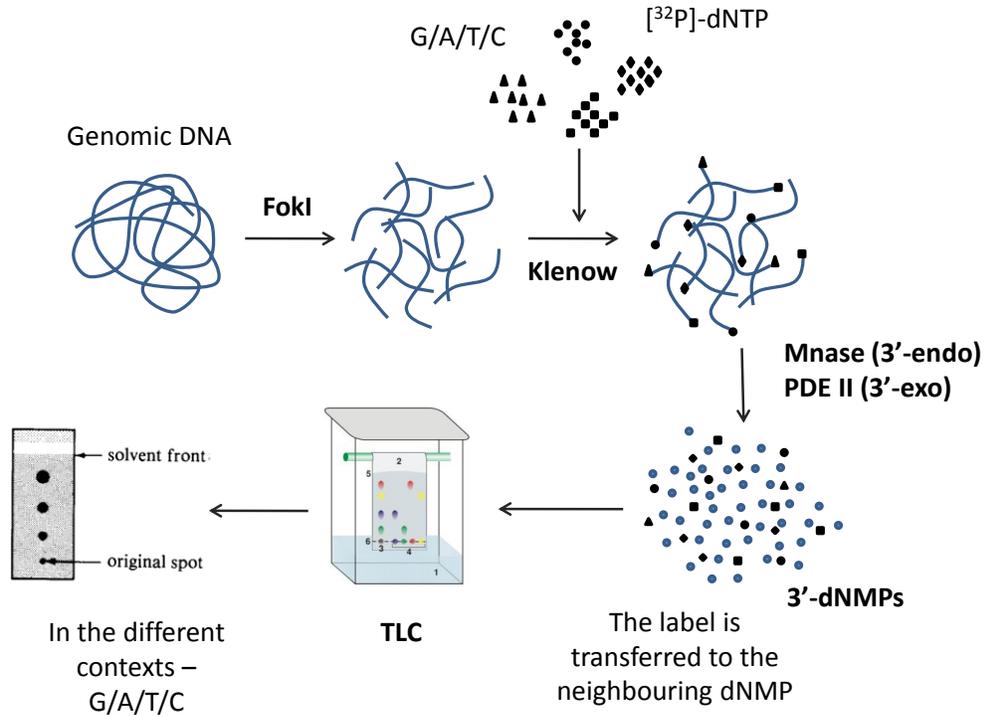


Figure 20. A schematic representation of the NNA method.

To cut the long story short, I have digested WT ES gDNA samples with FokI, and completed the procedure as described (Ramsahoye 2000) with a few modifications outlined in 2.3.11. In order to be able to identify which spot belongs to which base, including the modified cytosine bases, I also synthesised by PCR and digested a group of M13 fragments which after the digest with DpnII yielded one particular base and could serve as standards. However the results showed double spots rather than the expected single spot, both on 1D and 2D TLC (Figure 21). On the same figure it is also shown what the 2D TLC result from the NNA digestion should be, and even the positions of the fragments are not identical. Several discussions with TLC technology representatives and personal communication with Bernard Ramsahoye were unable to help resolve the reason for this result, and I was therefore not able to utilise this classical technique for the analysis of global levels of mCH methylation.

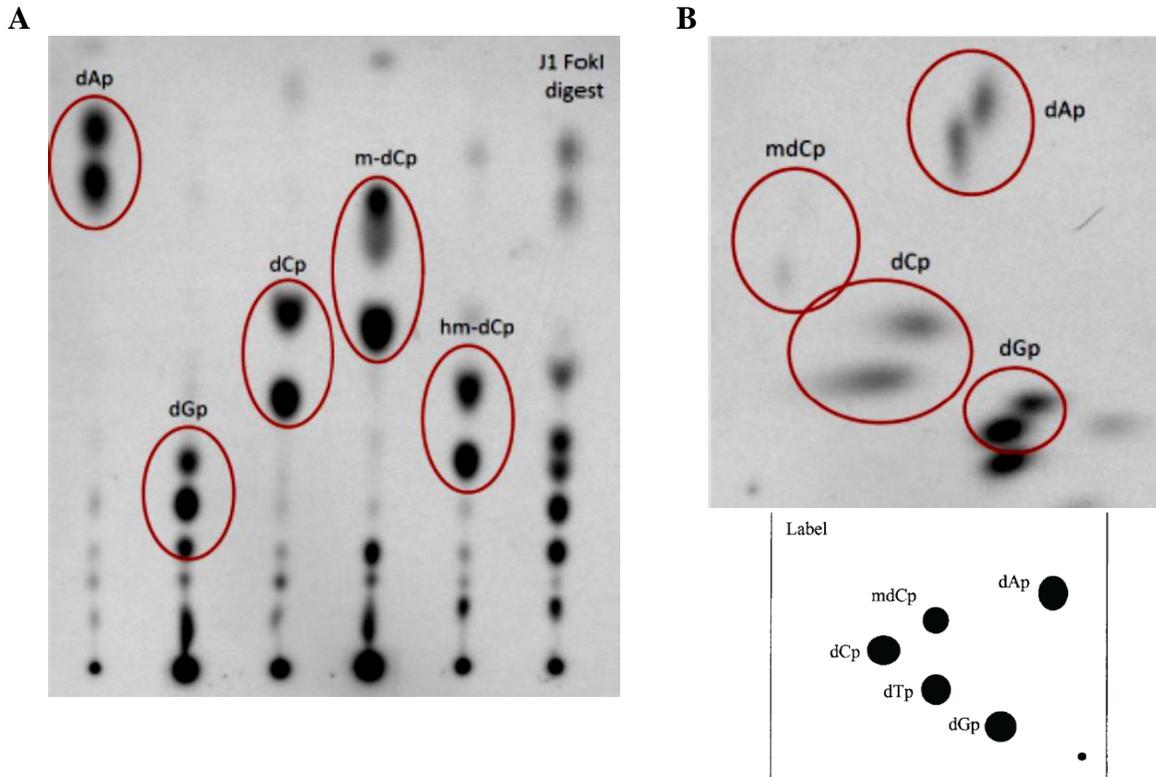


Figure 21. Results from the nearest neighbor analysis. **A.** 1D TLC, **B.** 2D TLC (upper) and a representation of what the result should look like (lower panel).

### 3.3.5.2 LUMA

The CCWGG version of LUMA was published recently (Yan et al. 2011), and the possibility to use this assay for non-CG context looked quite promising, given the possibility for high throughput analysis of multiple samples and the simplicity of the assay. A downside of this technique is the limitation that only CHG methylation can be studied, and in particular in the CCWGG context, but taking in mind that the CHG methylation predominates over CHH methylation in ES cells according to several reports (Lister et al. 2009; Ziller et al. 2011; Lister et al. 2011), this technique seemed quite appropriate. Another potentially limiting point is the use of a Pyrosequencer, but this equipment nowadays is available in many research institutions.

Together with the classical assay based on the MspI/HpaII CCGG digestion (Karimi et al. 2006) I tried the CCWGG version with Ajnl/Psp6I and also a Bsp19I – a CCATGG cutter, sensitive to 5mC according to SibEnzyme (personal communication). The principle of the LUMA assay is shown in Figure 22. Digestions were carried out as described in 2.3.12 closely following the published protocol (Karimi et al. 2011).

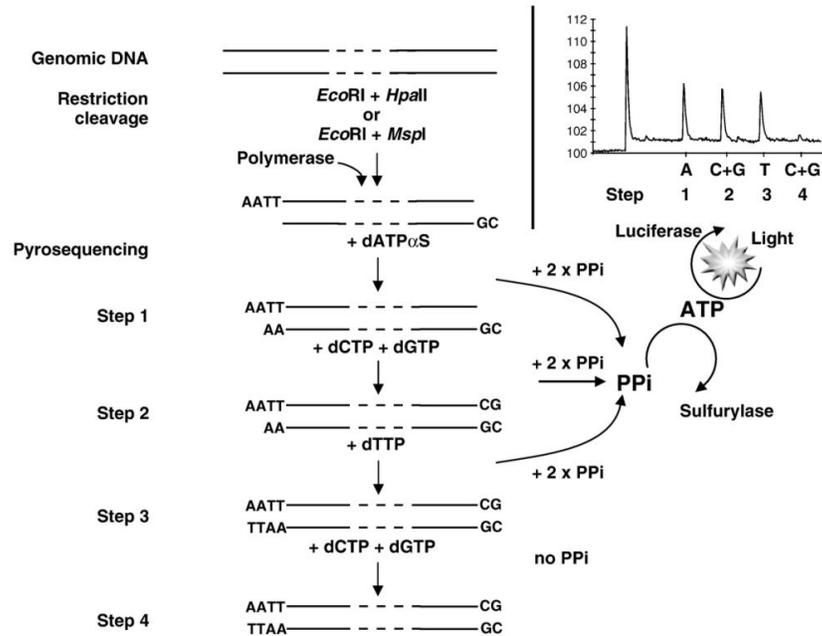


Figure 22. A schematic principle of the LUMA assay. Borrowed from Karimi et al. 2006.

The assay worked well overall – the pyrograms showed the expected profiles, the chosen enzymes worked well and showed clean signals (peaks) in individual test runs (not shown) and in double digests (Figure 23A). Calculating the CG methylation worked as described and gave good results, both for mouse and human gDNA samples (Figure 23B). A standard curve for CG methylation, produced with a M.SssI methylated PCR fragment, gave good linearity (Figure 23C), although it did not reach 100% methylation, because of an inefficient M.SssI activity – the maximum methylation of the fully methylated fragment was measured repeatedly around 65% (not shown).

The assay for CCWGG however showed inconsistent results, despite changing enzymes (PspGI vs Psp6I) and digestion buffers. A panel measurement of a selection of ES cell samples with varying methylation levels is shown on Figure 24A. The values unfortunately do not reflect the actual methylation levels of the samples according to expectations, and some values are highly negative, even for samples, which should be positive for CHG methylation. The result for the Bsp19I was similar (not shown). I therefore performed an *in vitro* digest with AjnI and Psp6I on a synthetically CCWGG methylated oligonucleotide sequence as in Yan et al. 2011. I tested all combinations of fully, hemi- and unmethylated fragments, and the result showed that AjnI did not

fully digest unmethylated or methylated DNA, with even less digestion of the hemimethylated DNA. It is not clear why it inefficiently digests the hemi-methylated fragments, but it may be due to the strand asymmetry. It is interesting to note that there also seems to be a difference in the extent of digestion between the two hemi-methylated fragments, the forward ‘F’ one being digested only slightly more by Psp6I. Both enzymes did not fully digest unmethylated DNA either.

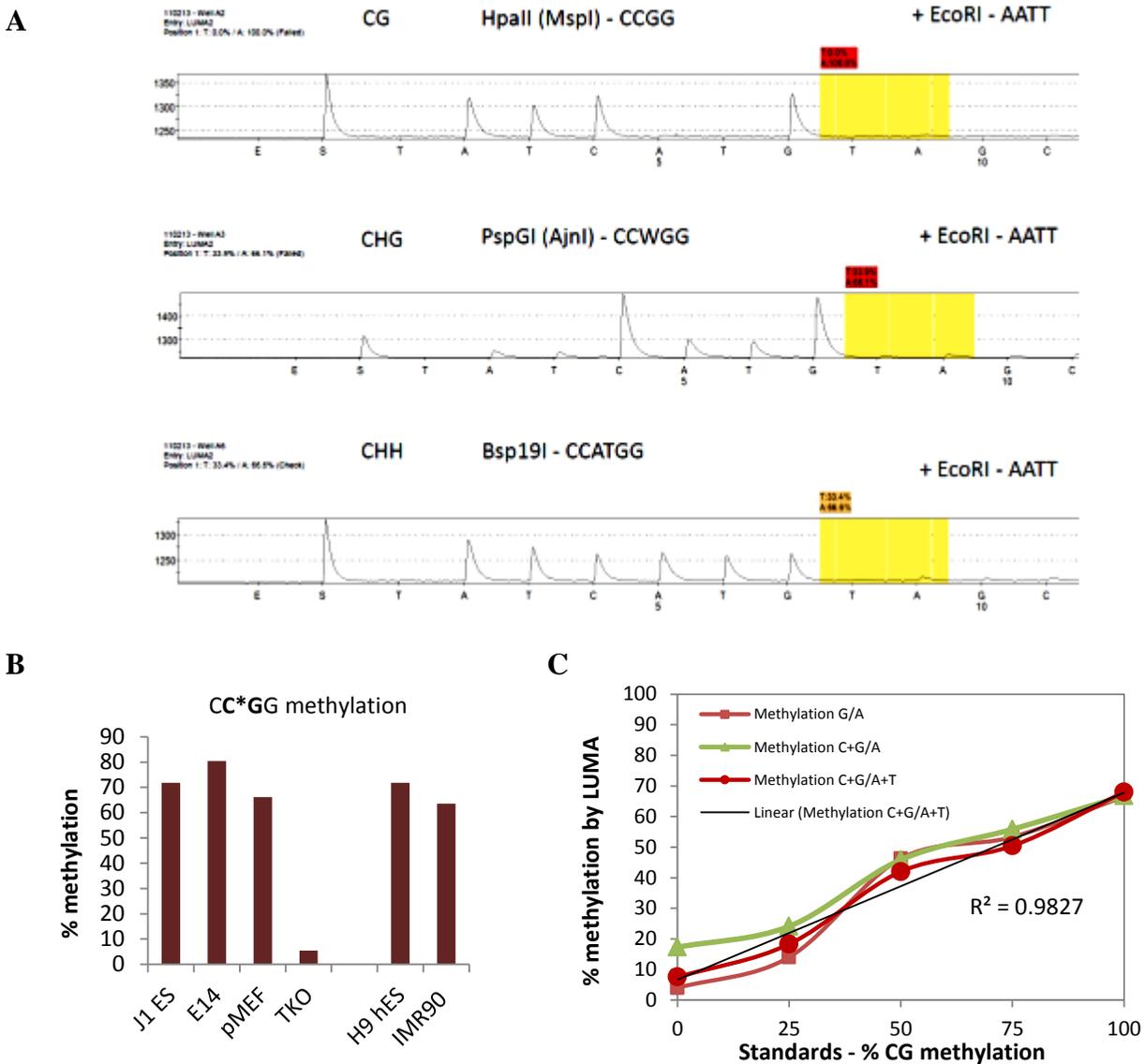
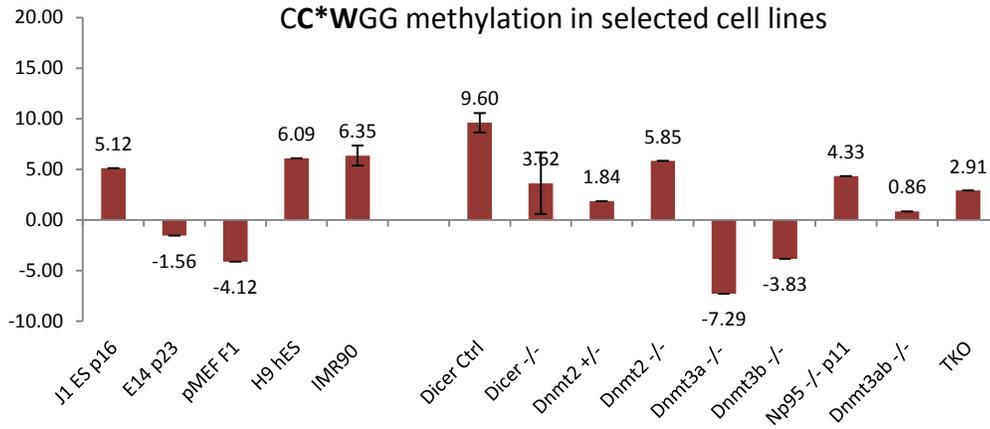


Figure 23. LUMA assay results. **A.** Pyrograms for the different digestions tested; each peak corresponds to a ‘filled up’ by the DNA polymerase position in the restriction sequence, which releases light and generates quantifiable signal. **B.** CCGG methylation in mouse (J1, E14) and human (H9) ES cells and fibroblasts (pMEF and IMR90). **C.** Measurement of *in vitro* M.SssI-methylated DNA fragment (65% 5mC), mixed with increasing amounts of unmethylated DNA to form a standard curve.

**A**



**B**

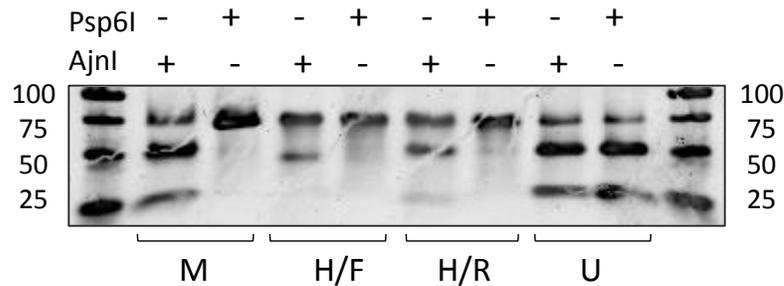


Figure 24. CCWGG LUMA trial. **A.** Methylation measurement of a panel of mouse and human cell lines with variable amounts of 5mC; the resulting percentages however do not correspond to methylation levels. **B.** In vitro digestion of oligonucleotides with Ajnl and Psp6I, visualized on a 6% TBE polyacrylamide gel. U is 'unmethylated', H is 'hemimethylated' respectively on the forward or reverse strand, M is 'methylated'.

We know that CHG methylation in mammals is mostly asymmetric, therefore situations in actual gDNA would be more similar to the hemi- and unmethylated oligos, and not the fully methylated. It therefore seems that the enzyme pair Ajnl/Psp6I is not as reliable and appropriate for CCWGG as MspI/HpaII is for CCGG, especially that in CCWGG context we are looking for much lower amounts of methylation than in CCGG. Yan et al. did not demonstrate a digest of hemi-methylated fragments and they also use different equipment, which might have a better sensitivity (Yan et al. 2011). In conclusion, the CCWGG version of the LUMA technique was not useful for the quantification of non-CG methylation levels in ES cells and differentiating cells in my hands.

### 3.4 Discussion

The results from the genomic analysis so far show, that even with the well established ‘classical’ techniques, the process of genome-wide or global analysis of non-CG methylation is not as straightforward as expected.

The MeDIP-seq technique, although very appropriate for the enrichment of lower levels of methylation and could be ideal to enrich for non-CG methylated DNA, is not useful at present due to the wide 5mC recognition properties of the current commercially available antibody. As shown here and in (Ficz et al. 2011), although the asymmetric peaks can contain information on non-CG context, it is not possible to clearly distinguish the real signal from the artefactual noise due to the many pull-down artefacts and amplification biases. It is indeed possible that the observed feature enrichment pattern is representative of non-CG methylation and it naturally follows closely the pattern of the symmetric peaks in lower levels. It has already been reported that the non-CG methylation colocalises with CG methylation (Lister et al. 2009; Ziller et al. 2011) meaning there is a chance that we wouldn’t observe anything different from CG even without the noise. However those assumptions are not sufficient to validate the current MeDIP-seq approach.

Other well established techniques such as BS conversion and sequencing appear to have some limitations in regard to non-CG context, especially used together with NGS, which make them less than ideal for the analysis of this type of methylation. The high depth and genomic coverage datasets are still the best option for whole-genome non-CG analysis (Lister et al. 2013; Shirane et al. 2013). On one hand they benefit from the filtering of potentially unconverted reads (with >3 unconverted non-CG C’s) and on the other, the standard analysis of those datasets includes a certain depth requirement for the analysed cytosines, typically 10-fold depth, which would exclude all highly degraded low coverage regions, and thus will have less false positives. Those datasets however require a lot of sequencing effort, which is both labour intensive and costly, especially if numerous samples are analysed, and at present, this cannot be a routine approach for many laboratories.

Lower coverage whole-genome datasets accumulate conversion artefacts, which cannot be filtered out bioinformatically because of the insufficient depth and, as shown, filtering itself can introduce additional biases and cannot be currently validated as an approach. In this sense, the historical doubts about the validity of non-CG methylation and its artefactual nature (Harrison et

al. 1998; Araujo et al. 1998; Laird 2010), are still valid for the current low coverage NGS datasets, because the BS conversion has not been improved as a process, and only high sequencing depth enables us to validate a potential non-CG call. The most common way to estimate conversion efficiency through unmethylated spiked-in DNA, which gets bisulphite treated together with the whole sample, is useful only to an extent, given that the BS-conversion resistance is by nature sequence and context specific (Warnecke et al. 2002). Therefore, the conversion rates of one genomic sequence cannot be a guarantee for the same levels of conversion in another sequence or genome. To address this and also the existing speculations that the TKO mESC could actually possess very low levels of methylation from Dnmt2 (investigated in detail in Chapter 5 and in Raddatz et al. 2013), I attempted sequencing a TKO library, PCR amplified before the bisulfite conversion, to represent a whole genome amplified (WGA) and truly unmethylated control. This attempt was not successful, however, due to the specificities of the Illumina library preparation technology, and was not repeated again for mouse BS-seq libraries. Nevertheless, using WGA was subsequently affirmed as an approach for all ELISA experiments (in Chapters 5, 6 and 7), and was also applied as a validation strategy for low methylation level WGBS in Patalano et al. 2015 (in preparation), where WGA was achieved via a Qiagen kit, rather than PCR amplification at a library level.

In addition, the contribution of BS-degradation and amplification biases had not been studied until now, but also affects the analysis of CG and especially non-CG methylation. Other known downsides of the NGS BS-seq analysis, have been the mapping issues, especially of C-rich fragments, because of the lower complexity of the endogenous sequence after BS conversion (Krueger et al. 2012). Roughly, a quarter of the raw sequencing data, and often more than that, is lost and unable to be analysed just because of mapping issues. As seen in 3.3.4, although the C-rich fragment is only 3-4 times less recovered after BS-treatment without amplification, after amplification and sequencing such a C-rich fragment is many-fold less represented in the dataset, in comparison to C-poor fragments. Thus, DNA degradation, amplification bias and mapping issues contribute altogether to the depletion of CH-rich sequences from WGBS datasets.

It is useful to know well the limitations of each technique, and the current analysis shows that with appropriate controls and good understanding for the underlying biases, one can nevertheless derive meaningful data from each technique, despite its limitations. The MeDIP-seq technique was able to give a direction towards the presence of non-CG methylation in the major

satellite repeat, an observation confirmed later by (Arand et al. 2012) and reported as early as 1975 in (Harbers et al. 1975). The low coverage BS-seq datasets proved very useful for high depth repeat analysis, and the presence of non-CG methylation in SINEs and LINEs has also been confirmed by other groups (Guo et al. 2013; Woodcock et al. 1997; Woodcock et al. 1998). IAP LTRs represented by the most active LTR1a did not show significant levels of non-CG methylation, and the lowest mCH levels were detected in the minor satellite. Although spatially very close to the major satellite and sharing functional roles, their methylation status is remarkably different – the major satellite showing high methylation in both CG and CH context, while the minor satellite is much less methylated in either context.

Another aspect of the current results addresses the presence of non-CG methylation in pluripotent versus differentiated cells. Although all major papers analysing mCH methylation (Ramsahoye et al. 2000; Lister et al. 2009; Ziller et al. 2011; Lister et al. 2011) have declared much higher levels in pluripotent cells, and very low or virtually no presence in fibroblasts or adult tissues, here we see examples of regions where this generalization is not true. The major satellite sequence preserves non-CG methylation also in lineage committed cells, albeit decreased, and strikingly the levels in LINE UTRs were even higher in MEFs than in mES cells. It is important to note that many earlier reports have also claimed non-CG methylation in adult tissues, and these observations might be correct, at least for particular regions, or particular tissues. Therefore, although I was unable to measure directly global levels of mCH methylation in various cell and tissue types at this stage, my results so far do not exclude the presence of non-CG methylation in differentiated tissues, and do not limit it to the pluripotent state. This was also evident from the replication origin analysis. The existence of cell fate ‘barcodes’ has been hypothesised by the authors who published the origin coordinates of chromosome 11 (Cayrou et al. 2011) and the non-CG methylation would have been an obvious candidate for pluripotent state barcodes. However this doesn’t seem to be the case, at least in chromosome 11 which we have analysed.

In cases like the major satellite where we have a region of high CH methylation, it would be interesting to know if only a selection of the tandem repeats, or those in a particular chromosome or nuclear area are very CH methylated and the rest are not, or rather all satellite repeats on all chromosomes have an even distribution of low mCH methylation. Another option could be related to the intrinsic heterogeneity of mouse pluripotent cells, in that some subpopulations might have

mCH methylation on the satellite and others not. In the DNA from a mixed pluripotent cell population this will appear as a lower percentage of mCH methylation for the locus, as we normally see non-CG methylation. It is interesting that in 3 out of the 19 sequenced clones of the major satellite we find 3 positive calls for mCH methylation, and such reads will be filtered out from the widely used filtering approach of high depth datasets, while the methylation can be entirely biological and not an artefact – all those three clones are very well converted elsewhere. This means that even with the high depth WGBS datasets real signal is inevitably being filtered out. On the other hand, in our M13 spike in conversion controls, we have observed BS-conversion resistant sequences appearing with more than 80% methylation (results not shown), and they would pass those filters, contributing to artefacts even in the high depth WGBS datasets. No doubt, defining the real and complete patterns of non-CG methylation currently remains a big challenge.

In summary, the evaluation of the established techniques for their feasibility to deliver valid results for non-CG context, modifying them if necessary, or developing novel techniques, would be crucial for identifying its distribution, and paving the way towards more complex questions about the potential function of this type of methylation.



## 4 Novel tools and techniques for the analysis of non-CG methylation

### 4.1 Introduction

As shown in Chapter 3 the classical techniques for non-CG DNA methylation analysis did not provide satisfactory results characterising the genomic distribution and developmental dynamics of mCH methylation. This highlighted the need to modify the existing techniques or, even, develop novel tools and techniques. As discussed in 1.6 (Introduction), such tools would involve 1) restriction enzymes specific for the modification – either inhibited, or activated by it, 2) modification-specific antibodies or binding proteins, 3) enzymes or chemical treatment creating, modifying or removing it. However, those are much more straightforward with a chemically distinct modification, rather than a context specific, and many of the approaches applied for 5hmC and its derivatives 5fC and 5caC would not be applicable for methodological distinction of 5mCG from 5mCA.

Restriction enzymes inhibited by methylation in non-CG context like the CCWGG have already been used in Chapter 3, with limited success. Nevertheless, it would be worth looking for others and testing more, if possible. The situation with restriction enzymes activated by non-CG context methylation, or at least not specifically by CG-context, seems more promising in the light of the recently described MspJI family of enzymes discussed in 1.6.3 (Cohen-Karni et al. 2011). One of the family members –RlaI, recognizes specifically non-CG context – S(V)<sup>m</sup>CW, where W is A or T. Half of the other members, including FspEI, LpnPI and SgrTI, recognize a <sup>m</sup>CDS motif (D being A, G or T) and are therefore biased towards mCG because this will be the dominating methylation context in their recognition configuration. The remaining two enzymes – MspJI and AspBHI, potentially including FspEI (for its optional C<sup>m</sup>C preference), cut all methylated cytosines, without a preference for the immediate neighbouring base (<sup>m</sup>CNNR for MspJI and YS<sup>m</sup>CNS for AspBHI). MspJI is commercially available and this makes it a potential candidate for the analysis of non-CG methylation, particularly for methylation enrichment methods where it will enrich for 5mC in both CG and non-CG context – an improvement of the current MspI-based RRBS method for CG enrichment. Moreover, this enzyme was already used for the generation of a 5mC enriched library for a NGS genome-wide analysis without BS-conversion (Cohen-Karni et al. 2011).

Antibodies capable of not only recognising the modified base, but also distinguishing the surrounding neighbouring bases, have not been generated to date. A priori, it is questionable whether this would be possible, although there is one very early report on the generation of antiserum against the (phospho)AT dinucleotide (Khan et al. 1977). While this is a promising achievement, our criterion is much more demanding, with the need to recognise both the nucleotide context and modification state of the cytosine as key. Nonetheless the advantage of such a tool justifies an attempt to generate the antibody.

In terms of non-restriction enzymes acting on CH context rather than CG or vice versa, the situation is again very challenging in comparison to the other cytosine modifications. The mammalian Dnmt3 family of methyltransferases, which are responsible for the *de novo* methylation, do not possess a single context preference, although they do have a target preference for CG (Gowher & Jeltsch 2001; Aoki et al. 2001). It would therefore be best to explore the possibility to find bacterial non-CG context methyltransferases, to use either as an *in vitro* CH methylation tool like the CG-specific M.SssI, or in transfection assays in the quest for a biological function of mCH (see 1.6.3). In terms of enzymatic modifications or the physical removal of the mCH mark and its stability, as discussed in 1.4 and 1.7, those have not yet been studied and are a matter of a separate investigation in this project, which will be discussed in more detail in Chapter 7. In terms of non-enzymatic chemical modifications, indeed there are ways to chemically modify 5mC (Tanaka et al. 2007), but any chemistry which works on CG context would work on non-CG context as well, which excludes this approach as an option for the current study.

## 4.2 Aims

1. To investigate the possibility for utilising new restriction enzymes sensitive to non-CG context methylation
2. To develop a MspJI-based method for reduced representation genome-wide analysis which will enrich for methylated sequences in non-CG context (or will at the least not exclude them)
3. To investigate the possibility for the development or utilisation of novel affinity tools for the functional analysis of non-CG methylation (antibodies) or methyltransferases
4. To look into the possibility for developing novel methods for measuring global amounts of context-specific methylation with those tools

## 4.3 Results

### 4.3.1 Enzymes specific for non-CG context methylation

#### 4.3.1.1 Restriction enzymes sensitive to non-CG context methylation

Although hundreds of restriction enzymes have been characterised, and many are commercially available, the enzymes recognizing cytosine-containing sequences are a small proportion, and most of those recognize the CG context. Table 3 shows a summary of my investigation, where several enzymes reported as sensitive to 5mC in non-CG context were tested for this project. Individual activities were tested on PCR fragments derived from the 2 kb M13 fragment used previously (all sequences are listed in Appendix Table 13). Each digest was performed according to the manufacturer's recommendations for the relevant enzyme, and analysed on a 2% agarose gel stained with SYBR®Safe. Most of the enzymes were sensitive to both 5mC and 5hmC, although many show leakiness (less than 100% sensitivity), and some were sensitive only to symmetric methylation on both strands, which is not likely to be physiologically relevant, as discussed in 3.3.5.2. The search for isoschizomers similar to MspI/HpaII, used widely for the analysis of CG-methylation, was unsuccessful, apart from the already known enzyme BstNI (Woodcock et al. 1987; Woodcock et al. 1988) which can be paired with PspGI, specific for CCWGG. However, as shown in 3.3.5.2, the isoschizomer counterparts of the same pair used for LUMA analysis (AjnI and Psp6I, SibEnzyme), failed to produce robust results for CCWGG. Two enzymes showed potential to form a pair for CATG, however one of them (NlaIII) showed leakiness and the other - some sequence preference ambiguity (CviAII), meaning they can potentially be used for a proof of principle analysis, but not for quantitative assays like the MspI/HpaII pair.

Although all of these enzymes were analysed for their potential to use in different enzyme-dependent applications, as seen from Table 3, the ones marked for LUMA did not prove useful when tested in the actual assay (3.3.5.2). The enzymes marked for MS-REA, however, can be used to validate locus specific methylation provided it is within the relevant recognition sequences, as well as the ones for NNA.

Table 3. Restriction enzymes sensitive to 5mC in non-CG context (in red) and their non-sensitive isoschizomers, where available (in black). Results are summarised based on real restriction digests and agarose gel electrophoresis for each individual enzyme.

	Cuts	Result from test	Application	Similar
<b>CCWGG</b>				
BstNI	CC↑WGG	Cuts well C, 5mC and 5hmC	MS-REA	
MvaI	CC↑WGG	Not tested myself		
<b>PspGI</b>	↑CCWGG	does not cut, not leaky	LUMA, MS-REA	EcoRII
<b>Psp6I</b>	↑CCWGG	does not cut, not leaky	LUMA, MS-REA	EcoRII
Ajnl	↑CCWGG	does not cut C, 5mC or 5hmC to 100%	LUMA, MS-REA	
<b>CCATGG</b> very rare combination, rarely methylated				
<b>Bsp19I</b>	C↑CATGG	weak activity	LUMA	
NcoI	C↑CATGG	weak activity	LUMA	
<b>CATG</b>				
<b>NlaIII</b>	CATG↑	LEAKY in 5mC, less in 5hmC	MS-REA	FaeI
CviAII	C↑ATG	slightly resistant to 5hmC	NNA, MS-REA	FatI
<b>CTAG</b>				
<b>BfaI</b>	C↑TAG	weak activity		FspBI
<b>MaeI</b>	C↑TAG	LEAKY in 5mC and 5hmC	MS-REA	
<b>XspI</b>	C↑TAG	Good, slightly leaky in 5hmC; but cuts hemi-methylated CTAG	NNA, MS-REA	FspBI

Enzyme name	Species	Recognition site without activator	Recognition site with activator
MspJI	<i>Mycobacterium</i> sp. JLS	<sup>m</sup> CNNR(G > A)	<sup>m</sup> CNNR
FspEI	<i>Frankia</i> sp. EAN1pec	C <sup>m</sup> C	C <sup>m</sup> C or <sup>m</sup> CDS
LpnPI	<i>Legionella pneumophila</i> Philadelphia 1	S <sup>m</sup> CDS(G ≫ C)	<sup>m</sup> CDS
AspBHI	<i>Azoarcus</i> sp. BH72	YS <sup>m</sup> CNS	YN <sup>m</sup> CNS
RlaI	<i>Ruminococcus lactaris</i>	S <sup>m</sup> CW	V <sup>m</sup> CW
SgrTI	<i>Streptomyces griseoflavus</i> Tu4000	C <sup>m</sup> CDS	B <sup>m</sup> CDS

N = A or C or G or T; D = A or G or T; B = C or G or T; V = A or C or G; R = A or G; S = C or G; W = A or T; Y = C or T.

Figure 25. MspJI family of restriction enzymes specifically cutting 5mC-containing DNA (figure from Cohen-Karni et al, 2011).

Another promising group of recently described enzymes, is the MspJI group of restriction enzymes (Cohen-Karni et al. 2011) – they cut only if 5mC is present in the DNA, therefore they can be used for selective targeting of methylated sequences. One of them recognises specifically 5mC in non-CG context (RlaI), the others are not non-CG context specific. Of this group of enzymes, only MspJI, FspEI and LpnPI are commercially available.

#### 4.3.1.2 Synthesis, cloning and validation of RlaI

Since RlaI is the only enzyme in the MspJI group, which recognises 5mC in non-CG context, and specifically CA or CT, which are the most common non-CG contexts in mammals, I decided to obtain it as a tool for further analysis of non-CG methylation.

The *RlaI* gene was synthesised as described in 2.3.13, the sequence was specifically codon-optimised for expression in *E. coli* by algorithms in the GeneArt® platform. It was sub-cloned from the original entry vector into a bacterial expression vector pQLinkHD, as described in 2.3.14. Our collaborator Tomasz Jurkowski (University of Stuttgart) performed bacterial expression and purification of the protein.

The activity of RlaI was tested in a digest of the synthetically methylated oligonucleotides used for the LUMA Ajnl/Psp6I test because the  $\text{C}^{\text{m}}\text{CWGG}$  sequence represents an RlaI restriction site (Figure 26). However, it only partially digested the hemimethylated oligos, which represent the most likely *in vivo* scenario, and therefore further optimisation of reaction conditions would be necessary in order to use RlaI for mCA analysis, which could not be performed within the timeframe for completing the current thesis work.

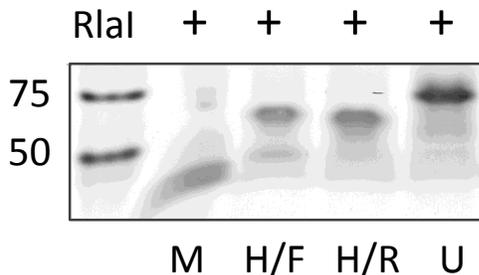


Figure 26. In vitro digestion of  $\text{C}^{\text{m}}\text{CWGG}$ -containing oligonucleotides with RlaI, visualized on a 6% TBE polyacrylamide gel. U stands for 'unmethylated', H for 'hemimethylated' respectively on the forward or reverse strands, M is 'methylated'.

### 4.3.1.3 Context specific bacterial methyl-transferases

In collaboration with Tamir Chandra (Reik lab) we identified bacterial methyl-transferases which show a strict DNA context specificity unlike the broader specificity of the mammalian native Dnmt proteins. The search for CA-specific methyl-transferase, as the most common mCH context in mammals (Ziller et al. 2011), led us to M.TspRI, which recognises and exclusively methylates the sequence motif \*CASTG (where S is G or C). The *M.TspRI* gene was synthesised in the same way as described for RlaI in 4.3.1.2 but optimised for mammalian expression, and cloned into a vector for eukaryotic inducible gene expression as described in 2.3.14.

Due to delays with the cloning and time constraints to complete the PhD, the M.TspRI ultimately was not used for transfection experiments and no further work with it is presented here.

### 4.3.2 *MspJI*-based methylation enrichment RRBS (*meRRBS*)

The results in chapter 3 revealed that low-resolution BS-seq datasets and MeDIP analysis against 5mC could not be used for genome-wide analysis of non-CG context methylation, either because of high conversion error masking real mCH signal, or because of the inability to discriminate between the 5mC contexts.

Next generation sequencing of an *MspJI*-based CG-enriched library has already been demonstrated successfully without bisulphite conversion (Cohen-Karni et al. 2011). This approach does not introduce a large proportion of false positives for the CG context, due to its high methylation levels, but it would for non-CG context and therefore it was not an option for us. We first quantitated the genomic frequency of the *MspJI* site (CNNR) to confirm that it was representative for a large fraction of genomic cytosines, including CH (Figure 27A). The high specificity of *MspJI* for 5mC and 5hmC, and not for unmodified cytosine, was also tested and confirmed experimentally (Figure 27B). I performed a modified protocol for classical *MspI* RRBS (Smallwood & Kelsey 2012), with a few modifications as described in Materials and methods (2.3.8). WT ES cell, pMEF and an unmethylated control DNA were used for pilot libraries. As a modification of the original protocol, special care was taken to preserve fragments smaller than 100bp, which are not normally preserved in the classical library preparation protocols. This was necessary due to the specific digestion pattern of *MspJI* in generating small size fragments from symmetrically methylated regions (Cohen-Karni et al. 2011).

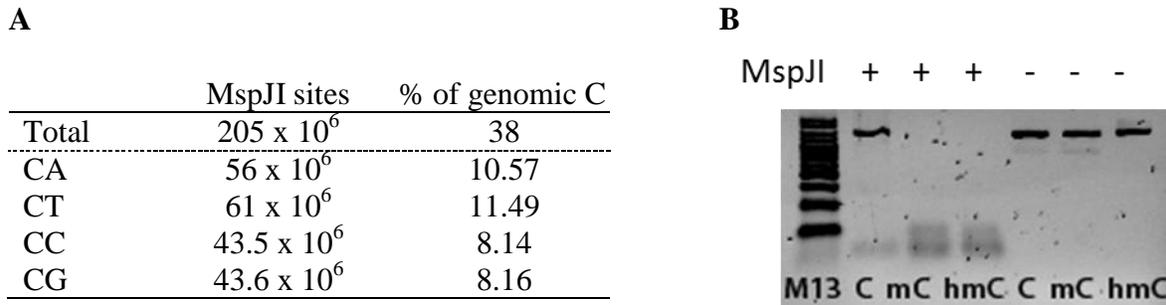


Figure 27. Characterisation of MspJI enzyme's restriction specificity. **A.** Frequency of the MspJI site in the genome and its context distribution. **B.** Specificity of MspJI validated on fully unmethylated, methylated and hydroxymethylated substrates.

The mapping efficiency of the three pilot datasets, however, was unexpectedly low (Table 4). For comparison, the mappability of the originally published non-bisulphite converted dataset was also very low, most likely due to the short length of reads. For this reason the further decrease of mappability after conversion is not surprising, due to the highly reduced sequence complexity of bisulphite-converted DNA (Krueger et al. 2012).

Table 4. Mapping efficiency of the MspJI meRRBS datasets, compared to the published IMR90 methylation enrichment dataset (Cohen-Karni et al. 2011).

	Mapping efficiency	Read number	Mapped reads number	Total analysed cytosines
J1	16.1%	21.24 M	3.43 M	64,526,866
pMEF	19.8%	19.18 M	3.81 M	79,940,702
TKO	8.5%	14.58 M	1.24 M	25,399,827
IMR90/NEB	32.7%	71.8 M	23.5 M	?

The analysis of the meRRBS datasets included mapping of the original MspJI site, upstream of the end of the sequence read. For this purpose both 3' and 5' ends of reads were analysed, and the probability of a valid methylation call was calculated based on expected digestion patterns (Figure 28A). The prediction was that each digestion would generate two DNA ends - one 'specific' 3' end, with the target 5mCNR site upstream, and one non-specifically cut 5' end, belonging to another fragment. Thus, half of the reads will be enriched for 5mC, and the other half will be entirely random, thus contaminating the signal (Figure 28B).



Thus analysed, the results showed high specificity, unlike the previously used approaches. The negative control TKO line showed barely detectable methylation ‘levels’, and the pMEF ES line possessed significantly decreased levels of non-CG context methylation (Figure 29).

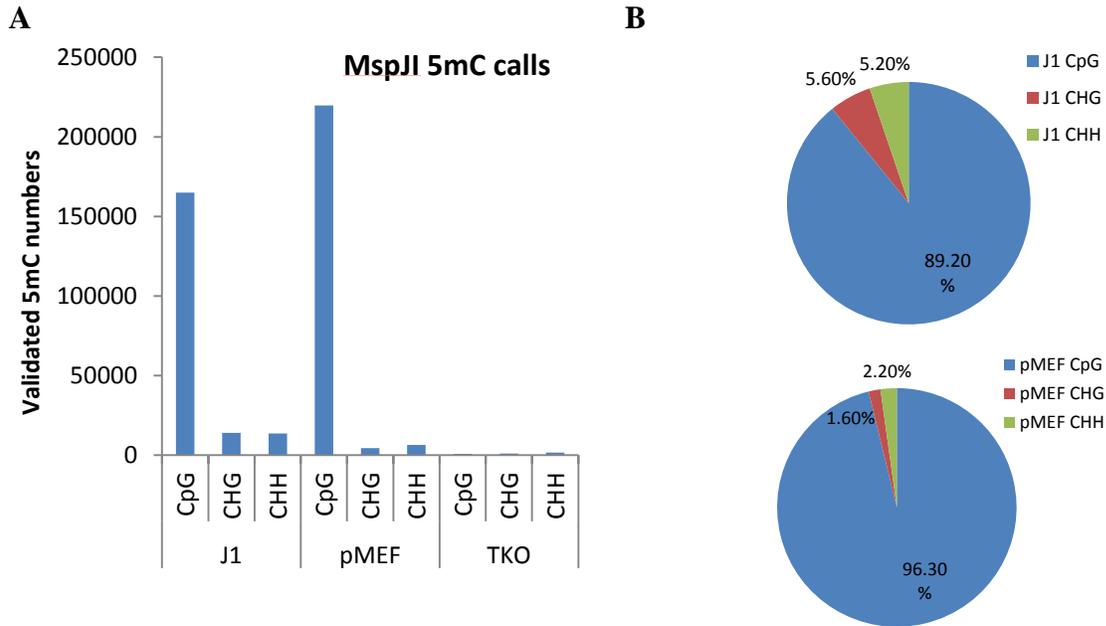


Figure 29. MspJI-RRBS positive methylation calls. (A) Total numbers in the three sequenced samples, and (B) proportionally in mES and pMEF.

The very low mapping efficiency, however, precluded using this approach for other biological samples of interest.

### 4.3.3 Development of context-specific antibodies for 5-methylcytosine

At present, all commercially available antibodies recognise the methylated cytosine itself, irrespective of the surrounding context, and they cannot be used to analyse context-specific methylation. To address this deficit and to enable the use of all antibody-dependent methods listed in Table 1, I initiated collaboration with Active Motif Inc for the production of antigens against the two most common methylation contexts - 5mCA and 5mCG. The company has presently established themselves as the best on the market for producing antibodies against DNA modifications. Unlike the common antibody, raised against 5-methylcytosine, these two had to be

raised against the two full dinucleotides, including the modified cytosine base, the neighbouring base and the sugar-phosphate backbone (Figure 30).

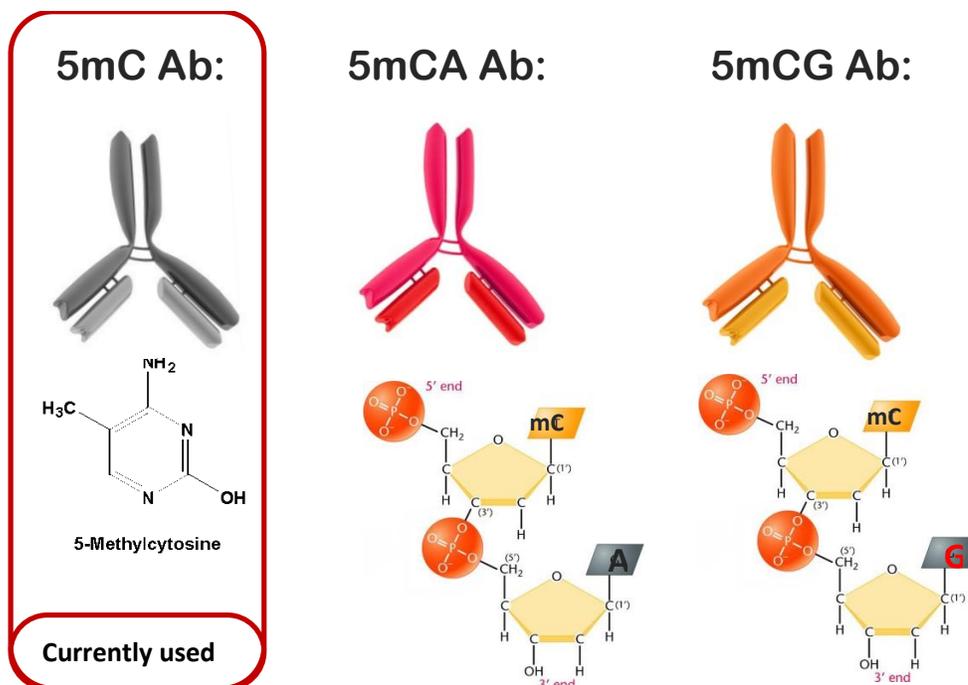


Figure 30. The specificity of the currently available antibody and the difference with the context-specific antibodies.

#### 4.3.3.1 Antigen synthesis and monoclonal antibody production

The antigen was synthesised by Active Motif employing the same linkage strategy, which was used for raising their very successful DNA modification-specific antibodies, some of which are listed in Table 6. Keyhole limpet hemocyanin (KLH) was used as a carrier protein.

The production of the antibodies was carried out by the Technology Development Lab (TDL) at the Babraham Bioscience Technologies (BBT), Babraham Research Campus. Five mice were subjected to immunisations for each antigen – labelled in figures as NO01 for mCA and NO02 for mCG. The specificities of the sera and subsequent hybridoma clones were tested with a quantitative ELISA assay, specifically developed for this purpose (see 4.3.3.3).

The mouse with the highest serum titre for each antigen was selected for splenectomy and hybridoma clones and sub-clones were subsequently screened for specificity.

### 4.3.3.2 Validation strategy

The initial clonal screen was carried out by the TDL in accordance with the collaboration agreement; it involved a single positive control and a single negative control. Due to the non-conventional nature of the antigens, it was necessary to perform further tests after the initial screen with more controls in order to assess antibody quality and choose the least cross-reactive clones. Two types of tests were performed – ELISA and immunofluorescence (IF).

Since all eukaryotic cell lines contain predominantly CG context DNA methylation, or mixed context methylation, those cell lines cannot be used to evaluate the context specificity. None of the existing mES Dnmt-KO lines have only CA or only CG context methylation to make their DNA appropriate as a control. Therefore, the initial specificity validation step was performed *in vitro* with a panel of context-specific methylated synthetic oligos. The sequences chosen for the oligos are native genomic DNA regions from the major satellite repeat (oligos #1-#8) and the LINE1 5' UTR (L1 #9-#12), which have shown high levels of native 5-methylcytosine in CH context as shown in Chapter 3. The panel consists of oligos containing 5mC in all four contexts, plus non-methylated oligos in the CA or CG context, as well as two oligos with inverted neighbouring sequences – the dinucleotide next to the methylated cytosine is inverted to the original positive control. Thus, I ended up with two positive control sequences and ten negative control sequences per context (Table 5).

Table 5. Oligonucleotide sequences used for the validation of mCA and mCG sera.

#	Oligo specificity	Name	Sequence
1	mCA positive control	mCA oligo #1	TGAGAAATG[mC]A <u>C</u> ACTGAA
2	mCG positive control	mCG oligo #2	TGAGAAATG[mC]G <u>C</u> ACTGAA
3	CA native sequence, no mC	CA oligo #3	TGAGAAATGC <u>C</u> ACTGAA
4	mCC negative control	mCC oligo #4	TGAGAAATG[mC]C <u>C</u> ACTGAA
5	mCT negative control	mCT oligo #5	TGAGAAATG[mC]T <u>C</u> ACTGAA
6	inverted dinucleotide for mCA	mCCA oligo #6	TGAGAAATGA[mC]C <u>A</u> CTGAA
7	inverted dinucleotide for mCG	mCTG oligo #7	CTGTAGGAG[mC]T <u>G</u> GAAATAT
8	CG native sequence, no mC	CG oligo #8	TGAGAAATGC <u>G</u> CACTGAA
9	mCA positive control	L1 mCA	CAAGCTCTCCTCTTG[mC]AGGGGAAGGTGCA
10	mCG positive control	L1 mCG	CAAGCTCTCCTCTTG[mC]GGGGGAAGGTGCA
11	CA native sequence, no mC	L1 CA	CAAGCTCTCCTCTTG <u>C</u> AGGGGAAGGTGCA
12	CG native sequence, no mC	L1 CG	CAAGCTCTCCTCTTG <u>C</u> GGGGGAAGGTGCA

The context specific oligos were first used to screen a large number of supernatants provided by our contractor via the ELISA assay (see 4.3.3.3). The successful candidates from this screen were further tested in a biological context via IF.

#### 4.3.3.3 Avidin-biotin DNA ELISA

In order to evaluate properly the specificity of the context-specific animal sera and hybridoma supernatants, a more quantitative assay than the standardly used dot blot had to be in place. Direct immobilisation of DNA on the polystyrene surface did not produce satisfactory results with the context-specific oligonucleotides for specificity validation (Figure 31A) – 400 ng of immobilised oligo gave very low signal, only 0.2 OD points above background. This is probably due to the very small size of the fragment and a low total amount of antigen: 1 x 5mC in an 18 bp long oligo, i.e. 4.5-fold less than in the M13 R2L PCR fragments). Therefore, I tried a sandwich ELISA, using an avidin-biotin system. I ordered the same oligos with a biotin modification on the 5' end, and attached them to an avidin-coated plate. For coating the plate I used a novel deglycosylated modification of avidin by Pierce, Thermo Scientific called NeutrAvidin®, which has the same affinity, but higher specificity and much lower background binding than its predecessors. The assay worked really well, it produced three-fold stronger signal with only 25 ng oligo per well, which means this system needs 20-fold less oligo (Figure 31B). Two pH buffers were tested in case pH mattered in this assay as well, and better results were achieved with 1 x PBS, pH ~7.0 (Figure 31B). The NeutrAvidin (Nav) coating was tested with dropping amounts of Nav per well at a constant antibody dilution (1:2000) and saturation of the well was reached at around 250 ng of protein, after which the signal plateaued and did not change (Figure 31C). The strength of the signal was repeatedly the same with 20 ng and 100 ng of oligo showing that the binding capacity of the Nav used in this assay is already saturated with 20 ng of oligonucleotide (Figure 31C, D), is in agreement with the binding kinetics revealed in Figure 31B. All optimisation experiments were done with the 33D3 5mC mAb clone (Eurogentec). Figure 31D shows that this commercial clone has a higher affinity preference for mCA context in comparison to mCG in certain DNA sequences.

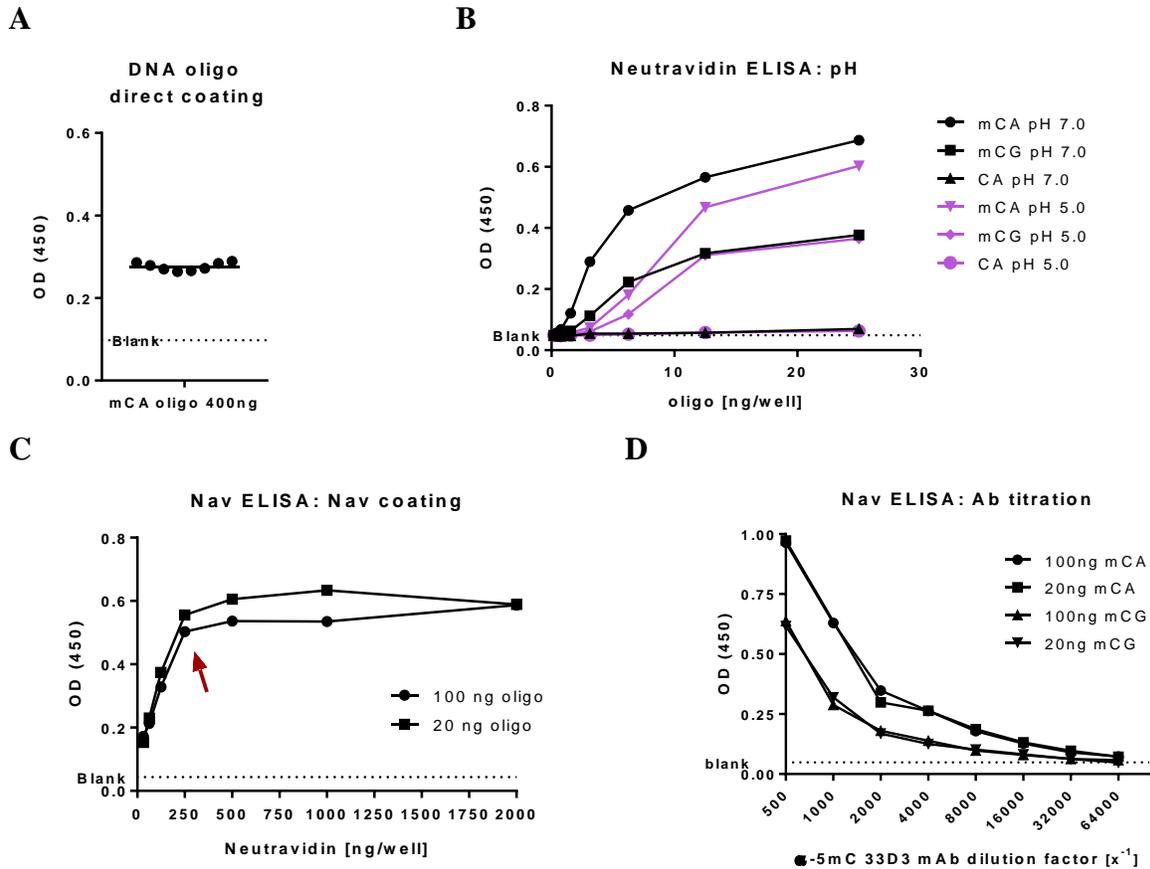


Figure 31. Optimisation of the avidin-biotin ELISA assay: **A**. oligo usage and signal intensity compared between direct immobilization in pH 5.0 and biotinylated oligos on the avidin-coated plate; **B**. Oligo binding titration with 500 ng/well Nav and 1:2000 5mC 33D3 mAb; **C**. Neutravidin titration with fixed amounts of oligo (details in figure legend) and 1:2000 5mC 33D3 mAb; **D**. Antibody titration with two fixed oligo concentrations and two methylation contexts at 300 ng Nav/well

#### 4.3.3.4 Results: ELISA screen

Ten supernatants for NO01 (mCA) were provided by our subcontractor to evaluate with the full panel of oligonucleotides, together with the mouse serum. The standardly used anti-5mC antibody was included in the analysis for reference.

Six of the tested mCA supernatants showed cross reactivity for mCCA context of unmethylated CA (7E1, 7A3, 7A9, 8H8, 8G8, 8G2), one was very specific although weak (2C8), and three demonstrated very weak to undetectable activity (5H6, 9H5, 4C10) (Appendix Table 19). The clone 2C8 was chosen for further evaluation with IF (Figure 32, all mCA supernatants are shown in). Five clones were tested for mCG – because of their higher overall specificity and

avidity (Figure 32, all mCG supernatants are shown in and Appendix Table 20). The two top candidates from mCG were 10E2 and 6F5 due to their high affinity and avidity, although all clones showed remarkable specificity and no background in the negative controls. However the 6F5 clone was lost in subsequent subcloning and only 10E2 was used for further assessment and experiments.

The 33D3 anti-5mC mouse monoclonal antibody again showed preference towards mCA in one of the oligo sequences (major satellite, see first report in Figure 31), but not the other (L1), showing that its affinity is affected by the flanking sequences.

Antibody species

33D3 mouse mAb  
Eurogentec

Mouse monoclonal supernatant 2C8 - mCA

Mouse monoclonal supernatant 10E2 - mCG

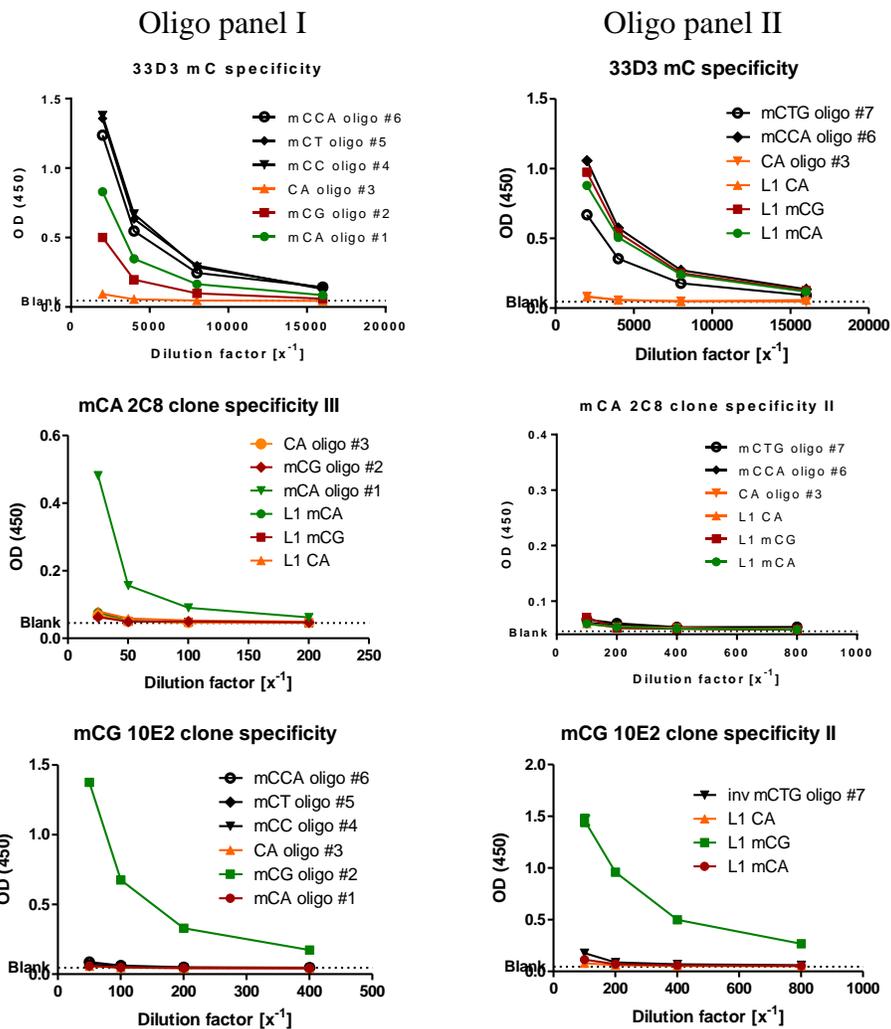


Figure 32. Specificity of mCA (2C8) and mCG (10E2) supernatants in comparison to the commercially available 5mC antibody. Graph colour coding: **Green** – positive control (mCA or mCG respectively), **Red** – opposite negative methylation control: mCG for mCA, and mCA for mCG, **Orange** – unmethylated DNA, **Black** – other negative methylation controls – mCC and mCT (not biologically relevant)

#### 4.3.3.5 Results: immunofluorescence (IF)

In order to verify that the specificity will remain the same in a more complex environment and to test for potential cross reactivity in isolated genomic DNA as well as in a chromatinised DNA template, it was important to test the monoclonal supernatants in biological applications. Mouse ES cell lines, and all the known MTase KOs would have both CG and non-CG mixed methylation, therefore they are not an ideal model for this validation. The best candidate should have either only mCG or only mCA, or even better if the two contexts are present in the same cell but can clearly be differentiated, spatially or structurally. Such an ideal candidate cell actually exists, and this is the mouse zygote. It has been reported even in the early days of non-CG methylation studies, that oocytes have high levels of this type of modification and that it is predominantly in CA context (Imamura et al. 2005; Haines et al. 2001). These early observations were later confirmed and investigated in more detail, when it became clear that the non-CG methylation in the oocyte is as much, if not more, than the CG methylation (Tomizawa et al. 2011; Shirane et al. 2013). Sperm on the other hand is very highly methylated, but exclusively in CG context (Tomizawa et al. 2011; Haines et al. 2001). Therefore, after fertilisation, it is expected that the maternal pronucleus will be still very high in non-CG methylation, but will also have mCG, while the paternal will have only CG methylation. This has been experimentally shown by (Haines et al. 2001) for individual imprinted loci, which retain high levels of mCA methylation even during 2-cell stage, and lose the mCA methylation quickly after that, while the paternally imprinted loci do not acquire mCA methylation at any stage. We have therefore stained mouse fertilised oocytes (performed by Fátima Santos as described in 2.3.17) and the two candidate clones were tested alongside the conventional 5mC antibody. The resulting patterns from the two context-specific antibodies show remarkable specificity, and behave entirely as anticipated (Figure 33). The mCA clone does not show any cross-reactive staining on the paternal pronucleus, while the mCG clone stains both pronuclei.

This test demonstrates that the new antibodies are capable of recognising the right target in a chromatinised *ex vivo* environment and are perfectly suitable for experiments requiring direct visualisation. It also confirms that the conventional 5mC antibody recognises both mCG and mCA methylation in IF experiments, as it has shown with the synthetic ELISA probes (4.3.3.3). The maternal pronucleus shows somewhat stronger signal with the commercial 5mC antibody,

combining the mCA and mCG signals in one. More importantly, this experiment also validates the observations reported by (Haines et al. 2001) that the maternal pronucleus retains its high levels of mCA methylation after fertilisation, when the paternal pronucleus undergoes quick active demethylation (the paternal demethylation is also seen in the middle and right panels of Figure 33). Our observations further this report by showing that mCA methylation is preserved in the entire maternal pronucleus, and not only in selected imprinted loci. It has been well documented that the majority of 5mC in the mouse maternal pronucleus does not actively demethylate after fertilisation (Nakamura et al. 2007; Santos et al. 2013), and our data shows that this applies to both mCG and mCA methylation, possibly via the same protection mechanism.

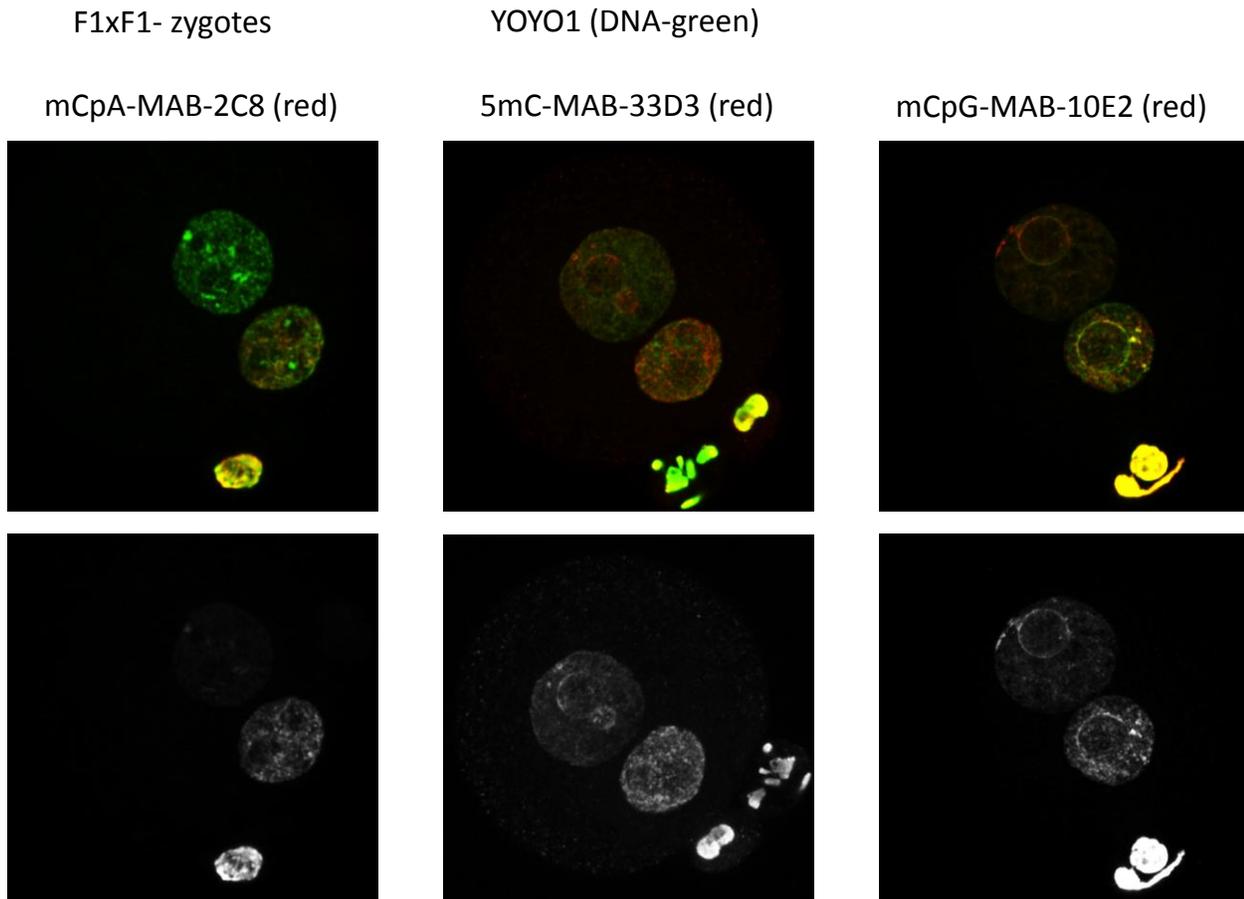


Figure 33. Immunofluorescence of mouse zygotes: left panel shows mCA, middle panel shows 5mC (all contexts) and the right panel shows mCG. DNA is in green (top coloured panels only), and the methylation is in red (top panels) or black and white (bottom panels). IF by Fátima Santos.

### 4.3.3.6 Antibody concentration and purification

After successfully validating the specificity of the two clones, they had to be prepared for further use in experiments. Media from hybridoma culturing (here referred to as ‘supernatant’) contain very diluted amounts of antibody. They can therefore be used straight or in low dilution ratios for ELISA or IF assays, but cannot be used for assays like MeDIP, where the high concentration of antibody in a smaller reaction volume is important for its optimal efficiency. Therefore, it was necessary to concentrate the supernatant, as described in 2.3.15.

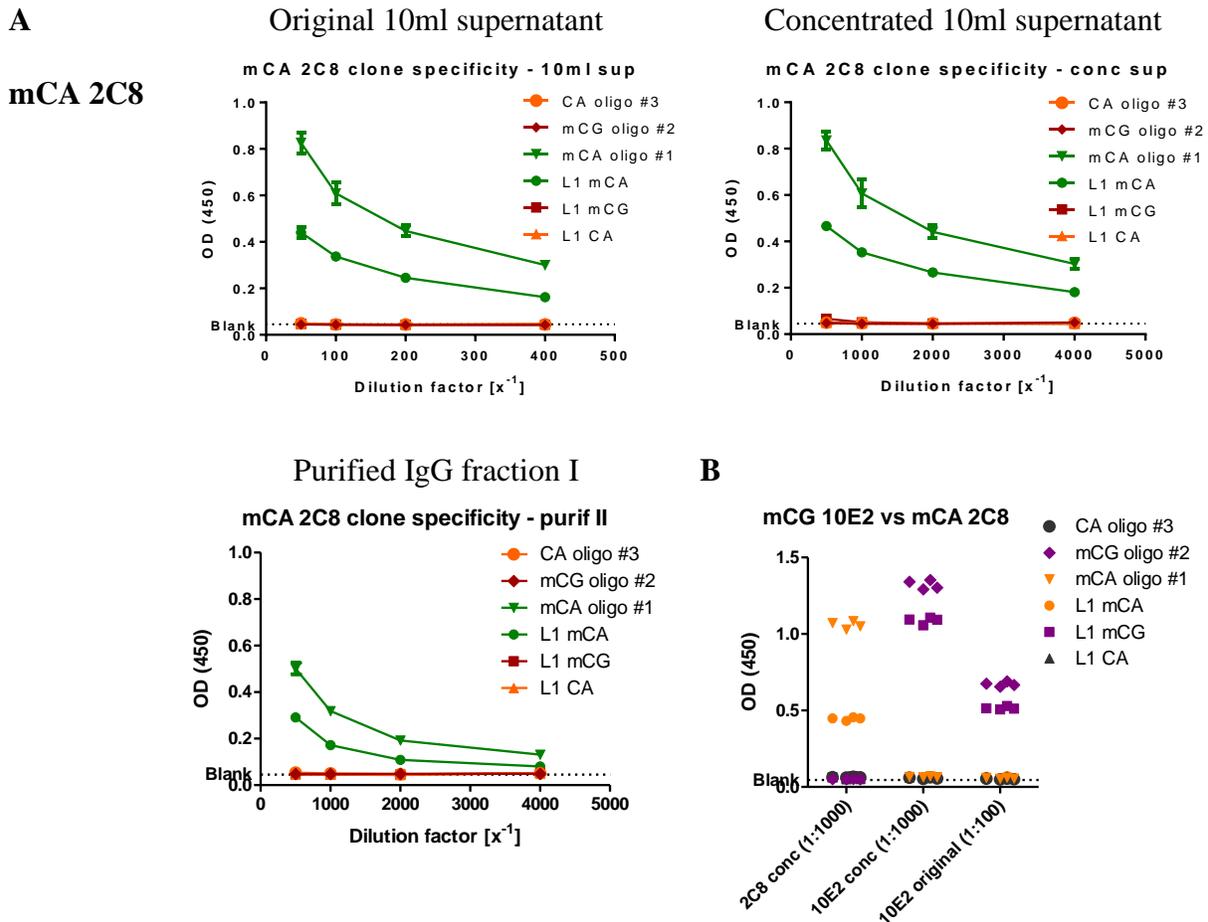


Figure 34. Concentration of the supernatant of 2C8 mCA and 10E2 mCG clones and subsequent Protein G resin purification (for mCA only). **A**. mCA, **B**. mCG.

The purified and concentrated supernatant was tested with an ELISA assay (Figure 34). Two fractions were tested – one concentrated supernatant, and one concentrated and subsequently purified with a G-resin. The concentrated supernatant showed ~10x higher activity than the

original supernatant, slightly lower than expected due to potential loss of protein during purification. The purified fractions have lower activity than the concentrated supernatant but higher than the original supernatant. Since the concentrated-only supernatant does not show any background and is in very high concentration, it seems most appropriate for use directly, including methods like MeDIP.

#### **4.3.4 DNA ELISA**

As shown in Table 1, introduction section 1.6.1, there are no easy to use high-throughput techniques for the fast measurement of global genomic DNA methylation. In addition, my results from chapter 3 make it clear that this is even more complicated when it comes to measuring global context specific methylation. For this reason, the development of context-specific antibodies provided an impetus to look into quantitative antibody-based methods, which could be applied to measure global DNA methylation more accurately than the dot blot technique. Such a method, well-established in the research and clinical practice, widely used for the highly sensitive and specific detection or quantitation of a given antigen, is the Enzyme-Linked Immuno-Sorbent Assay (ELISA) (Engvall & Perlmann 1971; van Weemen & Schuurs 1971). The technique is used routinely for both protein analysis and for the detection of pathogens or biomarkers, and has a very good quantitative capacity (Lequin 2005). An ELISA protocol for the analysis of DNA modifications has not been published in the scientific literature to date, although an antibody against 5mC has been available since the early 1980s (Achwal & Chandra 1982). Several companies offer ELISA-based kits for the detection of global DNA methylation, but they deliver inconsistent results and have therefore struggled to become widely used in the community (observations from our lab and personal communication to other scientists). The dot blot on nylon or nitrocellulose membranes is still used for the detection of DNA modifications, although it lacks a lot on the quantitative aspect for comparisons of varying levels of DNA methylation between samples.

A potential reason for the absence of an ELISA protocol adapted for DNA applications might be sought in the first step of the method - the immobilisation of the molecule. An essential aspect of the method principle requires that one of the interacting phases (antibody or antigen) is immobilised onto a solid surface – the polystyrene bottom of a 96-well microtitre plate. The attachment of proteins is due to passive adsorption via hydrophobic interactions, a result of full or

partial denaturation of the protein of interest in the alkaline salt-free ELISA coating buffer, and its subsequent nonspecific adsorption to the well surface. High-binding surface plates with an improved binding capacity have been developed after X-ray irradiation, leading to the exposure of negatively charged carboxyl groups on the polystyrene surface. Those groups attract electrostatically the charged proteins, in addition to the unspecific hydrophobic interactions, and make the adsorption more efficient. DNA is an acid with a significant amount of negative charges packed in close proximity, which could potentially affect its binding to the plate in two ways: 1) lack of true hydrophobic interactors from the DNA molecule, and 2) repulsion between the individual DNA strands leading to less efficient binding. I nevertheless set out to investigate the DNA adsorption possibilities.

#### *4.3.4.1 DNA adsorption to the 96-well polystyrene plate*

My search for positively charged plates, which could attract electrostatically the negatively charged DNA molecules lead to no success. One technology developed by Corning (DNA-BIND™) is utilising an N-oxysuccinimide esters modified surface, which can bind covalently to terminally aminated DNA oligos. Since it requires a chemical modification of the DNA strands prior to the assay, this technology would not be useful for non-amino-modified native genomic DNA. I nevertheless decided to try a direct binding of DNA fragments to the ELISA plate surface. To address the potential problem with the negative charge of DNA, I tested three types of coating buffers: 1) the standard ELISA protein coating 50 mM Carbonate-bicarbonate buffer, pH 9.6; 2) 50 mM Tris-HCl, pH 7.0, and 3) 50 mM Na Acetate buffer, pH 5.0. I chose pH 5.0 specifically as the value pointed out as the isoelectric point (pI) of DNA (Cai et al. 2006) at which the DNA molecule will be expected to have a total charge of zero, while it will be negatively charged at pH 7.0 and even more so at pH 9.6. I chose to test the direct binding on plain non-irradiated ELISA plates (i.e. so called ‘medium binding plates’) because the negative charge of the irradiated plates would repulse the DNA molecules.

Double stranded M13-derived DNA fragments obtained in PCR (2.3.6) were used for the adsorption tests, with or without DNA modifications in order to test against the relevant antibodies (see Table 6). Variable DNA quantities (100 ng and 200 ng) were added to each of the buffers, denatured for 5 min at 99°C and incubated on ice for 10 minutes. The DNA was then added to a medium binding plate (two replicate wells and two concentrations), including two

control wells with no DNA, and incubated for 1 hour at 37°C. I first detected the presence of attached DNA with a fluorescent single stranded DNA (ssDNA) dye – OliGreen® (Invitrogen) by washing and incubating 5 minutes with TE buffer plus 1:200 OliGreen, and reading on a CytoFluor II microplate reader. The result showed that DNA was successfully immobilised to the plate surface in the buffers with pH 7.0 and pH 5.0 but not with pH 9.6 (Figure 35A left). This indicated that the charge of DNA plays a role in its ability to attach to polystyrene. The immobilisation of DNA at pH 5.0 and pH 7.0 was further confirmed with an antibody against 5hmC (the DNA fragment contained 5hmC) as described in 2.3.16, and the wells with pH 9.6 were again negative (Figure 35A, right).

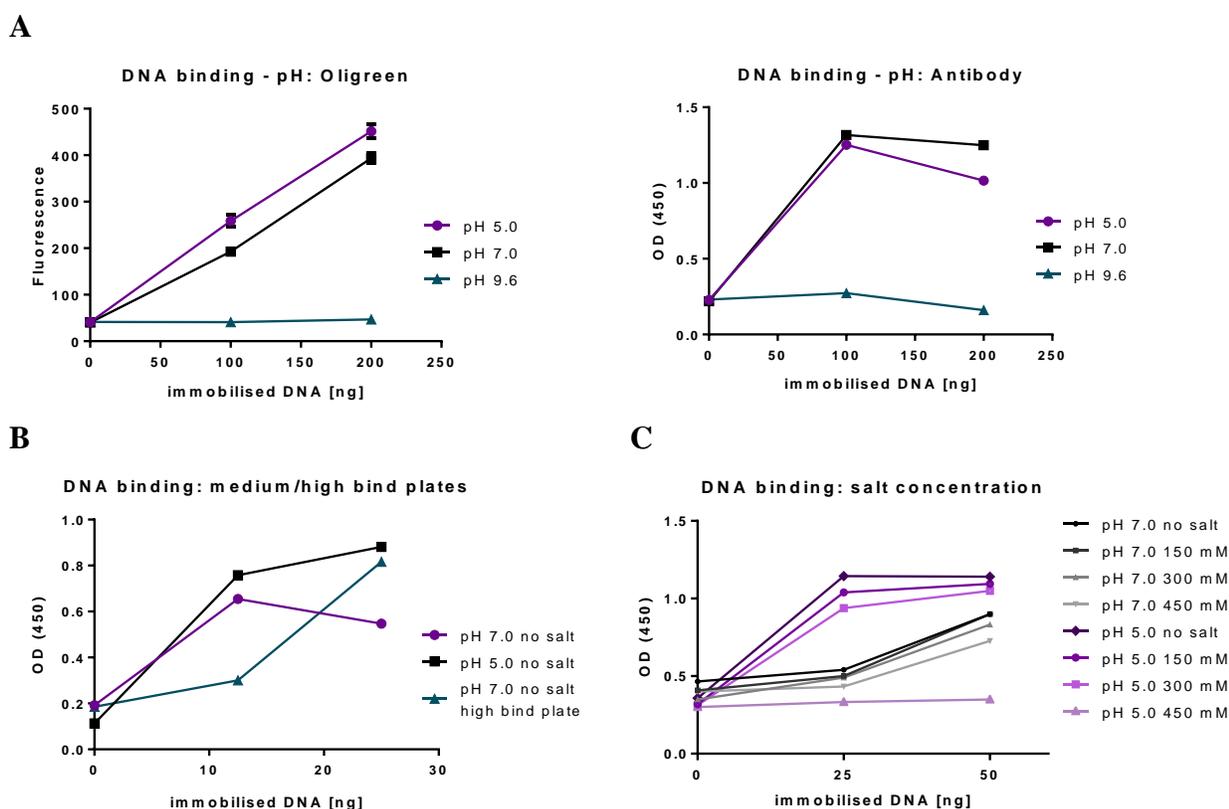


Figure 35. Optimisation of DNA adsorption. **A.** three different pH conditions were tested and measured with OliGreen (left) and antibody (right); **B.** Importance of charge for the adsorption was further demonstrated by using a negatively charged ‘high bind’ surface plate; **C.** higher salt was shown not to benefit the adsorption of DNA on the hydrophobic surface

In order to verify that indeed the charge makes the difference in the observed results, I used a ‘high bind’ negatively charged plate to test if the binding would be weaker. For this, I used the

pH 7.0 Tris-HCl buffer in which both the DNA and the ‘high bind’ plate will be negatively charged, while the ‘medium bind’ plate will remain uncharged. The result showed that the binding of DNA was weaker on the ‘high bind’ surface in comparison to the untreated neutrally charged surface and this must be a direct effect of the repulsion between the negative charges of the plate and DNA (Figure 35B).

I tried further optimisation using different salt concentrations. Charged ions can influence molecular interactions and therefore can facilitate or decrease the binding affinities. I prepared four different NaCl concentrations with both successful pH 7.0 and pH 5.0 buffers. The results were assessed only via the complete ELISA protocol with an antibody, as in 2.3.16. Higher salt concentrations (300 and 450 mM NaCl) showed marginally less binding at pH 7.0, but a very strong negative effect in acetate buffer: 300 mM NaCl leads to a 25% decreased attachment, dropping down to zero at 450 mM NaCl (Figure 35C). The highest signal was obtained from either no salt at all, or at physiological salt concentration (150 mM NaCl), both for pH 7.0 and 5.0. The DNA attachment in pH 5.0 was again stronger than in pH 7.0, making the no salt acetate buffer the top choice for DNA ELISA coating.

#### 4.3.4.2 *Optimising assay conditions for genomic DNA*

After establishing the DNA immobilisation conditions, I tested a panel of antibodies against DNA and its modifications to check if they could be used in an actual quantitative assay. Again, I used the M13-derived DNA fragments with incorporated 5mC, 5hmC, 5fC or unmodified cytosine. Since the incorporation of modified cytosine bases with PCR creates either 100% modified or 100% unmodified DNA fragments, one can be confident about the high representation of each antigen – in fact many fold higher than in native genomic DNA, which has only 3.5% 5mC and 0.1% 5hmC on average. A survey of antibodies revealed that not every antibody was appropriate for the assay (Table 6).

The successful antibodies showed good titration and a very high specificity with very low background (Figure 36A-E). None of the tested commercial antibodies showed cross reactivity to unmodified DNA, or a different DNA modification (Figure 36A-C). Four of the tested antibodies – the 5mC, 5hmC, mCA and mCG also showed activity towards native genomic DNA, albeit lower, which is expected due to the lower content of modified cytosines.

Table 6. Antibodies tested for their feasibility for the ELISA assay and the results of their performance

Antibody	Company	Appropriate for ELISA
5mC BI-MECY-100 (33D3)	Eurogentec	Yes
5mC pAb #61255	Active Motif	No (results not shown)
5m-cytidine mAb	Cayman Chemical	No (results not shown)
5hmC pAb #39769	Active Motif	Yes
5fC pAb #61223	Active Motif	Yes
anti-ssDNA IgM MAB3299	Chemicon (Millipore)	No (results not shown)
anti-DNA MAB3034	Chemicon (Millipore)	Yes

Since this type of assay is strongly dependent on a correct estimation of DNA quantity, I have also looked for an appropriate DNA antibody, which can be used for either loading control or in the course of assay optimisation for DNA quantity (Figure 36D).

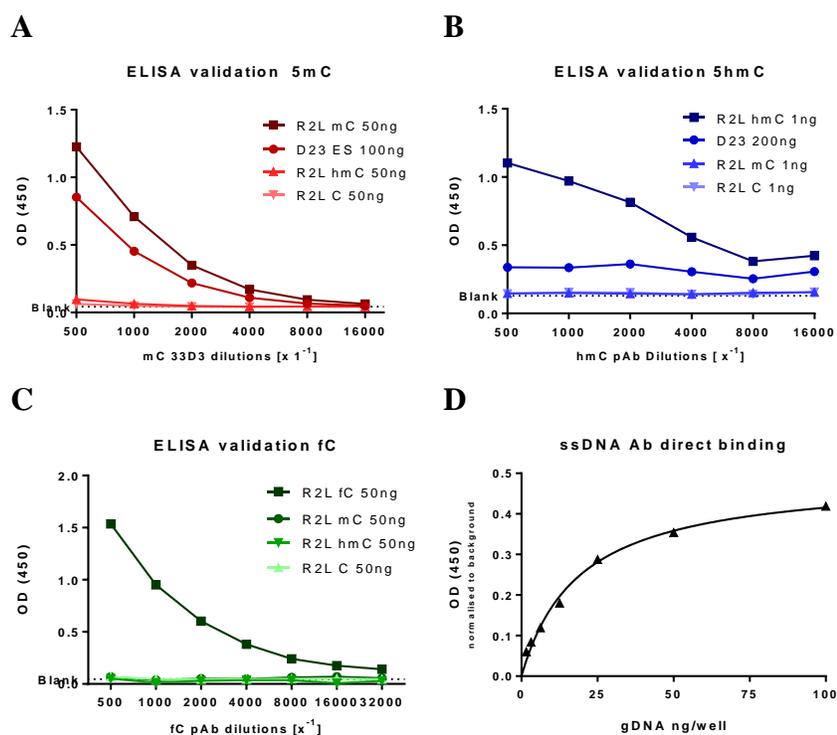


Figure 36. Validation of the ELISA direct DNA binding assay for the different DNA modifications. **A.** 5mC, **B.** 5hmC, **C.** 5fC and **D.** Validation of the use of an ssDNA antibody for loading control. All assays are performed with synthetic DNA fragments and trial genomic DNA.

Next, I performed calculations for the expected amounts of gDNA substrate for each modification in accordance with the antibody sensitivity, to get an idea about the sensitivity of the assay in comparison to dot blot. I compared the amount of cytosine per 100 ng of an R2L M13 fragment to the amount of each modification in 100 ng of genomic ES cell DNA (Table 7). M13 values were taken from its sequence (see Appendix Table 13) and the values for mouse gDNA have been taken from the annotated mouse genome (NCBI37/mm9). WT ES cell mass spectrometry results were: 5mC measured as 3.5% and 5hmC as 0.11% of total cytosine content, 5fC as 0.1 x 5hmC (Ito et al. 2010), mCA calculated as ¼ of all methylcytosines and mCG as ¾ of all methylcytosines in WT ES cells according to (Stadler et al. 2011) (Table 7):

Table 7. Mass comparison of antigen between genomic DNA and the DNA fragment controls for ELISA

Modification	R2L 100 ng	WT ES gDNA 100 ng	Fold difference gDNA vs R2L	Dot blot ratio gDNA vs R2L	ELISA ratio
C	25.9 ng	20.88 ng	-	-	-
mC	25.9 ng	0.7308 ng	35	10 x	10-40x
hmC	25.9 ng	0.023 ng	1126	1000 x	100-1000x
fC	25.9 ng	0.0023 ng	11260 ?	n/a	10 000x
mCA	7.19 ng	≤ 0.1827 ng	≥ 35	n/a	≥ 10x
mCG	6.11 ng	≥ 0.5481 ng	≤ 12	n/a	≥ 10x

These calculations helped to optimise and standardise the amount of genomic DNA and antibody dilution for each assay. For simplicity, and to allow usage of lowest possible amounts of gDNA, all antibodies in further chapters were used at a fixed 1:2000 dilution, while the gDNA amounts varied. I did not obtain any gDNA signal for 5fC (not shown), most likely due to the low representation of the antigen in gDNA and insufficient sensitivity of the antibody. Additional optimisation of the sensitivity of the assay might need to be sought for low concentration signals like mCA and 5fC modifications, but the current conditions are sufficient for 5mC and 5hmC (the Active Motif 5hmC antibody is highly sensitive and useful in this assay despite the lower representation of the 5hmC modification in the genomes).

### 4.3.4.3 Quantitative capacity of the DNA ELISA assay

I next tested the quantitative aspect of the assay. I used genomic DNA samples with known absolute amounts of 5mC and 5hmC, measured by mass spectrometry (kindly provided by Gabriella Ficz (Ficz et al. 2013)) – I chose four different samples with different amounts of 5mC and 5hmC. The relative amounts of 5mC and 5hmC were correctly measured with the ELISA assay for both modifications (Figure 37A-B). When the ELISA optical density (OD) signal was plotted against the molar 5mC or 5hmC percentage measured by mass spectrometer (LC-MS), the correlation coefficient was excellent in both cases ( $R=0.9888$  and  $R=0.999$ ) (Figure 37C). It is therefore possible to compare ELISA 5mC and 5hmC measurements with published LC-MS or HPLC absolute values by including a mass spec-measured standard in the assay. Thus, the query samples can be assigned absolute 5mC and 5hmC molar percentages and can be compared to any samples measured by HPLC/LC-MS.

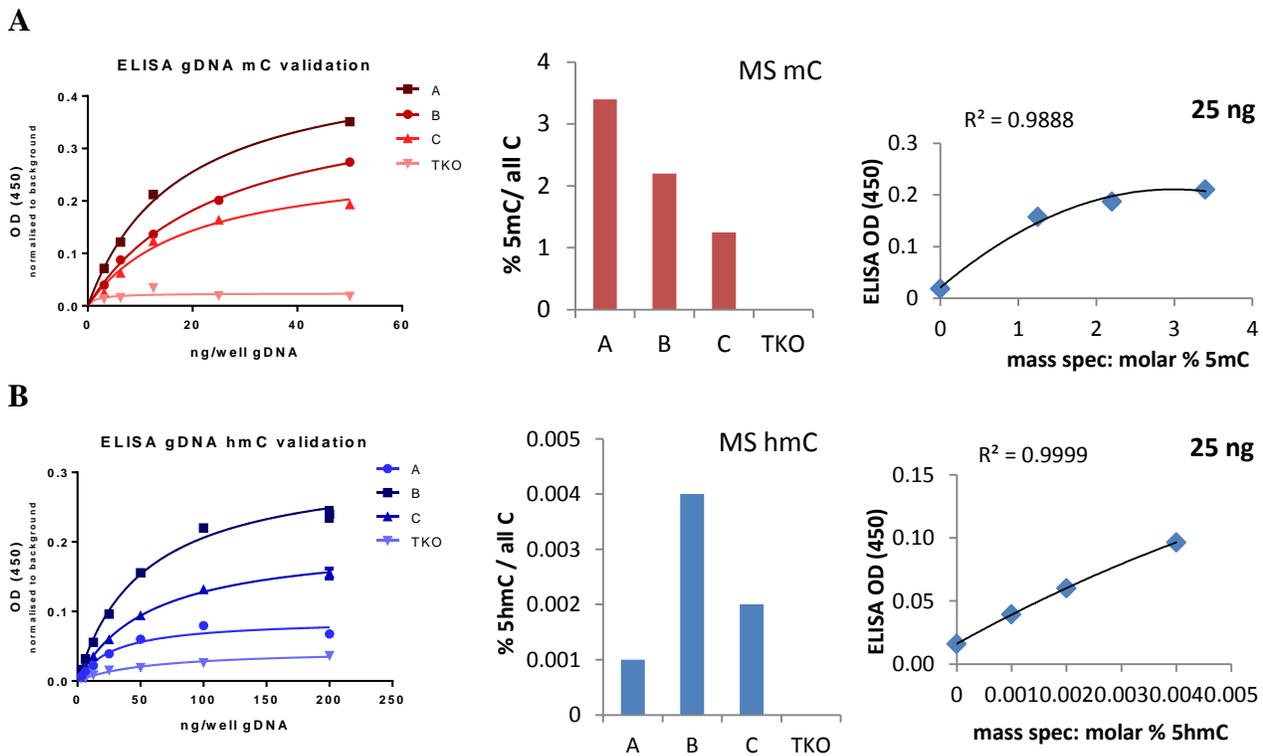


Figure 37. Mass spec validation of the quantitative potential of direct DNA binding ELISA assay. Four gDNA samples were measured by ELISA and compared to their LC-MS measurements for: **A.** 5mC; **B.** 5hmC; **C.** A direct correlation between ELISA OD (450) and mass spectrometry molar signal, demonstrating the quantitative properties of the ELISA assay. The curve fit is non-linear saturation curve.

As seen in Figure 37C, four data points can be enough for 5hmC but more points would be better to use for 5mC. This clearly shows that the quantitative power of each ELISA assay depends on the quality of the antibody – the higher the avidity and sensitivity of the antibody, the better correlation with LC-MS the assay will have.

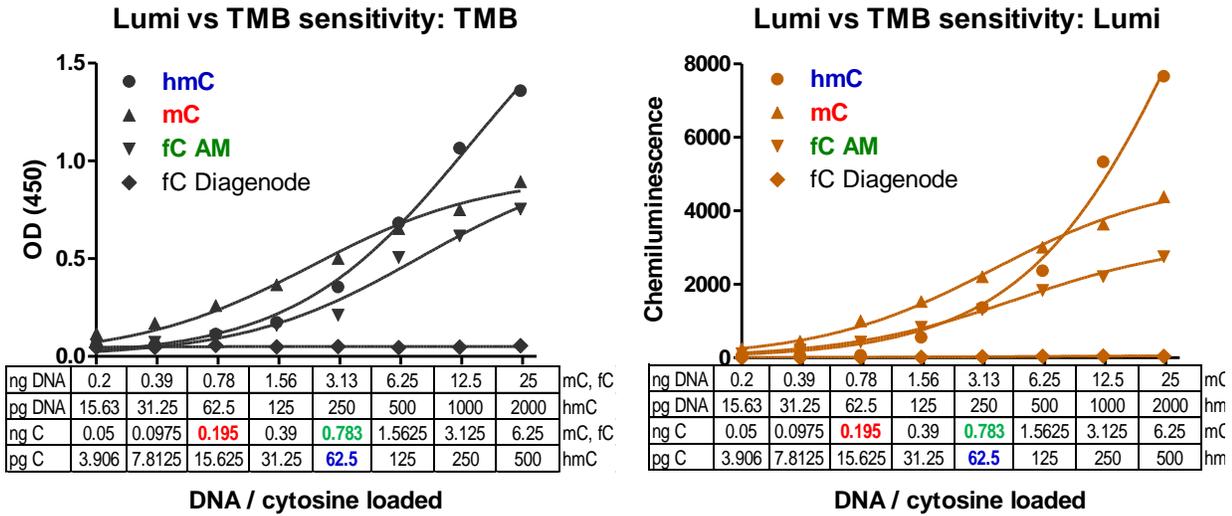
#### 4.3.4.4 Enhancing sensitivity and dynamic range of the mCA ELISA assay

From those first few ELISA tests, it became evident that unless the samples have very similar levels of DNA modifications, it was difficult to capture highly variable concentrations with one assay, without risking to overexpose or underexpose a sample and thus skew the result. It would also be valuable to look into enhancing the sensitivity for mCA, which would be quite low in many of the investigated samples in this project, while quite high in the WT ES cell controls. One possibility was to use more sensitive HRP substrates, that have a higher dynamic range and higher sensitivity than TMB, such as the fluorescent or chemiluminescent substrates. After a brief market research, the Thermo Scientific chemiluminescent SuperSignal ELISA Femto Maximum Sensitivity Substrate was brought to my attention as the most sensitive, with an estimated 10-fold sensitivity increase above TMB's capacity.

To compare both substrates I performed the same assay with M13 PCR fragments containing 5mC, 5hmC and 5fC at serial dilutions (Figure 38A). Two antibodies for 5fC were tested, one of them with weaker sensitivity than the antibody previously tested, in order to challenge the method's sensitivity limits. The overall values for 5hmC, 5mC and the more sensitive 5fC antibody were very similar with both substrates, although the  $R^2$  values of non-linear correlation were slightly more consistent (around 0.99) for the luminescent substrate (referred to as "Lumi", Figure 38B). The lower sensitivity 5fC antibody however failed to produce any signal with the TMB substrate even at 25 ng DNA, equalling background noise, while the luminescent substrate showed good titration curve with an  $R^2$  fit of 0.95 (Figure 38C) with a clear signal down to 0.8 ng DNA, accounting for more than 30-fold sensitivity. A remarkable feature of this substrate was the ability to expose correctly very low and very high signal at the same time, without reaching saturation.

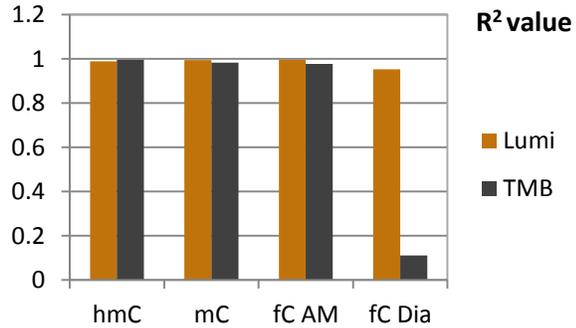
A similar test was done for mCA with PCR M13 DNA and genomic DNA, where TMB failed to register signal at 30 ng DNA while the other substrate was titrating down to 3 ng DNA, showing 10-fold higher sensitivity (Figure 38D).

A

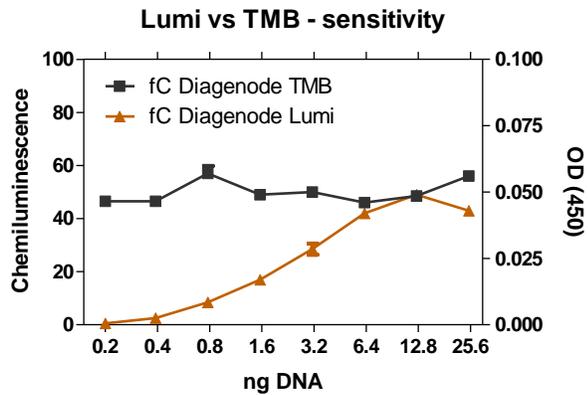


B

R <sup>2</sup>	hmC	mC	fC AM	fC Dia
Lumi	0.9890	0.9949	0.9964	0.9534
TMB	0.9957	0.9829	0.9772	0.1117



C



D

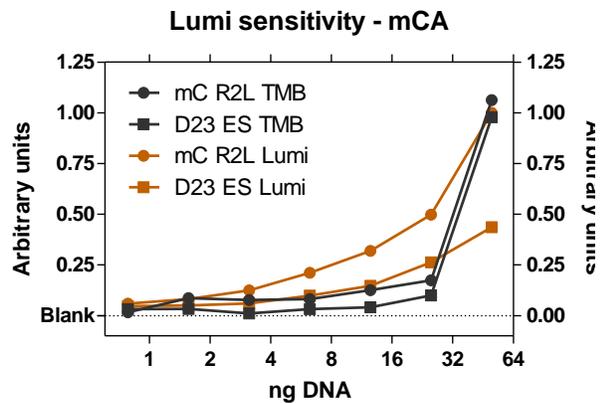


Figure 38. Enhancing the sensitivity of DNA ELISA. **A.** A comparison between three different antigens and four different antibodies performed with both substrates: TMB and Lumi. Detection sensitivities shown in corresponding colour for each modification. On both graphs the signal from the Diagenode fC antibody looks zero at this scale; **B.** R square values compared between the two substrates; **C.** The fC Diagenode antibody result scaled down - TMB against the Lumi substrate showing 50-fold increase in sensitivity; **D.** mCA genomic and synthetic DNA fragments showing 10-fold higher sensitivity with the Lumi substrate.

In conclusion, the chemiluminescent substrate provided more than 10-fold higher sensitivity and a much wider dynamic range for the simultaneous and accurate measurement of highly variable samples. This would be very useful for samples with lower abundance of target antigen like the mCA, and will enable the simultaneous analysis of samples of variable amounts of modification.

#### **4.4 Discussion**

Various possibilities for developing new tools and techniques for the analysis of non-CG methylation were addressed. No equivalent of MspI and HpaII enzyme pair was identified for non-CG context, although a number of enzymes could be used for selected techniques.

The MspJI methyl-RRBS technique can deliver very accurate results on CG or non-CG context methylation. However, it suffers from low mapping efficiency, which unfortunately decreases its applicability. There is a possibility to use the restriction enzyme enrichment technique without applying bisulphite conversion, as originally published, although this method is not accurate, and would find a better use to validate results obtained from another technique as in Shirane et al., 2013.

Methylation context-specific antibodies were successfully developed, despite the initial risks associated with this undertaking. They were shown to have both high specificity and strong affinity, which makes them useful for a number of techniques.

Importantly, a protocol for DNA ELISA has been developed to utilise the quantitative capacity of these antibodies for a quick and accurate measurement of CG and CA methylation.

The molecular tools and methods developed here will be useful to the broad scientific community and will have a beneficial impact on the study of non-CG context methylation.



## **5 Genomic distribution of non-CG methylation: results from novel techniques**

### **5.1 Introduction**

In the previous chapter, I outlined the challenges and current limitations that precluded a comprehensive genome-wide analysis of non-CG methylation in the reprogramming cycles during mammalian development, according to my general Aims outlined in 1.7. Having developed novel tools to allow for this analysis, I proceeded to investigate the genomic distribution of non-CG methylation in ES cells, a model of the ICM component of the preimplantation stage mammalian blastocyst (Nichols et al. 2009).

It has been suggested that mCH in mammals is characteristic particularly for ES cells (Lister et al. 2009; Ramsahoye et al. 2000; Ziller et al. 2011). However, the multiple reports of such methylation in other tissues, as discussed in the Introduction, leave open the question whether or not non-CG methylation is present in other tissues, or exclusively in mammalian ES cells, and whether ES cells indeed contain the highest genomic content of non-CG methylation. It would also be valuable to know more precisely what the dynamics of mCH is throughout early development, and whether it is subject to erasure and reprogramming, as its counterpart in CG. It is important to find out if it always follows CG dynamics, and is therefore a possible ‘bystander’ of the canonical CG methylation activity, or, more importantly, has its own regulated dynamics.

Establishing quantitative and potentially qualitative values for non-CG methylation would lend support to the idea that this genomic signature has the potential for biological significance. As the dominant type of CH methylation is actually in CA context (Lister et al. 2009; Ziller et al. 2011; Chen et al. 2011), in some of the experiments I have investigated CA methylation in particular using our newly derived anti mCA antibody, instead of total mCH levels, as the two should be tightly linked.

### **5.2 Aims**

1. To directly compare global CG and CH methylation levels between ES and differentiated cells from mouse and human
2. To address the distribution of CH methylation throughout genomic features and compartments in those cells

3. To follow the dynamics of CH methylation throughout the ES cell cycle and its relation to mCG dynamics
4. To investigate the levels of non-CG methylation in mouse tissues and map its dynamics during early development

## 5.3 Results

### 5.3.1 CH methylation levels in ES and differentiated cells

To assess the global levels of non-CG methylation I used my context specific antibodies and DNA ELISA protocol. To address mouse strain variability I have evaluated both 129-derived commercial mES cell lines (J1 and E14), and also B6 mES cells with B6 pMEFs derived in our institute. Human cells lines evaluated here have already been subject to genome-wide analysis and the results published (Lister et al. 2009). In this assessment, I compared the absolute levels of mCG and mCA between the cell lines, but not their proportional levels within each cell line. My DNA ELISA-based results showed in all cases that the mCA levels were lower in the differentiated cell lines in comparison to ES cells (Figure 39A), ranging between 25 % and 50 % of the ES amount for mouse, and more than 60 % for human. It is possible that the difference between human and mouse resides with the IMR90 line, which is not equivalent to mouse pMEFs (derived from primary tissue), but is a long-standing commercial cell line. Significant variation of absolute mCA levels has been observed between the mouse strains, with B6-derived mES cells having around 50 % higher mCA levels than the 129-derived lines (Figure 39B). We did not have 129-derived pMEFs, but I used mixed background pMEFs instead (outlined in Materials and methods). However, CG methylation did not vary between mES and pMEF lines (Figure 39C), although it was globally lower in mES cells from the B6 background (Figure 39D). mCG difference was observed in the human lines, the IMR90 having lower mCG (around 70 % of hES value) (Figure 39C). Human DNA is slightly less methylated globally, and my ELISA results confirm this observation (also shown by mass spec data from Tamir Chandra, Reik lab). It is interesting to note, however, that while human mES cells have slightly above half the CG methylation of mES cells, they possess only a third of the mCA methylation of mES cells (Figure 39B and D). This suggests that the ratio of mCG-mCA is different for mouse and human, the mouse ES cells having higher proportion of mCA methylation than the human ESCs, and the

difference between context methylation levels in pluripotent cells and foetal fibroblasts is less pronounced for the human.

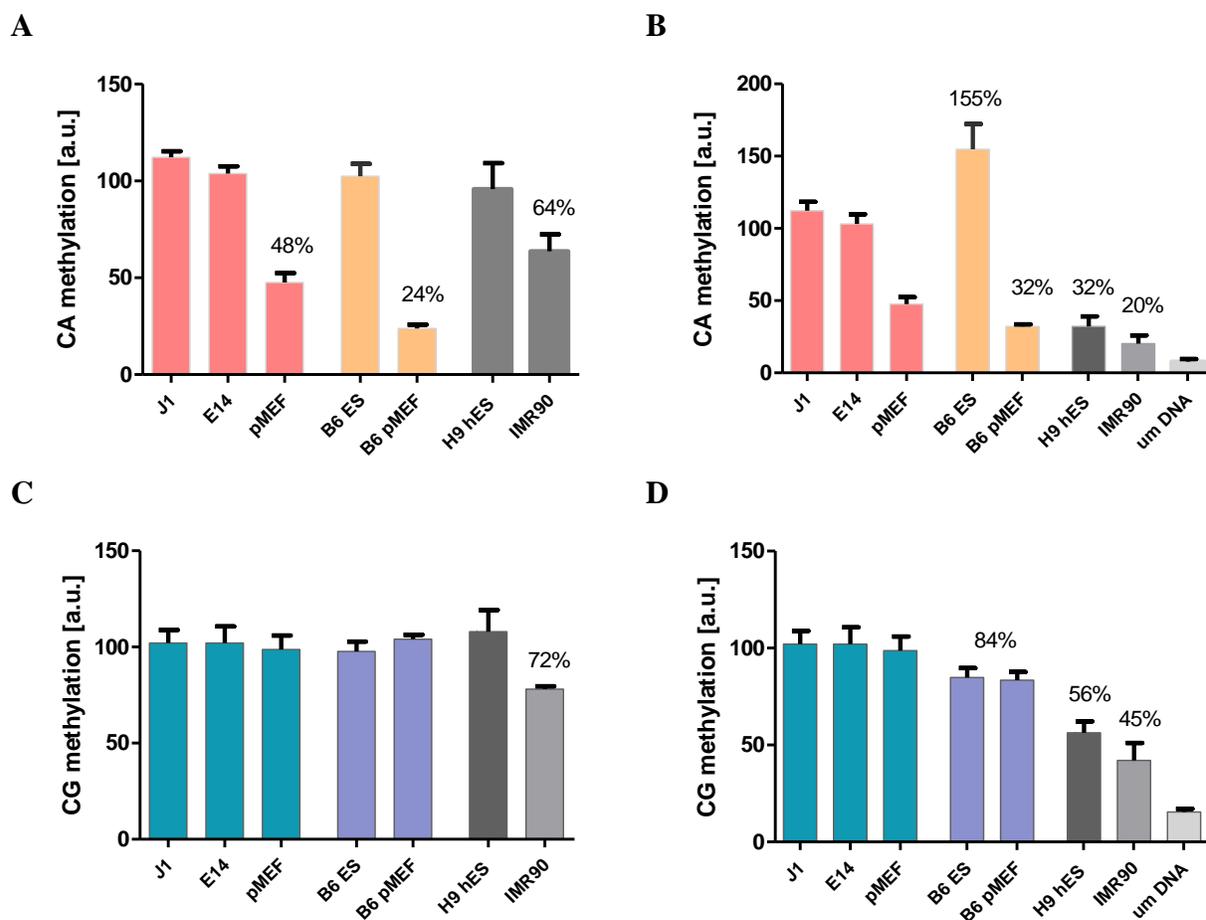


Figure 39. DNA ELISA for global mCA and mCG levels in pluripotent and differentiating mammalian cells. **A** and **C**. A comparison between ES cells and differentiated cell (embryonic fibroblasts), normalised to the ES cells within each sample group for better illustration of the percentage difference. **B** and **D**. Same samples but normalized to the mouse E14 line, to illustrate the differences in global amounts between the cell lines.

### 5.3.2 Genomic distribution of non-CG methylation

I next asked what the distribution of non-CG methylation was across the genomic features and compartments. I mapped a panel of features against my MspJI mERRBS positive calls, as I had done previously for the low coverage BS-seq in Chapter 3. Interestingly, the pMEFs showed slightly higher CG methylation in all features, and markedly lower non-CG methylation (Figure 40). The CGH methylation clearly followed the pattern of CG methylation across the different features, while CHH was more variable, although broadly very similar. The relative enrichment of

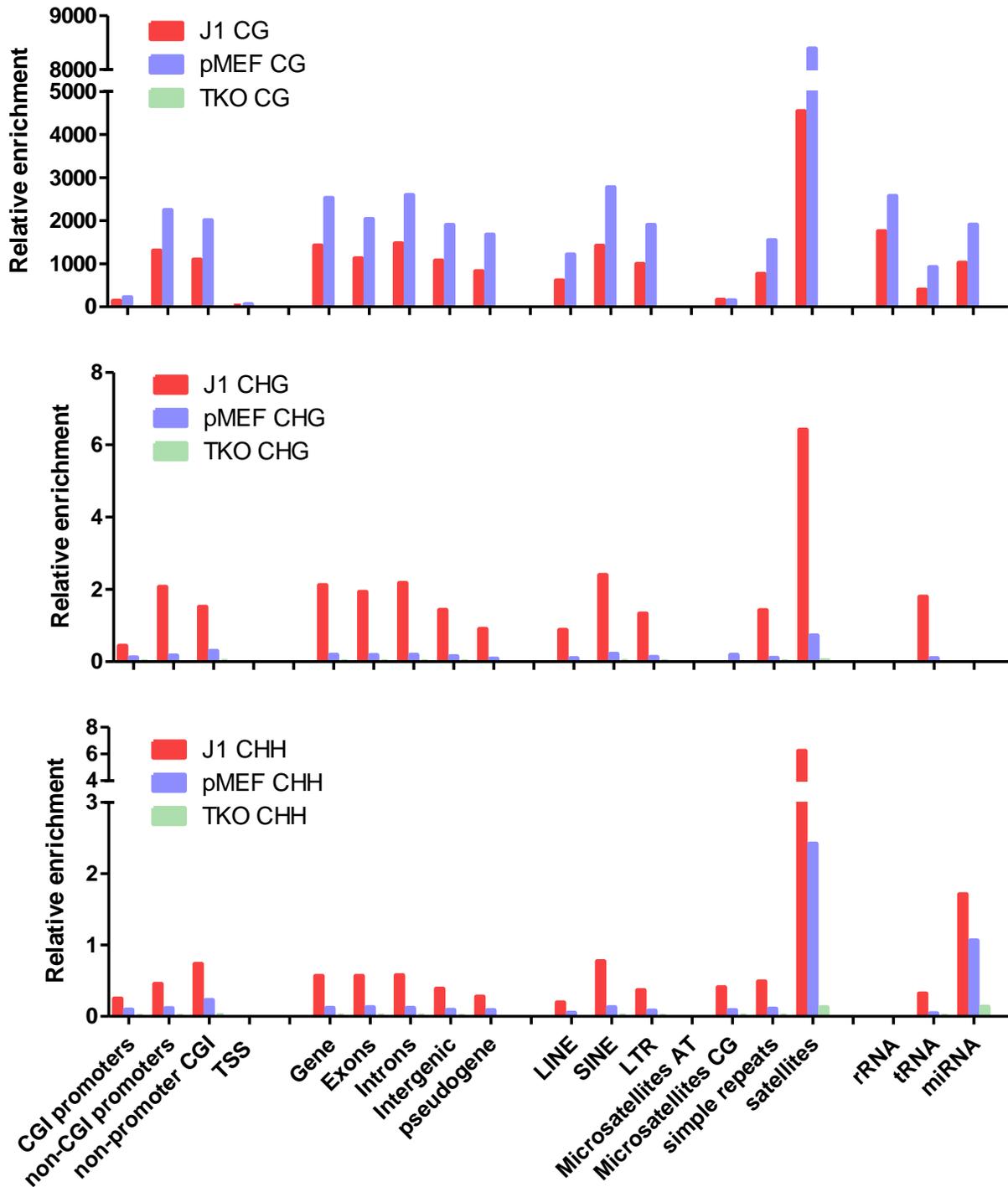
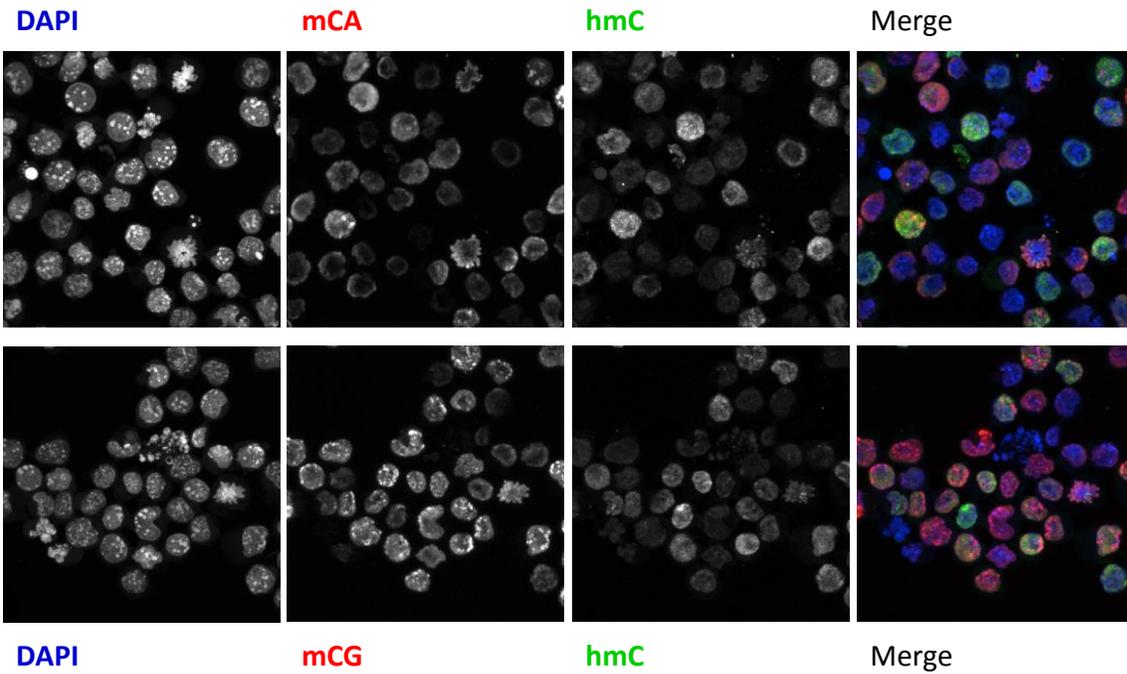


Figure 40. Feature enrichment of CG (upper panel), CHG (middle panel) and CHH (lower panel) methylation in mES and pMEFs; unmethylated TKO line is used as a background control. The major satellite sequences were not filtered out as in previous BS-seq analysis, and the high value reflects the high number of calls from this repeat mapping to individual sites in the genome, thus overestimating the enrichment.

CHG methylation was also higher than CHH in ES cells, which is in accord with the published data for mammalian ES cells measured by high depth RRBS or BS-seq (Lister et al. 2009; Ziller et al. 2011; Chen et al. 2011). CH methylation shows high enrichment in gene bodies (both introns and exons) and CGIs outside of promoter regions. Interestingly, tRNA and miRNA sequences also showed high values, but not rRNA loci which show higher CG methylation instead (Figure 40). From chapter 3 we already knew that mCH is enriched in some repeats and here it seemed particularly high in SINE elements. The major satellite sequence showed the highest enrichment, as in Chapter 3, although in this case the high value reflects its high repeat numbers in the genome as a whole (there was no filtering of the satellite repeats in this analysis), meaning it should not actually be so high per monomer, given the large fraction of the genome it occupies. As a matter of fact, to validate the amounts of satellite non-CG methylation as shown in Chapter 3 from the low BS-seq data, this value should have been much higher, provided it occupies 5-8 % of the mouse genome (Waterston et al. 2002; Abdurashitov et al. 2009). Finally, the background from the unmethylated TKO line was virtually zero, which confirmed the very high accuracy of the positive call validation strategy, developed in Chapter 4, without any signal contamination or BS-conversion errors.

The low percentage or low enrichment of non-CG methylation in genome-wide studies of pooled DNA would mean that either each cell in the population has very low levels of this methylation, homogeneously distributed, or it is variable among the cells within the population. To assess the variability within the cellular populations I next made use of the new antibodies in immunofluorescence (IF). This technique would also help to visually validate the observed feature enrichment patterns and similarities between mCG and mCH for mouse and human ES cells and embryonic fibroblasts (pMEF or IMR90). Interestingly, mouse ES cells exhibited greater variability in their mCA staining in comparison to their CG methylation (Figure 41, upper panel). Few cells showed complete nuclear mCA staining, while in most cells the nuclei were fully stained for mCG. Few cells did not have any staining, and the rest had partial nuclear staining. This effect was not due to the choice of a confocal plane, but was consistent for the whole nuclei. Neither of the signals followed the pattern of 5hmC, thus confirming that there is no cross-reactivity of either antibody for 5hmC. In pMEFs the mCA signal was visibly weaker than the signal for mCG (Figure 41, lower panel). A small group of ES cells was included in the image to visualise that the microscope settings were the right intensity. The pMEFs had no 5hmC signal as expected, and their CG methylation pattern followed the DAPI staining, as for mES cells.

J1 mES cells



pMEF cells

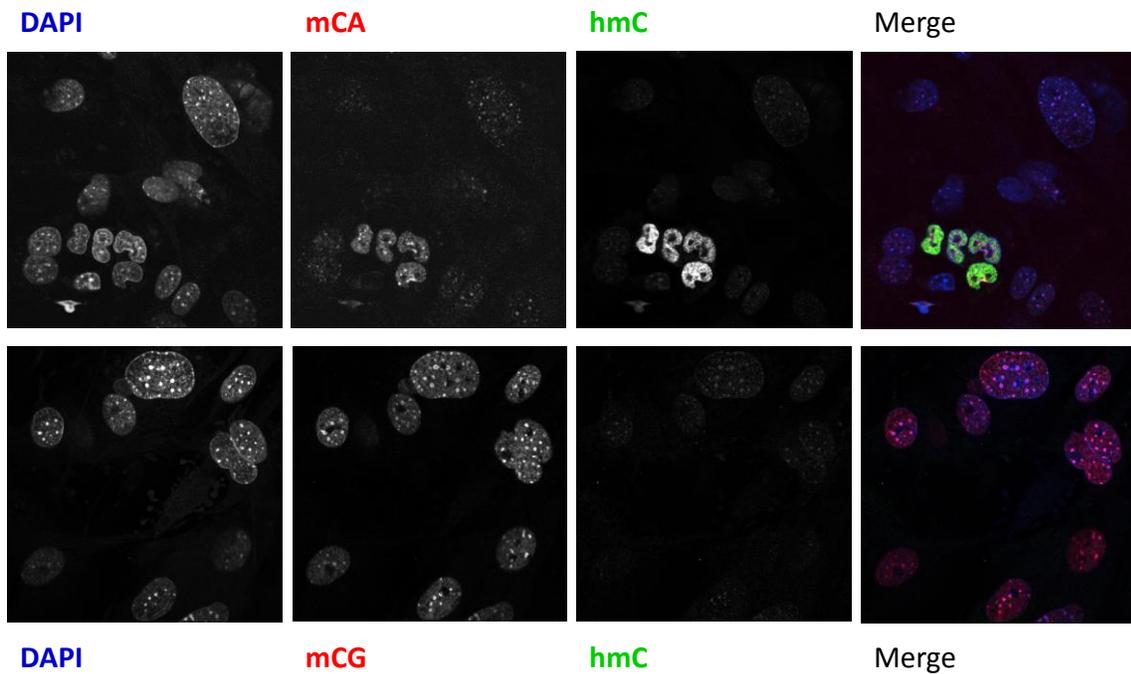


Figure 41. Immunofluorescence imaging of mouse ES cells (upper panel) and embryonic fibroblasts (lower panel). The colours in the merged image are indicated above and below each greyscale panel. Images taken by Fátima Santos.

Interestingly, the only nuclear compartment in pMEFs staining for mCA signal were the chromocentres. The chromocentres are characteristic of mouse nuclear organisation, and represent the dense heterochromatic foci of the chromosomes (Probst & Almouzni 2011). This result correlates with our earlier BS-seq observation that non-CG methylation was preserved in pMEF major satellites, although in lower levels (3.3.2). In the mESCs, however, very few of the cells actually had signal in their chromocentres, a surprising observation after our findings about major satellite in Chapter 3. Quantitation of around 800 cells revealed that between 15 and 25 % of mESCs showed chromocentre staining for mCA, while between 70-95 % showed chromocentre staining for mCG. This result answers an important point about the mCH distribution in the satellite, raised in Chapter 3, which could not be answered by BS-seq: the satellite mCH methylation is not homogeneously low in each cell, it is very heterogeneous and not present in every ES cell, not even in every chromocentre in an ES cell. Why the satellites are homogeneously methylated in CA in pMEFs, however, is not clear, although it seems possible that their CH “methylation” might be merely a background signal of both the BS-seq and IF techniques.

Human ES cells do not have chromocentric foci, and it is therefore more difficult to judge the genomic localisation of CA methylation. It was striking that these cells seemed much more homogenous for both mCA and 5hmC, unlike the mouse ES cells (Figure 42, upper panel). There was no obvious difference in intensity of mCA and mCG signal in the hES cells, similarly to the mES cells, and due to the lack of strong cell-to-cell variability, the mCA and mCG signal looked overall very similar. A higher magnification view revealed that the mCG signal was denser at the DAPI dense regions, while the mCA seemed more homogenous, showing a similar tendency to the mouse mESCs. In the IMR90s however, there was clearly lower signal for mCA, and also 5hmC, as observed with the pMEFs (Figure 42, lower panel). In all cells few dense ‘spots’ were staining for mCA, mCG or 5mC (not shown), and provided those resemble heterochromatic regions similar to the mouse chromocentres, this would make their pattern very similar to the pMEFs’.

The whole cell approach for antibody labelling of DNA cytosine modifications relies on DNA depurination with HCl, and therefore the adenines’ and guanines’ integrity is globally affected. The use of this established protocol with our context-specific antibodies, which are selected to recognise intact adenines and guanines, unlike the general 5mC antibody, thus raises concerns. To confirm our observation, specifically about the major satellite, and thus validate the HCl approach for the context-specific antibodies, I prepared metaphase chromosome spreads and

subjected those to IF with our antibodies (Figure 43).

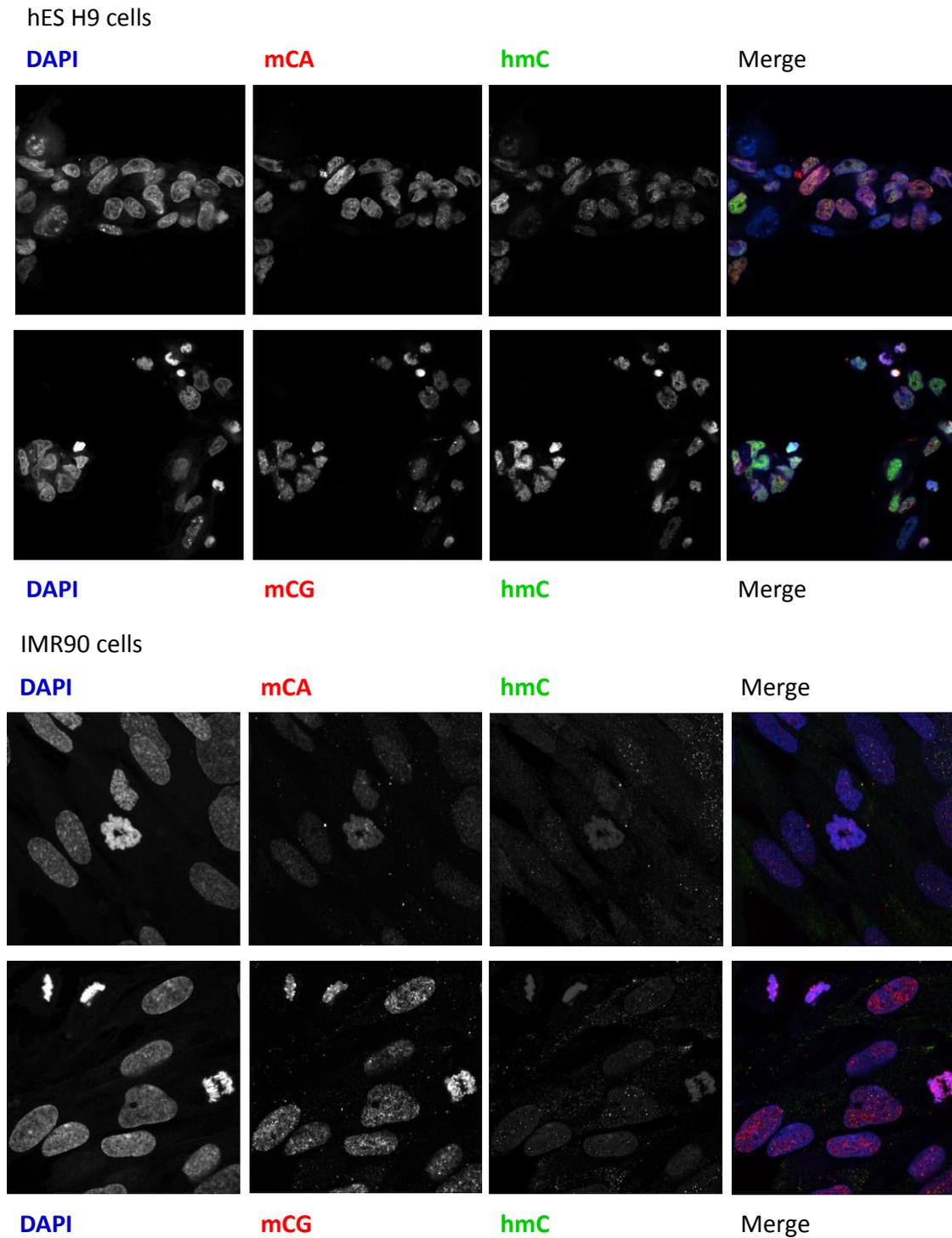


Figure 42. Immunofluorescence imaging of human ES cells (upper panel) and embryonic fibroblasts (lower panel). The colours in the merged image are indicated above and below each greyscale panel. Images taken by Fátima Santos.

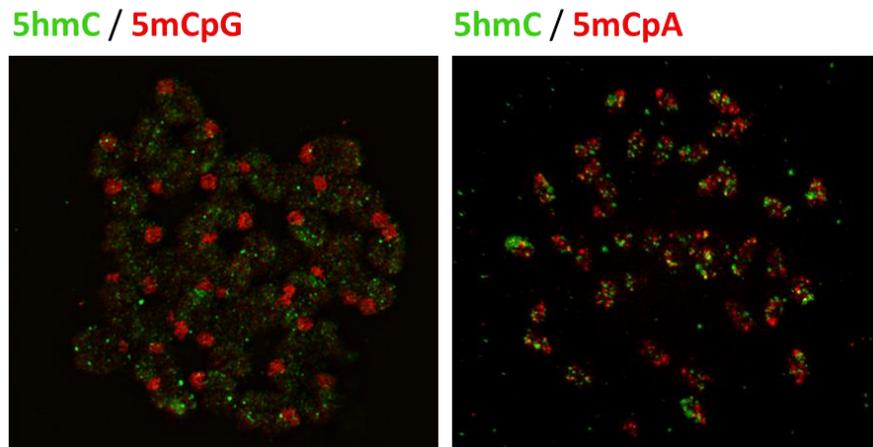


Figure 43. Immunofluorescence imaging of mouse ES metaphase spreads. The satellites are the condensed tips of the chromosomes, staining in red with mCG. The rest are the arms of the chromosomes, showing staining for both mCG or mCA, and 5hmC. Images taken by Simon Walker, BI Imaging Facility.

The metaphase spreads protocol does not depurinate DNA but denatures it, which ensures access to the methylation sites. This result confirmed the strong staining of satellite in CG, and the lack of staining for mCA – in this case none of the satellites stained for mCA, which also was observed in metaphase chromosomes with the whole cell approach. This observation validated our results obtained with the depurination approach, meaning that the overall damage to purines is not very high with the milder (2N vs 4N) HCl protocol which has been used.

### 5.3.3 Dynamics of non-CG methylation throughout the stages of cell cycle

In order to investigate at a higher level of resolution the connection of non-CG methylation to the overall cellular methylation patterns associated with CG methylation, ES cells were sorted according to their phase of the cell cycle by fluorescent activated cell sorting (FACS) (Figure 44A). To better control for S phase cells, I briefly pulse labelled the culture before harvesting with the nucleotide analogue EdU (5-ethynyl-2-deoxyuridine), which incorporates in replicating DNA and can be visualised by fluorescence. Due to the high rate of ES cell division the FACS success was partial, and both G1 and G2 phases contained S-phase cells, although the EdU  $+/+$  cells in G2/M could have been cells exiting S within the duration of the EdU pulse (Figure 44B). The DNA methylation was then measured in a few different ways for ultimate accuracy and validation – dot blot (Figure 44C) and LC-MS (Figure 44D) for total cytosine methylation, and ELISA for mCA and mCG methylation (Figure 44E). All results showed consistently a trend for slightly

higher methylation in G1 and G2, including both mCA and mCG. Because the differences were small overall, due to measurement variability there was no clear-cut statistical significance for the individual measurements, but the fact that all methods show the same trend confirms the validity of this dynamics. In this regards the mCA seemed to follow closely the mCG accumulation dynamics. This result is in agreement with an earlier study in HeLa cells, where the amounts of accumulated 5mC were measured during each cell cycle, and was observed that non-CG methylation follows the accumulation dynamics of mCG (Volpe 2005).

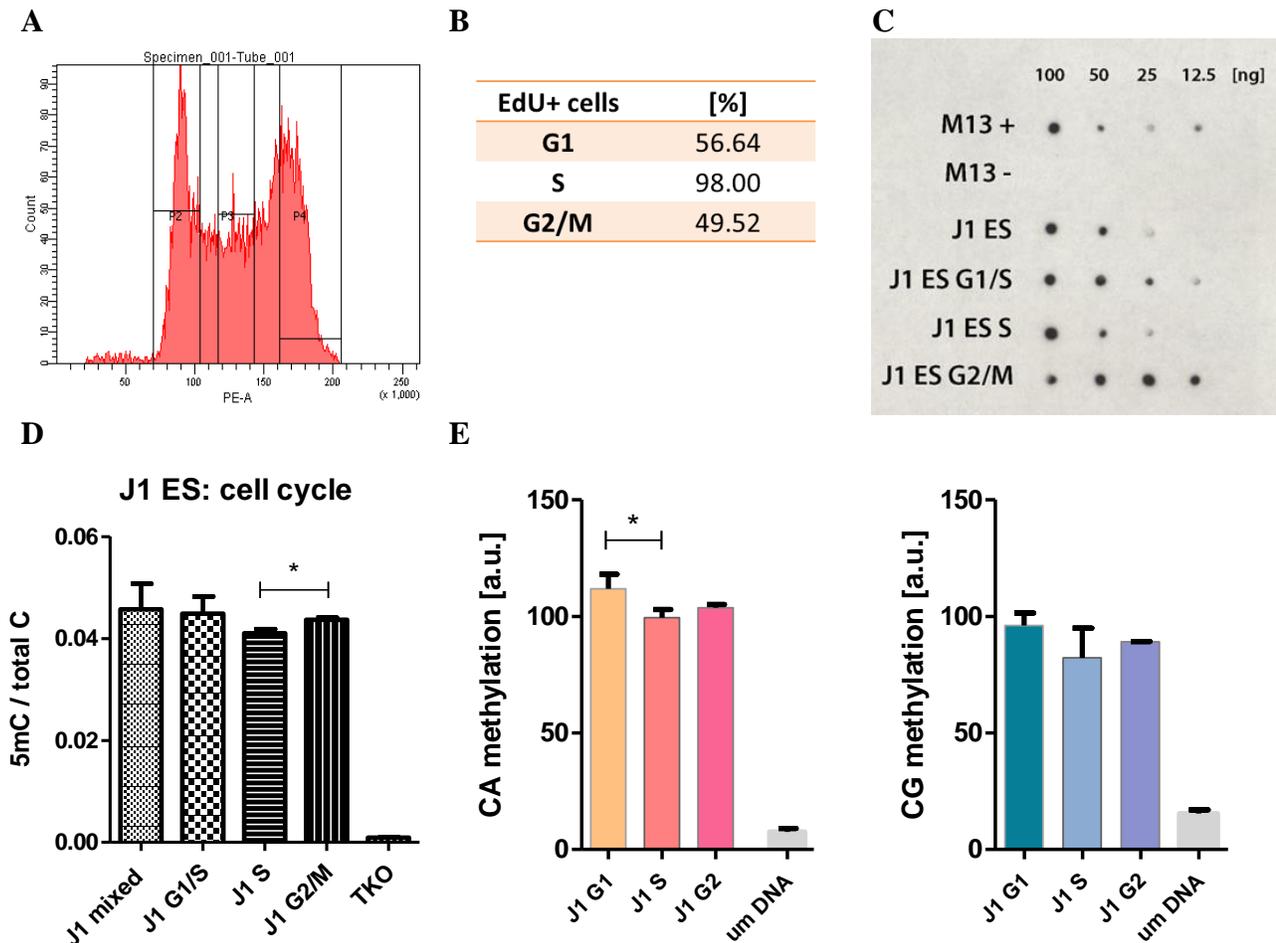


Figure 44. Methylation dynamics throughout the mES cell cycle. **A**. FACS sorting of mES cells showing the gating over the three cell cycle phases and the extent of phase separation. **B**. S-phase cells were pulse-labeled with EdU for 10 min before fixation and subsequently counted in each fraction as a control over the FACS-sort. **C**. Dot blot for 5mC, including M13-derived positive and negative control, and an unsorted J1 population. **D**. Global methylation levels as measured by LC-MS. **E**. ELISA measurement of relative methylation in CA (right) and CG (left) context. Two biological replicates were used, separated in two independent sorts.

### 5.3.4 Dynamics of non-CG methylation throughout mammalian development

The next step was to investigate the dynamic changes of non-CG versus CG methylation within the wider framework of mouse development. For this purpose, I first measured mCG and mCA methylation in a panel of mouse tissues. As expected, most mouse tissues showed low levels of CA methylation, and the brain samples, in line with recent reports (Lister et al. 2013; Guo et al. 2014), were the only ones which showed high levels of mCA, comparable with the levels in B6 mES cells (Figure 45 upper panel). The CG methylation was much less variable, with tissues like placenta and testis having the lowest levels, again in line with published reports (Oda et al. 2013; Senner et al. 2012; Oakes et al. 2007) (Figure 45, middle panel). Interestingly, the global 5mC methylation (Figure 45 lower panel) showed an averaged value between mCA and mCG for all samples, with a couple of exceptions, which might be due to technical errors. This result confirms that non-CG context methylation is generally lower in foetal or adult tissues, and drops with differentiation.

My next step was to assess the methylation dynamics before the pluripotency stage and during the phases of global erasure and *de novo* methylation establishment after fertilisation. Since IF is not fully reliable for quantification, I used published WGBS datasets instead, a number of which had already become available at this stage of my project. To avoid the problem of conversion errors and overestimated 5mC levels, I based my quantitation mainly on datasets prepared with the amplification-free PBAT library-preparation technique and its modifications (Miura et al. 2012; Smallwood et al. 2014). Where this wasn't possible, I used BS-seq datasets with very good conversion values, namely 0.25 for mCG and 0.45 for mCH (L. Wang et al. 2014). In order to be able to compare the mCG with mCA, which have very different levels within their own context and are hard to compare, I calculated their global levels instead, estimated as a percentage from the total cytosine content. In this way, their levels can be comparable within the same scale. This is because, despite the fact that CG context is highly methylated, it occupies only 4 % of total cytosine, and thus its global levels cannot ever be higher than 4 % (this is, if CG is 100 % methylated, which is biologically impossible). The remaining 96 % of cytosines are in non-CG context, and even if this context is only 1 % methylated, this already makes a quarter of the 4 % of global mCG. In reality, mCG is never 100 %, and mCH is often 1-2 % in cells, therefore, if presented as global levels, both types of methylation are in essence comparable levels, and can be plotted onto the same scale.

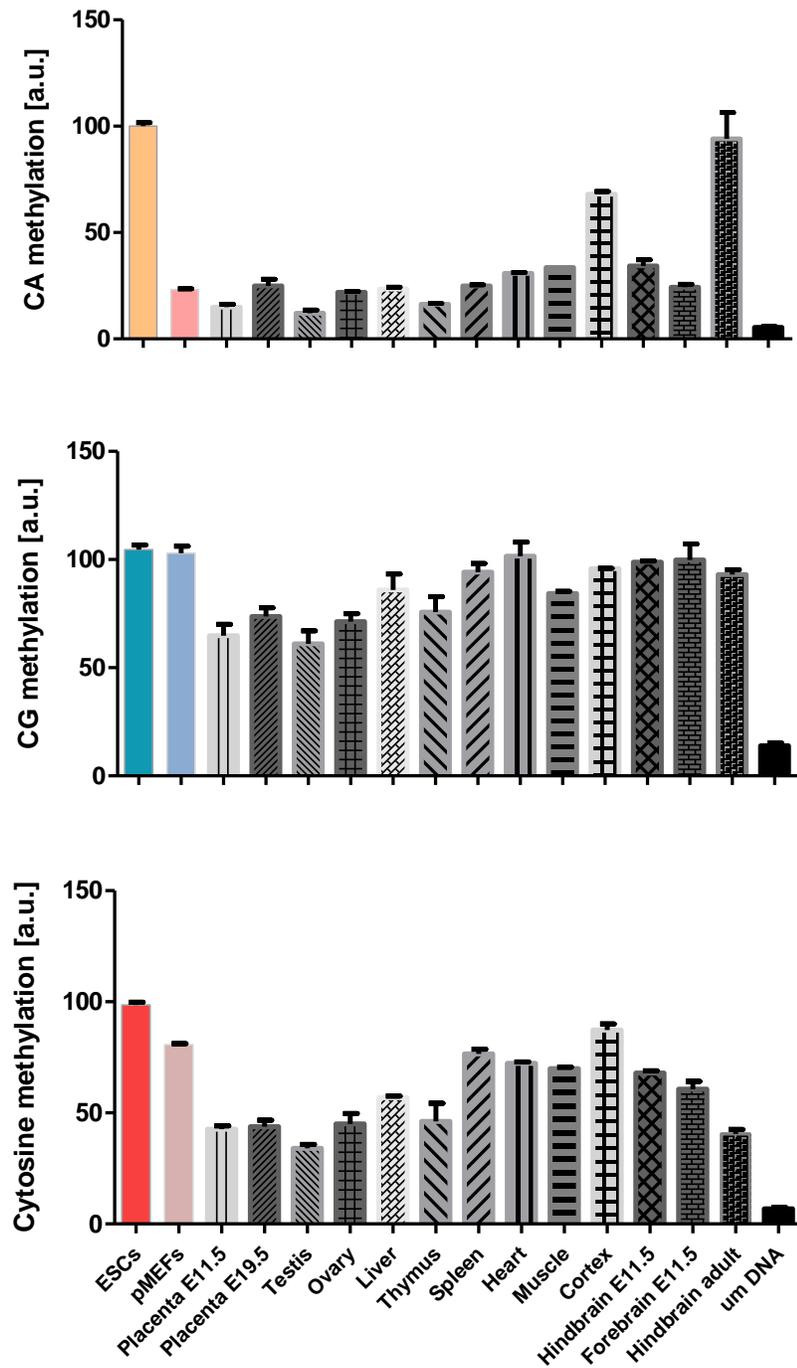


Figure 45. Global methylation in a panel of B6 mouse adult tissues (unless specifically marked otherwise): in CA context (upper panel), CG context (middle panel) and global cytosine without context (lower panel). All values were normalised to B6 ES cells as a reference.

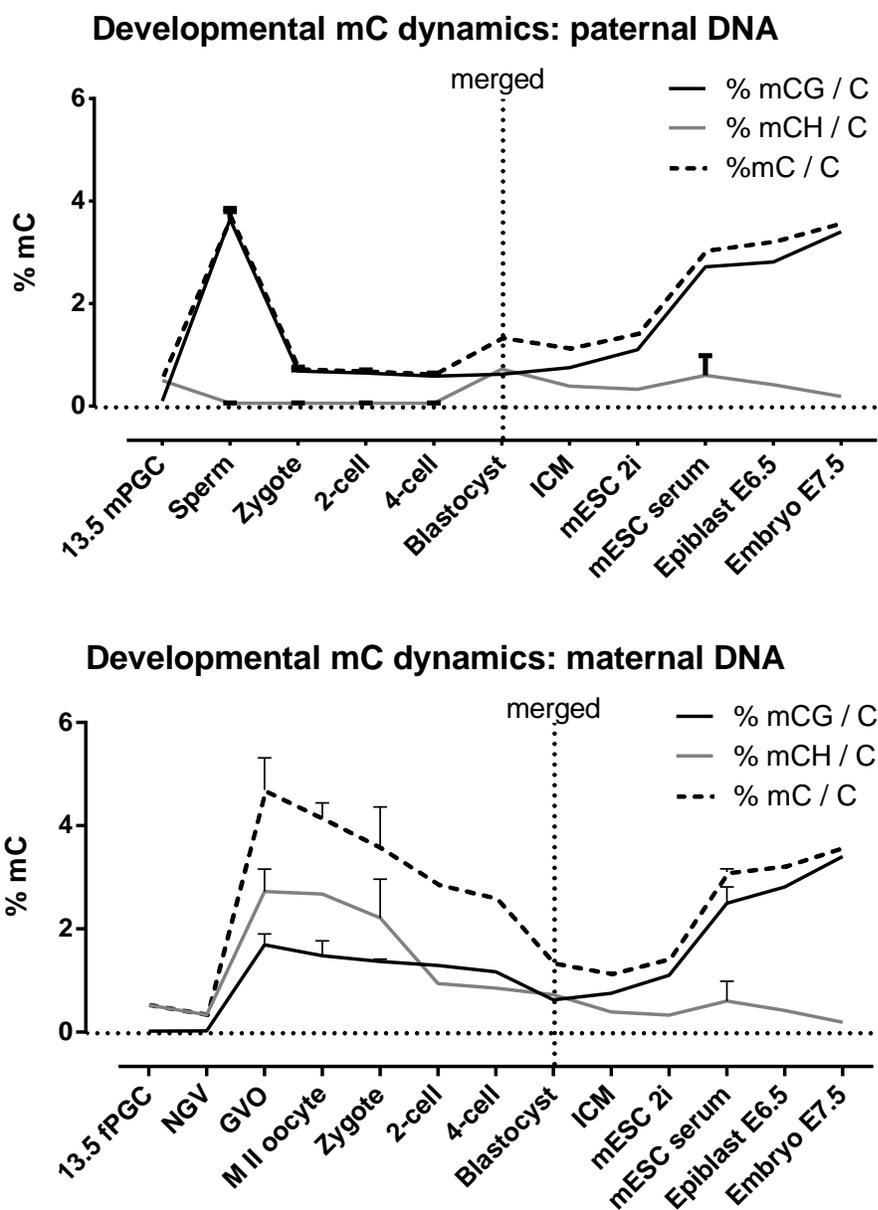


Figure 46. Global levels of methylation during mouse early development for paternal (upper) and maternal (lower) DNA. From the stage of blastocyst into embryonic development, the plot is the same and does not distinguish maternal from paternal DNA changes. The datasets used for this plot are listed in Appendix Table 21. For the zygote, I have estimated the paternal and maternal levels of 5mC based on the ratios given in Santos et al. 2013.

In this analysis, from the lowest methylation point in PGCs (E13.5) up to 4-cell stage, I have presented separately the levels of modification for paternal and maternal DNA (Figure 46). From zygote to 4-cell stage this separation was not directly obtained from the WGBS-data, from which only the global mCG and mCH levels were taken. There is therefore also a certain level of

assumption, based on other published data (Santos et al. 2013). Since BS-Seq cannot distinguish between a methylated and a hydroxymethylated cytosine as both are resistant to the bisulfite treatment (Huang et al. 2010), these values present the levels of both modifications. The levels of hmC, however, are not very high outside the zygote, and hence may be overlooked for the purpose of this analysis.

The results for the paternal pronucleus are taken as identical to that for sperm – very high in CG methylation, and rapidly demethylated after fertilisation (Figure 46, upper panel). The results for the maternal DNA however, were striking for the level of global mCH methylation in the late oocyte and zygote, not for its presence in itself, which has been reported, but for the levels, which exceed significantly the levels in ES cells (Figure 46, lower panel). While the brain samples showed similar or lower levels of mCH in comparison to mES cells, the oocyte has 3-4-fold higher levels than mES cells, making it the cell type with the highest amount of CH methylation. This argues against the hypothesis that non-CG methylation is a mark of pluripotency, and is related to the state of pluripotency (Lister et al. 2009; Chen et al. 2011; Ramsahoye et al. 2000). To verify the BS-seq data with our own BS-independent techniques we performed mCA and mCG IF on growing oocytes (work by Fátima Santos) (Figure 47). The oocyte indeed showed strong staining for both mCA and mCG, but with this technique it is not possible to determine which one has higher levels, or how they compare to the levels in ES cells.

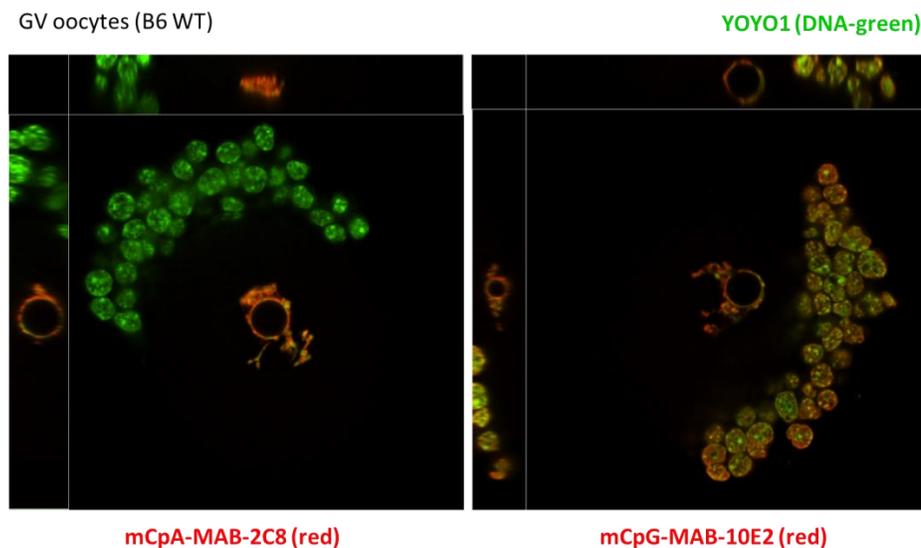


Figure 47. Immunofluorescence of growing oocyte (GVO), stained for mCA (left panel) and mCG (right panel). The cells surrounding the oocyte are cumulus cells, and they stain only for mCG like other tissues. IF by Fátima Santos

### 5.3.5 Discussion

After experiencing serious technical difficulties at the early stage of this project, the subsequently developed tools and techniques were finally able to help elucidate the levels and distribution of non-CG context methylation in early mammalian development. While ES cells were confirmed to have high levels of CH methylation in comparison to most types of differentiated tissues, they were also shown to possess far from the highest mCH level in mammalian development. The cell possessing highest levels of CH methylation is the mature oocyte, where the global levels of CH methylation are higher than its global levels of CG methylation. Unlike the rapidly dividing ES cells, this is a mitotically arrested cell with a very different function, DNA organisation, and DNA methylation machinery. It is perhaps in a way more ‘similar’ to the neuronal cells, which, I also confirmed, have very high levels of CH methylation (Figure 45). They are post-mitotic and non-dividing long-lived cells, much like the oocyte, although with very different DNA organisation and functions. It is an open question whether mCH methylation could then be more beneficial in a non-mitotic cell, where the problem of its post-replication maintenance is not specifically relevant. Certainly, however, the ES cells remain the most accessible model to study non-CG methylation, unlike the oocyte, which is significantly more difficult to access, handle and investigate. Therefore, this result could perhaps add another perspective into this work, to include questions or analyse further results in a way that could be of relevance to the oocyte and the zygote, if not for the ES cell pluripotent system.

Regarding its cellular localisation and dynamics, non-CG methylation seems both similar and different from CG methylation. It is similar in its levels of enrichment within genomic features, and dynamics throughout the cell cycle. Nevertheless, the immunofluorescent staining highlighted differences, which have been masked by the limitations of the previous molecular approaches (Table 8). First, it revealed that mCH methylation has much more heterogeneous distribution in the ES cell population than the CG methylation, at least globally. This is a very interesting finding, as it has now been well documented that the ES cell population is very heterogeneous, at least when grown in classical serum + LIF conditions, and encompasses different groups of subpopulations. (Macfarlan et al. 2012; Hayashi et al. 2008). Some of these subpopulations are closer to the ground state of pluripotency, while others are primed for differentiation (Smallwood et al. 2014). A third group show similarities for even earlier stages,

expressing genes characteristic for the 2-cell and 4-cell stages (Macfarlan et al. 2012). It would be interesting to know where the ES cells which are high in mCH stand, and whether the mCH mark is characteristic for the ground-like (naïve) state or the 2-cell-like state, or is a mark of a primed ES cell state, already set on a path of differentiation.

Table 8. Comparison of global differences of mCA and mCG in mESCs revealed by IF.

Nuclear localisation	mCA	mCG
1	homogenous	homogenous
2	in foci (~20%)	in foci (~80%)
3	peripheral (subtelomeric)	peripheral (subtelomeric)
	<b>1&amp;3: dominant, 2: minor</b>	<b>2: dominant, 1&amp;3: minor</b>
Nuclear distribution	15-25% ESC population	70-95% ESC population

Another aspect of the differences between mCA and mCG relates to the genomic localisation (Table 8). Although whole genome feature analysis shows similar enrichment patterns, the IF staining of the two types of methylation is very different. First, a limited amount of mCA occupies the heterochromatic foci, while they are very dense in CG methylation. Second, the overall pattern in both human and mouse ES cells is very homogenous within the nucleus and resembles much more the patterns of the transcriptionally active marks like histone acetylation, rather than the repressive ones, which cluster around chromocentric foci. This shows that the whole genome approach in NGS datasets can potentially give very generalised results, and not capture essential differences, unless they are correlated to other functionally meaningful traits like histone marks or expression data. The gene body feature for example includes both actively transcribed and repressed genes, and the pure association of the methylation marks with gene bodies therefore does not give enough information about the actual functional association. While CG methylation is generally associated with repressed transcriptional states, the mCH methylation, particularly in ES cells, has been associated with actively transcribed genes (Lister et al. 2009). Further work will be necessary in order to establish whether there is a clear role of mCH in active transcription, whether it is a consequence or a cause of the actively transcribed state, or its correlation with it is a result of an unrelated parallel side process.

## 6 Establishment and maintenance of non-CG methylation

### 6.1 Introduction

As discussed in the main Introduction, non-CG context methylation has so far been attributed mainly to the *de novo* Dnmts, although a possible *de novo* role for Dnmt1 *in vivo* is still disputed. While it is clear that both Dnmt3a and 3b have a low activity towards non-CG contexts (mainly CA), both *in vitro* and *in vivo*, it is not yet clear whether this activity is specifically directed (targeted) or merely a nonspecific (sporadic) activity, which occurs concomitantly with the main dominant targeting of CG. The latter possibility is supported by the fact, that the presence of non-CG methylation during development occurs within the two phases of establishment of methylation marks – in male PCGs and growing oocytes (Shirane et al. 2013; Kobayashi et al. 2013), and in pre-implantation embryo (where the ES cells are the *ex vivo* equivalent) (Ziller et al. 2011; Lister et al. 2009). The Dnmt enzymes are ‘error prone’ and as they do not have highest fidelity for maintaining the CG methylation state (Jeltsch & Jurkowska 2014; Goyal et al. 2006), it is possible that they also exhibit nonspecific side activities, especially when expressed at their highest levels and *de novo* establishment of marks is on the way.

Additionally, Dnmt2 has remained enigmatic in regard to its role in DNA methylation, and its function in methylating tRNAs does not exclude a possibility to also work on a DNA substrate. Its preference for methylating non-CG context makes it an appealing candidate for a ‘non-CG methyltransferase’ in mammals. It is possible to speculate that its DNA activity could need activation from an external signal, and during the *de novo* methylation establishment, it could be receiving the necessary stimulus to switch its substrate towards DNA. Dnmt2 is highly expressed in ES cells and Dnmt2 knockout cells are available.

It has already been suggested that non-CG methylation can be actively removed, potentially in a faster and more regulated fashion than mCG (Fuso et al. 2010). In light of recent reports regarding the predominance of non-CG methylation in the oocyte (Shirane et al. 2013), depicted also in my calculations in Chapter 5 (Figure 46), the mechanisms for demethylation of mCH need a closer attention. In the original reports arising from IF, the maternal pronucleus has been characterised largely by its passive demethylation post fertilisation, unlike the paternal pronucleus, which is an arena of very active demethylation processes. In the absence of a proper maintenance mechanism for mCH, a passive removal mechanism seems the most likely pathway

for describing the fate of mCH marks. It has been shown that the maternal pronucleus is protected by the developmental protein Stella (also known as Dppa3) and the H3K9me2 mark (Nakamura et al. 2007; Nakamura et al. 2012). From my calculations in Chapter 5 (Figure 46), it becomes clear that the maternal pronucleus has higher global methylation levels than the paternal and indeed higher than the levels in subsequent developmental stages, including ESCs and differentiated embryonic tissues. Thus, to protect all those methylated sites would constitute a significant challenge for any molecular mechanism in place, and even more so when DNA in very close proximity, in the same cell, is not afforded this protection, but apparently requires the opposite activities leading to DNA methylation loss. The stringent restriction of the activity of any molecular mechanism to one pronucleus and not the other would need, in the first place, a very well defined ‘labelling’ of identity, which could be provided by mCH.

It therefore seems justified to investigate the activity of Tet enzymes on mCH context cytosines. Interestingly, the human Tet2 enzyme does not show catalytic activity on a mCA-containing substrate *in vitro* (Hu et al. 2013). A similar observation has been reported for a highly conserved Tet-like enzyme from the amoeboid flagellate *Naegleria gruberi*, which shares high similarity with mammals (Hashimoto et al. 2014). If this would be true also for mammalian Tet1 and Tet3, then the maternal genome could actually be intrinsically ‘resistant’ to active demethylation by hydroxylation at the majority of its methylated sites, which are in CH. Whether such a property would affect the repair mechanisms participating in active demethylation in the zygote, is not known.

## 6.2 Aims

In accord with the highlighted open questions, the tasks to be addressed in this chapter are:

1. To investigate if Dnmt2 has DNA methylating activity in mammals and whether it is the enigmatic ‘CH methyltransferase’
2. To elucidate which of the canonical Dnmts is responsible for CH methylation in early development and if its activity to non-CG context is unspecific or regulated
3. To investigate the activity of mouse Tet enzymes on mCA and a potential resistance of non-CG context methylation to the mechanisms of active demethylation

## 6.3 Results

### 6.3.1 Investigating a role for *Dnmt2*

We have measured global levels of methylation in *Dnmt2*<sup>-/-</sup> and *Dnmt2*<sup>+/-</sup> ES cells, against a WT control. We first tested this in IF, which did not show any difference in the methylation levels of those three cell lines (Figure 48A). Small differences as we would expect, however, are a challenge for this technique and its quantitative capacity, and it is possible that they cannot be captured. Therefore, we next measured this with an LC-MS. Surprisingly, our results showed around 15 % of decrease in 5mC in the two *Dnmt2*<sup>-/-</sup> replicates, in comparison to the *Dnmt2*<sup>+/-</sup> and the WT control (Figure 48B). The global 5hmC levels did not differ between the <sup>-/-</sup> and <sup>+/-</sup> clones, although they differed from the J1 WT ESC, and thus they did not follow the pattern on 5mC.

In order to confirm that this effect was due to the *Dnmt2* and not a decrease of the other *Dnmts*, I surveyed the functional *Dnmts* in mESCs by RT-qPCR – *Dnmt1*, *Dnmt3a2* and *Dnmt3b*. All three showed normal mRNA levels in the *Dnmt2*-KO and the TKO ES cell line was used as a negative control (Figure 49A). As a verification, *Dnmt2* mRNA levels, on the other hand seemed normal in a selection of WT cell lines, including MEFs and the *Dnmt2* TKO ES line, but they were reduced by 50% in the *Dnmt2*<sup>+/-</sup> line, and further reduced in the *Dnmt2*<sup>-/-</sup> lines (Figure 49B). Because the 5mC decrease observed in our LC-MS data was highly significant, in order to exclude any doubt, we did IF staining of the protein levels of the canonical *Dnmts* in *Dnmt2*-KO cells. The IF showed normal levels of the *Dnmts* and the variable cellular levels of the *Dnmt3s* are usual (Figure 49C). These results confirmed that the only methyltransferase, which was decreased in the *Dnmt2*<sup>-/-</sup> cells, was *Dnmt2* with no concomitant decrease of canonical *Dnmts*.

To confirm if indeed the *Dnmt2* protein has an activity on DNA in mES cells, and contributes towards non-CG context methylation, we used an alternative approach. In order to find small genome-wide differences in the methylation pattern of the *Dnmt2*-KO lines, high depth WGBS has to be performed. The cost of this approach could not be justified for a trial experiment, and the differences could still be beyond the detection limit when conversion artefacts are taken into account. Therefore, I focused my further efforts on the TKO ES cell line, being the only mammalian biological system, in which *Dnmt2* is the only DNA methyltransferase present. Other

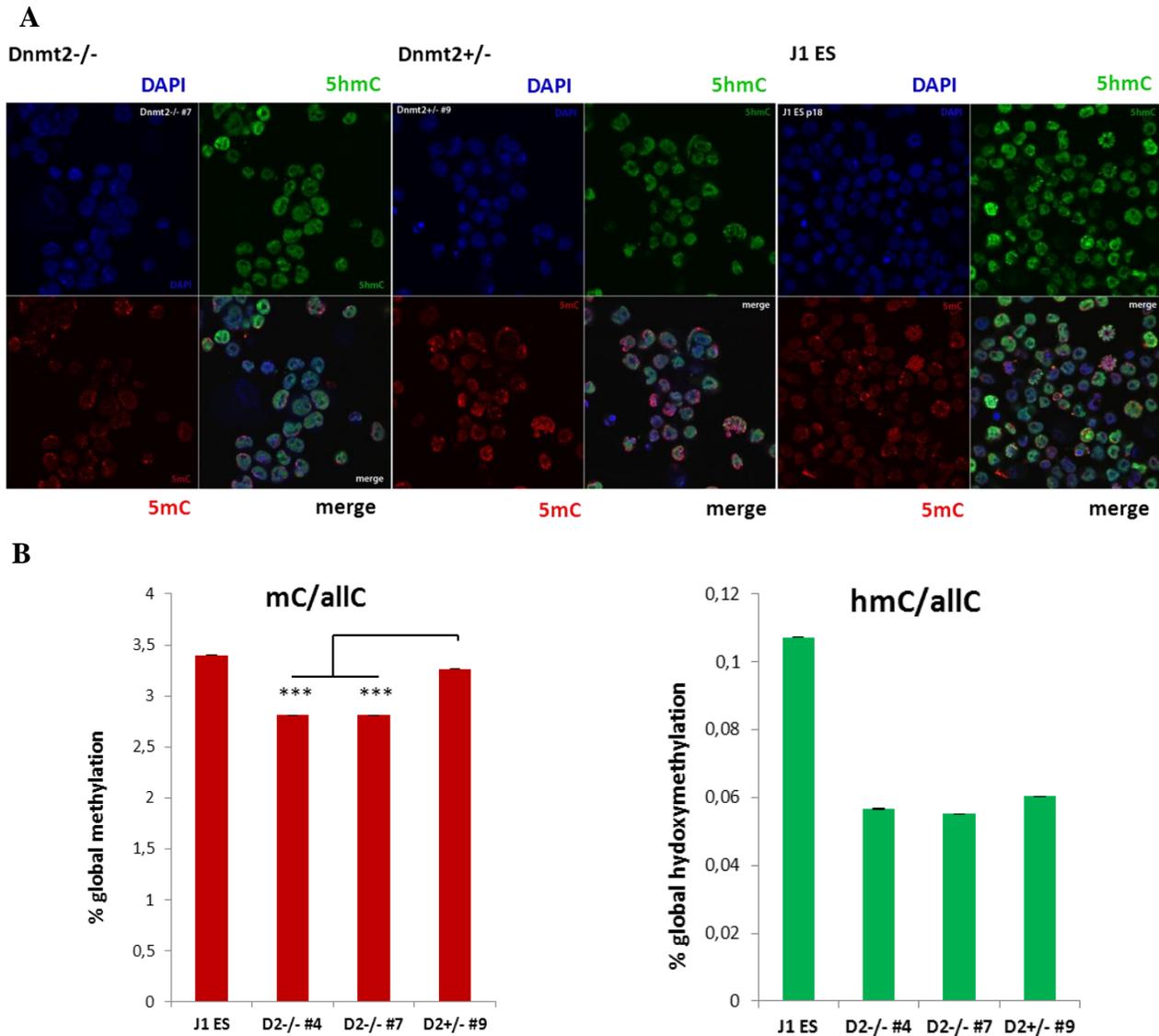


Figure 48. Estimation of global 5mC and 5hmC levels in Dnmt2-KO mESCs. **A.** Measured by IF. **B.** measured by LC-MS. D2-/-#4 and #7 are two replicates of Dnmt2-/- cells, and D2+/-#9 is their control Dnmt2+/- cell line.

*mansoni*, *Entamoeba histolytica*, and others (see Introduction). I performed low depth WGBS on the TKO ES line (already discussed in Chapter 3). The levels of unconverted cytosine found in the TKO line, however, did not differ from the levels of measured conversion artefacts (see Figure 11C). I then took a more targeted approach and examined the major satellite sequence with a classical targeted sequencing, as it was done in Chapter 3 for J1 ES. Due to its importance for the overall genomic stability and mitotic fidelity, the pericentric repeat sequence often remains highly methylated even during the natural phases of global DNA demethylation in early development, as

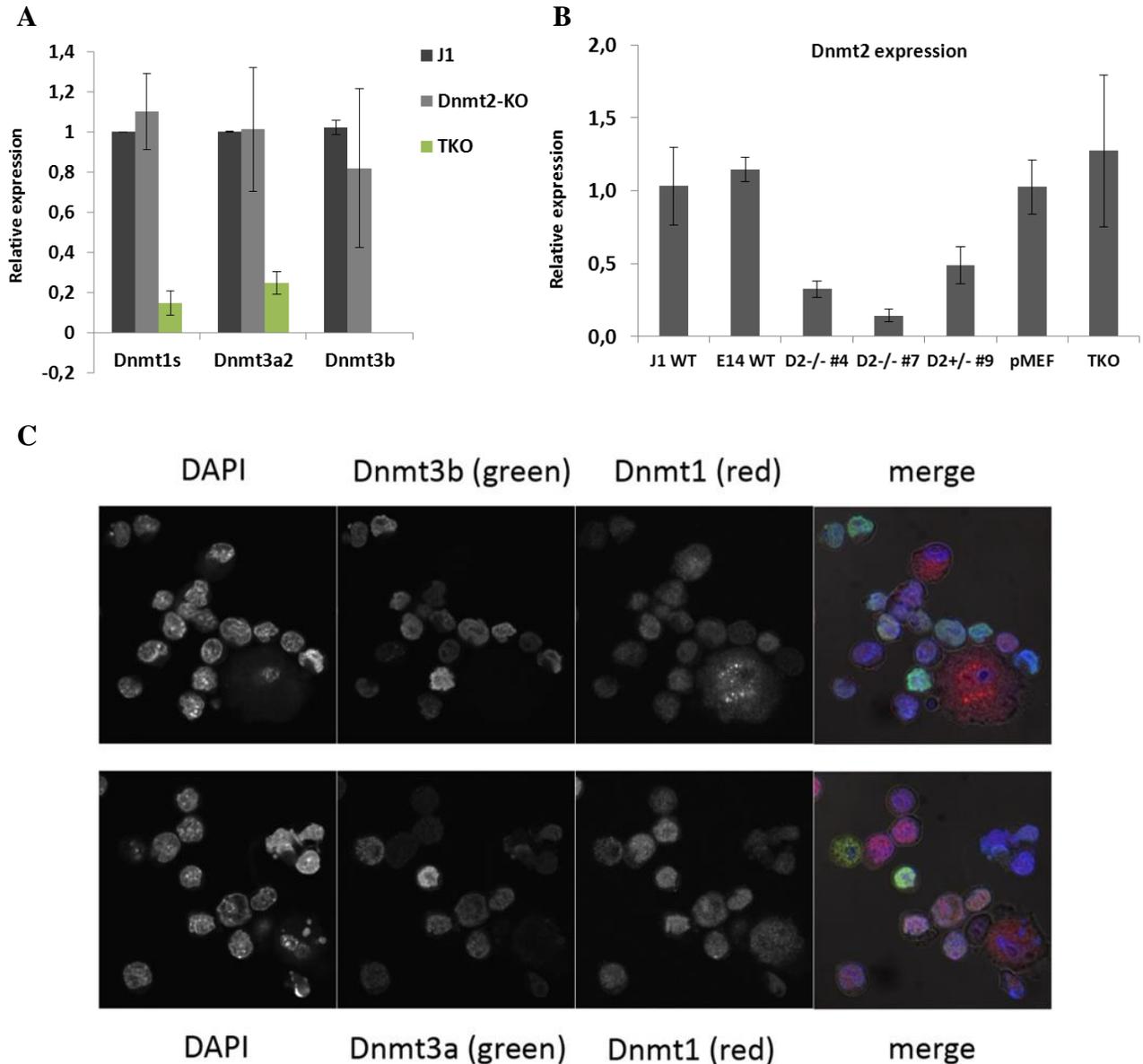


Figure 49. Expression of canonical Dnmts in the Dnmt2-KO ES cell lines. **A.** RT-qPCR of Dnmt1, Dnmt3a2 and Dnmt3b, the values are normalised to J1 WT. **B.** Expression of Dnmt2 in a selection of WT cell lines and Dnmt2-KO lines. The TKO line, which lacks the canonical Dnmts expresses WT levels of Dnmt2. **C.** IF for Dnmt1, Dnmt3a (below) and Dnmt3b (above) in Dnmt2-KO mES cells, colours are marked on the image.

reported for PGCs (Seisenberger et al. 2012), the paternal pronucleus in the zygote (Salvaing et al. 2012), and the transition from primed to ground state of pluripotency in mESCs (Ficz et al. 2013). If any methylation were remaining in the TKO, then one would imagine it should occupy those methylation ‘fortresses’ which are preserved methylated by the cell even in extreme demethylation conditions. Moreover, the major satellite has been shown to have a considerable

degree of non-CG methylation relative to the genomic average. My results, however, did not reveal any methylation of the TKO line in the major satellite, on either of the strands (Figure 50A). Few unconverted cytosines were present on the reverse strand; however, they did not occupy any of the positions methylated in the J1 WT line, and fall within the expected level of conversion artefacts.

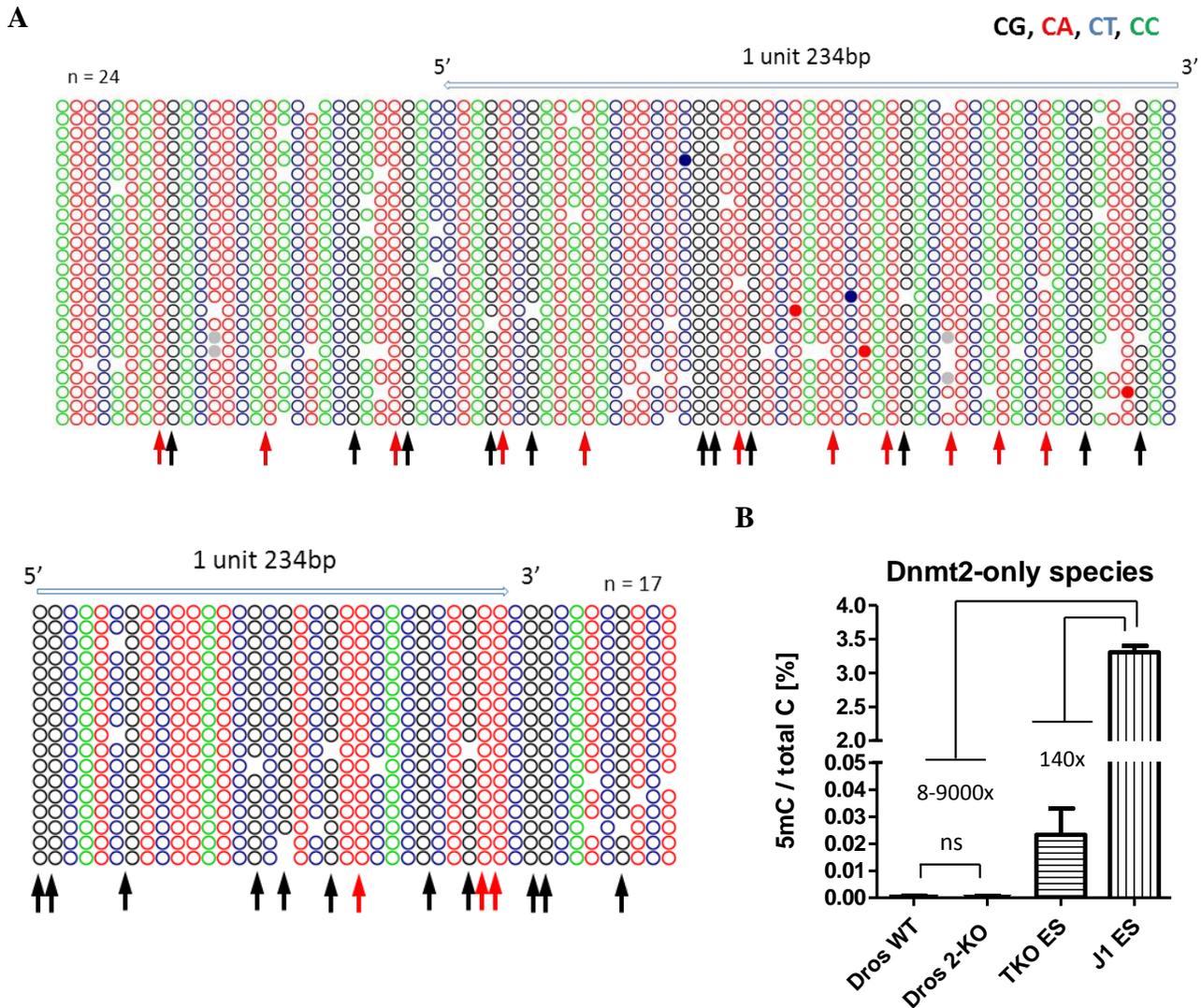


Figure 50. Methylation in the Dnmt-TKO mES cell line and other Dnmt2-only species. **A**. Targeted sequencing of the major satellite repeat in TKO mESC. Upper panel is reverse strand, lower panel is forward strand. Arrows mark where the 5mC signal is observed in the WT J1 ESC line (Black = CG, red = CA). **B**. Global genomic methylation of Dnmt2-only species measured by LC-MS. Fold-difference in global 5mC signal is shown for TKO or *Drosophila* samples in comparison to J1. The signal of *Drosophila melanogaster* WT and Dnmt2-KO DNA is not significantly different.

A final attempt to detect any methylation in other Dnmt2-only organisms, was performed in collaboration with Frank Lyko, who provided DNA samples from *D. melanogaster* and *Schistosoma mansoni*. LC-MS on whole genome DNA from fruit fly and the TKO line revealed very low signal, within the detection error of the equipment (B). *Schistosoma mansoni* had significantly higher signals, relative to zero, however the DNA from this parasite is known to be contaminated with bacteria and therefore the result could not be accepted as positive, alongside the signal from the uncontaminated Dnmt2-only DNA from mouse and fruit fly (Raddatz et al. 2013). At the same time the Dnmt2 enzymes in all three species were physiologically active and successfully methylated its tRNA target as shown in Raddatz et al, 2013. My results confirmed that Dnmt2 is not capable of delivering any *bona fide* detectable levels of DNA methylation. Even if very low levels of sporadic methylation occurred as a result of Dnmt2 activity, they would not exceed any biologically significant threshold, and would not explain the levels of native CH methylation.

### 6.3.2 *mCH contribution of the canonical Dnmts*

My next step was to evaluate the contribution of each of the canonical Dnmts towards CH methylation in my mES model system. I used a panel of constitutive knockout cells for each of the Dnmts (1<sup>-/-</sup>, 3a<sup>-/-</sup>, 3b<sup>-/-</sup>) and Np95<sup>-/-</sup>, including a double *de novo* methylase knockout (3a<sup>-/-</sup> 3b<sup>-/-</sup>, or DKO) and TKO (1<sup>-/-</sup>3a<sup>-/-</sup>3b<sup>-/-</sup>). A Dicer<sup>-/-</sup> mESC line was also included, to evaluate a possible role of RNAi pathways in the *de novo* methylation of CH context. My results confirmed a clear contribution of the *de novo* Dnmt3a and Dnmt3b MTs, which show around 50 % decrease of global mCA in both mutants, without a significant quantitative difference between the two (Figure 51 left panel). The DKO on the other hand was completely depleted of mCA, reaching the absolute zero and background values of the TKO and the unmethylated control DNA. The maintenance methylation knockouts Dnmt1<sup>-/-</sup> and Np95<sup>-/-</sup> also did not differ significantly between each other, fluctuating between 60 % and 100 % of WT mCA values. The Dicer<sup>-/-</sup> mES cells did not show any change in mCA methylation, contrary to expectations, following the downregulation of Dnm3a and Dnm3b in Dicer-KO cells (Kanellopoulou et al. 2005; Nesterova et al. 2008). The high clonal variability, however, might account for that. The results for mCG were in good agreement with previous reports, serving to validate the mCA/mCG antibody ELISA approach. The 3a and 3b single knockouts showed a trend for a slight decrease (~15 %), although

not quite statistically significant, due to replicate variation; the DKO and TKO mESC lines again showed background signal value (our DKO line is high-passage) (Figure 51 right panel). Dnmt1<sup>-/-</sup> and Np95<sup>-/-</sup> showed a marked decrease in mCG levels in comparison to their mCA levels, fluctuating between 30 % and 60 % of the WT mCG value. The Dicer<sup>-/-</sup> line did not show any statistical difference in mCG, again due to clonal variation (Figure 51).

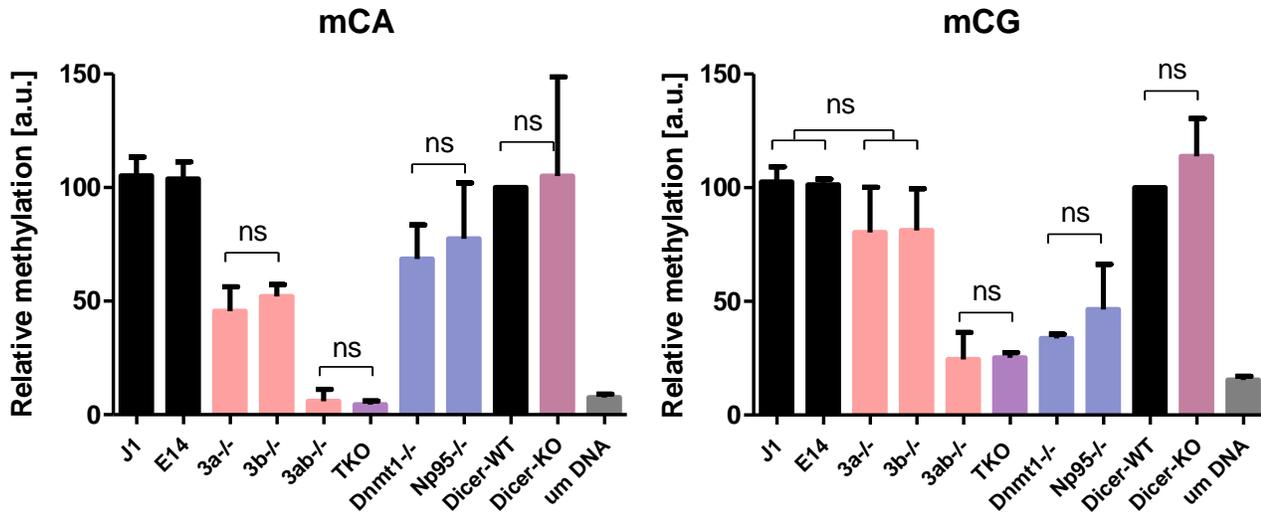


Figure 51. ELISA for global mCA and mCG levels in a panel of Dnmt KO mESCs and Dicer-KO. Left: mCA, right: mCG. All column comparisons not marked as non-significant are significantly different.

In summary, these experiments showed:

1) That both Dnmt3a and Dnmt3b contribute to non-CG methylation, although it is not clear whether this result is due to the direct enzymatic involvement of both enzymes, or there is a secondary effect due to their tight structural cooperation and functional dimer formation;

2) Dnmt1 and its partner Np95 do not seem to be directly involved in setting mCA marks, although there is a tendency of a decrease in mCA, which might be a secondary effect owing to Dnmt1's cooperation with the *de novo* Dnmts, as discussed in 1.3.1.

3) The previously reported decrease of Dnmt3a and Dnmt3b expression in Dicer<sup>-/-</sup> mES cells does not seem to lead to any detectable global decrease in either mCG or mCA levels. If RNAi has a role in targeting *de novo* DNA methylation, and disruption of this pathway would lead to decreased global methylation levels, then my results did not lend support to such a model.

It has been suggested that the presence of non-CG methylation is highly linked with the expression of Dnmt3a or 3b, in the sense that its occurrence could be a side effect of their high expression, rather than a regulated and biologically meaningful mark. In order to obtain a better understanding about the correlation of mCA levels and the occurrence of each Dnmt in the cell population, and investigate this possibility, I measured the mRNA expression levels of each of the KOs in my panel, and correlated those to global mCA and mCG levels. This correlation should replicate the result in Figure 51 to show directly which Dnmt is responsible for which type of methylation, since those systems are genetically manipulated to lack particular Dnmts. In addition, I have included non-manipulated biological samples, which express variable levels of the Dnmts, and have variable levels of methylation in each context. The results from this analysis will not necessarily show which Dnmt is responsible for which type of methylation. Their correlation, however, could show to what extent the mCA is directly linked to Dnmts' expression levels; whether it is an inevitable, and most likely a nonspecific consequence of the high expression of the *de novo* Dnmts, merely a function of their abundance and high activity at any point, or it is, on the contrary, a specific and regulated event.

The panel of WT samples included the tissues used in Chapter 5, as well as a time course of serum > 2i pluripotency ground state transition of an E14 WT mESC line. In this ground state transition, the *de novo* Dnmts quickly reduce their expression levels as a function of Prdm14 increase (Ficz et al. 2013), which could in turn be correlated with changing (or not) global mCA and mCG levels.

Interestingly, the highest significance correlation for mCA was achieved with Dnmt3b with the KO panel of ES cells ( $p=0.0008$ ), but there was also a clear trend for the entire WT panel of samples ( $p=0.0275$ , not shown on figure) (Figure 52, upper right). The ES KO panel showed the expected link between mCA and Dnmt3a2 as in Figure 51, although significantly lower than with Dnmt3b ( $p=0.0224$  with Pearson correlation and  $p=0.0087$  with Spearman correction for non-Gaussian distribution). There was no correlation for Dnmt3a1, which has low expression in ES cells. Dnmt3a1 however showed borderline significance with the WT group of tissues ( $p=0.0458$ ), suggesting it may play a role in mCH levels observed in adult tissues. However, because very few tissue types have any significant mCH levels, and for both male PGCs and oocyte it is known that the shorter Dnmt3a2 is the *de novo* Dnmt3a variant (Sakai et al. 2004; Ooi et al. 2009), this correlation has to be further verified with more replicates and tissue types. Neither Dnmt3L, Dnmt1, nor Np95 showed any correlation with the global levels of mCA, although a Spearman

correction for non-Gaussian sample distribution gave a strong correlation for Dnmt3L with the group of WT samples ( $p=0.0091$ , not shown in figure) (Figure 53, lower panels). Although a Dnmt3L knockout line was not included in this study, the rest of the samples showed a variety of high and low levels of Dnmt3L expression, but none of those correlated with measured mCA levels for the individual groups. In conclusion, these results indicate that the presence of Dnmt3b is most tightly related with presence of non-CG methylation, suggesting a lower level of regulation, while Dnmt3a and Dnmt3L could have a more highly regulated targeting for non-CG context.

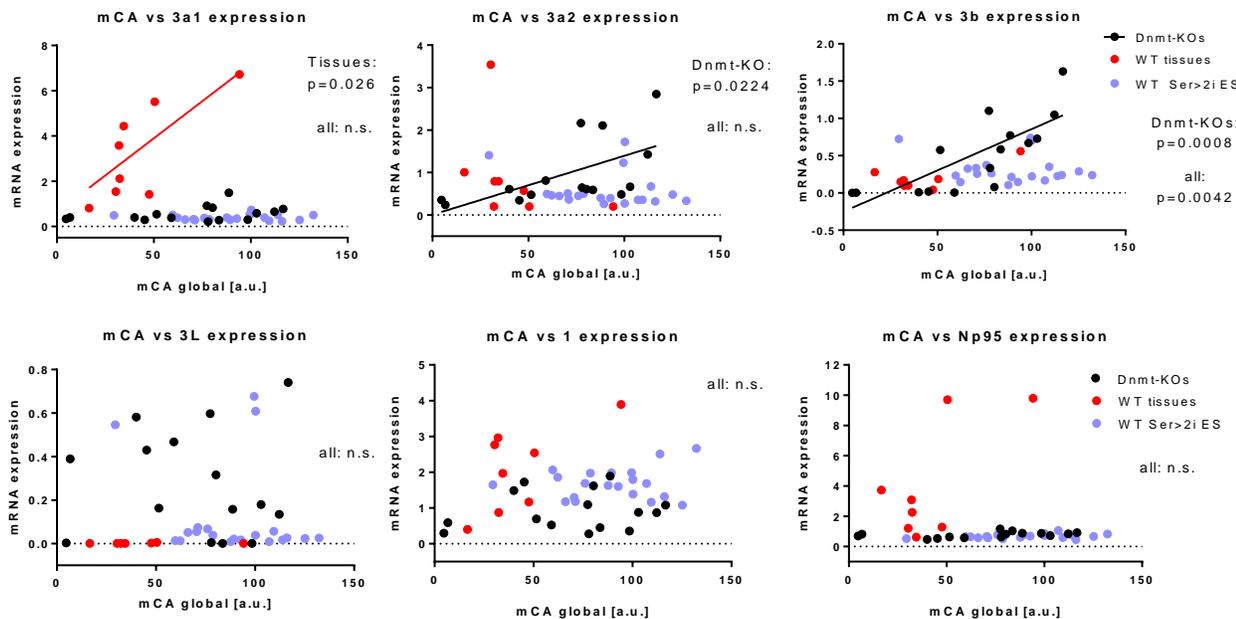


Figure 52. Direct correlation between Dnmts' expression levels and global mCA methylation. Three groups of samples were analysed: a panel of Dnmt KO mESCs (Dnmt-KOs), a panel of adult mouse tissues (WT tissues), and a serum > 2i reprogramming time course of E14 mESC (WT Ser>2i ES). The group 'WT' includes both groups of WT samples (tissues and 2i ESCs). The lines represent linear regression; unless otherwise stated, the correlation coefficients were calculated with a Pearson two-tailed test in Prism 6.0.

For mCG the results were also quite interesting. Due to the high redundancy between the different Dnmts in the maintenance and establishment of CG methylation, it is difficult to get as clear correlations for mCG as for mCA and the correlations were not as strong as for mCA. Unlike mCA, all *de novo* Dnmts showed borderline correlations with mCG in all samples combined – Dnmt3a1 (the adult Dnmt3a,  $p=0.0337$ ), Dnmt3a2 (the early development Dnmt3a,  $p=0.0269$  with Spearman non-Gaussian correction) and Dnmt3b ( $p=0.054$ ), while Dnmt1, Np95

and Dnmt3L showed no significant correlation for the pooled samples (Figure 53). Within the broad correlation, Dnmt3b correlated with the pool of ES cells ( $p=0.0192$ ) and the pool of WT samples ( $p=0.0192$ ), while Dnmt3a2 correlated with ES cells only ( $p=0.0334$ ) and Dnmt3a1 with WT samples only, including tissues ( $p=0.312$ ), which matches the known distributions of both isoforms of Dnm3a. This could suggest again a less regulated activity for Dnmt3b in comparison to the other two members of the *de novo* Dnmt family. In addition, Dnmt3b showed a correlation with CG methylation in the serum to 2i transition of E14 ESCs ( $p=0.0109$ ), but no such correlation was observed for Dnmt3a2, consistent with reported observations that their global decrease of methylation is due to a drop in the Dnmt3b expression (Figure 53, upper left panel). Dnmt1 is the only enzyme, which showed a correlation for the group of mES Dnmt-KO cells ( $p=0.414$ ) (Figure 53, lower middle panel), potentially due to the lack of a strong redundancy in its maintenance role. Interestingly, Dnmt3L showed correlation with mCG in ES cells ( $p=0.0358$ ) unlike mCA, while Np95 showed no correlation with mCG. This might be explained by the low number of tissue samples, where Np95 is highly expressed, and points to a lesser importance of the maintenance pathway in mES cells.

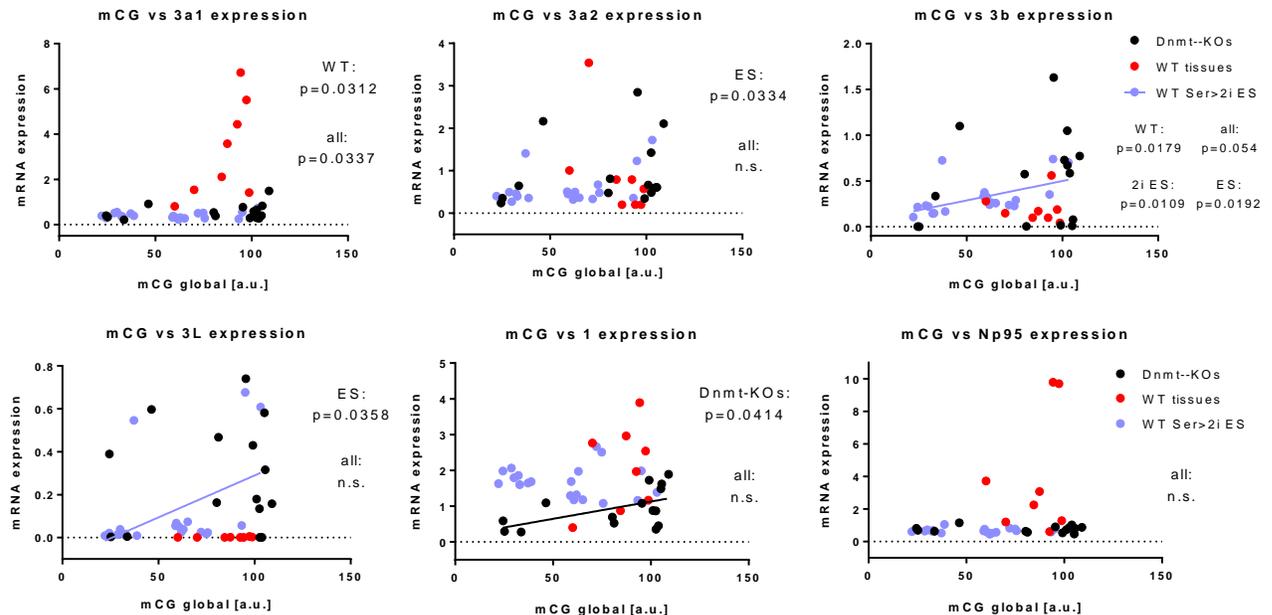


Figure 53. Direct correlation between Dnmts' expression levels and global mCG methylation. Three groups of samples were analysed: a panel of Dnmt KO mESCs (Dnmt-KOs), a panel of adult mouse tissues (WT tissues), and a serum > 2i reprogramming time course of E14 mESC (WT Ser>2i ES). The group 'WT' includes both groups of WT samples (tissues and 2i ESCs). The lines represent linear regression; unless otherwise stated, the correlation coefficients were calculated with a Pearson two-tailed test in Prism 6.0.

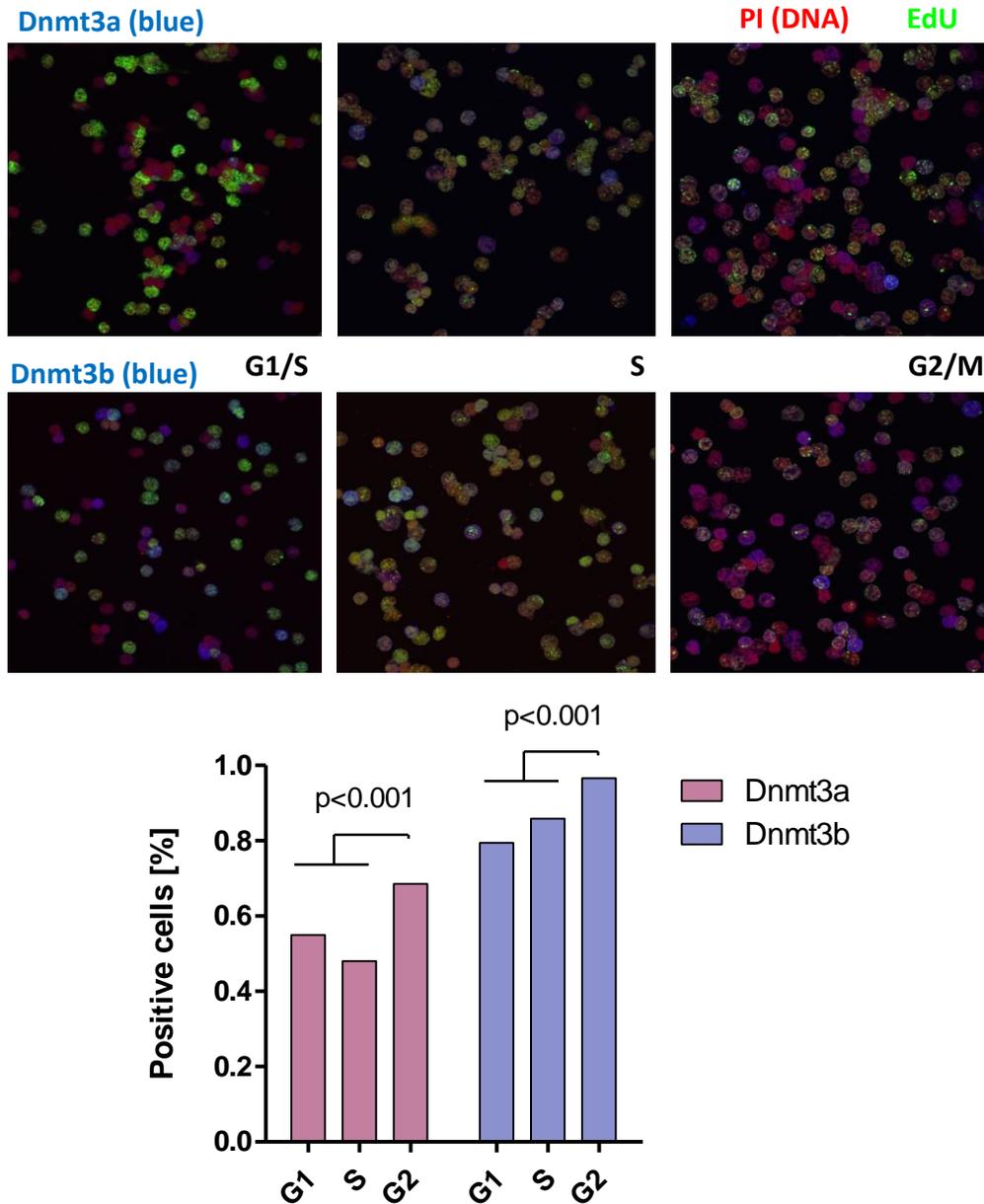


Figure 54. Expression of Dnmt3a and Dnmt3b in the different phases of the cell cycle of WT mESCs. Upper panel: IF of FACS-sorted cell populations, pulse labeled with thymidine analogue EdU (5-ethynyl-2-deoxyuridine) as a marker of DNA replication (green), PI (propidium iodide) as an intercalating DNA marker (red) and either Dnmt3a or Dnmt3b (blue) via IF. Significance of peaks is measured with Chi square test.

Another way to assess the direct link of mCA methylation and the abundance and activity of Dnmt3s was to follow their expression throughout the cell cycle, since I had already measured the global mCA and mCG values of cell cycle sorted cell fractions (Figure 44). For this purpose, I stained FACS-fractionated mES cell populations for Dnmt3a and Dnmt3b and estimated the

relative quantity of stained cells in each population (Figure 54). The EdU pulse labelling was again useful to discriminate S-phase cells in the G1 and G2/M fractions. Although the *de novo* Dnmt3s were present in each cell fraction, my results showed a significantly higher expression of both Dnmt3a and Dnmt3b in G2/M phase, following DNA replication (Figure 54, lower panel). Such a result would make sense in the light of the more recent understandings that the *de novo* Dnm3s take an active part in the maintenance of DNA methylation marks. These results also highlight the heterogeneity of ES cell populations, in the observation that only 50-70 % of the ES cells stained positively for Dnmt3a, while 80-95 % stained for Dnmt3b. This observation supports my result, that Dnm3a2 is a more tightly regulated enzyme than Dnmt3b (Figure 52 and Figure 53). The dynamics of mCA and mCG from my previous measurement in cell cycle sorted ES cells does show a potential global methylation increase in the G2/M fraction. However, this result is for the global 5mC and the result for mCA in particular shows a significant increase in the G1 phase, and not in G2/M. This again would show, that mCA methylation does not strictly follow the peaks in expression of the Dnmt3 enzymes, as previously suggested (Ziller et al. 2011), but has an independent and regulated dynamics.

### 6.3.3 Active demethylation of non-CG context

In order to investigate the sequence specificity of the Tet family enzymes in regard to their ability to oxidize 5mC to 5hmC, I have focused on Tet1, due to the commercial availability of catalytically active enzyme. As shown in Figure 6 (Introduction, 1.4), the structure of the catalytic domain of all Tet family members is highly conserved. Therefore, results obtained for Tet1 in a simplified *in vitro* system, where no potential allosteric regulation by nuclear components would take place, is likely to extend to the other Tet members under similar conditions.

I have performed an *in vitro* Tet1 oxidation of biotinylated dsDNA fragments (Table 5) which contain a single methylated site each, in variable cytosine context, and analysed the resulting oxidation products with an avidin-biotin sandwich ELISA system. My results show very high Tet1 activity on the symmetrically CG-methylated DNA substrate ('mCGm'), but much less on a hemi-methylated CG substrate ('mCG'), or on the mCA substrate ('mCA') (Figure 55A). The amount of 5mC had also changed accordingly on the same fragments (Figure 55B). To independently verify these results with a greater degree of stringency, and assess for further oxidation products, I also measured the oxidation fragments upon Tet1 treatment with LC-MS

analysis. The result was very similar to that of the ELISA in absolute numbers (Appendix Figure 74. The oxidation kinetics for 5fC accumulation for each substrate followed precisely the acquisition of 5hmC (Figure 55C and D). Only a small proportion (6 – 8 %) of 5fC was detected from the pool of oxidation products, in similar ratios for each substrate) (Figure 55E).

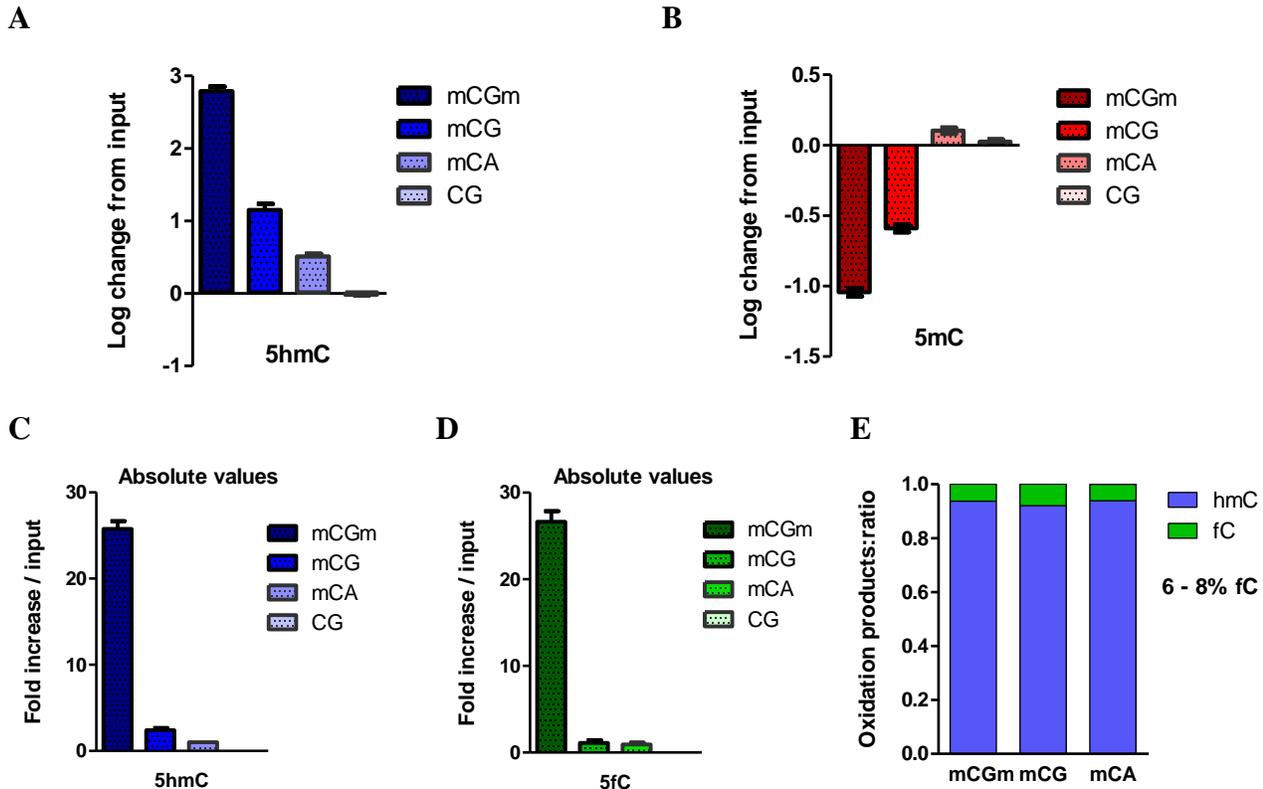


Figure 55. Tet1 activity on 5mC in different context. **A** and **B**. Log<sub>2</sub> acquisition of 5hmC signal on oxidised dsDNA-biotin fragments measured by biotin-avidin ELISA for 5hmC (**A**) and 5mC (**B**). **C** and **D**. LC-MS measurements of absolute oxidation products in same context-specific oxidized dsDNA fragments, 5hmC (**C**) and 5fC (**D**). Ratio of oxidation products 5hmC and 5fC in each substrate reaction. 5caC was not measured with LC-MS, and both 5fC and 5caC were beyond the sensitivity of the ELISA technique.

I next looked at the possibility that non-CG context methylation was able to escape or interfere with the repair-based mechanisms of active demethylation. The main proteins, implicated in initiating the short- or long-patch BER pathway, are TDG and AID, although for TDG it seems more likely to participate downstream of the oxidation pathway, as discussed in 1.4 of Introduction. It has been shown, that AID-KO fertilised eggs experience less demethylation of their paternal pronuclei (Santos et al. 2013), which are exclusively methylated in CG (Figure 33). This means that if AID contributes to active removal of mCH as well, then in a mCH-high AID-

KO system where active demethylation is happening, we could expect to observe an accumulation of mCH. The paternal pronucleus is not a good model because of its virtual lack of non-CG methylation, however, mESCs, which have relatively high amounts of mCH, could be an appropriate system. AID is expressed in mESCs and is involved in active demethylation during the reprogramming towards pluripotency in iPSCs (Bhutani et al. 2010; Bhutani et al. 2013; Kumar et al. 2013). I therefore measured global mCG and mCA amounts in iPSCs from WT and AID-KO pMEFs (generated in the Reik lab by Inês Milagre). There was no increase of global mCA in the AID-KO iPSCs, or the AID-KO pMEFs from which they were reprogrammed (Figure 56A). The same applied for mCG, however, which is a known target of demethylation by AID-mediated pathways (Figure 56B).

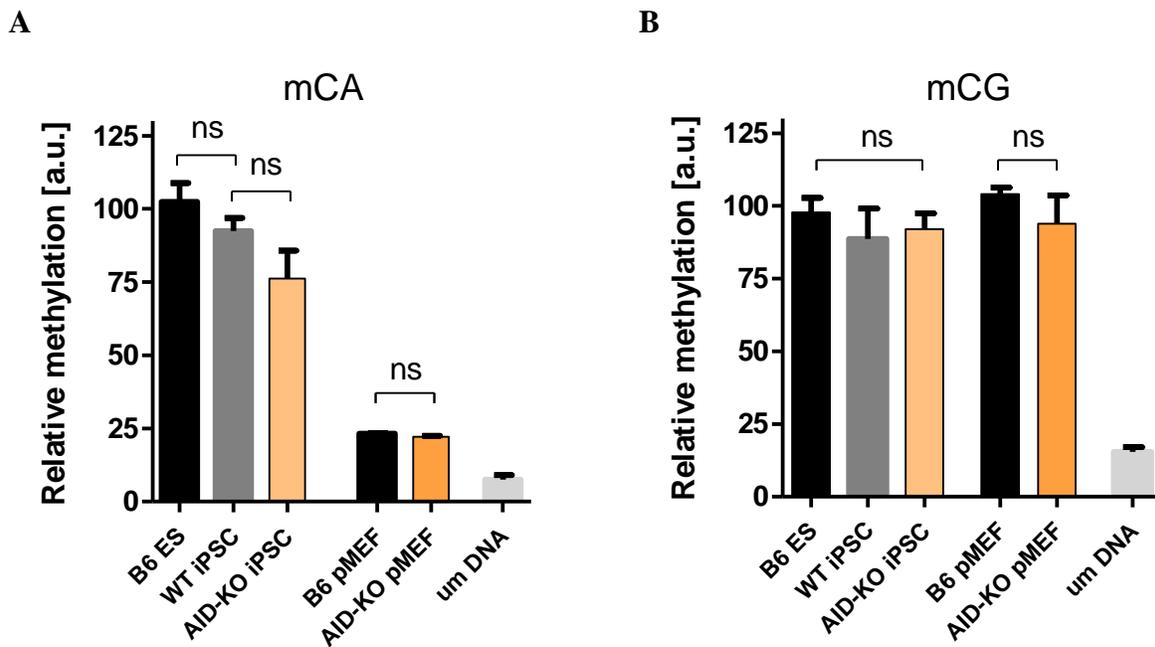


Figure 56. CA and CG methylation in AID-KO cells. The global methylation in CA (**A**) and CG (**B**) context, normalized to either B6 WT mESCs or pMEFs. None of the differences between the AID-KO cells and their corresponding WT were statistically significant. (note: there is a significance between AID-KO iPSC and WT B6 mESC, but not with the BS iPSC control line). umDNA is unmethylated DNA control.

This result could potentially indicate that the iPSC reprogramming system was not an appropriate approach for assessing the contribution of AID towards 5mC demethylation. None of the other systems where active demethylation is taking place however, like the early PGCs and paternal pronucleus, have enough mCH levels to be appropriate for this analysis.

I therefore undertook a different approach, based on bioinformatic analysis of WGBS datasets and prediction of outcomes. It was a step forward in the idea that the highly methylated maternal genome is ‘protected’ or ‘resistant’ to active demethylation by virtue of its high mCH levels. Here it was assessed from the perspective of AID and the repair pathways, and not the Tet enzymes, which we already know mCH is not substrate to. The approach was based on the zygote as a model, and the known specificity of AID. The consensus sequence of AID is WRC and it is known that its affinity for 5mC substrate (i.e. WR<sup>m</sup>C) is around 10-fold lower than towards its natural target – the unmethylated cytosine (Nabel et al. 2012). I reasoned that cytosine methylation of this sequence, might be a way to hinder the activity of AID, meaning that the more methylated this sequence motif is, the less AID-mediated active demethylation will occur. It seems plausible that such a mechanism might be employed by the maternal pronucleus in order to ‘escape’ active demethylation, especially if this is done by a type of methylation, which is in itself resistant to hydroxymethylation. I therefore measured methylation levels of the WRC motif in the genomes of oocyte and sperm, from published datasets (L. Wang et al. 2014; Kobayashi et al. 2012; Shirane et al. 2013; Smallwood et al. 2014). The methylation of this sequence was indeed higher in the oocyte genome, as expected, but surprisingly in the paternal pronucleus it was actually lower than the global 5mC genomic averages of both MII and sperm (Figure 57A). The context composition of the WRC motif is the same as the genomic cytosine composition, meaning this effect is not due to underlying sequence biases (Figure 57B). This suggests there might be a regulatory mechanism in place, which controls the level of methylation of the WRC sequence motif in the oocyte and the sperm.

Although on a global scale, the methylation of the WRC motif seems low (5.43 % for MII) and the majority of WRC sites (> 90 %) remain unmethylated, we should not forget that the global methylation levels for the mouse genome range between 3-4 %, and 2-3 % for the human, and the CG context alone is around 3% for the mouse genome. Next to those values, which undoubtedly have a significant effect on the function of the mammalian genomes, the 5.43 % seem a very high value for a single sequence motif. For comparison, the global 5mC change between primed (serum + LIF grown) and naïve (2i grown) mESCs is slightly higher than 2-fold (3.10 % 5mC primed vs 1.43 % 5mC naïve, Figure 46) and this leads to profound phenotypic changes.

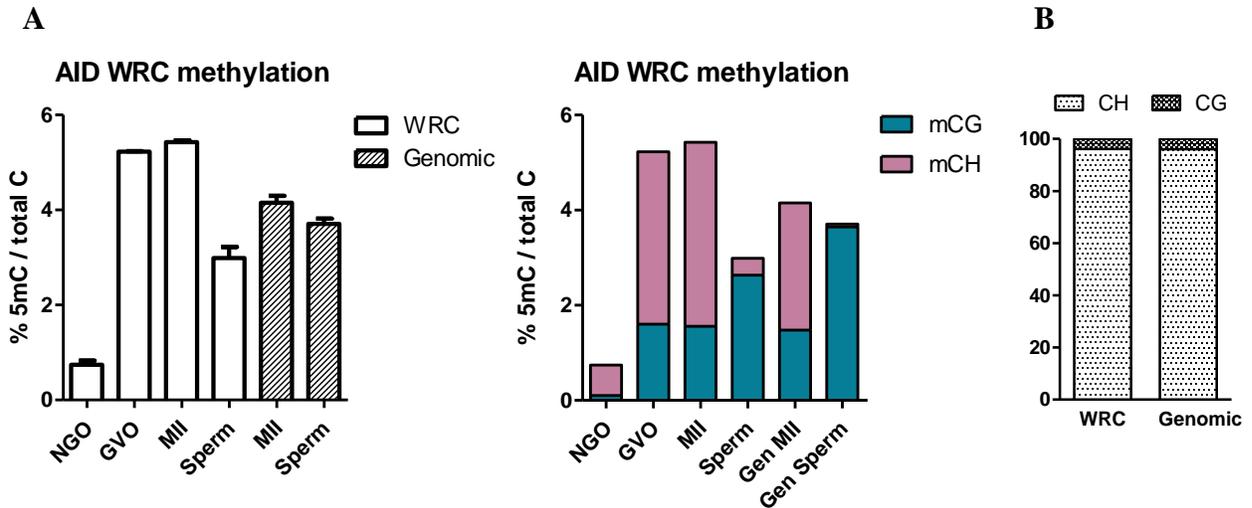


Figure 57. Methylation in the AID targeting motif WRC. **A.** Global methylation in the WRC motif for oocyte and sperm, in comparison to the global genomic values (left); right: the same figure with sequence context proportional values (mCG and mCH). **B.** Context composition of the mouse WRC motif in comparison to the genomic cytosine composition.

## 6.4 Discussion

In this chapter, my experiments addressed the individual roles of all mammalian Dnmts in their contribution to non-CG context methylation. Although it might seem that this topic has already been addressed substantially in the existing literature, there are a few important findings from my work.

Dnmt2 has been controversial in its role in DNA methylation. Although a number of papers have claimed in the past (Lyko, Whittaker, et al. 2000; Kunert et al. 2003), and more recently (Takayama et al. 2014; Capuano et al. 2014) such a role, my results could not support such a claim (Raddatz et al. 2013). I generated the first WGBS dataset of a Dnmt2-only mammalian model organism, and with the conversion rates from my internal spike-in control I could not justify that the observed low signal is real. Moreover, data from highly sensitive mass spectrometry also could not support the presence of any methylation above background and beyond the equipment's thresholds of detectability. The fact that WT and Dnmt2-KO from *Drosophila melanogaster* DNA did not show any differences, clearly shows that Dnmt2 cannot have a role in DNA methylation in the fruit fly. Moreover, the measured fruit fly samples ranged between different tissues and developmental stages, but showed equally negative results, and were

pooled in the figure for clarity. None of the recent papers included a Dnm2-KO DNA in their studies, and the compared measurements of global 5mC in the mouse within the range of  $7.6 \pm 0.8$  % are highly overestimated (Capuano et al. 2014). In addition, one of the studies attributes the DNA methylation activity in the fruit fly to a novel unidentified enzyme (Takayama et al. 2014).

The ability of the two major *de novo* Dnmts to act both *in vitro* and *in vivo* on non-CG context has been known for a long time (Aoki et al. 2001; Gowher & Jeltsch 2001; Mund et al. 2004; Lyko et al. 1999). Later, CH methylation has been associated most commonly with the presence of Dnmt3a (Lister et al. 2009; Shirane et al. 2013; Lister et al. 2013) or both Dnm3a and Dnmt3b (Ziller et al. 2011). Those associations are mainly based on the availability (or highest expression) of a particular *de novo* methyltransferase in the relevant study tissues, and a comparable cross-tissue and cross-developmental stages study so far has not been done. Also, the fact that under the term ‘Dnmt3a’ there are actually two distinct enzymes is rarely acknowledged. On one hand, for some studies this is understandable, because the Dnmt3 isoform in the oocyte (Shirane et al. 2013) and mES cells (Stadler et al. 2011; Ziller et al. 2011) is of the same type (Dnmt3a2), but it does not become clear which is the main isoform in the mouse brain, when the activity is attributed to Dnmt3a (Lister et al. 2013).

My study evaluated the two isoforms of Dnmt3a independently. Dnmt3a2 showed a strong correlation with the presence of mCA, and at the same time suggests a higher level of regulation in WT samples, where its presence does not always predict higher mCA levels. Dnmt3b is indisputably equally responsible for establishing non-CG methylation marks, and its activity seems less regulated for the establishment of both mCG and mCA marks – its presence highly correlates with both marks. The fact that Dnm3a2 seems to be under a more stringent regulatory control, is also in line with its tightly regulated expression in particular windows during mammalian development. It is interesting that Dnmt3a2 is a truncated form of Dnmt3a1, and it lacks additional unique regulatory domains that could support an alternative allosteric regulation. Therefore it remains an open question how the activity of this isoform is regulated, although a likely possibility is that this happens through posttranslational modifications or in an ncRNA dependent manner (Jeltsch & Jurkowska 2014). A recent study reported that the activity of Dnmt3a2 can be downregulated via phosphorylation by CK2 (Cycline kinase 2), leading to a decrease in overall CG methylation levels within particular genomic regions (Deplus et al. 2014). It would be interesting to find out if such regulation would also affect non-CG methylation. The

lack of some regulatory domains in Dnmt3a2 might also explain its close cooperation with Dnmt3L, which acts as an allosteric regulator. Indeed, it has been shown that the activity of Dnmt3a2 rises 20-fold upon cooperation with Dnmt3L (Kareta et al. 2006). Nevertheless, the expression of Dnmt3L is even more tightly regulated, and the fact that Dnmt3L showed correlation with mCG levels rather than mCA suggests that Dnmt3a2 could also function independently, or in a response to stimuli and regulators different from Dnmt3L. Nevertheless, the targeting towards imprinting loci in the oocyte and towards repetitive sequences in the male PGCs, both of which contain non-CG methylation, has been shown to depend strictly on its cooperation with Dnmt3L (Bourc'his et al. 2001; Bourc'his & Bestor 2004; Hata et al. 2002; Hata et al. 2006).

Dnmt3a1 has borderline correlation with non-CG methylation, although this result is grounded only in its very high expression in brain cortex (Appendix, Figure 73). On the other hand, it is also very highly expressed in other tissues, which do not have high mCH levels, suggesting it may not be the factor responsible for high mCH in brain. Clearly, more samples are needed to validate this observation. Moreover, it has been reported that the mouse hippocampus expresses Dnmt3a2 and its presence is crucial for the proper cognitive function, which declines with aging along with the decline in Dnmt3a2 expression (Oliveira et al. 2012). It therefore remains an open question whether the high mCA levels in the brain are due to Dnmt3a2, or the ubiquitous Dnmt3a1. Interestingly, the same study shows that Dnmt3a2 has been robustly and transiently activated by neuronal receptor stimulus, and was partially responsive to nuclear calcium signalling, supporting a high level of regulation for this enzyme (Oliveira et al. 2012).

For Dnmt3b my results indicate that the higher the levels in a cell population or a tissue, the higher the presence of the mCH and mCG marks will be. However, the expression of Dnmt3b is much more ubiquitous and homogenous within the mES cell population, while the distribution of mCA signal is highly variable (Figure 54). This fact may also point towards cooperation between Dnm3a and Dnmt3b in establishing the mCH marks. It has been reported that the two work synergistically in ES cells and potentially form heterodimers (Li et al. 2007), and while the Dnmt3b activity is necessary, it might be modulated by the presence of Dnmt3a2 to drive the heterogeneous distribution of mCH. It would be important to co-stain and compare mCA and Dnmt3a2 patterns, and determine if the variation in mCA in a mES cell population follows the fluctuations of Dnmt3a2 or Dnmt3b expression.

Regarding the maintenance methyltransferase Dnmt1, my results show a borderline

correlation for mCG only. This seems due to the high redundancy with Dnmt3s and it has been known, that the maintenance methylation, especially of repeats, is highly dependent on the Dnmt3 enzymes (Kim et al. 2002; Chen et al. 2003; Liang et al. 2002). In my observations, Dnmt3s indeed seem the main driver for DNA methylation. Moreover, it has been shown that Dnmt1 does not copy effectively the moderately methylated regions like imprints, and its activity is highest in densely methylated sequences, which makes it unreliable at regions with intermediate methylation (Lorincz et al. 2002). This suggests that the role of Dnmt3s in maintenance is essential and this is not the domain of Dnmt1 only, as has also been proposed recently (Jeltsch & Jurkowska 2014). Interestingly, my result argues against any independent purely *de novo* activity of Dnmt1.

Dnmt3L shows a definite correlation with mCG in ES cells, and potentially with mCA. No correlation with methylation levels was found for Np95, pointing to a weaker role in mES cells, which is surprising given their high replication rates, in comparison to adult tissues. However, Np95 is a large enzyme, an E3 ubiquitin ligase, with roles outside of DNA methylation, including chromatin modifications, heterochromatin reorganisation and DNA damage repair (Mistry et al. 2008; Nishiyama et al. 2013; Papait et al. 2007). This is a possible reason why its levels could not be directly linked with DNA methylation levels.

My findings about the potential resistance of mCH marks to active demethylation are intriguing and open new questions and perspectives for follow up work. This is a property that currently clearly distinguishes the mCH from the mCG mark, and if it could be validated *in vivo* during the process of active demethylation in the early embryo, this would finally constitute a functional significance of non-CG context methylation.

Although promising, the possibility that mCH methylation affects AID activity in the zygote is not clear, given that more than 90 % of the WRC sites remain unmethylated in the maternal genome. There are a few ways in which this could affect globally the AID activity, beyond the direct physical blockage of preferred target site. It is possible that the AID enzyme, similarly to Dnmt1, is sensitive to overall DNA methylation quantities for a wider region, and responds with a change in activity if a number of AID sites are methylated above a certain threshold. Another option is that the high CA methylation could be changing the physical properties of DNA, by making it less accessible for AID. This effect could be achieved with a small number of modifications and will affect a wider span of sequences on a global scale. It is known that AID works only on single stranded DNA, and hence its activity is high when DNA strand separation is

possible and DNA is in an open and accessible state as is the case during transcription (Brar et al. 2008; Parsa et al. 2012). On the other hand, methylation has been shown to modify very dramatically the mechanical properties of DNA, by changing flexibility, stabilising strand duplexes and affecting DNA breathing dynamics (Severin et al. 2011; Shimooka et al. 2013). Moreover, it is possible that mCH context methylation induces conformational changes (kinks) in the nuclear DNA, as it has been shown to do in solution *in vitro* (Shimooka et al. 2013; Johannsen et al. 2014). Such changes will certainly affect the overall DNA properties and local micro-environment, including strand coiling and tension. Many proteins' binding capabilities are affected by such factors (Alexandrov et al. 2012), and it has recently been shown that AID in particular relies on DNA breathing and negative supercoiling to access its single strand substrate (Parsa et al. 2012). It is therefore very likely that this scenario of high mCH methylation on the maternal pronucleus, but not the paternal, will have an effect on the activity of AID. The experimental approaches to address this directly though, remain very challenging due to the difficulty to generate a mCH-KO cell without affecting CG methylation.

In summary, I was able to address in higher depth the question about the role of Dnmts in the establishment of CH methylation, and showed that this process is regulated rather than unspecific. In addition, I uncovered an area of potential distinction between CG and CH methylation in their capacity to be (or not) actively removed, which may have profound impact on the dynamics of active demethylation in the early embryo.



## 7 *Functional significance of non-CG methylation*

### 7.1 Introduction

In this chapter, following the insights from my work on the overall characteristics of non-CG methylation, I will address the question of its potential biological significance. Since the process of its establishment by the Dnmts seems highly regulated, it is also likely that there is a functional significance associated with non-CG methylation. So far, CH methylation in ESCs has been associated with highly expressed genes (Lister et al. 2009; Ficz et al. 2011), although exactly the opposite has been reported for the mammalian brain (Lister et al. 2013; Guo et al. 2014). My findings indicated a strong correlation with Dnmt3a2, the so called ‘euchromatic’ methyltransferase, supporting an association with gene expression in mESCs, and so do my IF results for its nuclear distribution in ES cells (Figure 52 and Figure 41). It would therefore be justified to investigate this further, for a validation as well as a deeper insight into this association, to correlate the genomic distribution of non-CG methylation with histone marks. It is currently not known whether CH methylation is a consequence, an unrelated concomitant event, or a factor determining active transcription.

Because of the strong connection of CH methylation with ES cells, to the extent that the re-acquisition of non-CG context methylation is seen as a hallmark of a completed pluripotent state reprogramming for iPS cells (Lister et al. 2011), it will be important to find out how depletion of mCH affects stem cell pluripotency. This is of course very challenging due to the lack of conventional pathways to create mCH knockout systems. If mCH is indeed related to pluripotency, this also opens the question of protein binders, which can read this mark, as there are such for CG methylation.

I have therefore incorporated those aspects into my last aims of the project, to address as much as technically possible subject to time constraints for this thesis.

### 7.2 Aims

5. To correlate the mCH mark with histone modifications and protein binders using published ChIP-seq datasets
6. To look for protein readers of CH context methylation via direct nuclear pull-down

7. To analyse the effect of reduced non-CG methylation levels on the pluripotency status of mouse ES cells

## 7.3 Results

### 7.3.1 Correlation of mCH with histone marks and pluripotency factors

In order to gain further insights into the functional significance of non-CG context methylation, and be able to add more meaning to my previous findings, I first wanted to correlate non-CG methylation with histone marks, and also compare its genomic overlap with the binding profiles of proteins of interest. For this purpose, I performed MeDIP-seq with my mCG and mCA antibodies for J1 and E14 WT mESCs. Building on the experience from Chapter 3, I modified the MeDIP analysis and library preparation procedures in a couple of ways. First, by incorporating additional normalisation controls – one ‘Input’ sample (non-precipitated mouse gDNA), and one sample of an unmethylated *in vitro* amplified mouse genomic DNA. Secondly, all samples were TrueSeq barcoded and pooled in equal quantities, so that they were subjected to the immunoprecipitation together. The idea was to avoid the accumulation of artefacts in the less methylated samples, similar to the case with TKO in Chapter 3, by introducing a competition between the samples with varying amounts of 5mC. This modified approach ensured that both the sequence specificities and antibody pull-down artefacts will be taken into account and subtracted by the peak calling algorithm. Additionally, specifically for the mCA MeDIP data, to be extra safe Simon Andrews (Babraham Bioinformatics) filtered out the high CA genomic content, which seemed to be giving higher false positive signal (> 200 CA per 3kb). During the analysis, the mCG data was filtered in the same way, to ensure that I will be comparing the same sequences in both datasets, and no technical variability will arise from that.

I first explored the correlation with histone marks. I retrieved datasets from published studies on a selection of histone modifications, for which there were available datasets for mESCs. I mapped the localisation of my mCG or mCA peaks by the widely used MACS peak-caller (Zhang et al. 2008) and saved those genomic locations as genomic coordinates, which were then overlaid by the ChIP data. Where possible, more than one dataset was used to ensure a higher validity. ChIP-seq peak calling was performed in the same way, with Input or whole cell extract (WCE) used as a reference. The histone marks I analysed and the genomic signature they

were associated with are listed in Table 9, and the source datasets are listed in Table 22, Appendix.

Table 9. Histone modifications analysed and their corresponding genomic signatures.

Active	Signature
H3K4me3	Actively transcribed gene bodies, around TSS and transcriptional initiation
H3K4me1	Inactive/poised enhancers
H3K27ac	Active enhancers
H3K36me3	Gene bodies, transcriptional elongation
H3K79me2	Epithelial to mesenchymal transition genes (MEF), some housekeeping genes (ES)
Repressive	Signature
H3K9me3	Constitutive heterochromatin
H3K27me3	Facultative heterochromatin, bivalent domains in ESC
H4K20me3	Pericentric heterochromatin, imprinted regions and IAP retrotransposons
H3K36me3	Some constitutive heterochromatic loci

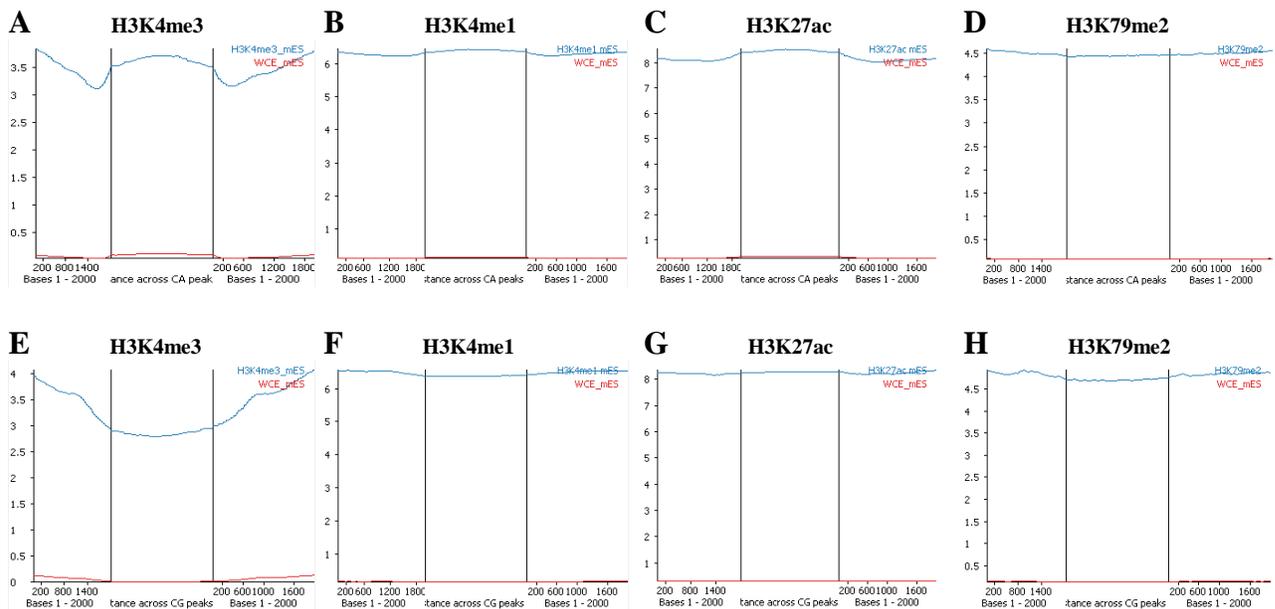


Figure 58. Correlation of mCA (A-D) and mCG (E-H) methylation with active histone modifications. 2kb flanking regions are displayed on both sides of the mCA or mCG peak, which is in the middle panel of each graph. **A, E:** H3K4me3; **B, F:** H3K4me1; **C, G:** H3K27ac; **D, H:** H3K79me2. The blue line represents the coverage trend plot of each histone mark, and the red line is the control (WCE) for comparison of background.

From the active marks, the strongest correlation for mCA was observed with H3K4me3 (Figure 58A). There was a very slight trend for H3K4me1, as well as H3K27ac, and absolutely no

correlation with H3K79me2 (Figure 58B, C and D). Interestingly, for mCG the trends were exactly the opposite: strong negative correlation with H3K4me3, slight negative trend for H3K4me1 and H3K79me2, and not for H3K27ac (Figure 58E-H). The H3K36me3 mark, which marks both active and repressive chromatin, is analysed with the repressive marks.

From the repressive marks, there was no correlation for mCA with H3K9me3, a marked negative correlation with H3K27me3, and no correlation with either H3K20me3 (borderline negative) or H3K36me3, both localising with pericentric heterochromatin (Hemberger et al. 2009; Chen et al. 2008) (Figure 59A-D). This result was surprising given the partial colocalisation of mCA with heterochromatic foci in both primed and naïve mESCs. For mCG there was no correlation with H3K9me3, nor H3K27me3 and H3K36me3, and a borderline positive one for H3K20me3 (Figure 59A-D).

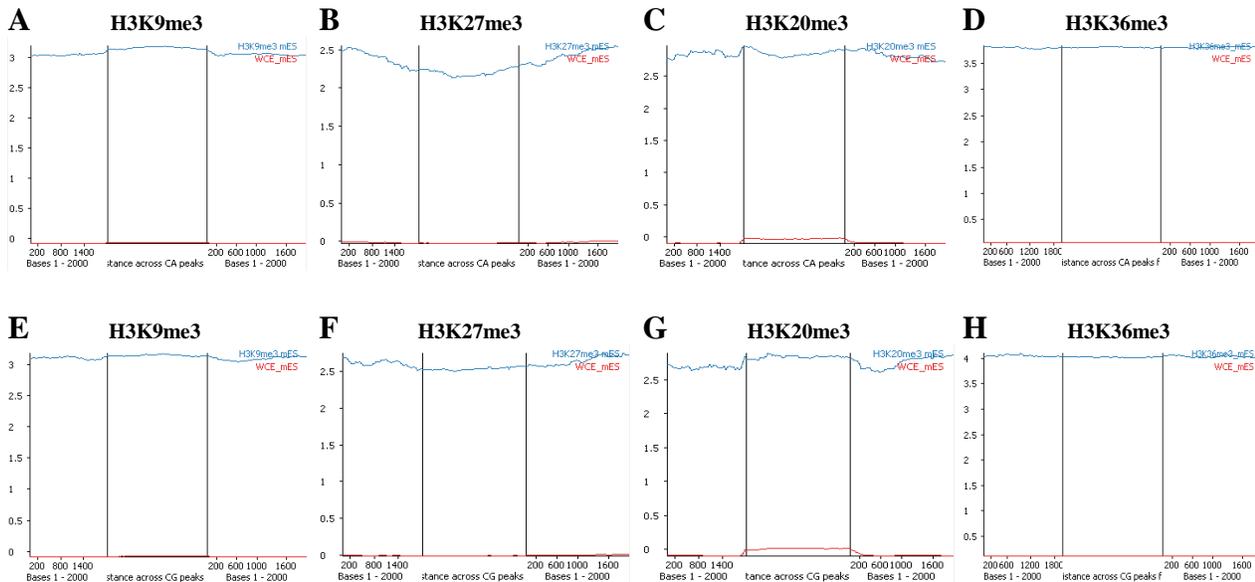


Figure 59. Correlation of mCA (A-D) and mCG (E-H) methylation with repressive histone modifications. 2kb flanking regions are displayed on both sides of the mCA or mCG peak, which is in the middle panel of each graph. **A, E:** H3K9me3; **B, F:** H3K27me3; **C, G:** H3K20me3; **D, H:** H3K36me3. The blue line represents the coverage trend plot of each histone mark, and the red line is the control (WCE) for comparison of background.

To validate this result, I co-stained J1 WT mES cells with mCA and the same panel of histone marks and DNA binders. Unfortunately, the co-immunostaining worked for few of the histone marks, and for none of the proteins. This is because the IF procedure for DNA modifications requires treatment with concentrated HCl which denatures all proteins. For this

reason the co-staining procedure is in fact two independent IF protocols - first for labelling the intact proteins/histone marks and subsequently for labelling the DNA modification. In order to avoid degradation of the antibody-labelled protein from the first round, before starting the protocol with HCl the fixed cells on the slide are cross-linked with 2% PFA a second time. This works in some cases but for many cases it does not give enough protection against HCl and the signal from the labelled proteins is destroyed before the final imaging. In other cases, the over-fixation does not allow the DNA methylation antibody any more access to the chromatin. I managed to successfully co-stain with H3K4me3, H3K9me3 and H3K36me3, and the results are presented in Figure 60.

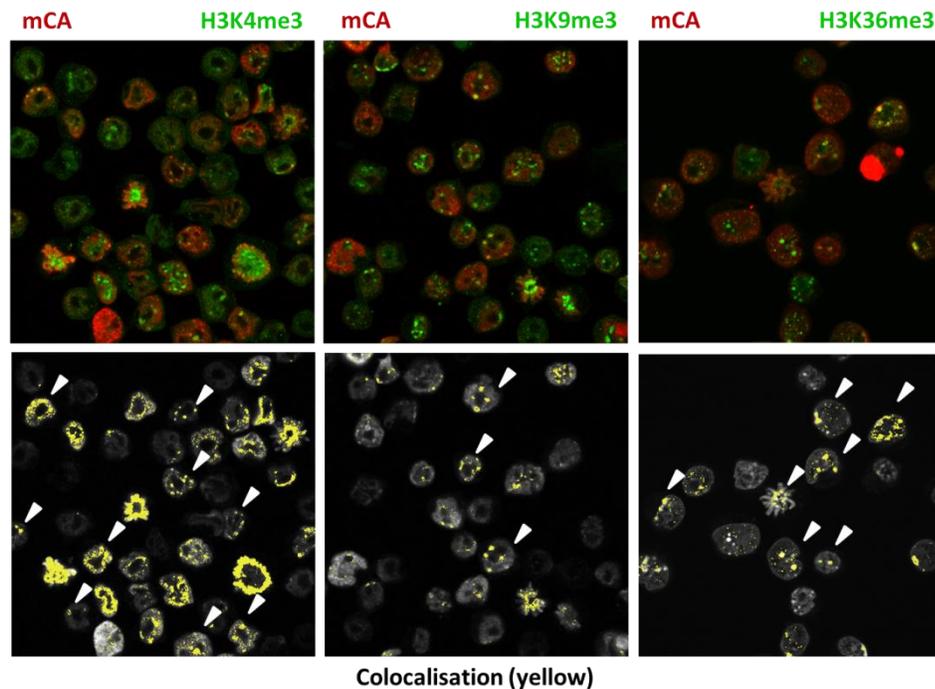


Figure 60. Immunofluorescence of WT J1 mES cells. Red: mCA, Green: histone mark, as labelled (left panels). The colocalisation of both marks was visualised in ImageJ and plotted in yellow over the pattern for mCA (right panels). White arrows indicate heterochromatic sites of colocalisation of mCA and the histone mark. IF by Fátima Santos.

mCA showed much less overlap with the H3K9me3 and H3K36me3 marks than with the H3K4me3 (Figure 60, lower panels). It was interesting to note, that while few of the chromocentres which stained for mCA overlapped with H3K9me3, most of them overlapped with H3K4me3 or H3K36me3, which are also partially localised in the chromocentres, although not as pronounced as H3K9me3. The localisation of H3K4me3 in pericentric foci might explain why the

ChIP data did not show strong overlap between mCA, clearly present in the major satellites, and other pericentric marks like H3K20me3, pointing out that the localisation of those two marks on the pericentric foci could be mutually exclusive.

Because of the association of mCA with active transcription, I plotted Pol II binding sites, and also nascent mRNA transcript coordinates (nuclear run-on, GRO-seq), which correspond to sites where Pol II is engaged in nascent transcription or paused (Min et al. 2011). Pol II showed an interesting enrichment pattern, on the borders on mCA peaks, and no enrichment for mCG (Figure 61A-B). Surprisingly, the GRO-seq nascent RNA transcripts data gave negative correlation with mCA, and slightly negative for mCG (Figure 61C-D).

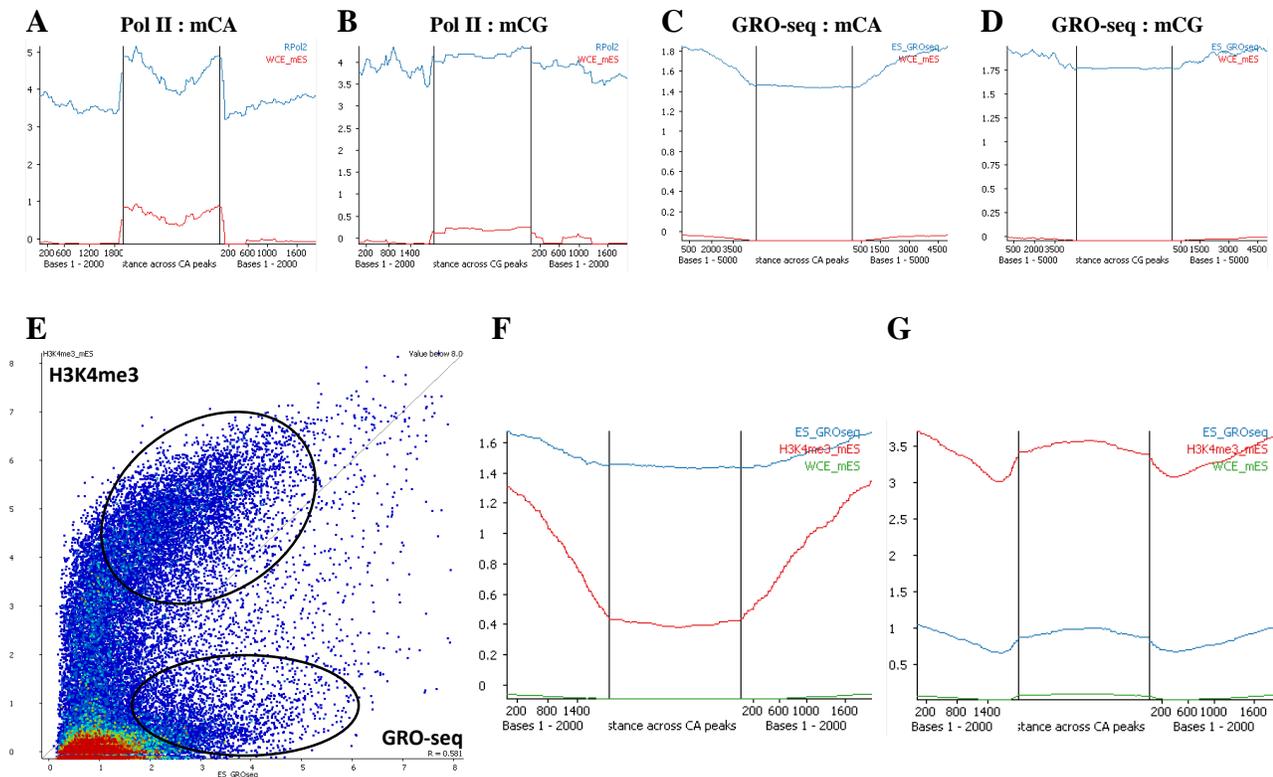


Figure 61. Correlation of mCA and mCG methylation with transcription. **A, B:** Pol II; **C, D:** GRO-seq; The blue line represents the coverage trend plot of each histone mark, and the red line is the control (WCE) for comparison of background. **E.** Scatter plot between H3K4me3 enrichment and peaks of nascent transcription. **F, G.** Enrichment of low-H3K4me3 transcripts (F) and high-H3K4me3 transcripts (G) over mCA peaks. The blue line represents the coverage trend plot of the transcripts, and the red line is the H3K4me3 enrichment, the green is background control (WCE).

The GRO-seq data should correlate quite well with the H3K4me3 mark at the sites of active transcription and I looked further into this. Plotting the H3K4me3 enrichment over the nascent

RNA transcript coordinates, split the group into two types of nascent transcripts – a smaller group with high transcription and no colocalisation with H3K4me3, and a bigger highly transcribed group, which correlated with this mark (Figure 61E). It is possible that those groups represent the paused (no H3K4me3) and actively transcribed (with H3K4me3) PolIII sites, which the GRO-seq dataset is composed of. The small transcription group without H3K4me3 showed negative enrichment for mCA (Figure 61F), while the transcripts marked by H3K4me3 also showed enrichment for mCA (Figure 61G). This interesting observation brought further insight with which types of transcription the mCA mark specifically correlated.

I next plotted the mCA and mCG marks against different protein binding sites. I first looked at four of the major pluripotency related proteins – Nanog, Oct4, Sox2 and Stella (PGC7), the latter also implicated in early development and in PGCs. Interestingly, the three of four pluripotency factors showed enrichment for mCA but not for mCG, while Stella did not show any correlation with either mark (Figure 62A-D). None of the four factors showed any enrichment for the mCG mark (Figure 62E-H).

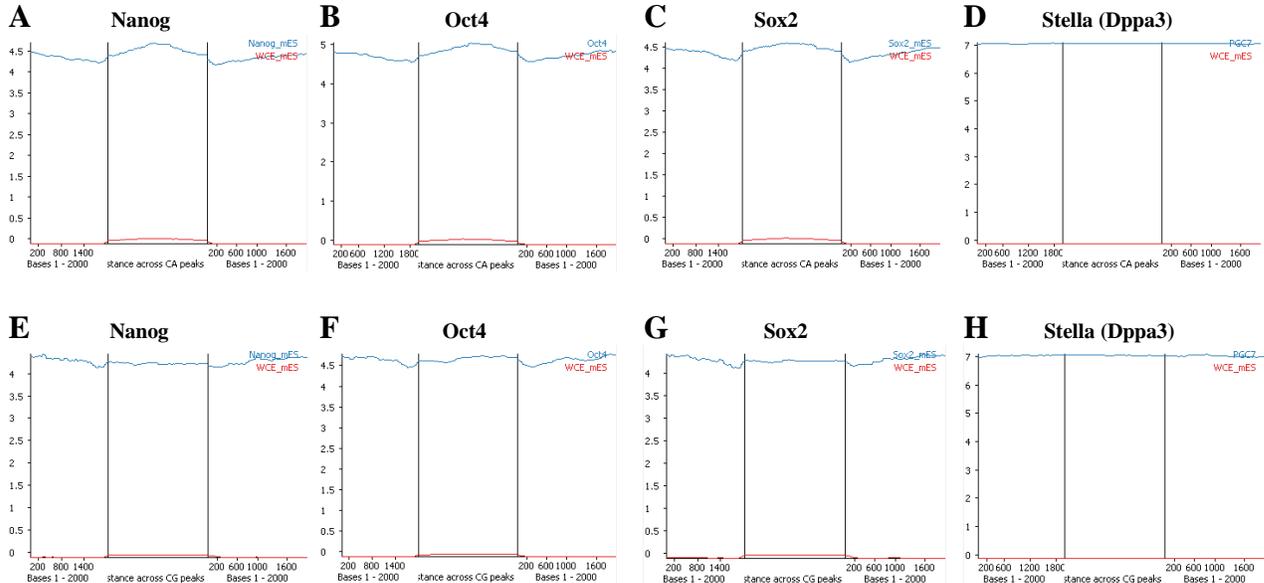


Figure 62. Correlation of mCA (A-D) and mCG (E-H) methylation with pluripotency factor binding sites. 2kb flanking regions are displayed on both sides of the mCA or mCG peak, which is in the middle panel of each graph. **A, E:** Nanog; **B, F:** Oct4; **C, G:** Sox2; **D, H:** Stella. The blue line represents the coverage trend plot of each histone mark, and the red line is the control (WCE) for comparison of background.

It is worth noting that the consensus recognition motifs for the positively enriched Nanog and Sox2 contain a CA site (Nanog: G\_CATT\_C, Sox2: CATTGT (Whyte et al. 2013)), but so does the consensus for Stella (CCYCAGSCTSS (Bian & Yu 2014)), verifying that the observed enrichments cannot be a sequence-based artefact.

For comparison, I next analysed two factors unrelated to pluripotency – Tet3, with a role in the zygote, and a limited number of tissues such as brain, and the pro-differentiation factor Tcf3, which plays a major role in determining tissue-specific cell fate during embryogenesis (Wray et al. 2011). Tet3 did not show any correlation with neither mCG nor mCA (Figure 63A, E), while the Tcf3 showed enrichment on mCA peaks, and a lower enrichment for mCG peaks (Figure 63B, F). While Tet3 is a known CG binder, the binding site of Tcf3 also happens to have a CA motif (E-box motif: 5'-CANNTG-3', Uniprot). This time the control showed slight enrichment meaning that some of the signal might be sequence background, although the real signal is clearly stronger. This result pointed out that mCA correlates not only with pluripotency factors, but also with differentiation factors, and this might be related primarily to their consensus binding motifs in addition to each protein's sensitivity towards methylation.

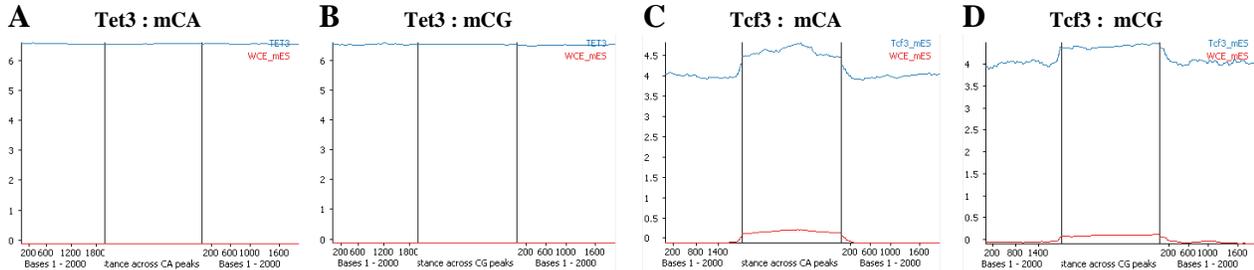


Figure 63. Correlation of mCA and mCG methylation with early development (Tet3) and differentiation factors (Tcf3). The blue line represents the coverage trend plot of each histone mark, and the red line is the control (WCE) for comparison of background.

### 7.3.2 Identifying protein binders for non-CG context methylation

The identification of protein binders for mES cells was done in collaboration with Michiel Vermeulen (Utrecht University). They have recently optimised a quantitative SILAC-based technique for identification of protein binders from nuclear extracts, explained in detail in Appendix Figure 75. The concept is that the nuclear extracts are labelled isotopically prior to their incubation with a DNA bait of a chosen sequence, and compared against a control DNA incubated

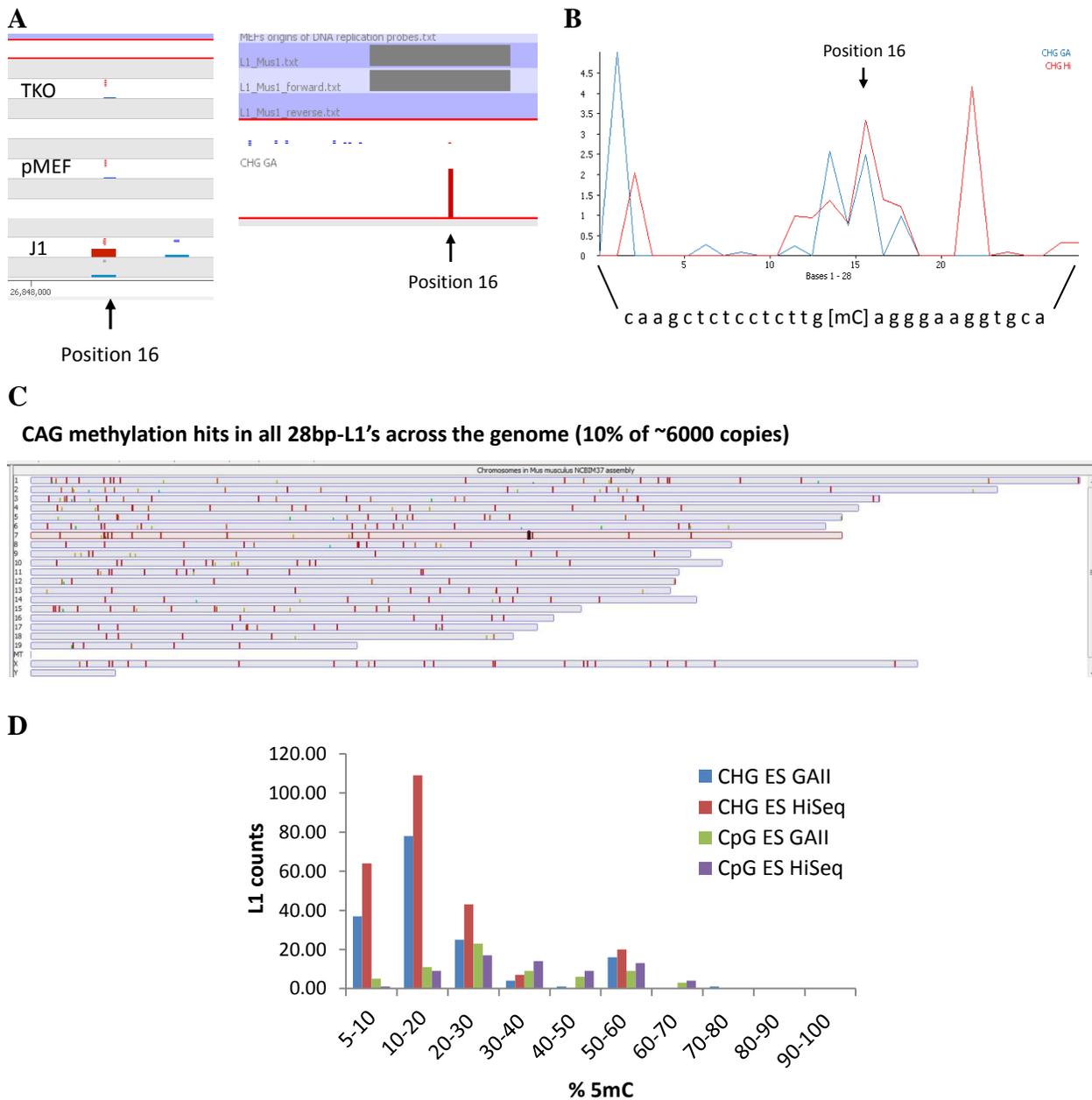


Figure 64. Validation of a second bait for a protein binders experiment. **A**. Identification of L1\_Mus1 in the MspJI meRRBS datasets (left) and validation in high depth WGBS dataset. **B**. Methylation rate of the same CAG position #16 in all L1\_Mus1 motifs within the mouse genome. **C**. A genomic distribution of instances of L1\_Mus1 which show methylation at #16 CAG. **D**. Methylation percentage binning for the L1\_Mus1 CAG position, a comparison with CG is shown, demonstrating a predominant CHG methylation of this region. GAll and HiSeq are two parallel datasets for WT mESC (Stadler et al. 2011).

with unlabelled extracts (forward reaction). This is done twice per bait, and in the second reaction the isotope labelling is swapped between the bait and the control DNA (reverse reaction). This

allows a precise quantitative comparison of the bound proteins to the bait, and also the identification of proteins repelled by the bait (bound to the control) with high statistical confidence.

For statistical validity, our experiment was performed with two DNA baits – one sequence from the major satellite repeat, with a previously validated *in vivo* methylated CHH (CAC) position, used also in the validation of the mCA monoclonal antibody in Chapter 4. The second one was chosen specifically for this study. For this purpose I used my MspJI meRRBS dataset and identified top positions with highest number of methylation calls (>20). Not surprising, one of them was the major satellite, as discussed also in Chapter 5 (Figure 40), and several more candidates, out of which the LINE L1\_Mus1 element was the only one which belonged to a known genomic feature and was ES cell specific (not enriched in pMEFs) (Figure 64A, left). It was also in the CHG (CAG) context, which was desirable provided my other bait was in CHH, especially since CHG is more highly methylated in ES cells (Lister et al. 2009; Ziller et al. 2011). Moreover, I was able to successfully validate the methylation of this position on the same locus in a high depth WGBS from mESCs (Stadler et al. 2011) (Figure 64A, right).

Additionally, I used the BS-seq data to also analyse the genomic occurrence of the L1\_Mus1 element and calculate the genome-wide methylation rate of this repeat, on that same CHG position (Figure 64B). More than 6000 sequences containing this 28 bp consensus were present in the mouse genome (Figure 64C) and more than 10 % were methylated on this CHG position, reaching a maximum of 60 % methylation in a selection of ~20 motifs (Figure 64D).

Next, the protein pull down was performed by our collaborators with the two mCA methylated baits, and their corresponding unmethylated controls (Table 5, rows 1,3, 9 & 11). We were hoping that both baits would show, to some extent, similar results, so we could identify universal mCA binders, alongside the sequence-specific readers, which will have preference for the flanking sequence. At this first round of pull down, we identified a few sequence specific binders, enriched for the methylated control (readers), and a few which were clearly repelled by the methylation (repellers). Interestingly, the proteins differed for the L1 and major satellite probe, demonstrating that most proteins would have flanking sequence preferences, and, most likely, there were no universal mCA readers. For the major satellite probe the definite binders were MeCP2, Foxk1, Foxk2 and an unknown zinc-finger protein Zfp646, and the repellers were three zinc-finger proteins (Zscan4f, Zbtb43 and Zbtb22) (Figure 65, upper panel). The primary function

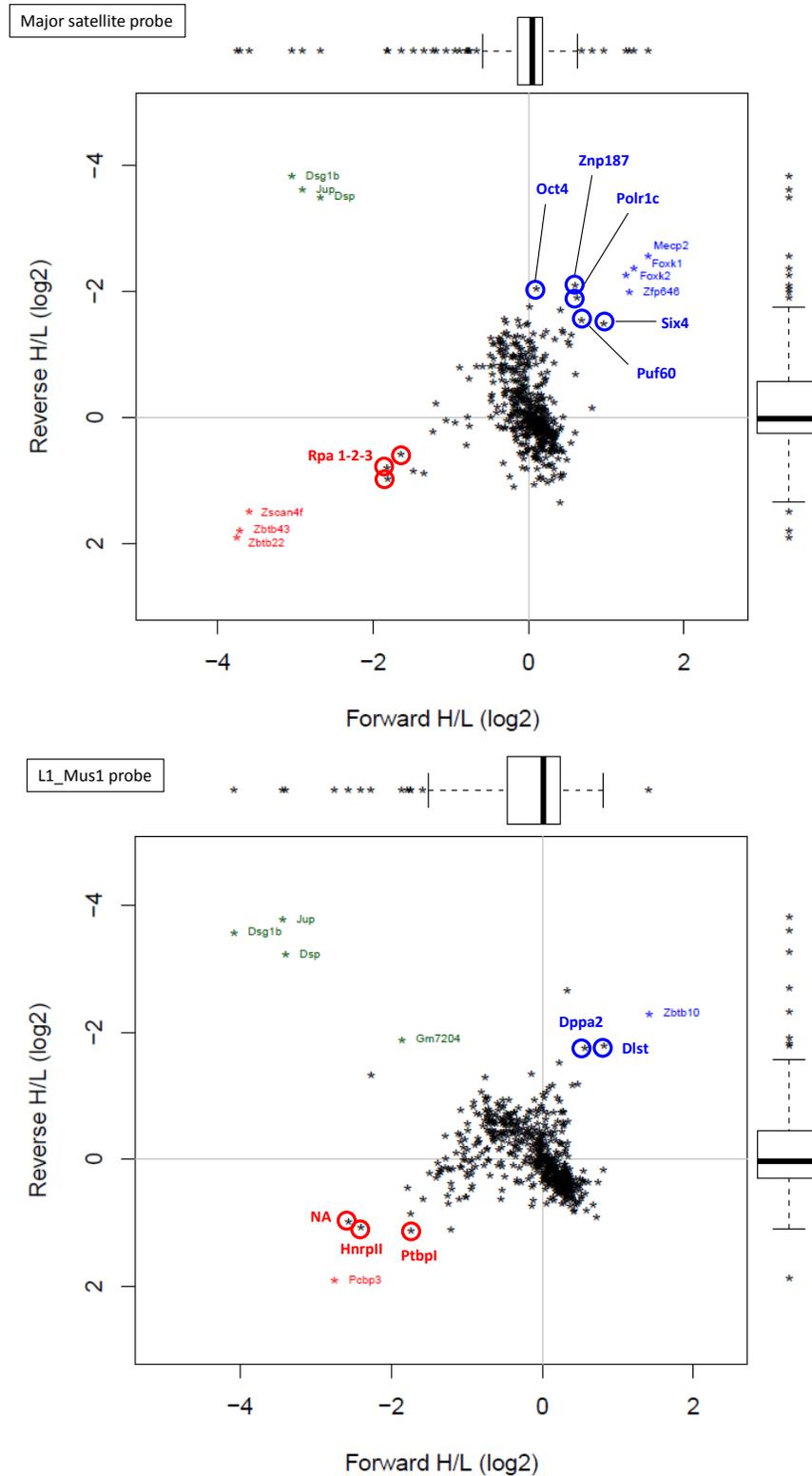


Figure 65. Protein pull-down result for the major satellite (upper panel) and L1\_Mus1 probe (lower panel), in an 'mCA against CA' experiment. The top right corner shows the specific mCA readers, while the bottom left corner shows the proteins repelled by mCA. 95 % confidence intervals are shown by the box plots above and to the right of the main plot.

and information about those proteins are summarised in Appendix, Table 23 and Table 24. With marginal statistical significance among the binders were the pluripotency factor Oct4, an RNA Pol I and II subunit Polr1c, a splicing factor Puf60 (Poly-(U)-binding-splicing factor), and a homeobox protein Six4. This result was interesting and validated my bioinformatic observation for mCH enrichment over Oct4 binding sites, while Sox2 was also among the mCA binders although it did not reach statistical significance. Interestingly, as statistically marginal repellers we found all three subunits of RPA (Replication protein A), the well-described cofactor of AID, which facilitates its deamination function during the somatic hypermutation in B cells (Chaudhuri et al. 2004; Weill & Reynaud 2004). RPA has an important role in UNG2-mediated BER also as a direct recruiter of UNG2 to facilitate the fast removal of uracil from AID-mediated deamination sites (Otterlei et al. 1999; Dianov et al. 1999; Doseth et al. 2012). This finding is very appealing in the light of my hypothesis for the role of mCH in inhibiting active demethylation through BER.

The L1\_Mus1 probe gave much fewer interactors, Zbtb10 as a single reader above the significance threshold (unknown function), and Pcpb3 (Poly-(C)-binding protein 3), as a single repeller (Figure 65, lower panel). With marginal significance were the pluripotency factor Dppa2 (Developmental pluripotency-associated protein 2), and the mitochondrial Dlst (function in oxoglutarate-succinate metabolism). A marginally significant repeller was the splicing factor HnrpII (Appendix, Table 23 and Table 24).

I then did a functional association analysis of the reader and repeller groups of proteins for both probes, including the statistical marginals, and performed clustering on the DAVID web server (<http://david.abcc.ncifcrf.gov/>) (Huang et al. 2007) for the given gene ontology (GO) and association terms. Interestingly, the mCA readers grouped in one big cluster, strongly associated with positive transcriptional regulation and alternative splicing (Figure 66, left panel; raw data in Table 25, Appendix). In addition, there was a small cluster for DNA binding-associated GO terms without any particular known function. The annotations of the repellers however clustered into more groups – the two highly scored groups were for replication and DNA repair, and transcriptional regulation (not specifically positive in this one) (Figure 66, right panel; raw data in Table 26, Appendix). Two smaller groups formed for non-DNA related GO terms in cluster #3 and one group of unclustered terms, mainly associated with DNA and RNA binding.

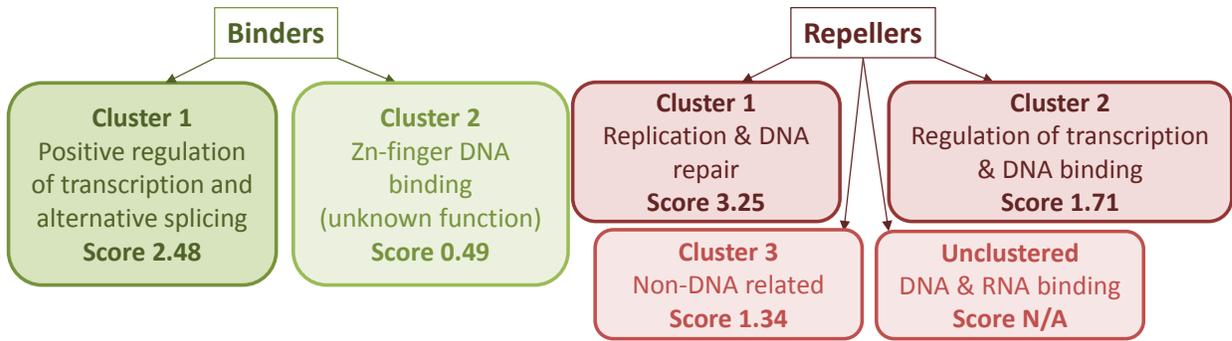
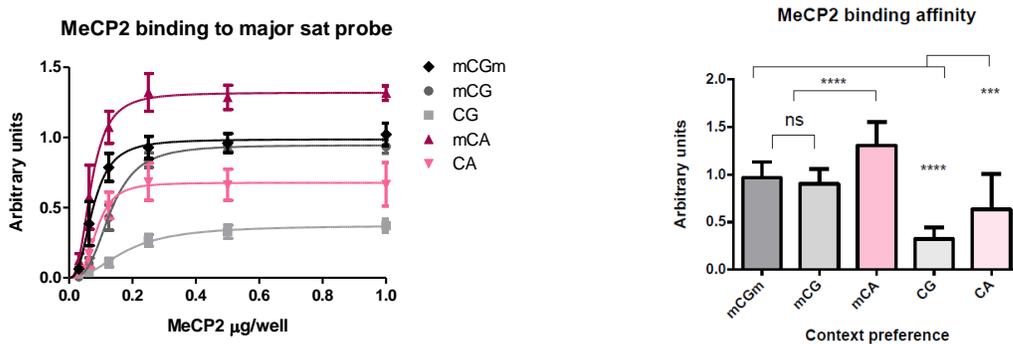


Figure 66. A schematic representation of biological function associations for the mCA binders and repellors. The full lists of GO terms and associations can be found in Appendix Table 25 and Table 26.

It was curious that MePC2 was pulled down as a mCA binder, being also one of the main mCG binders. In order to validate this result I performed an *in vitro* binding assay between MeCP2 (methyl-binding domain only) and the same major satellite probe, in an avidin-biotin ELISA format.

A



B

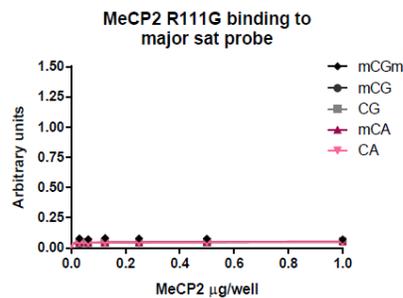


Figure 67. *In vitro* binding assay of the recombinant MeCP2 protein and DNA fragments with varying methylation context. Only the methyl-binding domain (MBD) of MeCP2 was used. **A.** Kinetic curve of the binding to the WT MeCP2 in variable amounts, and a section at 0.5  $\mu\text{g}/\text{well}$  with statistical significance of the different contexts. **B.** No binding with a mutant R111G MeCP2 was detected (Free et al. 2001).

My result confirmed the binding towards mCA, and importantly, this was the preferred substrate in comparison to symmetrically and hemi-methylated mCG substrate (Figure 67A). The background binding to unmethylated CG-containing fragment was low, but the binding towards CA-containing unmethylated oligo was higher. To test the validity and specificity of this interaction an inactive R111G MeCP2 mutant (Free et al. 2001) was used in the assay and found not to show any residual background binding to any of the fragments (Figure 67B). This would suggest that the CA vs CG underlying sequence might be the reason, and not the methylation itself, towards the highest affinity to mCA. For that matter, the crystal structure of MeCP2 bound to DNA shows that MeCP2 introduces bending to the DNA upon its binding, and the possibility that CA context might be slightly curved (Shimooka et al. 2013) could give an explanation for this result and MeCP2 binding preference (Ho et al. 2008).

### ***7.3.3 Effect of global mCH decrease on pluripotency in mESC***

In order to test how the loss of mCH would affect ES cells, we would need to abolish the function of the enzyme(s) responsible for its establishment, according to the classical approach. However, in the case of non-CG methylation this poses an immense challenge because abolishing the activity of Dnmt3a or Dnmt3b, the enzymes responsible for its establishment, results in a strong decrease of CG methylation. However, it is known that Dnmt3a/3b-DKO cells lose their methylation progressively and early passage cells have considerable CG methylation (Chen et al. 2003). Based on my results in Chapter 6, I hypothesised that the CH methylation will decrease faster upon their simultaneous knockout, than the CG methylation, because of the mCG maintenance role of Dnmt1. This means, that there will be a window immediately after knockout where the CG methylation is unchanged but the non-CG is decreased or completely abolished. We therefore obtained Dnm3a and Dnmt3b conditional DKO mES cells, inducible upon tamoxifen treatment. After knockout induction, the cells grew very slowly and were terminated at day nine post knockout, with three time points taken in total – at day 3, day 6 and day 9. The global methylation estimates for CG and CA at each time point revealed that the mCG indeed did not change even until day #9 (Figure 68A). mCA, on the contrary, decreased and reached a stable but not zero level at day #6 after induction of deletion. Provided there is no maintenance of mCA, and the ES cells divide at their usual rates of every 16 hours, by day #3 (72hr) CA methylation should have reached levels less than 10% of its original amount. The cells, however, grew much

slower, and they reached their lowest point at day #6, which is not too surprising given the very slow growth rate. It is disappointing that the decrease did not continue after day #6. Nevertheless, for my purposes at days #6 and #9, these ES cells are already knocked down more than 60 % for mCH.

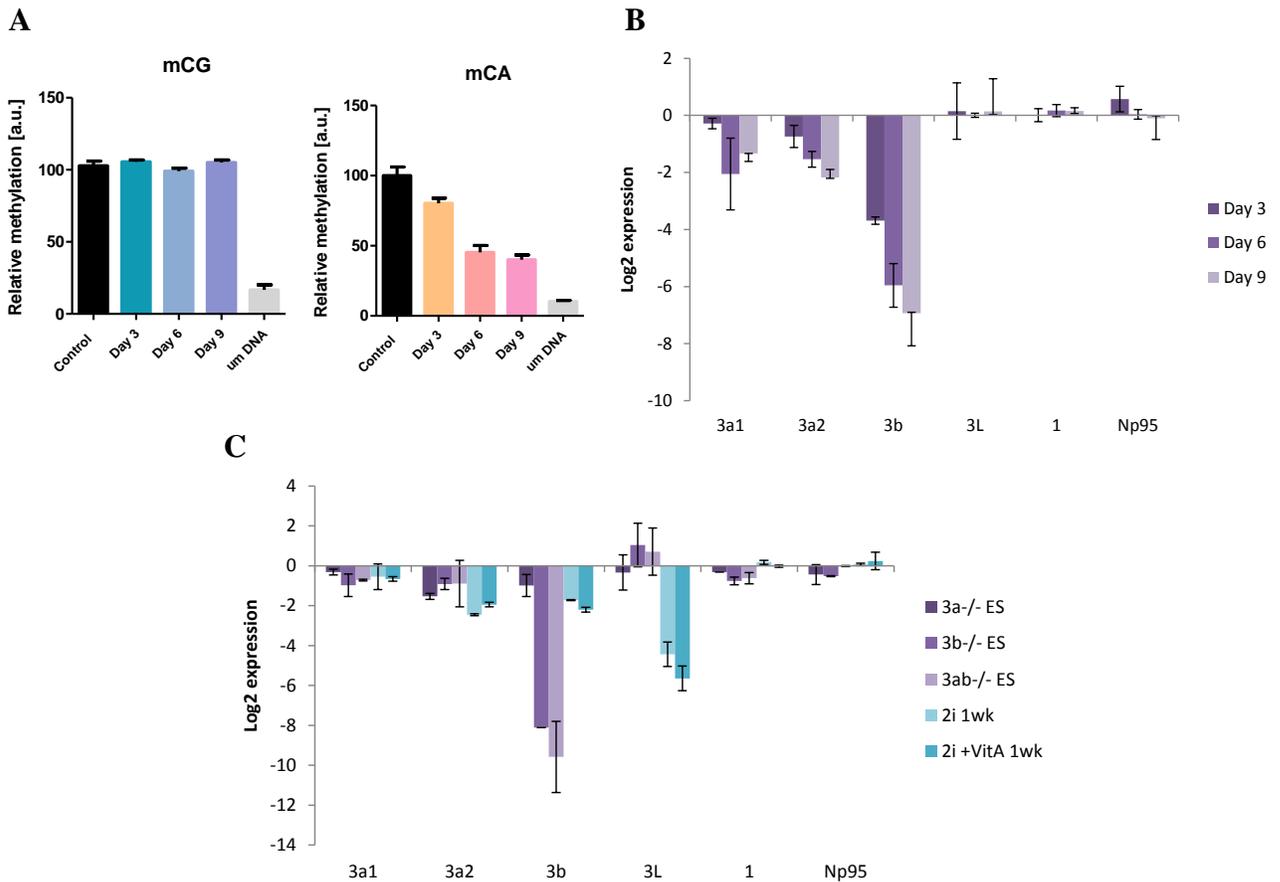


Figure 68. Dnmt3a and Dnmt3b double knockout (DKO) inducible mES cell line – time course post KO-induction. **A.** Global mCG (left panel) and mCA (right panel) levels, compared with a control from non-induced cells grown in parallel. **B.** RT-qPCR of the three time course points for all Dnmt enzymes and Np95 (without Dnmt2). **C.** RT-qPCR for all Dnmt enzymes and Np95 (without Dnmt2) of constitutive Dnmt3a and Dnmt3b single- or double-KO, together with E14 mES cells grown in 2i (with or without VitA).

I next measured the expression of all Dnmts and the other proteins involved in DNA methylation, including Dnmt3L and Np95. As expected, both Dnmt3s were downregulated, although this was much more pronounced for Dnmt3b (Figure 68B). This might mean that there is still mRNA remaining from Dnmt3a, since both transcripts (3a1 and 3a2) did not drop more than 4-fold from WT levels. I therefore compared with the expression levels in our constitutive Dnmt-KOs, which revealed that the Dnmt3a transcripts in general do not drop significantly (Figure

68C), provided we have all evidence that the proteins are not there (results not shown). Interestingly, Dnmt3b levels were slightly higher than in the constitutive knockouts, ranging between  $2^{-(6\div 8)}$  log scale values, rather than  $2^{-(8\div 10)}$ , which might explain why the decrease in mCA levels was not 100% complete. In my comparison I also included ground state ES cells, grown for a week in 2i, which have been shown to downregulate both *de novo* Dnmts upon ground state reprogramming (Ficz et al. 2013; Leitch et al. 2013; Habibi et al. 2013; Hackett et al. 2013) (Figure 68C). Reprogramming from the primed ES cell state to ground (naïve) pluripotency is achieved after swapping culture media from classical serum + LIF ES medium, to a standardised medium containing two inhibitors for the Erk1/2 and Gsk3 $\beta$  signalling pathways (henceforth termed ‘2i’) (Nichols et al. 2009; Wray et al. 2011).

The original 2i medium contains Vitamin A, which generates retinoic acid, one of the known differentiation and growth factors. To assess whether or not this makes a difference, I also included in the comparison also cells grown in medium without Vitamin A, together with cells grown in the original medium (marked as ‘+Vit A’). The levels of decrease of Dnmt3a2 but not Dnmt3b transcripts in my conditional 3ab-KO on days #6 and #9 seemed very similar to the levels of Dnmt3a2 in ground state ‘2i’ ES cells (~2-fold decrease on a log<sub>2</sub> scale). However, Dnmt3b was much more downregulated in the DKO, and the characteristic downregulation of Dnmt3L in the naïve state (Leitch et al. 2013) was not observed in the conditional DKO (neither in the constitutive KOs).

I next wanted to evaluate how a decrease of mCH would affect the pluripotency state of the conditional DKO mES cells. Because of the similarity with ground state ES cells in the downregulation of Dnmt3 expression, I measured the mCA and mCG levels in the naïve ES cells for comparison. It is known that naïve ES cells decrease mCG quite drastically during their transition to ground state, but it has not been clear whether this is the case with their mCH levels.

Since the demethylation of naïve ES cells is due to a downregulation of the *de novo* machinery and not the maintenance (Ficz et al. 2013; Leitch et al. 2013; Habibi et al. 2013; Hackett et al. 2013), which is also confirmed in my experiment (Figure 68C), the expectation was that mCH will also be down, if not completely abolished. However, my ELISA measurements showed, that while mCG was decreasing in the time course of serum to 2i, the mCA levels fluctuated but did not decrease (Figure 69A and B). There was a slightly more pronounced decrease of mCG in the cells grown in the original 2i medium, than the cells without Vit A, but

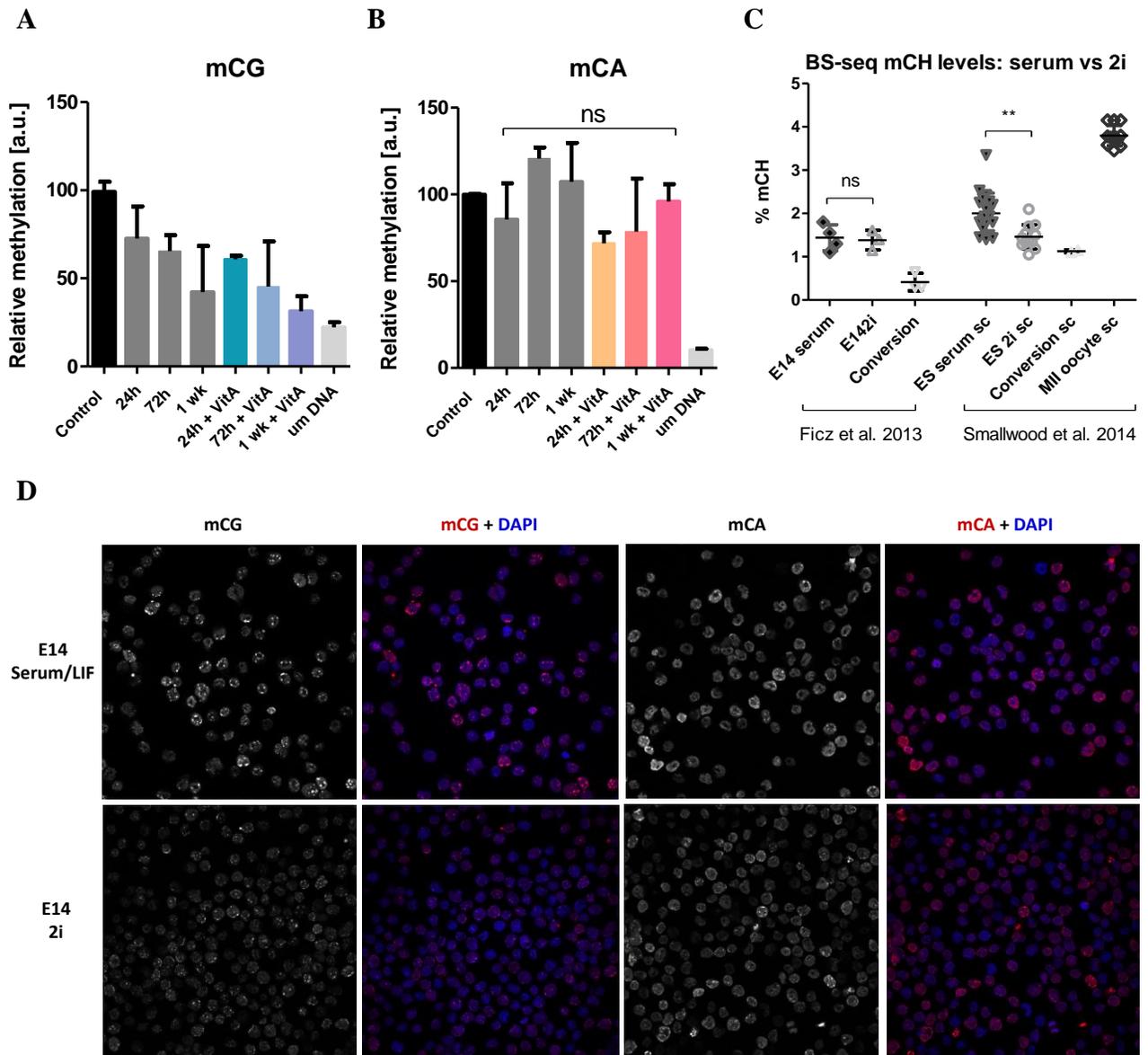


Figure 69. Global mCG (**A**) and mCA (**B**) levels in E14 WT cells' time course towards ground state reprogramming. The fluctuation in mCA was estimated as not statistically significant with 1way ANOVA applying Dunnett's Multiple Comparison Test for each value against the Control value. **C**. mCH levels measured in published WGBS datasets; the conversion controls and the MII oocyte values are given for reference. **D**. IF for mCG (left panel) and mCA (right panel) on E14 WT mESC grown in Serum/LIF and 2i (~1 week). IF by Fátima Santos.

there was no significant change in mCA in either case. I estimated mCH values in two sets of published datasets for WT mES cells in primed (serum) and ground (2i) state and although the datasets from pooled DNA did not show any difference (Ficz et al. 2013), the single cell methylomes (Smallwood et al. 2014) showed slight but statistically significant decrease in 2i

(Figure 69C). This difference might be due to culturing conditions or more likely to the nature of the methylome data – single cell versus pooled ES population, and nevertheless, due to a high inter-cell variability and value overlap, the overall difference in 2i vs serum is not high. I validated those results with IF, and again a decrease in mCA was not observed, on the contrary to mCG (Figure 69D). Notably, the pattern of mCA was as heterogeneous as for primed ES cells, observed also in the single cell methylomes in Figure 69C. The nuclear distribution was quite homogenous as in primed ES cells, again with only a proportion of the heterochromatic foci stained.

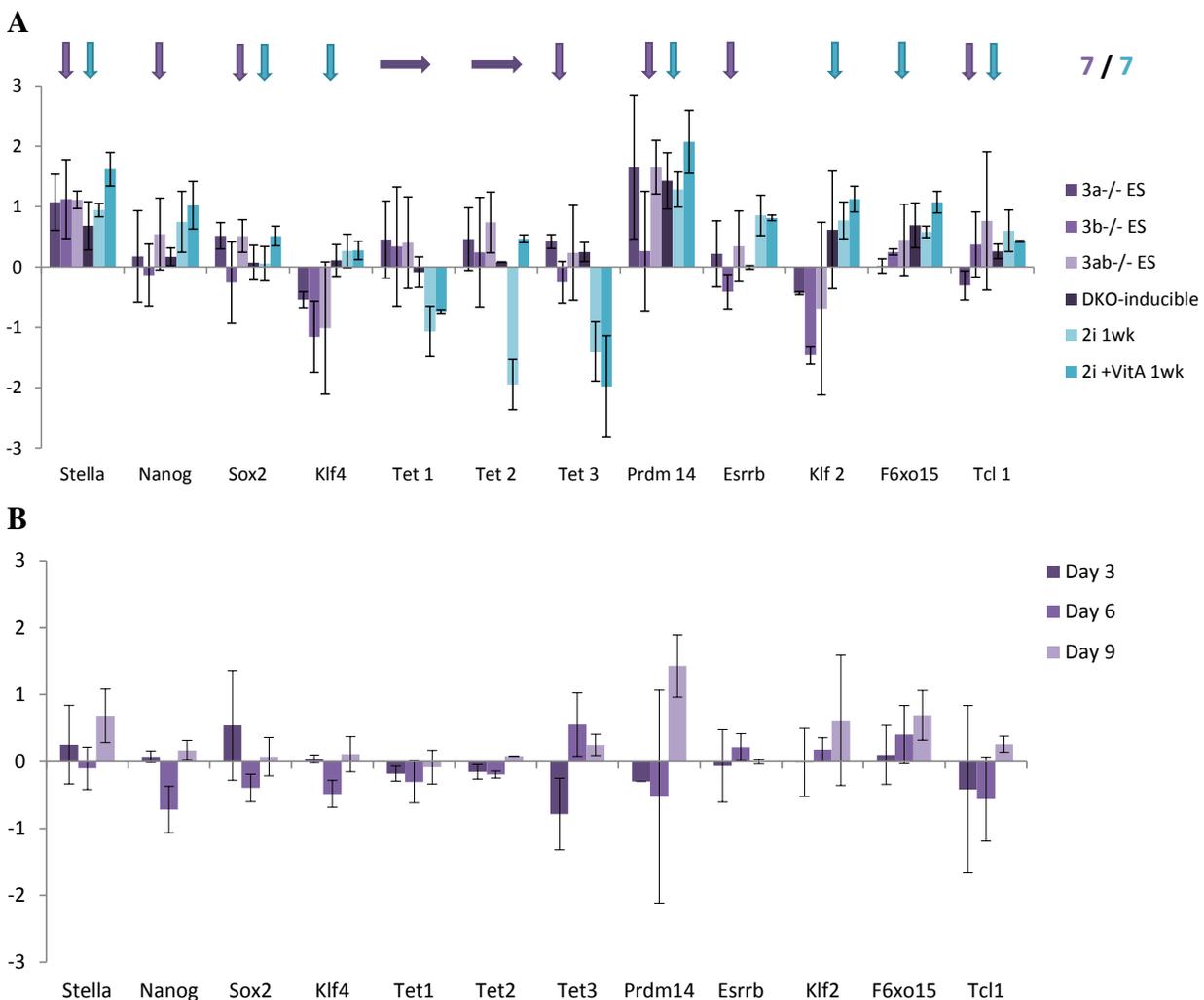


Figure 70. Expression of pluripotency factors and developmentally related genes. **A.** Expression in the panel of *Dnmt3a/b* constitutive KOs, the inducible 3ab-DKO (day #9) and naïve ES cells grown with or without Vitamin A. Arrows point to genes in which the inducible DKO behaves like 2i cells (blue arrows) or like the constitutive *de novo* KOs (purple arrows). **B.** Changes in expression during the three time points after 3ab-DKO induction.

I next analysed the expression of a panel of transcription factors in my conditional DKO time course and all of the constitutive KOs, and compared those against the 2i (1 week) cells. I chose a selection of pluripotency related genes, including a group of Tet1/2-regulated genes (Ficz et al. 2011). All constitutive Dnmt-KOs showed similar patterns of expression within their group, but there were also clear similarities to the 2i ground pluripotency state, mainly in the upregulation of *Stella* and *Prdm14* (Figure 70A). The rest of the pluripotency factors, however, showed some major differences arising from the overall upregulation in 2i but not in the constitutive Dnmt-KOs (like *Nanog*, *Klf4*, *Esrrb*, *Klf2*, *F6xo15*), while the Tet enzymes were generally downregulated in 2i but not in the constitutive KOs.

It was therefore interesting to observe, that the long-term conditional DKO (day #9) showed more similarities to the 2i cells than the other Dnmt-KOs (Figure 70B). It was probably due to the fact that its mCG levels were completely normal (at least equal to primed pluripotent WT ES cells) and therefore gene expression will not be aberrantly affected as in the constitutive KOs. However it was very intriguing to observe that the 60 % overall downregulation of mCH alone is driving this effect. In terms of global cellular methylation levels, this decrease amounts to roughly 15 % overall decrease of methylation; for comparison, the global decrease of methylation in long term 2i cells (>3 weeks) is around 50 % (Ficz et al. 2013). However, the main difference here, in addition to the 3-fold difference in overall levels, is the fact that in 2i this decrease comes from the CG context, while in my conditional DKO it is solely due to non-CG context.

## 7.4 Discussion

In this chapter, I undertook a few different approaches in order to investigate the functional significance of non-CG methylation. Each approach was successful in revealing a different aspect of the functional potential of the mCH mark.

My findings from the correlation with histone marks confirmed the strong connection between mCH methylation and transcription. Moreover, they revealed that there are functional sub-regions within the mCH ‘domain’ and they correlate with different and potentially non-overlapping marks. mCH has been previously associated with gene bodies (Lister et al. 2009; Ficz et al. 2011). However, the strongest correlation according to my MeDIP versus ChIP comparison was not with the typical gene body modification H3K36me3, but with H3K4me3. H3K4me3 is

also an active mark, and it localises to transcription start sites of active genes (Okitsu et al. 2010; Lauberth et al. 2013). Together with H3K27me3 it defines the so called ‘bivalent’ loci characteristic for ES cells only, which define genes as temporarily paused for transcription (Bernstein et al. 2006). However, mCH shows a strong negative correlation with H3K27me3, suggesting that it is not characteristic for the bivalent genes, but only the actively transcribed genes with H3K4me3. In addition, the overlap with the nascent strand enrichment showed that not all H3K4me3 associated genes overlap with CH methylation either, but only the actively transcribed group, and not the paused genes. The IF co-staining profiles revealed that large sections of mCA remain not correlated with H4Kme3. Some of the pericentric staining colocalised with H3K36me3, which is known to localise to pericentric chromatin in addition to gene bodies.

There was no positive correlation observed for mCA with neither of the repressive marks, apart from a slight overlap with H3K9me3 at some of the pericentric loci, as shown by IF (Figure 60). This is interesting, given that in plants non-CG methylation correlates most strongly with H3K9 methylation, and inversely with active marks like histone acetylation, as discussed in 1.5 (Stroud et al. 2014). For these features and its overall correlation with active transcription, non-CG methylation in mammals, or at least in ES cells, seems unrelated to its counterpart in plants.

Positive correlation was observed for major pluripotency factors, as well as proteins involved in differentiation, potentially governed by the consensus recognition motif of each protein. Indeed, our IF co-staining attempts failed to produce meaningful results for a validation of these observations, although the positive link between our protein pull down experiment and the MeDIP/ChIP overlap was very encouraging and strengthens the observations. Also, the two IF co-staining experiments which worked entirely support the MeDIP/ChIP result. However, it is undoubtedly necessary to validate further those correlations experimentally either by more attempts for co-localisation in IF, co-IP experiments or *in vitro* binding assays.

With the help of our collaborators, we further identified experimentally a panel of proteins, which either bind with high affinity or are repelled by mCA. The strongest amongst the binders was the known methyl-cytosine reader MeCP2, which surprised with its high *in vitro* affinity to mCA. A recent study confirmed its binding to non-CG context methylation in neuronal cells, where MeCP2 is very highly expressed and important for differentiation (Guo et al. 2014). It is important to note that MeCP2 is implicated in various pathological conditions, if mutated, which

opens the intriguing possibility that its binding to mCA contributes to any of those conditions. In addition, MeCP2 has been very well investigated and several mutants have been characterised, which affect its DNA binding capacity or its downstream effector functions (Free et al. 2001; Baker et al. 2013). It will be valuable to analyse those mutants for their ability to bind mCA, with the possibility to find a mutant, which binds only to mCA but not mCG or vice versa, and could serve as a very valuable tool for the functional study of non-CG methylation.

The rest of our binders were, like MeCP2, strongly related to tissue differentiation (Foxk1, Foxk2, Six4, Creb1), or pluripotency, like the classical pluripotency marker Oct4 and the pluripotency and cancer associated Dppa2. It is interesting that all binders clustered within one group of active transcriptional regulators, supporting my numerous previous associations of mCA with active transcription. A number of proteins, both among the binders and the repellers (Puf60, HnrpII, Ptbp1), were related to splicing, which is interesting in the light of reports, which connect CH methylation to mRNA splicing (Guo et al. 2013). As a whole, all of these proteins fall into the wider area of transcriptional control, including the PolII and III subunit Polr1c, which implies a role of the mCH mark in transcriptional regulation. Among the repellers, three showed a connection to telomere maintenance and a role in telomere elongation in early development (Zscan4f, Hnrnpd and RPA), which is globally related to the fine-tuning of the maintenance of the pluripotency state and developmental progression. All of these point to a potential role of mCH in the regulation and diversification between pluripotency and differentiation.

It is curious to note, that the major satellite probe yielded a higher number of both readers and repellers, despite its shorter size, which demonstrated that the major satellite is a potentially stronger recognition target for many proteins. This is probably due to its high abundance in the nucleus, and importance for global genomic organisation and stability. It also revealed that DNA binding proteins are in general very sequence specific, even if they are sensitive to DNA methylation.

Among the repellers, the presence of the AID cofactor RPA with all of its three subunits was undoubtedly the most exciting result. This result strengthens the hypothesis of mCH-dependent resistance to demethylation by AID-mediated BER, and sheds another perspective, by showing that AID itself might not be the only protein in the pathway which is affected by high levels of mCH. If mCH prevents DNA binding of more than one of the initiator proteins for BER, then it is very likely that the process will be affected. Moreover, RPA has been shown to assist the BER helicases in the unwinding of a short stretch of nucleotides upon long-patch BER initiation

(Ahn et al. 2009), which means that not just a methylated AID recognition WRC sequence could cause an inhibitory effect, but also if the mCA site is in a close proximity to the BER initiation site. Also, the RPA was repelled by an 18-bp long DNA fragment with an mCA site in the middle, meaning that the effect of one methylation site spreads across a longer stretch of sequence. It is intriguing to know how far this effect spreads on the neighbouring sequence. This finding already supports a previously predicted pathway (in Chapter 6), by which ~5 % of mCH in the female pronucleus could affect BER in a global genome-wide scale, by also affecting the accessibility of neighbouring DNA sequences. However, RPA has not been the only repeller associated with replication and repair in the nuclear pull-down, and the second largest repeller functional cluster was associated with transcriptional regulation, although not specifically in positive way as the result was for the mCA binders.

My experiment with the conditional 3ab-DKO revealed that changes of mCH methylation on a global scale trigger phenotypes similar to phenotypes resulting from the global changes in CG context methylation. This overall loss triggered changes in gene expression, which are very similar to the effects resulting from global mCG loss, and in fact, it is very likely that the equivalent loss of 15 % mCG would have the same effect, if not even less pronounced. It has previously been reported that hypomethylation of ES cells leads to pluripotency ground state (Leitch et al. 2013), but it is novel to reveal that loss of mCH methylation alone can trigger such an effect. This demonstrates, that non-CG context methylation is not a ‘second class’ methylation mark, neither a mere meaningless result of a Dnmt3-mediated sporadic activity. For the ES cell, mCH methylation seems to be as meaningful as CG methylation, and the global loss of one or the other, leads to similar consequences. The notion of lack of function for mCH comes from the fact that direct locus-specific effects cannot be observed in the same way as for mCG, because of its very low sequence-specific enrichment levels (typically 2-10% vs 80% for mCG). This makes it very challenging to study in a cell population context, but for an individual cell each cytosine signal is binary – it is either methylated (100%) or not (0%), and this is irrespective of context. Whether the observed effect will also be true for the other mCH-high cell types, like oocyte and neurons, is not known.

In addition, my observations show that naïve mES cells have a heterogeneous distribution of mCA. It has been reported that ground state mES cells represent a mixed population of pluri- and toti-potent cells (Morgani et al. 2013) and it would be important to find out if the mCA-high cells

fall exclusively within one of those two groups, or belong to both sub-populations.

Lastly, it has been proposed that Prdm14 is the main negative regulator of Dnmt3b, which drives the global demethylation in serum>2i transition (Leitch et al. 2013). My result shows that the opposite could also be true and the depletion of Dnmt3b can lead to upregulation of Prdm14 and trigger a naïve-like state. Certainly, more extensive genome-wide work would have to be carried out in order to compare comprehensively the mCH-knockdown state with the pluripotency ground state, including analysis on imprinted loci. For this purpose, it would be helpful to obtain a higher passage culture of the 3ab-inducible KO, at a point where mCH drops down to zero while mCG remains constant.



## **8 General discussion and future perspectives**

Asymmetric non-CG methylation has long been recognised in the mammalian genome but its distribution, quantity and biological significance have remained very elusive. While a lot of work has been published in recent years, the majority of the focus has addressed the distribution and abundance of non-CG methylation. A lot has also been done in regard of the factors involved in its establishment, with most of the published data indicating Dnmt3a, one of the *de novo* methylases, as the main enzyme responsible for its occurrence. My work attempted to build on the current knowledge, by addressing the occurrence, quantity and its establishment mainly from the perspective of justifying a functional significance.

Non-CG methylation is a very challenging biological phenomenon, veiled in mystery due to technical difficulties in studying it, masked by BS-conversion artefacts, and with a tendency to be ignored because of its very low context numbers, which for the most part look insignificant. Furthermore, the information which has been revealed so far, describes some conflicting characteristics, which have made it difficult to assign one clear function. Its highest distribution in mammals has been confirmed for four cell types: early maturation stage of paternal PGCs (E16.5) (Seisenberger et al. 2012; Kobayashi et al. 2013), maturing (late) and ovulated MII oocyte (Shirane et al. 2013; Tomizawa et al. 2011), embryonic stem cells (ESCs) (Lister et al. 2009; Stadler et al. 2011; Ziller et al. 2011) and the adult mammalian brain (Lister et al. 2013; Guo et al. 2014). Among those, the ES cells have the lowest absolute and relative mCH amounts (Table 10). In addition, while in mESCs mCH has been associated with actively transcribed genes, in the brain it has been associated with transcriptional repression, in both cases as a function of its gene body distribution. In the ESCs and E16.5 male PGCs the mCH is a highly transient mark, while in non-replicating cells like the oocyte and the adult neuron it is a lasting long-term mark (if not permanent). It is not clear how this mark can be of potential equal importance both to the non-replicative and to the highly dividing cells, especially when in the latter it is a very transient phenomenon, observed within a very short developmental window. A possible answer to this could be that there are a few different functional roles of non-CG methylation, a direct consequence of its molecular characteristics, which distinguish it from CG methylation.

Table 10. Summary of main characteristics of non-CG methylation according to cell type.

Feature \ Cell type	ESCs	16.5 PGCs	Oocyte	Brain (Neuron)
<b>Cell cycle positioning</b>	Replicating	Replicating	Pre-replicative	Post-replicative
<b>Potency</b>	Pluripotent	Differentiating	Totipotent	Terminally differentiated
<b>mCG (% CG)</b>	60-70%	50%	40%	>90%
<b>mCH (% of all mC)</b>	25 %	50%	65 %	35 %
<b>mCH mark stability</b>	Transient	Transient	Time limited/long term	Permanent
<b>Methylase</b>	Dnmt3a2 + Dnmt3b	Dnmt3a2 + 3L	Dnmt3a2 + 3L	Dnmt3a
<b>Gene expression</b>	Positive correlation	Not known	Not known	Negative correlation

## 8.1 A role in transcriptional regulation

In agreement with the previously published observations (Lister et al. 2009), my results strongly point to a direct association of non-CG context methylation with active transcription, at least in mESCs.

First, in both the germline and in mESCs, and most likely in the brain, mCH is established by the ‘euchromatic’ isoform of Dnmt3a – the shorter Dnmt3a2. Its nuclear localisation in mESCs looks also very euchromatic, much like the nuclear distribution of Dnmt3a2, spread throughout the whole nucleus with a very small proportion associated with heterochromatic foci (Figure 41). It remains to be established whether the mCH heterogeneity in the mESC population follows the heterogeneity of Dnmt3a2. Moreover, a comparison of genome-wide ChIP-seq data sets against my results links mCH most strongly with H3K4me3, including in the majority of instances where it localises to heterochromatic foci (Figure 60). It is somewhat surprising that it does not show any significant correlation to the four repressive histone marks I looked into. In contrast, there is a positive correlation with the two modifications associated with active enhancers - H3K27ac and H3K4me1. This behaviour is the opposite of what has been shown for non-CG methylation in plants, which is quite surprising, and could be interpreted as both did not have a common evolutionary origin and drivers.

These associations are strongly supported by our protein binders assay, where all of the identified readers and repellors are related to transcriptional regulation, and the readers - to

positive regulation in particular (Figure 66). This is an important point, demonstrating that the CH methylation may not be merely a consequence of a more accessible to the MTases open DNA structure during transcription, but could be a factor in itself affecting the gene activity.

More specifically, our mCA readers and repellers are related to the transcriptional control of either pluripotency or differentiation-related developmental processes, pointing towards a role in the diversification between the two states. These observations are functionally exemplified with our mCH-knockdown system, where a 60 % loss of global mCH triggered changes in gene expression, which resemble elements of a transition towards the ground state of pluripotency. This result is not very intuitive, given that naïve mESCs, reprogrammed in the classical way by switching serum/LIF to 2i conditions, do not lose their CH methylation. Its levels remain almost as high, and as similarly heterogeneous as in primed mESCs. This in itself is very surprising, given that naïve mESCs are known to be more homogeneous in their phenotype than the primed mES cell population. My results show that, while other aspects of the ground state favour a more uniform behaviour and quantitative distribution of pluripotency factors, they do not become homogenous in regard to their mCH methylation levels and distribution. The mCH values from a recent single cell whole-genome bisulphite analysis of ESCs grown in serum and 2i (Smallwood et al. 2014) confirm this observation (Figure 69C). The single cell analysis, however, shows a slight decrease in global levels of mCH methylation in the ground state, which will be beyond the sensitivity of the ELISA and IF employed here, and is not observed in other WGBS reports (Ficz et al. 2013).

These observations together with earlier associations of mCH methylation with transcription would suggest a role in creating epigenetic heterogeneity and diversity in the ES cell population, similar to the heterogeneity of primed state mESCs, but occurring at an earlier developmental window. Whether this initial mCH heterogeneity contributes to a transition into the primed mESC state, later defined by changes in their CG methylation marks, is not known and could be the next question to address. This finding would suggest that an initial state of ‘priming’ could exist also in the naïve pluripotent state, and therefore may be part of the very definition of pluripotency, underlying the intrinsic ability of pluripotent cells to undertake diverse differentiation pathways.

Naïve pluripotency is characterised by protection from *de novo* methylation (Leitch et al. 2013). This definition could be adapted to include a low but important level of *de novo* methylation, maintained by the remaining low Dnmt3 levels, which create initial diversity and set the ground state for subsequent full-scale priming for future developmental events.

## 8.2 Global methylation buffer

While the individual mCH position seems an entirely valid methylation signal for the ES cell, due to the very low level of mCH methylation per single locus it seems that the effect of this methylation on a population level seems minimal, in comparison to the effect CG methylation could have. This suggests that one of the effects of non-CH methylation is in its ‘bulk’ levels, contributing to a globally higher state of methylation. On a genome-wide scale, the saturation capacity of the CG context is very low, since it occupies just 4 below % of the total cytosine pool in the mouse and human genomes. This means that the greatest quantitative effect of CG methylation could not exceed 4 %. In fact, this value would be somewhat less than a theoretical maximum owing to the need to factor in the majority of CGIs which need to remain unmethylated. The rest of the individual or non-promoter CGIs are normally methylated in most cell types, and therefore there is very little capacity for creating a change in global methylation levels, if the cell is restricted only to its CG context. In addition, these specificities of the CG context, together with its evolutionarily controlled genomic distribution, strongly limit the possibilities of creating diversity through the CG context methylation, except during large-scale reprogramming phases. Therefore, the nature and ‘power’ of non-CG methylation could reside in its absolute numbers, and its ubiquitous and unrestricted genomic distribution. As my IF (Figure 41) and meRRBS feature enrichment results show (Figure 40), non-CG methylation is ubiquitous, and apart from CGIs and transcription start sites, it occurs everywhere throughout the genome. As such, this may offer a few benefits to the cell, especially to serve as an epigenetic buffer in instances when DNA methyl-transferase activity may be particularly high, such as in the oocyte and in ESC and during the re-acquisition of DNA methylation in foetal male germ cells. In those situations, CH context can take up a high extent of methylation activity on a global scale, without affecting significantly the expression levels of individual genes.

Keeping global genomic methylation levels high is very important for the cell, with a few notable exceptions. The global decrease of methylation in somatic cells is implicated as an early event in carcinogenesis (Friso et al. 2013), and is a hallmark of all cancer cells, also serving as an accurate predictor of cancer aggression and metastatic capacity (Li et al. 2014). This is undoubtedly linked to their high capacity to proliferate, as a common feature cancer cells have with ESCs, especially in the ES naïve state, is the lower global methylation levels. Elevating non-

CG methylation and increasing overall genomic stability, while keeping CG methylation down and maintaining proliferation capacity, might be another way for the ES cells to keep the genome under check, same applying for E16.5 male PGCs. On the contrary, any levels of non-CG methylation have rarely been associated with cancer, and generally it is known as a non-cancer related phenomenon.

All developmental stages associated with *de novo* methylation maintain very high levels of methyltransferase expression. Their activity must be highly regulated, and while in the ESCs and E16.5 male PGCs those developmental windows are quite short, the timeframe for the maturing oocyte is very long, spread out over many years. To keep CG methylation low and avoid epigenetic ‘errors’ for such a long time, while maintaining the Dnmt3a2 levels very high, could be a real challenge. Therefore, ‘side-tracking’ Dnmts’ activity towards the CH context, which has a weak effect on individual loci and can act as a global methylation ‘sponge’, could be the long-term solution.

In this way, while the methylation in CG context can have a real effect on individual genes or repetitive sequences, it seems that the true asset of non-CG context methylation lies in its numbers and contribution to the body of ‘bulk’ methylation, rather than to the gene-specific or locus-specific methylation.

### 8.3 Resistance to active demethylation

An important and novel characteristic of non-CG context methylation, which has emerged from my work, is the potential to resist the known mechanisms of active demethylation.

My results for mouse Tet1 (Figure 55) and a published *in vitro* study for human Tet2 (Hu et al. 2013) clearly show, that loss of DNA methylation mediated by the Tet family of enzymes via the oxidative pathway does not happen outside of the CG context. In addition to this, I have explored the possibility if CH context is also resistant to the second major pathway implicated in active demethylation, which involves the components of base excision repair (BER) (Hajkova 2010; Santos et al. 2013; Popp et al. 2010). The AID deaminase is involved in the highly controlled somatic hypermutation (SHM), happening only in maturing B-cells, which leads to the immunoglobulin class switch, necessary for the progression towards an adaptive immune response during infection. In addition, it has been found important also for the phase of global methylation erasure in developing PGCs, as well as for the paternal pronucleus in the zygote

(Popp et al. 2010; Santos et al. 2013). Although 5mC is a known substrate of AID (Morgan et al. 2004), it has been reported that AID is 10-fold less active on methylated cytosines, within its recognition sequence WRC (Nabel et al. 2012). In addition, my protein pull down results show that the AID cofactor RPA is also strongly repelled by CH methylation. Given the unusually high level of global methylation on the maternal pronucleus, found both on CG and CH context, it is tempting to speculate that it can function to inhibit the BER activating mechanism, thus conferring intrinsic resistance of the maternal DNA to both major pathways of active demethylation. Such resistance would not exclude the currently accepted mechanism by which the maternal pronucleus escapes active demethylation, mediated by the developmental protein Stella (Nakamura et al. 2007), but will only reinforce it and act in parallel. Stella has recently been shown to interact directly with Tet3 to block its activity (Bian & Yu 2014), meaning it might have more relevance for CG context methylation, but remain irrelevant for mCA, which is the predominant form of methylation on the maternal DNA. In addition, it is possible that those two pathways could cooperate, although no association with either mCG or mCA was observed in my Stella ChIP colocalisation analysis.

In order to validate the model of mCH-conferred resistance to active demethylation, however, more details must be investigated about the long-patch BER-mediated mechanism in the zygote. It is intriguing to address also if RPA is indeed a factor in BER together with AID during the paternal reprogramming in the zygote. According to the most recently proposed mechanism for BER-mediated demethylation in the zygote (Santos et al. 2013), AID does not act on methylated cytosines, as previously presumed, but induces deamination on its original B-cell substrate – the unmethylated cytosines, thus creating uracils. This could potentially trigger long-patch BER, which removes longer stretches of DNA (up to 13bp), including methylated cytosines, rather than a single nucleotide via the short-patch BER. In this recent model, it has been proposed that AID initiates BER on its own, and its activity is coupled with replication, when single stranded DNA substrate becomes accessible on the paternal pronucleus. UNG2, on the other hand, has been pointed as the most likely enzyme, implicated in the uracil removal down the line (Santos et al. 2013). However, on its own AID has been shown to have activity towards single strand DNA substrate only *in vitro* (Pham et al. 2003; Brar et al. 2008). *In vivo* it has been established that it works together with its cofactor RPA (Chaudhuri et al. 2004), which can facilitate AID-mediated deamination on actively transcribing double-stranded DNA (Chaudhuri et al. 2003). Subsequently RPA recruits UNG2 (Otterlei et al. 1999), thus bridging the activity of

AID and UNG2 together in BER. RPA-independent activity of AID would depend on the availability of single strand substrates, which have been suggested to form spontaneously as a result of negative supercoiling (Parsa et al. 2012) or the activity of the RNA exosome (Basu et al. 2011). Whether any of those mechanisms or coupling with DNA replication, as suggested, are functional *in vivo*, however, is not known, but it has been shown that AID can function in an RPA-independent manner by associating with the stalled Pol II transcriptional complex (Yamane et al. 2011). This single *in vivo* example of an independent function of AID, however, does not trigger BER, but it is mutagenic and claimed to contribute to B-cell carcinogenesis. UNG2, however, can be recruited to DNA independently of RPA, to active replication foci through the replication factor PCNA for a post-replicative U:G mismatch repair. It has been shown, nevertheless, that this activity cannot substitute for the RPA-mediated UNG2 BER since the targeting sites by PCNA and RPA do not overlap (Torseth et al. 2012).

As any other replication, repair and recombination related protein, RPA is highly expressed in both human and mouse oocytes (Roig et al. 2004; Carofiglio et al. 2013), and is hence present in zygotes. It is therefore very likely that the oocyte RPA is involved in the proposed pathway, working as a cofactor for AID in the zygote, similarly to their concerted function in SHM. Unlike the situation in SHM in B-cells, there is no active transcription on the paternal pronucleus before its DNA undergoes replication. Therefore the AID and UNG2-mediated BER could indeed happen together with RPA during the replication, as suggested by the current model from Santos et al. However, it is important to also note that the AID-RPA-mediated SHM is an error-prone BER, while the proposed mechanism by Santos et al. regards an error-free BER, which might also imply a different mechanism. Nevertheless, RPA has also been known as a processivity factor for the BER helicases (Ahn et al. 2009), and it is therefore possible that long-patch BER cannot at all proceed without RPA. Moreover, the RPA-AID binding has been very evolutionary conserved, and precedes other regulating mechanisms in the SHM process (Basu et al. 2008), suggesting that the AID-RPA cooperation might be universal and extend beyond their role in SHM. Immunofluorescence for RPA and other BER components in the zygote could help elucidate this relationship, and methylation analysis of fertilised oocytes maternally knocked-out for RPA (if viable) could show if the extent of demethylation is affected as it is in the AID and UNG2 KO zygotes. More importantly, analysis of the AID and UNG2 KO zygotes with the context specific antibodies could elucidate if the amount of CA methylation on the maternal pronucleus remains indeed unchanged after the knockout, and if the increase in methylation on the paternal

pronucleus, observed by Santos et al., is due to mCG only. Similarly, quantitative IF of the Tet3 KO zygotes for mCA and mCG will also help validate the current *in vitro* observations that mCA is resistant to active demethylation by the Tet enzymes.

In addition, the association of mCH methylation with actively transcribed TSS and stalled polymerases as revealed by the GRO-seq analysis (Figure 61) could also interfere with the reported sporadic and mutagenic activity of AID on those exposed ssDNA strands, and therefore implicate an anti-mutagenic role of mCH. Similarly, if the methylated WRC sites are localised near the imprinted DMR regions in the oocyte, this could also ensure their targeted protection from active demethylation in the zygote. All of these possibilities could be investigated bioinformatically.

It is understandable that the biological implications of a mechanism for resistance to active DNA demethylation might be quite substantial. In addition to its importance in the zygote, it might explain the lack of methylation erosion in the oocyte during the long period of its maturation, a phenomenon which has been reported for other non-replicating mCH-high cells such as neurons (Oliveira et al. 2012). It could also explain the ‘necessity’ for high levels of mCH in the neurons themselves, which, unlike oocytes, also have very high levels of mCG and very active Tet-dependent hydroxylation mechanisms. The high Tet levels in neurons could be a factor in the reported neuronal methylation erosion, which makes the mCG a less long-lived or less ‘durable’ mark in non-dividing cells. The mCH on the other hand could be the virtually permanent molecular mark, which is not propagated through cell divisions into the cell population, but is a stable epigenetic memory of individual DNA molecules until the death of the cell, thus representing a ‘molecular epigenetic memory’. This would be especially relevant, if the presence of mCA affects permanently the binding properties of transcription factors or other chromatin regulators. Such property clearly differentiates mCH from mCG, which has the role to set the cellular identity, not allowing it to change through cell divisions, but can be erased and eroded in non-dividing cells with high Tet activity. This positions mCG and mCH in different planes, with virtually opposite properties, that could serve contrary roles in the cell.

In summary, the currently presented results characterising non-CG context methylation, provide insights into its specific properties and unique features, and demonstrate that it could have independent and very different, yet important functions, to those of the classical CG methylation.

## 9 References

- Aapola, U. et al., 2000. Isolation and Initial Characterization of a Novel Zinc Finger Gene , DNMT3L , on 21q22 . 3 , Related to the Cytosine-5- Methyltransferase 3 Gene Family. *Cell*, 298, pp.293–298.
- Abdurashitov, M. a et al., 2009. GluI digestion of mouse gamma-satellite DNA: study of primary structure and ACGT sites methylation. *BMC genomics*, 10, p.322.
- Achwal, C.W. & Chandra, H.S., 1982. A sensitive immunochemical method for detecting 5mC in DNA fragments. *FEBS letters*, 150(2), pp.469–72.
- Ahn, B. et al., 2009. Mechanism of Werner DNA helicase: POT1 and RPA stimulates WRN to unwind beyond gaps in the translocating strand. *PLoS one*, 4(3), p.e4673.
- Alexandrov, B.S. et al., 2012. DNA breathing dynamics distinguish binding from nonbinding consensus sites for transcription factor YY1 in cells. *Nucleic acids research*, 40(20), pp.10116–23.
- Aoki, a et al., 2001. Enzymatic properties of de novo-type mouse DNA (cytosine-5) methyltransferases. *Nucleic acids research*, 29(17), pp.3506–12.
- Arand, J. et al., 2012. In vivo control of CpG and non-CpG DNA methylation by DNA methyltransferases. *PLoS genetics*, 8(6), p.e1002750.
- Araujo, F.D. et al., 1998. Concurrent Replication and Methylation at Mammalian Origins of Replication. *Microbiology*, 18(6), pp.3475–3482.
- Aravin, A.A. et al., 2008. Article A piRNA Pathway Primed by Individual Transposons Is Linked to De Novo DNA Methylation in Mice. *Molecular Cell*, pp.785–799.
- Arita, K. et al., 2012. Recognition of modification status on a histone H3 tail by linked histone reader modules of the epigenetic regulator UHRF1. *Proceedings of the National Academy of Sciences of the United States of America*, 109(32), pp.12950–5.
- Baker, S.A. et al., 2013. An AT-hook domain in MeCP2 determines the clinical course of Rett syndrome and related disorders. *Cell*, 152(5), pp.984–96.
- Ball, M.P. et al., 2009. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nature biotechnology*, 27(4), pp.361–8.
- Barciszewska, A.-M., Nowak, S. & Naskręć-Barciszewska, M.Z., 2014. The degree of global DNA hypomethylation in peripheral blood correlates with that in matched tumor tissues in several neoplasia. *PLoS one*, 9(3), p.e92599.

- Barrès, R. et al., 2009. Non-CpG methylation of the PGC-1alpha promoter through DNMT3B controls mitochondrial density. *Cell metabolism*, 10(3), pp.189–98.
- Barski, A. et al., 2007. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4), pp.823–37.
- Basu, U. et al., 2011. The RNA exosome targets the AID cytidine deaminase to both strands of transcribed duplex DNA substrates. *Cell*, 144(3), pp.353–63.
- Basu, U., Wang, Y. & Alt, F.W., 2008. Evolution of phosphorylation-dependent regulation of activation-induced cytidine deaminase. *Molecular cell*, 32(2), pp.285–91.
- Beard, C., Li, E. & Jaenisch, R., 1995. Loss of methylation activates Xist in somatic but not in embryonic cells. *Genes & Development*, 9(19), pp.2325–2334.
- Bernstein, B.E. et al., 2006. A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell*, pp.315–326.
- Bhutani, N. et al., 2013. A critical role for AID in the initiation of reprogramming to induced pluripotent stem cells. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 27(3), pp.1107–13.
- Bhutani, N. et al., 2010. Reprogramming towards pluripotency requires AID-dependent DNA demethylation. *Nature*, 463(7284), pp.1042–7.
- Bhutani, N., Burns, D.M. & Blau, H.M., 2011. DNA demethylation dynamics. *Cell*, 146(6), pp.866–72.
- Bian, C. & Yu, X., 2014. PGC7 suppresses TET3 for protecting DNA methylation. *Nucleic Acids Research*, 42(5), pp.2893–2905.
- Biniszkiwicz, D. et al., 2002. Dnmt1 Overexpression Causes Genomic Hypermethylation, Loss of Imprinting, and Embryonic Lethality. *Molecular and cellular biology*, 22(7), pp.2124–2135.
- Bird, A.I. & Southern, E.M., 1978. Use of Restriction Enzymes to Study Eukaryotic DNA Methylation. *J. Mol. Biol.*, 118, pp.27 – 47.
- Bock, C. et al., 2010. Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nature biotechnology*, 28(10), pp.1106–14.
- Booth, M.J. et al., 2012. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science (New York, N.Y.)*, 336(6083), pp.934–7.
- Bostick, M. et al., 2007. UHRF1 Plays a Role in Maintaining DNA Methylation in Mammalian Cells. *Science*, 317, pp.1760–1764.

- Bourc'his, D. et al., 2001. Dnmt3L and the establishment of maternal genomic imprints. *Science (New York, N.Y.)*, 294(5551), pp.2536–9.
- Bourc'his, D. & Bestor, T.H., 2004. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature*, 431(September), pp.96–99.
- Brar, S.S. et al., 2008. Activation-induced deaminase, AID, is catalytically active as a monomer on single-stranded DNA. *DNA repair*, 7(1), pp.77–87.
- Brinkman, A.B. et al., 2010. Whole-genome DNA methylation profiling using MethylCap-seq. *Methods (San Diego, Calif.)*, 52(3), pp.232–6.
- Cai, P. et al., 2006. Microcalorimetric studies on the adsorption of DNA by soil colloidal particles. *Colloids and surfaces. B, Biointerfaces*, 49(1), pp.49–54.
- Cam, H.P. et al., 2005. Comprehensive analysis of heterochromatin- and RNAi- mediated epigenetic control of the fission yeast genome. *Nature Genetics*, 37(8), pp.809–819.
- Capuano, F. et al., 2014. Cytosine DNA methylation is found in *Drosophila melanogaster* but absent in *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and other yeast species. *Analytical chemistry*, 86(8), pp.3697–702.
- Carofiglio, F. et al., 2013. SPO11-independent DNA repair foci and their role in meiotic silencing. *PLoS genetics*, 9(6), p.e1003538.
- Cayrou, C. et al., 2011. Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. *Genome research*.
- Cedar, H. et al., 1979. Direct detection of methylated cytosine in DNA by use of the restriction enzyme MspI. *Nucleic Acids Research*, 6(6), pp.2125–2132.
- Cedar, H. & Bergman, Y., 2009. Linking DNA methylation and histone modification : patterns and paradigms. *Nature Reviews Genetics*, 10(May), pp.295–304.
- Chan, S.W. et al., 2004. RNA Silencing Genes Control de Novo DNA Methylation. *Science*, 303, pp.1336–1336.
- Chan, S.W.-L. et al., 2006. RNAi, DRD1, and histone methylation actively target developmentally important non-CG DNA methylation in arabidopsis. *PLoS genetics*, 2(6), p.e83.
- Chaudhuri, J. et al., 2003. Transcription-targeted DNA deamination by the AID antibody diversification enzyme. *Nature*, 422(April), pp.1–5.
- Chaudhuri, J., Khuong, C. & Alt, F.W., 2004. Replication protein A interacts with AID to promote deamination of somatic hypermutation targets. *Nature*, 430(7003), pp.992–8.

- Chen, F., Lieber, M.R. & Hsieh, C., 2002. The DNA methyltransferase-like protein DNMT3L stimulates de novo methylation by Dnmt3a. *PNAS*, 99(26), pp.16916–16921.
- Chen, C.-C., Wang, K.-Y. & Shen, C.-K.J., 2013. DNA 5-methylcytosine demethylation activities of the mammalian DNA methyltransferases. *The Journal of biological chemistry*, 288(13), pp.9084–91.
- Chen, E.S. et al., 2008. Cell cycle control of centromeric repeat transcription and heterochromatin assembly. *Nature*, 451(February), pp.4–7.
- Chen, P.-Y. et al., 2011. A comparative analysis of DNA methylation across human embryonic stem cell lines. *Genome biology*, 12(7), p.R62.
- Chen, R.Z. et al., 1998. DNA hypomethylation leads to elevated mutation rates. *Nature*, 395(September), pp.89–93.
- Chen, T. et al., 2002. A novel Dnmt3a isoform produced from an alternative promoter localizes to euchromatin and its expression correlates with active de novo methylation. *The Journal of biological chemistry*, 277(41), pp.38746–54.
- Chen, T. et al., 2007. Complete inactivation of DNMT1 leads to mitotic catastrophe in human cancer cells. *Nature genetics*, 39(3), pp.391–6.
- Chen, T. et al., 2003. Establishment and Maintenance of Genomic Methylation Patterns in Mouse Embryonic Stem Cells by Dnmt3a and Dnmt3b. *Society*, 23(16), pp.5594–5605.
- Cheng, X. & Blumenthal, R.M., 2008. Mammalian DNA Methyltransferases: A Structural Perspective. *Structure*, 16(3), pp.341–350.
- Chhibber, A. & Schroeder, B.G., 2008. Single-molecule polymerase chain reaction reduces bias: application to DNA methylation analysis by bisulfite sequencing. *Analytical biochemistry*, 377(1), pp.46–54.
- Cingolani, P. et al., 2013. Intronic non-CG DNA hydroxymethylation and alternative mRNA splicing in honey bees. *BMC genomics*, 14, p.666.
- Clark, S.J., Harrison, J. & Frommer, M., 1995. CpNpG methylation in mammalian cells. *Nature Genetics*, 10, pp.20–27.
- Cohen-Karni, D. et al., 2011. The MspJI family of modification-dependent restriction endonucleases for epigenetic studies. *Proceedings of the National Academy of Sciences of the United States of America*, 108(27), pp.11040–5.
- Cokus, S.J. et al., 2008. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, 452(7184), pp.215–9.

- Cooper, D.N. et al., 2010. Methylation-mediated deamination of 5-methylcytosine appears to give rise to mutations causing human inherited disease in CpNpG trinucleotides, as well as in CpG dinucleotides. , 4(6), pp.406–410.
- Coupland, P. et al., 2012. Direct sequencing of small genomes on the Pacific Biosciences RS without library preparation. *BioTechniques*, 53(6), pp.365–72.
- Crowther, P.J. et al., 1989. The effect of E.coli host strain on the consensus sequence of regions of the human LI transposon. *Nucleic Acids Research*, 17(18), pp.7229–7240.
- Dennis, K. et al., 2001. Lsh, a member of the SNF2 family, is required for genome-wide methylation. *Genes & Development*, 15, pp.2940–2944.
- Deplus, R. et al., 2014. Regulation of DNA Methylation Patterns by CK2-Mediated Phosphorylation of Dnmt3a. *Cell Reports*.
- Dhayalan, A. et al., 2010. The Dnmt3a PWWP domain reads histone 3 lysine 36 trimethylation and guides DNA methylation. *The Journal of biological chemistry*, 285(34), pp.26114–20.
- Dianov, G.L. et al., 1999. Replication protein A stimulates proliferating cell nuclear antigen-dependent repair of abasic sites in DNA by human cell extracts. *Biochemistry*, 38(34), pp.11021–5.
- Dodge, J.E. et al., 2002. De novo methylation of MMLV provirus in embryonic stem cells: CpG versus non-CpG methylation. *Gene*, 289(1-2), pp.41–8.
- Dodge, J.E. et al., 2005. Inactivation of Dnmt3b in mouse embryonic fibroblasts results in DNA hypomethylation, chromosomal instability, and spontaneous immortalization. *The Journal of biological chemistry*, 280(18), pp.17986–91.
- Dong, a et al., 2001. Structure of human DNMT2, an enigmatic DNA methyltransferase homolog that displays denaturant-resistant binding to DNA. *Nucleic acids research*, 29(2), pp.439–48.
- Dong, K.B. et al., 2008. DNA methylation in ES cells requires the lysine methyltransferase G9a but not its catalytic activity. *EMBO Journal*, (April), pp.2691–2701.
- Doseth, B. et al., 2012. Strikingly different properties of uracil-DNA glycosylases UNG2 and SMUG1 may explain divergent roles in processing of genomic uracil. *DNA repair*, 11(6), pp.587–93.
- Down, T.A. et al., 2008. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis Thomas. *Nature Biotechnology*, 26(7), pp.779–785.
- Dunican, D.S. et al., 2013. Lsh regulates LTR retrotransposon repression independently of Dnmt3b function. *Genome biology*, 14(12), p.R146.

- Dyachenko, O. V. et al., 2010. Human non-CG methylation: Are human stem cells plant-like? *Epigenetics*, 5(7), pp.569–572.
- Eads, C. a et al., 2000. MethyLight: a high-throughput assay to measure DNA methylation. *Nucleic acids research*, 28(8), p.E32.
- Eden, A., Waghmare, A. & Jaenisch, R., 2003. Chromosomal Instability and Tumors Promoted by DNA. , 300(April), p.2003.
- Ehrlich, M. et al., 1982. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. , 10(81382), pp.2709–2722.
- Engvall, E. & Perlmann, P., 1971. Enzyme-linked immunosorbent assay (ELISA). Quantitative assay of immunoglobulin G. *Immunochemistry*, 8(9), pp.871–4.
- Epsztejn-Litman, S. et al., 2008. De novo DNA methylation promoted by G9a prevents reprogramming of embryonically silenced genes. *Nature structural & molecular biology*, 15(11), pp.1176–83.
- Fatemi, M. et al., 2002. Dnmt3a and Dnmt1 functionally cooperate during de novo methylation of DNA. *European Journal of Biochemistry*, 269(20), pp.4981–4984.
- Fatemi, M. et al., 2005. Footprinting of mammalian promoters: use of a CpG DNA methyltransferase revealing nucleosome positions at a single molecule level. *Nucleic acids research*, 33(20), p.e176.
- Fatemi, M. et al., 2001. The Activity of the Murine DNA Methyltransferase Dnmt1 is Controlled by Interaction of the Catalytic Domain with the N-terminal Part of the Enzyme Leading to an Allosteric Activation of the Enzyme after Binding to Methylated DNA. *Online*.
- Feldman, N. et al., 2006. G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis. *Nature cell biology*, 8(2), pp.188–94.
- Feng, S. et al., 2010. Conservation and divergence of methylation patterning in plants and animals. *Nature*.
- Feng, S., Jacobsen, S.E. & Reik, W., 2011. Epigenetic Reprogramming in Plant. *Science*, 622(2010), pp.622–627.
- Ficz, G. et al., 2011. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature*, 473(7347), pp.398–402.
- Ficz, G. et al., 2013. FGF Signaling Inhibition in ESCs Drives Rapid Genome-wide Demethylation to the Epigenetic Ground State of Pluripotency. *Cell stem cell*, 13, pp.351–359.

- Fisher, O., Siman-tov, R. & Ankri, S., 2004. Characterization of cytosine methylated regions and 5-cytosine DNA methyltransferase ( Ehmeth ) in the protozoan parasite *Entamoeba histolytica*. *Construction*, 32(1).
- Flusberg, B.A. et al., 2010. Direct detection of dnA methylation during single-molecule , real-time sequencing. *Nature methods*, 7(6), pp.461–465.
- Franchina, M. & Kay, P.H., 2000. Evidence that cytosine residues within 5'-CCTGG-3' pentanucleotides can be methylated in human DNA independently of the methylating system that modifies 5'-CG-3' dinucleotides. *DNA and cell biology*, 19(9), pp.521–6.
- Franchini, D.-M., Schmitz, K.-M. & Petersen-Mahrt, S.K., 2012. 5-Methylcytosine DNA demethylation: more than losing a methyl group. *Annual review of genetics*, 46, pp.419–41.
- Free, a et al., 2001. DNA recognition by the methyl-CpG binding domain of MeCP2. *The Journal of biological chemistry*, 276(5), pp.3353–60.
- Freitag, M. & Selker, E.U., 2005. Controlling DNA methylation : many roads to one modification. *Current Opinion in Genetics & Development*, 15, pp.191–199.
- Friso, S. et al., 2013. Global DNA hypomethylation in peripheral blood mononuclear cells as a biomarker of cancer risk. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 22(3), pp.348–55.
- Fukagawa, T. et al., 2004. Dicer is essential for formation of the heterochromatin structure in vertebrate cells. *Group*, 6(8).
- Fuks, F. et al., 2003. The DNA methyltransferases associate with HP1 and the SUV39H1 histone methyltransferase. *Nucleic Acids Research*, 31(9), pp.2305–2312.
- Fuso, A. et al., 2010. Early Demethylation of non-CpG, CpC-rich, elements in the myogenin 5'-flanking region: A priming effect on the spreading of active demethylation. *Cell Cycle*, 9(19), pp.3965–3976.
- Gal-Yam, E.N. et al., 2006. Constitutive nucleosome depletion and ordered factor assembly at the GRP78 promoter revealed by single molecule footprinting. *PLoS genetics*, 2(9), p.e160.
- Gaudet, F. et al., 2003. Induction of tumors in mice by genomic hypomethylation. *Science (New York, N.Y.)*, 300(5618), pp.489–92.
- Genereux, D.P. et al., 2008. Errors in the bisulfite conversion of DNA: modulating inappropriate- and failed-conversion frequencies. *Nucleic acids research*, 36(22), p.e150.
- Goll, M.G. et al., 2006. Methylation of tRNA<sup>Asp</sup> by the DNA Methyltransferase Homolog Dnmt2. *Science*, 311, pp.395–398.

- Goll, M.G. & Bestor, T.H., 2005. Eukaryotic cytosine methyltransferases. *Review Literature And Arts Of The Americas*, (74), pp.481–514.
- Gonzalzo, M.L. & Jones, P. a, 2002. Quantitative methylation analysis using methylation-sensitive single-nucleotide primer extension (Ms-SNuPE). *Methods (San Diego, Calif.)*, 27(2), pp.128–33.
- Gonzalo, S. et al., 2006. DNA methyltransferases control telomere length and telomere recombination in mammalian cells. *Nature cell biology*, 8(4), pp.416–24.
- Gou, D. et al., 2010. SETDB1 is involved in postembryonic DNA methylation and gene silencing in *Drosophila*. *PloS one*, 5(5), p.e10581.
- Gowher, H. & Jeltsch, a, 2001. Enzymatic properties of recombinant Dnmt3a DNA methyltransferase from mouse: the enzyme modifies DNA in a non-processive manner and also methylates non-CpG [correction of non-CpA] sites. *Journal of molecular biology*, 309(5), pp.1201–8.
- Gowher, H., Leismann, O. & Jeltsch, a, 2000. DNA of *Drosophila melanogaster* contains 5-methylcytosine. *The EMBO journal*, 19(24), pp.6918–23.
- Goyal, R., Reinhardt, R. & Jeltsch, A., 2006. Accuracy of DNA methylation pattern preservation by the Dnmt1 methyltransferase. *Nucleic acids research*, 34(4), pp.1182–8.
- Grafstrom, R.H., Yuan, R. & Hamilton, D.L., 1985. The characteristics of DNA methylation in an in vitro DNA synthesizing system from mouse fibroblasts. *Nucleic Acids Research*, 13(8), pp.2827–2842.
- Grandjean, V. et al., 2007. Inheritance of an epigenetic mark: the CpG DNA methyltransferase 1 is required for de novo establishment of a complex pattern of non-CpG methylation. *PloS one*, 2(11), p.e1136.
- Gravina, G.L. et al., 2013. Increased levels of DNA methyltransferases are associated with the tumorigenic capacity of prostate cancer cells. *Oncology reports*, 29(3), pp.1189–95.
- Grunau, C., Clark, S.J. & Rosenthal, a, 2001. Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. *Nucleic acids research*, 29(13), pp.E65–5.
- Van der Gun, B.T.F. et al., 2010. Targeted DNA methylation by a DNA methyltransferase coupled to a triple helix forming oligonucleotide to down-regulate the epithelial cell adhesion molecule. *Bioconjugate chemistry*, 21(7), pp.1239–45.
- Guo, J.U. et al., 2014. Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nature neuroscience*, 17(2), pp.215–22.
- Guo, J.U. et al., 2011. Hydroxylation of 5-Methylcytosine by TET1 Promotes Active DNA Demethylation in the Adult Brain. *Cell*, 145, pp.1–12.

- Guo, W. et al., 2013. Characterizing the strand-specific distribution of non-CpG methylation in human pluripotent cells. *Nucleic acids research*, 42(5), pp.1–8.
- Habibi, E. et al., 2013. Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. *Cell stem cell*, 13(3), pp.360–9.
- Hackett, J. a et al., 2013. Synergistic Mechanisms of DNA Demethylation during Transition to Ground-State Pluripotency. *Stem cell reports*, 1(6), pp.518–31.
- Haines, T.R., Rodenhiser, D.I. & Ainsworth, P.J., 2001. Allele-specific non-CpG methylation of the Nf1 gene during early mouse development. *Developmental biology*, 240(2), pp.585–98.
- Hajkova, P., 2010. Genome-Wide Reprogramming in the Mouse Germ Line Entails the Base Excision Repair Pathway. *Science*, 78.
- Handa, V. & Jeltsch, A., 2005. Profound flanking sequence preference of Dnmt3a and Dnmt3b mammalian DNA methyltransferases shape the human epigenome. *Journal of molecular biology*, 348(5), pp.1103–12.
- Harbers, K., Harbers, B. & Spencer, H., 1975. NUCLEOTIDE CLUSTERS IN DEOXYRIBONUCLEIC ACIDS THE DISTRIEUTION OF 5-METHYLCYTOSINE IN PYRIMIDINE OLIGONUCLEOTIDES OF MOUSE L-CELL SATELLITE DNA AND MAIN BAND DNA. *Biochemical and Biophysical Research Communications*, 66(2), pp.738–746.
- Harrison, J., Stirzaker, C. & Clark, S.J., 1998. Cytosines adjacent to methylated CpG sites can be partially resistant to conversion in genomic bisulfite sequencing leading to methylation artifacts. *Analytical biochemistry*, 264(1), pp.129–32.
- Hashimoto, H. et al., 2014. Structure of a Naegleria Tet-like dioxygenase in complex with 5-methylcytosine DNA. *Nature*, 506(7488), pp.391–5.
- Hata, K. et al., 2002. Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. *Development (Cambridge, England)*, 129(8), pp.1983–93.
- Hata, K. et al., 2006. Meiotic and Epigenetic Aberrations in Dnmt3L-Deficient Male Germ Cells. *Molecular Reproduction and Development*, 122(April 2005), pp.116–122.
- Hawkins, R.D. et al., 2010. Distinct Epigenomic Landscapes of Pluripotent and Lineage-Committed Human Cells. *Cell Stem Cell*, 6(5), pp.479–491.
- Hayashi, K. et al., 2008. Dynamic equilibrium and heterogeneity of mouse pluripotent stem cells with distinct functional and epigenetic states. *Cell stem cell*, 3(4), pp.391–401.

- Hayatsu, H., Negishi, K. & Shiraishi, M., 2004. DNA methylation analysis: speedup of bisulfite-mediated deamination of cytosine in the genomic sequencing procedure. *Proceedings of the Japan Academy, Series B*, 80(4), pp.189–194.
- He, Y.-F. et al., 2011. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science (New York, N.Y.)*, 333(6047), pp.1303–7.
- Hemberger, M., Dean, W. & Reik, W., 2009. Epigenetic dynamics of stem cells and cell lineage commitment : digging Waddington ' s canal. *Nature Publishing Group*, 10(8), pp.526–537.
- Henderson, I.R. et al., 2010. Accurate sodium bisulfite sequencing in plants. *Epigenetics : official journal of the DNA Methylation Society*, 5(1), pp.47–9.
- Henderson, I.R. & Jacobsen, S.E., 2008. Tandem repeats upstream of the Arabidopsis endogene SDC recruit non-CG DNA methylation and initiate siRNA spreading. *Genes and Development*, 22(12), pp.1597–1606.
- Hengesbach, M. et al., 2008. Use of DNazymes for site-specific analysis of ribonucleotide modifications Use of DNazymes for site-specific analysis of ribonucleotide modifications. *Most*, pp.180–187.
- Herman, J.G. et al., 1996. Methylation-specific PCR : A novel PCR assay for methylation status of CpG islands. *Reactions*, 93(September), pp.9821–9826.
- Hermann, A., Gowher, H. & Jeltsch, A., 2004. Biochemistry and biology of mammalian DNA methyltransferases. *Cellular and molecular life sciences : CMLS*, 61(19-20), pp.2571–87.
- Ho, K.L. et al., 2008. MeCP2 binding to DNA depends upon hydration at methyl-CpG. *Molecular cell*, 29(4), pp.525–31.
- Holliday, R. & Pugh, J.E., 1975. DNA modification mechanisms and gene activity during development. *Science (New York, N.Y.)*, 187(4173), pp.226–32.
- Hu, L. et al., 2013. Crystal structure of TET2-DNA complex: insight into TET-mediated 5mC oxidation. *Cell*, 155(7), pp.1545–55.
- Huang, D.W. et al., 2007. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic acids research*, 35(Web Server issue), pp.W169–75.
- Huang, J. et al., 2004. Lsh , an epigenetic guardian of repetitive elements. *Pharmacia*, 32(17).
- Huang, Y. et al., 2012. The anti-CMS technique for genome-wide mapping of 5-hydroxymethylcytosine. *Nature protocols*, 7(10), pp.1897–908.
- Huang, Y. et al., 2010. The Behaviour of 5-Hydroxymethylcytosine in Bisulfite Sequencing. *America*, 5(1), pp.1–9.

- Illingworth, R.S. et al., 2010. Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS genetics*, 6(9), p.e1001134.
- Imamura, T. et al., 2005. Dynamic CpG and non-CpG methylation of the Peg1/Mest gene in the mouse oocyte and preimplantation embryo. *The Journal of biological chemistry*, 280(20), pp.20171–5.
- Imamura, T. et al., 2004. Non-coding RNA directed DNA demethylation of Sphk1 CpG island. *Biochemical and biophysical research communications*, 322(2), pp.593–600.
- Ioannou, A.K. et al., 2010. Analytica Chimica Acta Electroanalytical study of SYBR Green I and ethidium bromide intercalation in methylated and unmethylated amplicons. *Analytica Chimica Acta*, 657, pp.163–168.
- Ito, S. et al., 2010. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature*.
- Ito, S. et al., 2011. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science (New York, N.Y.)*, 333(6047), pp.1300–3.
- Jackson, J.P. et al., 2002. Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase. *Nature*, 416(6880), pp.556–60.
- Jackson-Grusby, L. et al., 2001. Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation. *Nature genetics*, 27(1), pp.31–9.
- Jeltsch, A. & Jurkowska, R.Z., 2014. New concepts in DNA methylation. *Trends in biochemical sciences*, 39(7), pp.310–318.
- Jia, D. et al., 2007. Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. *Nature*, 449(7159), pp.248–51.
- Johannsen, M.W. et al., 2014. Triplex-mediated analysis of cytosine methylation at CpA sites in DNA. *Chemical communications (Cambridge, England)*, 50(5), pp.551–3.
- Jurkowska, R. et al., 2011. Oligomerization and binding of the Dnmt3a DNA methyltransferase to parallel DNA molecules: heterochromatic localization and role of Dnmt3L. *The Journal of biological chemistry*, 286(27), pp.24200–7.
- Jurkowska, R.Z., Siddique, A.N., et al., 2011. Approaches to enzyme and substrate design of the murine Dnmt3a DNA methyltransferase. *Chembiochem : a European journal of chemical biology*, 12(10), pp.1589–94.
- Jurkowska, R.Z., Jurkowski, T.P. & Jeltsch, A., 2011. Structure and function of mammalian DNA methyltransferases. *Chembiochem : a European journal of chemical biology*, 12(2), pp.206–22.

- Jurkowski, T.P. et al., 2008. Human DNMT2 methylates tRNA(Asp) molecules using a DNA methyltransferase-like catalytic mechanism. *RNA (New York, N.Y.)*, 14(8), pp.1663–70.
- Jurkowski, T.P. & Jeltsch, A., 2011. On the evolutionary origin of eukaryotic DNA methyltransferases and Dnmt2. *PloS one*, 6(11), p.e28104.
- Kagiwada, S. et al., 2013. Replication-coupled passive DNA demethylation for the erasure of genome imprints in mice. *The EMBO journal*, 32(3), pp.340–53.
- Kaneda, M. et al., 2004. Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. *Nature*, 429(6994), pp.900–3.
- Kanellopoulou, C. et al., 2005. Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. *Genes & Development*, pp.489–501.
- Kareta, M.S. et al., 2006. Reconstitution and mechanism of the stimulation of de novo methylation by human DNMT3L. *The Journal of biological chemistry*, 281(36), pp.25893–902.
- Karimi, M. et al., 2006. LUMA (Luminometric Methylation Assay)--a high throughput method to the analysis of genomic DNA methylation. *Experimental cell research*, 312(11), pp.1989–95.
- Karimi, M., Luttrupp, K. & Ekström, T.J., 2011. Global DNA Methylation Analysis Using the Luminometric Methylation Assay T. O. Tollefsbol, ed. *Methods in Molecular Biology*, 791, pp.135–144.
- Katoh, M. et al., 2006. Developmentally Regulated DNA Methylation in Dictyostelium discoideum †. *Society*, 5(1), pp.18–25.
- Kawai, J. et al., 1993. Methylation profiles of genomic DNA of mouse developmental brain detected by restriction landmark genomic scanning (RLGS) method. *Nucleic acids research*, 21(24), pp.5604–8.
- Khan, S.A., Humayun, M.Z. & Jacob, T.M., 1977. Antibodies specific to a deoxyribodinucleotide sequence. *Nucleic Acids Research*, 4(9), pp.2997–3006.
- Kim, D.H. et al., 2006. Argonaute-1 directs siRNA-mediated transcriptional gene silencing in human cells. *Nature Structural & Molecular Biology*, 13(9), pp.793–797.
- Kim, G.-D. et al., 2002. Co-operation and communication between the human maintenance and de novo DNA (cytosine-5) methyltransferases. *The EMBO journal*, 21(15), pp.4183–95.
- Kobayashi, H. et al., 2012. Contribution of intragenic DNA methylation in mouse gametic DNA methylomes to establish oocyte-specific heritable marks. *PLoS genetics*, 8(1), p.e1002440.

- Kobayashi, H. et al., 2013. High-resolution DNA methylome analysis of primordial germ cells identifies gender-specific reprogramming in mice. *Genome research*, 23(4), pp.616–27.
- Kohli, R.M. & Zhang, Y., 2013. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature*, 502(7472), pp.472–9.
- Kouidou, S. et al., 2005. Non-CpG cytosine methylation of p53 exon 5 in non-small cell lung carcinoma. *Lung cancer (Amsterdam, Netherlands)*, 50(3), pp.299–307.
- Kouidou, S., Malousi, A. & Maglaveras, N., 2006. Methylation and repeats in silent and nonsense mutations of p53. *Mutation research*, 599(1-2), pp.167–77.
- Kriaucionis, S. & Heintz, N., 2010. The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain. *Science*, 929(2009), pp.929–930.
- Krueger, F. et al., 2012. DNA methylome analysis using short bisulfite sequencing data. *Nature methods*, 9(2), pp.145–151.
- Krueger, F. & Andrews, S.R., 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics (Oxford, England)*, 27(11), pp.1571–2.
- Kuhlmann, M. et al., 2005. Silencing of retrotransposons in Dictyostelium by DNA methylation and RNAi. *Homo*, 33(19), pp.6405–6417.
- Kumaki, Y., Oda, M. & Okano, M., 2008. QUMA : quantification tool for methylation analysis. *Nucleic Acids Research*, 36(Web Server issue), pp.170–175.
- Kumar, R. et al., 2013. AID stabilizes stem-cell phenotype by removing epigenetic memory of pluripotency genes. *Nature*, 500(7460), pp.89–92.
- Kunert, N. et al., 2003. A Dnmt2-like protein mediates DNA methylation in Drosophila. *Development*, pp.5083–5090.
- Kuramochi-Miyagawa, S. et al., 2008. DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. *Genes & Development*, 22, pp.908–917.
- Laird, P.W., 2010. Principles and challenges of genomewide DNA methylation analysis. *Nature reviews. Genetics*, 11(3), pp.191–203.
- Langmead, B. et al., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3), p.R25.
- Lauberth, S.M. et al., 2013. H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation. *Cell*, 152(5), pp.1021–36.

- Laurent, L. et al., 2010. Dynamic changes in the human methylome during differentiation. *Genome Research*, pp.320–331.
- Law, J. a & Jacobsen, S.E., 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature reviews. Genetics*, 11(3), pp.204–220.
- Lee, H.J., Hore, T. a & Reik, W., 2014. Reprogramming the Methylome: Erasing Memory and Creating Diversity. *Cell stem cell*, 14(6), pp.710–719.
- Lee, J., Jin, S., et al., 2010. Genomics Presence of 5-methylcytosine in CpNpG trinucleotides in the human genome. *Genomics*, 96(2), pp.67–72.
- Lee, J., Jang, S.J., et al., 2010. Presence of 5-methylcytosine in CpNpG trinucleotides in the human genome. *Genomics*.
- Lehnertz, B. et al., 2003. Suv39h-Mediated Histone H3 Lysine 9 Methylation Directs DNA Methylation to Major Satellite Repeats at Pericentric Heterochromatin. *Current Biology*, 13, pp.1192–1200.
- Lei, H. et al., 1996. De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells. *Analysis*, 3205, pp.3195–3205.
- Leitch, H.G. et al., 2013. Naive pluripotency is associated with global DNA hypomethylation. *Nature structural & molecular biology*, 20(3), pp.311–6.
- Leonhardt, H. & Page, A.W., 1992. A Targeting Sequence Directs DNA Methyltransferase to Sites of DNA Replication in Mammalian. *Cell*, 71, pp.865–873.
- Lequin, R.M., 2005. Enzyme immunoassay (EIA)/enzyme-linked immunosorbent assay (ELISA). *Clinical chemistry*, 51(12), pp.2415–8.
- Li, E., Beard, C. & Jaenisch, R., 1993. Role of DNA methylation in genomic imprinting. *Nature*, 366, pp.362–365.
- Li, E., Bestor, T.H. & Jaenisch, R., 1992. Targeted Mutation of the DNA Methyltransferase Gene Results in Embryonic Lethality. *Cell*, 69, pp.915–926.
- Li, H. et al., 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(28), pp.19489 –19500.
- Li, J. et al., 2014. The prognostic value of global DNA hypomethylation in cancer: a meta-analysis. *PloS one*, 9(9), p.e106290.
- Li, J.-Y. et al., 2007. Synergistic function of DNA methyltransferases Dnmt3a and Dnmt3b in the methylation of Oct4 and Nanog. *Molecular and cellular biology*, 27(24), pp.8748–59.

- Li, Y. et al., 2010. The DNA Methylome of Human Peripheral Blood Mononuclear Cells. *PLoS Biology*, 8(11).
- Liang, G. et al., 2002. Cooperativity between DNA Methyltransferases in the Maintenance Methylation of Repetitive Elements. , 22(2), pp.480–491.
- Lin, I.G. et al., 2002. Murine De Novo Methyltransferase Dnmt3a Demonstrates Strand Asymmetry and Site Preference in the Methylation of DNA In Vitro Murine De Novo Methyltransferase Dnmt3a Demonstrates Strand Asymmetry and Site Preference in the Methylation of DNA In Vitro. *Molecular and cellular biology*, 22(3), pp.704–723.
- Lippman, Z. et al., 2004. Role of transposable elements in heterochromatin and epigenetic control. *Nature*, 430(22 JULY 2004), pp.471–476.
- Lister, R. et al., 2013. Global epigenomic reconfiguration during mammalian brain development. *Science (New York, N.Y.)*, 341(6146), p.1237905.
- Lister, R. et al., 2008a. Highly integrated single-base resolution maps of the epigenome in Arabidopsis - supplement. *Cell*, 133(3), pp.523–36.
- Lister, R. et al., 2008b. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, 133(3), pp.523–36.
- Lister, R. et al., 2011. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature*, 470(7336), pp.68–73.
- Lister, R. et al., 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271), pp.315–22.
- Liu, K. et al., 2003. Endogenous Assays of DNA Methyltransferases : Evidence for Differential Activities of DNMT1 , DNMT2 , and DNMT3 in Mammalian Cells In Vivo. *Society*, 23(8), pp.2709–2719.
- Liutkeviciute, Z. et al., 2009. Cytosine-5-methyltransferases add aldehydes to DNA. *Nature Chemical Biology*, 5(6), pp.400–402.
- Lorincz, M.C. et al., 2002. DNA Methylation Density Influences the Stability of an Epigenetic Imprint and Dnmt3a / b-Independent De Novo Methylation. *Society*, 22(21), pp.7572–7580.
- Lucifero, D. et al., 2004. Gene-specific timing and epigenetic memory in oocyte imprinting. *Access*, 13(8), pp.839–849.
- Lyko, F., 2001. DNA methylation learns to fly. *Trends in genetics*, 17(4), pp.169–172.
- Lyko, F. et al., 1999. Mammalian (cytosine-5) methyltransferases cause genomic DNA methylation and lethality in Drosophila. *Nature genetics*, 23(3), pp.363–6.

- Lyko, F. et al., 2010. The Honey Bee Epigenomes : Differential Methylation of Brain DNA in Queens and Workers. *PLoS Biology*, 8(11).
- Lyko, F., Whittaker, A.J., et al., 2000. The putative *Drosophila* methyltransferase gene *dDnmt2* is contained in a transposon-like element and is expressed specifically in ovaries. *Mechanisms of Development*, 95, pp.215–217.
- Lyko, F., Ramsahoye, B.H. & Jaenisch, R., 2000. DNA methylation in *Drosophila melanogaster*. *Nature*, 408, pp.538–540.
- Macfarlan, T.S. et al., 2012. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature*, 487(7405), pp.57–63.
- Maiti, A. & Drohat, A.C., 2011. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. *The Journal of biological chemistry*, 286(41), pp.35334–8.
- Malone, C.S. et al., 2001. CmC(A/T)GG DNA methylation in mature B cell lymphoma gene silencing. *Proceedings of the National Academy of Sciences of the United States of America*, 98(18), pp.10404–9.
- Marhold, J. et al., 2004. The *Drosophila* MBD2 / 3 protein mediates interactions between the MI-2 chromatin complex and CpT / A-methylated DNA. *Development*, pp.6033–6039.
- Martens, J.H.A. et al., 2005. The profile of repeat-associated histone lysine methylation states in the mouse epigenome. *EMBO Journal*, 24(4), pp.800–812.
- Matzke, M. et al., 2009. RNA-mediated chromatin-based silencing in plants. *Current Opinion in Cell Biology*, 21, pp.367–376.
- Meilinger, D. et al., 2009. Np95 interacts with de novo DNA methyltransferases, Dnmt3a and Dnmt3b, and mediates epigenetic silencing of the viral CMV promoter in embryonic stem cells. *EMBO reports*, 10(11), pp.1259–64.
- Meissner, A. et al., 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454(7205), pp.766–70.
- Meissner, A. et al., 2005. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic acids research*, 33(18), pp.5868–77.
- Min, I.M. et al., 2011. in embryonic stem cells Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes & Development*, pp.742–754.
- Miranda, T.B. et al., 2010. Methylation-sensitive single-molecule analysis of chromatin structure. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, Chapter 21(January), p.Unit 21.17.1–16.

- Mistry, H. et al., 2008. Interplay between Np95 and Eme1 in the DNA damage response. *Biochemical and biophysical research communications*, 375(3), pp.321–5.
- Miura, F. et al., 2012. Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic acids research*, 40(17), p.e136.
- Morgan, H.D. et al., 2004. Activation-induced cytidine deaminase deaminates 5-methylcytosine in DNA and is expressed in pluripotent tissues: implications for epigenetic reprogramming. *The Journal of biological chemistry*, 279(50), pp.52353–60.
- Morgan, H.D. et al., 2005. Epigenetic reprogramming in mammals. *Human molecular genetics*, 14 Spec No(1), pp.R47–58.
- Morgani, S.M. et al., 2013. Totipotent embryonic stem cells arise in ground-state culture conditions. *Cell reports*, 3(6), pp.1945–57.
- Morris, K. V et al., 2004. Small Interfering RNA – Induced Transcriptional Gene Silencing in Human Cells. *Science*, 305, pp.1289–1292.
- Mund, C. et al., 2004. Comparative analysis of DNA methylation patterns in transgenic *Drosophila* overexpressing mouse DNA methyltransferases. *The Biochemical journal*, 378(Pt 3), pp.763–8.
- Nabel, C.S. et al., 2012. AID/APOBEC deaminases disfavor modified cytosines implicated in DNA demethylation. *Nature chemical biology*, 8(9), pp.751–8.
- Nagano, T., 2010. No Title. *Science*, 1717(2008).
- Nakamura, T. et al., 2012. PGC7 binds histone H3K9me2 to protect against conversion of 5mC to 5hmC in early embryos. *Nature*, 486(7403), pp.415–9.
- Nakamura, T. et al., 2007. PGC7/Stella protects against DNA demethylation in early embryogenesis. *Nature cell biology*, 9(1), pp.64–71.
- Neri, F. et al., 2013. Dnmt3L antagonizes DNA methylation at bivalent promoters and favors DNA methylation at gene bodies in ESCs. *Cell*, 155(1), pp.121–34.
- Nesterova, T.B. et al., 2008. Dicer regulates Xist promoter methylation in ES cells indirectly through transcriptional control of Dnmt3a. *Access*, 21, pp.1–21.
- Nichols, J. et al., 2009. Suppression of Erk signalling promotes ground state pluripotency in the mouse embryo. *Development (Cambridge, England)*, 136(19), pp.3215–22.
- Nishiyama, A. et al., 2013. Uhrf1-dependent H3K23 ubiquitylation couples maintenance DNA methylation and replication. *Nature*, 502(7470), pp.249–53.

- Noma, K. et al., 2004. RITS acts in cis to promote RNA interference – mediated transcriptional and post-transcriptional silencing. *October*, 36(11), pp.1174–1180.
- Novo, C. et al., 2013. The heterochromatic chromosome caps in great apes impact telomere metabolism. *Nucleic acids research*, 41(9), pp.4792–801.
- Nyce, J., Liu, L. & Jones, P.A., 1986. Variable effects of DNA-synthesis inhibitors upon DNA methylation in mammalian cells. *Nucleic Acids Research*, 14(10), pp.4353–4367.
- Oakes, C.C. et al., 2007. A unique configuration of genome-wide DNA methylation patterns in the testis. *Proceedings of the National Academy of Sciences of the United States of America*, 104(1), pp.228–33.
- Oda, M. et al., 2013. Regulation of lineage specific DNA hypomethylation in mouse trophectoderm. *PLoS one*, 8(6), p.e68846.
- Okano, M. et al., 1999. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3), pp.247–57.
- Okano, M. & Li, E., 2002. Genetic Analyses of DNA Methyltransferase Genes in Mouse. *Molecular Biology*, pp.2462–2465.
- Okano, M., Xie, S. & Li, E., 1998. Cloning and characterization of a family of novel mammalian DNA ( cytosine-5 ) methyltransferases Non-invasive sexing of preimplantation stage mammalian embryos. *Nature genetics*, 19(july), pp.219–220.
- Okano, M., Xie, S. & Li, E., 1998. Dnmt2 is not required for de novo and maintenance methylation of viral DNA in embryonic stem cells. *Nucleic acids research*, 26(11), pp.2536–40.
- Okitsu, C.Y., Hsieh, J.C.F. & Hsieh, C.-L., 2010. Transcriptional activity affects the H3K4me3 level and distribution in the coding region. *Molecular and cellular biology*, 30(12), pp.2933–46.
- Oliveira, A.M.M., Hemstedt, T.J. & Bading, H., 2012. Rescue of aging-associated decline in Dnmt3a2 expression restores cognitive abilities. *Nature neuroscience*, 15(8), pp.1111–3.
- Ooi, S.K.T. et al., 2007. DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature*, 448(7154), pp.714–7.
- Ooi, S.K.T., O'Donnell, A.H. & Bestor, T.H., 2009. Mammalian cytosine methylation at a glance. *Journal of cell science*, 122(Pt 16), pp.2787–91.
- Otani, J. et al., 2009. Structural basis for recognition of H3K4 methylation status by the DNA methyltransferase 3A ATRX-DNMT3-DNMT3L domain. *EMBO reports*, 10(11), pp.1235–41.

- Otterlei, M. et al., 1999. Post-replicative base excision repair in replication foci. *The EMBO journal*, 18(13), pp.3834–44.
- Oyola, S.O. et al., 2012. Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes. *BMC genomics*, 13(1), p.1.
- Papait, R. et al., 2007. Np95 Is Implicated in Pericentromeric Heterochromatin Replication and in Major Satellite Silencing □. *Molecular Biology of the Cell*, 18(March), pp.1098 –1106.
- Parsa, J.-Y. et al., 2012. Negative supercoiling creates single-stranded patches of DNA that are substrates for AID-mediated mutagenesis. *PLoS genetics*, 8(2), p.e1002518.
- Pekowska, A. et al., 2010. A unique H3K4me2 profile marks tissue-specific gene regulation. *Genome Research*, 20, pp.1493–1502.
- Penn, N.W. et al., 1972. The Presence of 5-Hydroxymethylcytosine in Animal Deoxyribonucleic Acid. *Biochemical Journal*, 126, pp.781–790.
- Peters, A.H.F.M. et al., 2003. Partitioning and Plasticity of Repressive Histone Methylation States in Mammalian Chromatin. , 12, pp.1577–1589.
- Phalke, S. et al., 2009. Retrotransposon silencing and telomere integrity in somatic cells of *Drosophila* depends on the cytosine-5 methyltransferase DNMT2. *Nature genetics*, 41(6), pp.696–702.
- Pham, P. et al., 2003. Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. *Nature*, 424(6944), pp.103–7.
- Piperi, C. & Papavassiliou, A.G., 2011. Strategies for DNA methylation analysis in developmental studies. *Development, growth & differentiation*, 53(3), pp.287–99.
- Ponger, L. & Li, W.-H., 2005. Evolutionary diversification of DNA methyltransferases in eukaryotic genomes. *Molecular biology and evolution*, 22(4), pp.1119–28.
- Popp, C. et al., 2010. Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. *Nature*, 463(7284), pp.1101–1105.
- Probst, A. V & Almouzni, G., 2011. Heterochromatin establishment in the context of genome-wide epigenetic reprogramming. *Trends in genetics : TIG*, 27(5), pp.177–85.
- Quail, M. et al., 2012. Optimal enzymes for amplifying sequencing libraries. *Nature methods*, 9(1), pp.10–1.
- Raddatz, G. et al., 2013. Dnmt2-dependent methylomes lack defined DNA methylation patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 110(21), pp.8627–31.

- Raiber, E.-A. et al., 2012. Genome-wide distribution of 5-formylcytosine in embryonic stem cells is associated with transcription and depends on thymine DNA glycosylase. *Genome biology*, 13(8), p.R69.
- Ramsahoye, B.H., 2000. Nearest-Neighbor Analysis. *Methods in molecular biology (Clifton, N.J.)*, 200(3), pp.9–15.
- Ramsahoye, B.H. et al., 2000. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proceedings of the National Academy of Sciences of the United States of America*, 97(10), pp.5237–42.
- Rauch, T. & Pfeifer, G.P., 2005. Methylated-CpG island recovery assay: a new technique for the rapid detection of methylated-CpG islands in cancer. *Laboratory investigation; a journal of technical methods and pathology*, 85(9), pp.1172–80.
- Reik, W., 2007. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, 447(7143), pp.425–32.
- Reik, W., Dean, W. & Walter, J., 2001. Epigenetic Reprogramming in Mammalian Development. *Science*, 293, pp.1089–1093.
- Rein, T., Zorbas, H. & Depamphilis, M.L., 1997. Active Mammalian Replication Origins Are Associated with a High-Density Cluster of mCpG Dinucleotides. *Molecular and cellular biology*, 17(1), pp.416–426.
- Renbaum, P. et al., 1990. Cloning , characterization , and expression in Escherichia coli of the gene coding for the CpG DNA methylase from Spiroplasma sp. strain MQI(M-SssI). *Nucleic acids research*, 249(4), pp.1145–1152.
- Rideout III, W.M., Eggan, K. & Jaenisch, R., 2011. Nuclear Cloning and Epigenetic Reprogramming of the Genome. *Science*, 293(2001), pp.1093–1098.
- Riggs, A.D., 1975. X inactivation, differentiation, and DNA methylation. *Cytogenetic and Genome Research*, 14(1), pp.9–25.
- Robertson, K.D. et al., 1999. The human DNA methyltransferases (DNMTs) 1, 3a and 3b: coordinate mRNA expression in normal tissues and overexpression in tumors. *Nucleic acids research*, 27(11), pp.2291–8.
- Robinson, M.D. et al., 2010. Evaluation of affinity-based genome-wide DNA methylation data: Effects of CpG density, amplification bias, and copy number variation. *Genome research*, (20), pp.1719–1729.
- Roig, I. et al., 2004. Female-specific features of recombinational double-stranded DNA repair in relation to synapsis and telomere dynamics in human oocytes. *Chromosoma*, 113(1), pp.22–33.

- Rollins, R. a et al., 2006. Large-scale structure of genomic methylation patterns. *Genome research*, 16(2), pp.157–63.
- Ross, J.P. et al., 2010. Recombinant mammalian DNA methyltransferase activity on model transcriptional gene silencing short RNA-DNA heteroduplex substrates. *The Biochemical journal*, 432(2), pp.323–32.
- Di Ruscio, A. et al., 2013. DNMT1-interacting RNAs block gene-specific DNA methylation. *Nature*, 503(7476), pp.371–6.
- Sado, T. et al., 2000. X inactivation in the mouse embryo deficient for Dnmt1: distinct effect of hypomethylation on imprinted and random X inactivation. *Developmental biology*, 225(2), pp.294–303.
- Sakai, Y. et al., 2004. Co-expression of de novo DNA methyltransferases Dnmt3a2 and Dnmt3L in gonocytes of mouse embryos. *Gene expression patterns : GEP*, 5(2), pp.231–7.
- Salle, S. La et al., 2004. Windows for sex-specific methylation marked by DNA methyltransferase expression profiles in mouse germ cells. *Developmental Biology*, 268, pp.403 – 415.
- Salomon, R. & Kaye, A.M., 1970. Methylation of mouse DNA in vivo: di- and tripyrimidine sequences containing 5-methylcytosine. *Biochim Biophys Acta*, 204(2), pp.340–351.
- Salvaing, J. et al., 2012. 5-Methylcytosine and 5-hydroxymethylcytosine spatiotemporal profiles in the mouse zygote. *PloS one*, 7(5), p.e38156.
- Santos, F. et al., 2013. Active demethylation in mouse zygotes involves cytosine deamination and base excision repair. *Epigenetics & chromatin*, 6(1), p.39.
- Schaefer, M. et al., 2009. Azacytidine Inhibits RNA Methylation at DNMT2 Target Sites in Human Cancer Cell Lines. *Cancer Research*, pp.8127–8132.
- Schaefer, M. & Lyko, F., 2010a. Lack of evidence for dnA methylation of Invader4 retroelements in Drosophila and implications for dnmt2-mediated epigenetic regulation. *Nature Publishing Group*, 42(11), pp.920–921.
- Schaefer, M. & Lyko, F., 2010b. Solving the Dnmt2 enigma. *Chromosoma*, 119(1), pp.35–40.
- Schmidt, C.S. et al., 2012. Global DNA hypomethylation prevents consolidation of differentiation programs and allows reversion to the embryonic stem cell state. *PloS one*, 7(12), p.e52629.
- Schmittgen, T.D. & Livak, K.J., 2008. Analyzing real-time PCR data by the comparative C T method. *Nature Protocols*, 3(6), pp.1101–1108.

- Seisenberger, S. et al., 2012. The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells. *Molecular cell*, 48(6), pp.849–62.
- Senner, C.E. et al., 2012. DNA methylation profiles define stem cell identity and reveal a tight embryonic-extraembryonic lineage boundary. *Stem cells (Dayton, Ohio)*, 30(12), pp.2732–45.
- Serre, D., Lee, B.H. & 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(2), pp.391–9.
- Severin, P.M.D. et al., 2011. Cytosine methylation alters DNA mechanical properties. *Nucleic acids research*, 39(20), pp.8740–51.
- Shanmuganathan, R. et al., 2013. Conventional and nanotechniques for DNA methylation profiling. *The Journal of molecular diagnostics : JMD*, 15(1), pp.17–26.
- Sharif, J. et al., 2007. The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. *Nature*, 450(7171), pp.908–12.
- Shen, L. & Zhang, Y., 2013. 5-Hydroxymethylcytosine: generation, fate, and genomic distribution. *Current opinion in cell biology*, 25(3), pp.289–96.
- Shimooka, Y., Nishikawa, J.-I. & Ohyama, T., 2013. Most methylation-susceptible DNA sequences in human embryonic stem cells undergo a change in conformation or flexibility upon methylation. *Biochemistry*, 52(8), pp.1344–53.
- Shiraishi, M. & Hayatsu, H., 2004. High-speed conversion of cytosine to uracil in bisulfite genomic sequencing analysis of DNA methylation. *DNA research : an international journal for rapid publication of reports on genes and genomes*, 11(6), pp.409–15.
- Shirane, K. et al., 2013. Mouse Oocyte Methylomes at Base Resolution Reveal Genome-Wide Accumulation of Non-CpG Methylation and Role of DNA Methyltransferases. *PLoS genetics*, 9(4), p.e1003439.
- Singh, A., Zubko, E. & Meyer, P., 2008. Cooperative activity of DNA methyltransferases for maintenance of symmetrical and non-symmetrical cytosine methylation in *Arabidopsis thaliana*. *The Plant journal : for cell and molecular biology*, 56(5), pp.814–23.
- Smallwood, S. et al., 2014. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature Methods*, (july), pp.2–7.
- Smallwood, S. & Kelsey, G., 2012. Genome-Wide Analysis of DNA Methylation in Low Cell Numbers by Reduced Representation Bisulfite Sequencing N. Engel, ed. *Genomic Imprinting: Methods and Protocols, Methods in Molecular Biology*, 925, pp.187–197.

- Sneider, T.W., 1980. Novikoff rat hepatoma and bovine liver DNAs were digested with *Map I* or *Hpa A* restriction fragments generated by the isoschizomers *Hpa I* of the sequence CCGG. In this case the rationale is that *Hpa II* will cleave -G but not CG where. *Biochemistry*, 8(17), pp.3829–3840.
- Song, J. et al., 2011. Structure of DNMT1-DNA complex reveals a role for autoinhibition in maintenance DNA methylation. *Science (New York, N.Y.)*, 331(6020), pp.1036–40.
- Song, J. et al., 2012. Structure-based mechanistic insights into DNMT1-mediated maintenance DNA methylation. *Science (New York, N.Y.)*, 335(6069), pp.709–12.
- Spruijt, C.G., Baymaz, H.I. & Vermeulen, M., 2013. Identifying Specific Protein–DNA Interactions Using SILAC-Based Quantitative Proteomics M. Bina, ed. *Methods in Molecular Biology*, 977, pp.137–157.
- Stadler, M.B. et al., 2011. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, 480(7378), pp.490–5.
- Straussman, R. et al., 2009. Developmental programming of CpG island methylation profiles in the human genome. *Nature Structural & Molecular Biology*, 16(5), pp.564–571.
- Stroud, H. et al., 2014. Non-CG methylation patterns shape the epigenetic landscape in *Arabidopsis*. *Nature structural & molecular biology*, 21(1), pp.64–72.
- Suetake, I. et al., 2003. Distinct Enzymatic Properties of Recombinant Mouse DNA Methyltransferases Dnmt3a and Dnmt3b. *Journal of Biochemistry*, 133(6), pp.737–744.
- Suetake, I. et al., 2004. DNMT3L stimulates the DNA methylation activity of Dnmt3a and Dnmt3b through a direct interaction. *The Journal of biological chemistry*, 279(26), pp.27816–23.
- Sugiyama, T. et al., 2005. RNA-dependent RNA polymerase is an essential component of a self-enforcing loop coupling heterochromatin assembly to siRNA production. *PNAS*, 102(1), pp.1–6.
- Sutherland, E., Coe, L. & Raleigh, E.A., 1992. *McrBC*: a Multisubunit Restriction Endonuclease. *Journal of molecular biology*, (225), pp.327–348.
- Tahiliani, M. et al., 2010. Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1. *Science*, 930(2009), pp.930–935.
- Takagi, H., Tajima, S. & Asano, a, 1995. Overexpression of DNA methyltransferase in myoblast cells accelerates myotube formation. *European journal of biochemistry / FEBS*, 231(2), pp.282–91.
- Takashima, S. et al., 2009. Abnormal DNA methyltransferase expression in mouse germline stem cells results in spermatogenic defects. *Biology of reproduction*, 81(1), pp.155–64.

- Takayama, S. et al., 2014. Genome methylation in *D. melanogaster* is found at specific short motifs and is independent of DNMT2 activity. *Genome research*, 24(5), pp.821–30.
- Tam, O.H. et al., 2008. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature*, 453(May).
- Tamaru, H. et al., 2003. Trimethylated lysine 9 of histone H3 is a mark for DNA methylation in *Neurospora crassa*. *Sciences-New York*, 34(may), pp.75–79.
- Tanaka, K. et al., 2007. Direct labeling of 5-methylcytosine and its applications. *Journal of the American Chemical Society*, 129(17), pp.5612–20.
- Tanaka, K. & Okamoto, A., 2007. Degradation of DNA by bisulfite treatment. *Bioorganic & Medicinal Chemistry Letters*, 17(December 2006), pp.1912–1915.
- Tasheva, E.S. & Roufa, D.J., 1994a. A mammalian origin of bidirectional DNA replication within the Chinese hamster RPS14 locus. *Molecular and cellular biology*, 14(9), pp.5628–35.
- Tasheva, E.S. & Roufa, D.J., 1994b. Densely methylated DNA islands in mammalian chromosomal replication origins. *Molecular and cellular biology*, 14(9), pp.5636–44.
- Thalhammer, A. et al., 2011. Hydroxylation of methylated CpG dinucleotides reverses stabilisation of DNA duplexes by cytosine 5-methylation. *Chemical communications (Cambridge, England)*, 47(18), pp.5325–7.
- Tomizawa, S.-I. et al., 2011. Dynamic stage-specific changes in imprinted differentially methylated regions during early mammalian development and prevalence of non-CpG methylation in oocytes. *Development (Cambridge, England)*, 138(5), pp.811–20.
- Torseth, K. et al., 2012. The UNG2 Arg88Cys variant abrogates RPA-mediated recruitment of UNG2 to single-stranded DNA. *DNA repair*, 11(6), pp.559–69.
- Toth, M., Müller, U. & Doerfler, W., 1990. Establishment of de novo DNA methylation patterns. Transcription factor binding and deoxycytidine methylation at CpG and non-CpG sequences in an integrated adenovirus promoter. *Journal of molecular biology*, 214(3), pp.673–83.
- Tschiersch, B. et al., 1994. The protein encoded by the *Drosophila* position- combines domains of antagonistic regulators of homeotic gene complexes. *EMBO Journal*, 13(16), pp.3822–3831.
- Tsumura, A. et al., 2006. Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, dnmt3a and Dnmt3b. *Genes to Cells*, 11, pp.805–814.

- Tufarelli, C. et al., 2003. Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. *Nature Genetics*, 34(2), pp.157–165.
- Verdel, A., 2004. RNAi-Mediated Targeting of Heterochromatin by the RITS. *Spring*, 672(2004).
- Vertino, P.M., Yen, R.C. & Gao, J.I.N., 1996. De novo methylation of CpG island sequences in human fibroblasts overexpressing DNA De Novo Methylation of CpG Island Sequences in Human Fibroblasts Overexpressing DNA ( Cytosine-5- ) -Methyltransferase. , 16(8).
- Vilkaitis, G. et al., 2005. Processive methylation of hemimethylated CpG sites by mouse Dnmt1 DNA methyltransferase. *The Journal of biological chemistry*, 280(1), pp.64–72.
- Vire, E. et al., 2006. The Polycomb group protein EZH2 directly controls DNA methylation. *Nature*, 439(February), pp.871–874.
- Volpe, P., 2005. The language of methylation in genomics of eukaryotes. *Biochemistry. Biokhimiā*, 70(5), pp.584–95.
- Volpe, T.A. et al., 2010. Regulation of Heterochromatic Silencing and Histone H3 Lysine-9 Methylation by RNAi. *Science*, 1833(2002).
- Vrbsky, J. et al., 2010. siRNA – Mediated Methylation of Arabidopsis Telomeres. *PLoS Genetics*, 6(6).
- Wang, J. et al., 2014. Genome-wide screen of DNA methylation changes induced by low dose X-ray radiation in mice. *PloS one*, 9(3), p.e90804.
- Wang, L. et al., 2014. Programming and inheritance of parental DNA methylomes in mammals. *Cell*, 157(4), pp.979–91.
- Warnecke, P.M. et al., 2002. Identification and resolution of artifacts in bisulfite sequencing. *Methods (San Diego, Calif.)*, 27(2), pp.101–7.
- Watanabe, T. et al., 2008. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature*, 453(May), pp.539–544.
- Waterston, R.H. et al., 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915), pp.520–62.
- Weber, M. et al., 2005. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature genetics*, 37(8), pp.853–62.
- Webster, K.E. et al., 2005. Meiotic and epigenetic defects in Dnmt3L-knockout mouse spermatogenesis. *PNAS*, 102(11), pp.4068–4073.

- Van Weemen, B.K. & Schuurs, A.H.W.M., 1971. IMMUNOASSAY USING ANTIGEN-ENZYM CONJUGATES. *FEBS Letters*, 15(3), pp.232–236.
- Weill, J.-C. & Reynaud, C.-A., 2004. RPA tightens AID to DNA...editing. *Nature immunology*, 5(9), pp.876–8.
- Weinberg, M.S. et al., 2006. The antisense strand of small interfering RNAs directs histone methylation and transcriptional gene silencing in human cells. *Spring*, pp.256–262.
- Weissmann, F. et al., 2003. DNA Hypermethylation in *Drosophila melanogaster* Causes Irregular Chromosome Condensation and Dysregulation of Epigenetic Histone Modifications †. *Society*, 23(7), pp.2577–2586.
- White, G.P. et al., 2002. Differential Patterns of Methylation of the IFN- $\gamma$  Promoter at CpG and Non-CpG Sites Underlie Differences in IFN- $\gamma$  Gene Expression Between Human Neonatal and Adult CD45RO- T Cells. *The Journal of Immunology* *the Journal of Immunology*, 168, pp.2820–2827.
- Whyte, W. a et al., 2013. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2), pp.307–19.
- Woodcock, D.M. et al., 1997. Asymmetric methylation in the hypermethylated CpG promoter region of the human L1 retrotransposon. *The Journal of biological chemistry*, 272(12), pp.7810–6.
- Woodcock, D.M. et al., 1988. Methylation at dinucleotides other than CpG: implications for human maintenance methylation. *Gene*, 74, pp.151–152.
- Woodcock, D.M., Crowther, P.J. & Diver, W.P., 1987. The majority of methylated deoxycytidines in human DNA are not in the CpG dinucleotide. *Biochemical and biophysical research communications*, 145(2), pp.888–894.
- Woodcock, D.M., Linsenmeyer, M.E. & Warren, W.D., 1998. DNA methylation in mouse A-repeats in DNA methyltransferase-knockout ES cells and in normal cells determined by bisulfite genomic sequencing. *Gene*, 206(1), pp.63–7.
- Wray, J. et al., 2011. Inhibition of glycogen synthase kinase-3 alleviates Tcf3 repression of the pluripotency network and increases embryonic stem cell resistance to differentiation. *Nature cell biology*, 13(7), pp.838–45.
- Wu, S.C. & Zhang, Y., 2010. Active DNA demethylation : *Nature Publishing Group*, 11(9), pp.607–620.
- Wu, Z. et al., 2008. Microarray-based Ms-SNuPE: near-quantitative analysis for a high-throughput DNA methylation. *Biosensors & bioelectronics*, 23(9), pp.1333–9.

- Xiang, H. et al., 2010. Single base – resolution methylome of the silkworm reveals a sparse epigenomic map. *Online*, 28(5), pp.516–520.
- Xiong, Z. & Laird, P.W., 1997. COBRA: a sensitive and quantitative DNA methylation assay. *Nucleic acids research*, 25(12), pp.2532–4.
- Yamane, A. et al., 2011. Deep-sequencing identification of the genomic targets of the cytidine deaminase AID and its cofactor RPA in B lymphocytes. *Nature immunology*, 12(1), pp.62–9.
- Yan, J., Zierath, J.R. & Barrès, R., 2011. Evidence for non-CpG methylation in mammals. *Experimental cell research*, pp.4–10.
- Yu, M. et al., 2012. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell*, 149(6), pp.1368–80.
- Yu, W. et al., 2014. Genome-wide DNA methylation patterns in LSH mutant reveals de-repression of repeat elements and redundant epigenetic silencing pathways. *Genome research*.
- Zemach, A. et al., 2013. The arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. *Cell*, 153(1), pp.193–205.
- Zhang, H. et al., 2014. Analysis of trichloroethylene-induced global DNA hypomethylation in hepatic L-02 cells by liquid chromatography-electrospray ionization tandem mass spectrometry. *Biochemical and biophysical research communications*, 446(2), pp.590–5.
- Zhang, X. et al., 2006. Resource Mapping and Functional Analysis of DNA Methylation in Arabidopsis. *Cell*, pp.1189–1201.
- Zhang, Y. et al., 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(13), pp.4246–53.
- Zhang, Y. et al., 2008. Model-based analysis of ChIP-Seq (MACS). *Genome biology*, 9(9), p.R137.
- Zhao, J. et al., 2008. Polycomb Proteins Targeted by a Short Repeat RNA to the Mouse X Chromosome. *Science*, 322, pp.750–756.
- Zilberman, D. et al., 2004. Role of Arabidopsis ARGONAUTE4 in RNA-Directed DNA Methylation Triggered by Inverted Repeats. *Current*, 14, pp.1214–1220.
- Zilberman, D. & Henikoff, S., 2007. Genome-wide analysis of DNA methylation patterns. *Development (Cambridge, England)*, 134(22), pp.3959–65.

Ziller, M. et al., 2011. Genomic Distribution and Inter-Sample Variation of Non- CpG Methylation across Human Cell Types. *PLoS Genetics*, 7(12).

## 10 Appendix

### 10.1 Additional tables

Table 11. Base composition of the mouse genome and contribution of each cytosine context

<b>Base composition of the mouse genomic sequence (NCBIM37, sequences containing N's were ignored)</b>				
Total number of bases:	2558525758	Total number of dinucleotides:	2558525109	
base	percent total	word	percent absolute	relative
T	29,13	CA	7,45	35,70
A	29,11	CT	7,35	35,22
G	20,88	CC	5,24	25,11
C	20,88	CG	0,83	3,98

Table 12. Table of used oligonucleotides

<b>qPCR primers</b>		
<b>Gene</b>	<b>Forward primer</b>	<b>Reverse primer</b>
<b>Hspcb</b>	GCTGGCTGAGGACAAGGAGA	CGTCGGTTAGTGGAAATCTTCATG
<b>Atp5b</b>	GGCCAAGATGTCCTGCTGTT	GCTGGTAGCCTACAGCAGAAGG
<b>Dnmt2al</b>	ATGACCGATCCAAGGACAAC	TTCAAACCTTTGACATTTTCT
<b>Dnmt2pro</b>	AACCAGGACGCCACAGTAGT	GCCATTGATCAGGCCTTAAA
<b>Dnmt1s</b>	GGGTCTCGTTCAGAGCTG	GCAGGAATTCATGCAGTAAG
<b>Dnmt3a1</b>	AGAACAGAAGCAGACCAACATCG	TGGAAGGTGAGTCTTGCCATG
<b>Dnmt3a2</b>	GGCTCACACCTGAGCTGTACTG	GCCTGGTTCTCTCCACAGC
<b>Dnmt3b</b>	TGGTGATTGGTGGAAGCC	AATGGACGGTTGTCGCC
<b>MajorSat</b>	GAAAATGAGAAATACACACTTT	GTCAAGTGGATGTTTCTCATT
<b>U1</b>	CTTACCTGGCAGGGGAGATA	CAGTCCCCACTACCACAAA
<b>Maj</b>	GACGACTTGAAAAATGACGAAATC	CATATCCAGGTCCTTCAGTGTGC
<b>Min</b>	CATGGAAAATGATAAAAACC	CATCTAATATGTTCTACAGTGTGG
<b>DNA primers</b>		
<b>Region</b>	<b>Forward primer</b>	<b>Reverse primer</b>
<b>Dnmt2genot</b>	TTTTAATGAGCCCTCCATGC	TTTTCCCTTCTCCTTTCTTTCC
<b>IRES-R (mDnmt2)</b>	-	GAAAGAGGGCTCTGTCTCCAGTCTC
<b>M13 2kb</b>	ATTTCCATGAGCGTTTTTCC	GCAAGGCAAAGAATTAGCAA
<b>M13 RT1</b>	GCTCCCCTCTGATTCTAAC	TTAGAGCTTGACGGGGAAAG
<b>M13 RT2</b>	GACAGGTTTCCCAGCTGG	GGGTAACGCCAGGGTTTT
<b>M13 R3</b>	CCAGACGCGAATTATTTTTG	GATGAACGGTAATCGTAAACTAGC
<b>M13-R5</b>	GGCGTTACCCAACTTAATCG	AAAGCGCCATTCGCCATT
<b>MSatFor BS</b>	AAATGAGAAATATATATTTTTAGGA	CAAATAAATATTTCTCATTTTCC
<b>MSatRev BS</b>	AAATAAAAAATACACACTTT	GTTAAGTGGATGTTTTTTATT
<b>Sequencing primers</b>		
<b>Region</b>	<b>Forward primer</b>	<b>Reverse primer</b>
<b>M13 (F-40 &amp; R)</b>	GTTTTCCCAGTCACGAC	CAGGAAACAGCTATGACC
<b>pQE (FOR-REV)</b>	CCCGAAAAGTGCCACCTG	GGTCATTACTGGAGTCTTG
<b>Cloning primers</b>		
<b>AttB-M.MpeI F</b>	GGGGACAAGTTTGTACAAAAAAGCAGGCTTCGGATCCGCCACCATGGATAGCAACAAG GACAAGA	
<b>AttB-M.MpeI R</b>	GGGGACCACTTTGTACAAGAAAGCTGGGTCTGAATTCTCAGTGGTGGTGGTGGTGGTCT CG	

Table 13. M13-derived PCR fragments. The sequences of the M13 fragments used for a variety of experiments such as NNA standards, fragments with variable cytosine content for BS-degradation or for the analytical test digest of variety of restriction enzymes. For NNA, the base for which the fragment was used is indicated in the 'Name' column as well as capitalised in the actual sequence, and the DpnII restriction site is underlined.

Name	Length	Sequence 5' > 3'
<b>R1</b>	211 bp	gctcccgctctgattctaacgaggaaagcacggtatacgtgctcgtcaaagcaacc atagtacgcgcctgtagcggcgcatataagcgcggcgggtgtggtggttacgcgca gcgtgaccgctacacttgccagcgccttagcgcggcctccttctcgctttctccct tcctttctcgccacggttcgcccggctttccccgctcaagctctaa
<b>R2L</b> <b>"G"</b>	278 bp	gacaggtttcccgactggaaagcgggcagtgagcgcaacgcaattaatgtgagtta gctcactcattagggcaccacaggctttacactttatgcttccggctcgtatggtgt gtggaattgtgagcggataacaatttcacacaggaaacagctatgaccatgattac gaattcgagctcggtaaccggGgatcctctagagtcgacctgcaggcatgcaagct tggcactggccgctcgttttacaacgctcgtgactgggaaaaccctggcgttacc
<b>R3</b>	201 bp	ccagacgcgaattatTTTTgatggcgTTCctattggTTaaaaaatgagctgattta acaaaaatttaatgcgaattTTaacaataattaacgTTTacaattTaaatatttg cttatacaatcttctgTTTTtggggctTTTctgattatcaaccggggtacatatg attgacatgctagTTTTacgattaccgTTCatc
<b>R3 end</b> <b>"A"</b>	547 bp	ccagacgcgaattatTTTTgatggcgTTCctattggTTaaaaaatgagctgattta acaaaaatttaatgcgaattTTaacaataattaacgTTTacaattTaaatatttg cttatacaatcttctgTTTTtggggctTTTctgattatcaaccggggtacatatg attgacatgctagTTTTacgattaccgTTCatcgcattctcttgtttgctccagact ctcaggcaatgacctgatagcTTTgtAgatcctcaaaaatagctaccctctccg gcattaatttatcagctagaacggTtgaatatcatattgatggTgatttgactgtc tccggcctttctcaccctTTTgaatctttacctacacattactcaggcattgcatt taaaatataatgagggttctaaaaattTTtatccttgcgTtgaataaaasggcttct cccgcaaaagtattacagggtcataatgTTTTtggTacaaccgatttagctttatg ctctgaggctttattgcttaattTTtgcTaatctttgccttgc
<b>R5</b> <b>"C"</b>	123 bp	ggcgttacccaacttaatcgccctgcagcacatcccccttctgccagctggcgtaa tagcgaagaggcccgcacCgatcgccttcccaacagttgcgagcctgaatggcg aatggcgcttt

Table 14. Table of used antibodies

Target (Antigen)	Company	Clone
<b>5-methylcytosine</b>	Eurogentec	BI-MECY-0100, clone 33D3
<b>5-hydroxymethylsytosine</b>	Active Motif	cat # 39769 rabbit polyclonal
<b>5-formylcytosine</b>	Active Motif	pAb #61223
<b>ssDNA</b>	Chemicon (Millipore)	MAB3034
<b>Dnmt1</b>	Santa Cruz	H-300 sc-20701 rabbit polyclonal
<b>Dnmt3a</b>	Abcam	ab13888 [64B1446] (Imgenex - IMG-268A)
<b>Dnmt3b</b>	Abcam	ab13604 [52A1018] (Imgenex - IMG-184A)
<b>Oct3/4</b>	Santa Cruz	sc-5279 (C10) mouse monoclonal
<b>H3K9me3</b>	Active Motif	mouse monoclonal
<b>Anti-His 6x</b>	Pierce	MA1-21315 (HIS H8)
<b>Anti-mouse HRP</b>	Southern Biotech	1030-05 Goat Anti Mouse IgG HRP 1.0 mL
<b>Anti-mouse HRP</b>	NA931VS	ECL Mouse IgG, HRP-linked whole Ab (from sheep)
<b>Anti-rabbit HRP</b>	NA934VS	ECL Rabbit IgG, HRP-linked whole Ab (from donkey)

Table 15. Illumina adapters and primers

Name	Sequence 5' > 3'
<b>PE adapter 1</b>	P-GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
<b>PE adapter 2</b>	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
<b>PE PCR Primer 1.0</b>	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCT TCCGATCT
<b>PE PCR Primer 2.0</b>	CAAGCAGAAGACGGCATAACGAGATCGGTCTCGGCATTCCTGCTGAACCGC TCTTCCGATCT
<b>PE PCR Primer 2.0 indexed</b> (Quail et al. 2012)	CAAGCAGAAGACGGCATAACGAGATXXXXXXXXXGAGATCGGTCTCGGCATT CCTGCTGAACCGCTCTTCCGATCT
<b>Sequencing primer PE 1.0</b>	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
<b>Sequencing primer PE 2.0</b>	CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
<b>Sequencing primer PE 2.0</b> <b>indexed</b> (Quail et al. 2012)	AAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTC

Table 16. Table of consensus repeat sequences

Name	Length	Sequence 5' > 3'
<b>Major satellite</b>	234 bp	aatgagaaatgcacactgaaggacctggaatatggcgagaaaactgaaaatcacggaaa atgagaaatacacactttaggacgtgaaatatggcgaggaaaactgaaaagggtggaaa atthagaaatgtccactgtaggacgtggaatatggcaagaaaactgaaaatcatggaaa atgagaaacatccacttgacgacttgaaaaatgacgaaatcactaaaaaacgtgaaaa tgagaaatgcacactgaa
<b>Minor satellite</b>	118 bp	ttgtagaacagtgtatatcaatgagttacaatgagaaacatggaaaatgataaaaacc acactgtagaacatattagatgagtgagttacactgaaaaacacattcgttggaaacg gg
<b>SINEB1</b>	161 bp	cgaggcctaataatagtataggcctcgggctgggtggcgacgcctttaatcccagcac tcgggaggcagaggcaggcggatctctgtgagttcgaggccagcctggctacatagc gagttccaggccagccagggtacatagtgagaccctgtctcaaa
<b>IAP LTR1a</b>	153 bp	tgttgggagccgcgccacattcgccgttacaagatggcgctgacagctgtgttctaag tggtaacaaataatctgcgcatatgccgaggggtgttctctactccatgtgctctgccc ttccccgtgacgtcaactcggccgatgggctgcag
<b>LINE1 3' UTR</b>	720 bp	ggatggacctggagagcatcatcctgagtgaggtaacacaatcacaagggaactcaca caatatgtactcactgataagtggatactagccaaaacctaggataccacgatata agatacaatttcttaaacacatgaaactcaagaaaaatgaagactgaagtgtggacac tatgccctccttagaagtgggaacaaaacccatggaaggagttacagaaacaaag tttgagctgagatgaaaggatggaccatgtagagactgccatatccagggatccacc ccataatcagcatccaaacgctgacaccattgcatatactagcaagattttatcgaaa ggaccagatgtagctgtctcttgtgagactatgccggggcctagcaaacacagaagt ggatgctcacagtcagctaattggatggatcacagggctcccaatggaggagctagaga aagtaccaaggagctaaagggatcttcaaccctataggtggaacaacactatgaact aaccagtaccctgagctcttgactctagctgcatatgtatcaaaagatggcctagtc ggccatcactggaaagagaggccattggacacgcagactttgtgtgccccggtagcag gggaacgccagggccaaaggggggagtggtgggtaggggagtggggtgggtgggt aagggggacttttggtatagcatt
<b>LINE1-5' UTR</b>	282 bp	atcctttaatcagacttgtccactctcccacgcggtctcttggactgtcgaagacctt gtccgtcttcgtgtctccccgactccgtcgtgggacacaccggcccctgtcggccgg tggaggcctggcctcctgtccacgggtgggcccagcccctccgccgattcgggtgctg tcgtcgccagcggtagaaccaggccctgaggttccttgaatccttaaatcagacgaa ttcactctcagacatggtggacccttgacggtttcgttgtgtcacagact

Table 17. Unspecific cutters (target CN). The enzymes in red colour are methylation-dependent (cut only methylated sites):

	<b>FokI</b>	<b>SgeI</b>	<b>MnlI</b>	<b>FspEI</b>	<b>MspJI</b>
Sites cut total (out of)	5.4 x 10 <sup>6</sup>	~10.5 x 10 <sup>6</sup> (203 x 10 <sup>6</sup> )	23 x 10 <sup>6</sup>	~10.6 x 10 <sup>6</sup> (213 x 10 <sup>6</sup> )	~20.5 x 10 <sup>6</sup> (410 x 10 <sup>6</sup> )
Genome occurrence	1 : 476	~1 : 252 (1 : 12.6)	1 : 111	~1 : 240 (1 : 12)	~1 : 125 (1 : 6)
C's detected	<b>0.22%</b>	<b>0.43%</b>	<b>0.87%</b>	<b>0.42%</b>	<b>0.81 %</b>
C-enrichment	<b>1 : 2135</b>	<b>1 : 1108</b>	<b>1 : 556</b>	<b>1 : 1163</b>	<b>1 : 588</b>

Table 18. Specific cutters (target a specific dinucleotide – CA, CT or CG)

	<b>FokI</b>	<b>BstNI</b> (CHG)	<b>CviAII</b> (CAH)	<b>XspI</b> (CTH)	<b>MspI/HpaII</b> (CG)
Sites cut total	5.4 x 10 <sup>6</sup>	7.65 x 10 <sup>6</sup>	12.5 x 10 <sup>6</sup>	8 x 10 <sup>6</sup>	1.6 x 10 <sup>6</sup>
% CHG/CHH/CG	1 : 476	6.79%	4.56%	2.96%	7.47%
C's detected	<b>0.22%</b>	<b>2.86 %</b>	<b>4.66 %</b>	<b>3.02 %</b>	<b>0.6 %</b>
C-enrichment	<b>1 : 2135</b>	<b>1 : 170</b>	<b>1 : 102</b>	<b>1 : 160</b>	

Table 19. Full screen of mCA monoclonal supernatants

Graph colour coding:			
<b>Green</b> – positive control (mCA or mCG respectively)			
<b>Red</b> – opposite negative methylation control: mCG for mCA, and mCA for mCG			
<b>Orange</b> – unmethylated DNA			
<b>Black</b> – other negative methylation controls – mCC and mCT (not biologically relevant)			
Antibody species	Oligo panel I	Oligo panel II	Good – Y/N
33D3 mouse mAb Eurogentec	<p><b>33D3 mC specificity</b></p>	<p><b>33D3 mC specificity</b></p>	> strong  > very specific
1. Mouse polyclonal serum NO01-2	<p><b>mCA poly mouse serum specificity</b></p>	<p><b>mCA poly mouse serum specificity II</b></p>	Y?  > strong  > not very specific
2. Mouse monoclonal supernatant 2C8	<p><b>mCA 2C8 clone specificity III</b></p>	<p><b>mCA 2C8 clone specificity II</b></p>	Y  > weak  > specific

<p>3. Mouse monoclonal supernatant 7E1</p>	<p><b>mCA 7E1 clone specificity</b></p>	<p><b>mCA 7E1 clone specificity II</b></p>	<p>Y? &gt; strong &gt; quite specific</p>
<p>4. Mouse monoclonal supernatant 7A3</p>	<p><b>mCA 7A3 clone specificity</b></p>	<p><b>mCA 7A3 clone specificity II</b></p>	<p>N &gt; weak &gt; not specific</p>
<p>5. Mouse monoclonal supernatant 7A9</p>	<p><b>mCA 7A9 clone specificity</b></p>	<p><b>mCA 7A9 clone specificity II</b></p>	<p>Y? &gt; strong &gt; quite specific</p>
<p>6. Mouse monoclonal supernatant 8H8</p>	<p><b>mCA 8H8 clone specificity</b></p>	<p><b>mCA 8H8 clone specificity II</b></p>	<p>N &gt; weak &gt; not specific</p>
<p><b>Oligo panels I and II together</b></p>			
<p>7. Mouse monoclonal supernatant 5H6</p>	<p><b>mCA 5H6 clone specificity</b></p>		<p>N &gt; weak</p>

<p>8. Mouse monoclonal supernatant 9H5</p>	<p><b>mCA 9H5 clone specificity</b></p>		<p>N &gt; weak</p>
<p>9. Mouse monoclonal supernatant 8G8</p>	<p><b>mCA 8G8 clone specificity</b></p>		<p>Y &gt; strong &gt; quite specific</p>
<p>10. Mouse monoclonal supernatant 8G2</p>	<p><b>mCA 8G2 clone specificity</b></p>		<p>N? &gt; weak &gt; quite specific</p>
<p>11. Mouse monoclonal supernatant 4C10</p>	<p><b>mCA 4C10 clone specificity</b></p>		<p>N &gt; weak</p>

Table 20. Full screen of mCG monoclonal supernatants

Antibody species	Oligo panel I	Oligo panel II	Good: Y/N?
1. Mouse polyclonal serum NO02-5	<p><b>mCG poly mouse serum specificity</b></p>	<p><b>mCG poly mouse serum specificity II</b></p>	Y? > strong > quite specific
2. Mouse monoclonal supernatant 5B11	<p><b>mCG 5B11 clone specificity</b></p>	<p><b>mCG 5B11 clone specificity II</b></p>	Y > weak > very specific
3. Mouse monoclonal supernatant 10E2	<p><b>mCG 10E2 clone specificity</b></p>	<p><b>mCG 10E2 clone specificity II</b></p>	YY > very strong > very specific
4. Mouse monoclonal supernatant 9F10	<p><b>mCG 9F10 clone specificity</b></p>	<p><b>mCG 9F10 clone specificity II</b></p>	Y > strong > very specific

<p>5. Mouse monoclonal supernatant 5G6</p>	<p><b>mCG 5G6 clone specificity</b></p>	<p><b>mCG 5G6 clone specificity II</b></p>	<p>N</p> <p>&gt; weak</p> <p>&gt; specific</p>
<p>6. Mouse monoclonal supernatant 6F5</p>	<p><b>mCG 6F5 clone specificity</b></p>	<p><b>mCG 6F5 clone specificity II</b></p>	<p>Y</p> <p>&gt; strong</p> <p>&gt; very specific</p>

Table 21. Datasets used for the calculation of global mCG and mCH levels in mouse early development.

Developmental stage	Dataset/s
13.5 fPGC	Wang 2014
NGV	Shirane 2013 PBAT
GVO	Shirane 2013 PBAT, Kobayashi 2012, Wang 2014
Sperm	Kobayashi 2012, Wang 2014
M II oocyte	Smallwood 2014 (scPBAT)
Zygote	Peat 2014 PBAT
2-cell	Wang 2014
4-cell	Wang 2014
Blastocyst	Kobayashi 2012
ICM	Wang 2014
mESC 2i	Smallwood 2014 (scPBAT)
mESC serum	Kobayashi 2012, Smallwood 2014 (scPBAT)
Epiblast E6.5	Wang 2014
Embryo E7.5	Wang 2014

Table 22. Datasets used for the ChIP analysis of histone marks and protein binders

Protein/mark	Dataset	Cell type
H3K9me3	Mikkelsen_2007	mES
	Karimi_2011	mES J1
	Marks_2012	mES E14
H3K4me1	Buecker_2014	mES
H3K4me3	Mikkelsen_2007	mES
	Marson_2008	mES
H3K20me3	Mikkelsen_2007	mES
H3K27me3	Mikkelsen_2007	mES
H3K27ac	Buecker_2014	mES
H3K36me3	Mikkelsen_2007	mES
	Marson_2008	mES
H3K79me2	Marson_2008	mES
Pol2	Mikkelsen_2007	mES
Nanog	Marson_2008	mES
Sox2	Marson_2008	mES
Oct 4	Marson_2008	mES
Tcf3	Marson_2008	mES
Stella (PGC)	Bian_2014	mES
Tet3	Bian_2014	mES
Tet 1	Williams_2011	mES
GRO-seq (active/paused PolII)	Min_2011	mES, pMEF
WCE/Input	Marson_2008	mES
	Mikkelsen_2007	mES
	Buecker_2014	mES

Table 23. Statistically significant and marginally significant mCA protein readers and their major functions of interest, summarised from Uniprot.

Binders		Function
<b>Major satellite probe</b>		
		Binds both 5-mC and 5hmC-containing DNA, no flanking preference. Mediates transcriptional repression; seems to regulate dendritic growth and spine maturation
MeCP2	Q9Z2D6	
Foxk1	P42128	Transcriptional regulator that binds to the upstream enhancer region (CCAC box) of myoglobin gene. Has a role in myogenic differentiation
Foxk2	F8VPY3	Recognizes the core sequence 5'-TAAACA-3'. At E12.5, expressed ubiquitously in the developing central nervous system.
Zfp646	Q6NV66	N/A
Oct4	P20263	Pluripotency factor
Six4	Q61321	Involved in skeletal muscle development. Also implicated in retina and kidney development
Puf60	Q3UEB3	Plol(U) binding splicing factor
Polr1c	F6ZXK9	Pol I and III subunit
Zscan26	F8WJ31	N/A
<b>L1 probe</b>		
Zbtb10	E9Q8X5	N/A
Dppa2	Q9CWH0	Developmental pluripotency-associated protein, may be involved in the maintenance of the active epigenetic status of genes
Dlst	Q9D2G2	Mitochondrial, 2-oxoglutarate dehydrogenase complex catalyzes the overall conversion of 2-oxoglutarate to succinyl-CoA and CO <sub>2</sub>
Creb1	Q01147	Cyclic AMP-responsive element-binding protein 1, transcription factor that stimulates transcription upon binding to the DNA cAMP response element (CRE), involved in the differentiation of adipose cells.
Rpl38	Q9JJI8	60S ribosomal protein L38

Table 24. Statistically significant and marginally significant mCA protein repellors and their major functions of interest, summarised from Uniprot.

Repellers	Function
<b>Major satellite probe</b>	
Zscan4f	Q3URS2 Acts as an activator of spontaneous telomere sister chromatid exchange (T-SCE) and telomere elongation during early development (ES, Blastocyst)
Zbtb43	G3X9N4 May be involved in transcriptional regulation.
Zbtb22	Q9Z0G7 May be involved in transcriptional regulation.
Rpa1	Q5SWN2 Replication protein A complex, binds and stabilizes single-stranded DNA intermediates upon replication and repair
Rpa2	Q3TE40 Plays also a role in base excision repair (BER) probably through interaction with UNG. May also play a role in telomere maintenance .
Rpa3	Q9CQ71 The RPA complex controls DNA repair and DNA damage checkpoint activation. Binds telomere ssDNA and even stronger telomere RNA, might have a role in
Hnrnpd	Q60668 telomere elongation.
Patz1	Q5NBY9 DNA binding protein, role in spermatogenesis, T-cell differentiation
<b>L1 probe</b>	
Pcbp3	P57722 Single-stranded nucleic acid binding protein that binds preferentially to oligo dC, iron chaperone for ferritin
Hnrnp11	Q921F4 RNA-binding protein that functions as regulator of alternative splicing for multiple target mRNAs
Ptbp1	Q922I7 Plays a role in pre-mRNA splicing and in the regulation of alternative splicing events. Activates exon skipping of its own pre-mRNA during muscle cell differentiation.

Table 25. GO and functional terms for nuclear pull-down identified mCA binders. Clustering was performed through the DAVID Functional Annotation Bioinformatics Microarray Analysis (<http://david.abcc.ncifcrf.gov/>)

<b>Binders</b>					
<b>Annotation Cluster 1</b>		<b>Enrichment Score: 2.48</b>	<b>Count</b>	<b>P_Value</b>	<b>Benjamini</b>
SP_PIR_KEYWORDS	nucleus		9	0,000031	0,001100
SP_PIR_KEYWORDS	dna-binding		6	0,000290	0,004900
GOTERM_MF_FAT	sequence-specific DNA binding		5	0,000790	0,028000
SP_PIR_KEYWORDS	Transcription		6	0,000850	0,009600
GOTERM_BP_FAT	positive regulation of transcription, DNA-dependent		4	0,001400	0,340000
GOTERM_MF_FAT	DNA binding		7	0,001500	0,026000
GOTERM_BP_FAT	positive regulation of RNA metabolic process		4	0,001500	0,190000
GOTERM_BP_FAT	transcription		6	0,001500	0,140000
GOTERM_BP_FAT	positive regulation of transcription		4	0,002100	0,140000
GOTERM_BP_FAT	positive regulation of gene expression		4	0,002300	0,130000
GOTERM_BP_FAT	positive regulation of nucleobase, nucleoside, nucleotide		4	0,002600	0,120000
GOTERM_MF_FAT	transcription factor activity		5	0,002700	0,032000
GOTERM_BP_FAT	positive regulation of nitrogen compound metabolic process		4	0,002800	0,110000
GOTERM_BP_FAT	positive regulation of macromolecule biosynthetic process		4	0,002900	0,100000
GOTERM_BP_FAT	positive regulation of cellular biosynthetic process		4	0,003200	0,100000
GOTERM_BP_FAT	positive regulation of biosynthetic process		4	0,003300	0,093000
GOTERM_BP_FAT	regulation of transcription		6	0,004300	0,110000
GOTERM_BP_FAT	positive regulation of macromolecule metabolic process		4	0,004700	0,110000
SP_PIR_KEYWORDS	transcription regulation		5	0,004900	0,033000
GOTERM_BP_FAT	regulation of transcription, DNA-dependent		5	0,006600	0,140000
GOTERM_BP_FAT	regulation of RNA metabolic process		5	0,007000	0,140000
SP_PIR_KEYWORDS	phosphoprotein		8	0,012000	0,065000
GOTERM_MF_FAT	transcription regulator activity		5	0,013000	0,110000
SP_PIR_KEYWORDS	repressor		3	0,016000	0,073000
GOTERM_BP_FAT	regulation of transcription from RNA polymerase II promot		3	0,048000	0,620000
SP_PIR_KEYWORDS	alternative splicing		6	0,050000	0,190000
UP_SEQ_FEATURE	splice variant		6	0,074000	0,990000
<b>Annotation Cluster 2</b>		<b>Enrichment Score: 0.49</b>	<b>Count</b>	<b>P_Value</b>	<b>Benjamini</b>
INTERPRO	Zinc finger, C2H2-like		3	0,074000	0,930000
INTERPRO	Zinc finger, C2H2-type		3	0,075000	0,730000
SMART	ZnF_C2H2		3	0,190000	0,910000
GOTERM_MF_FAT	zinc ion binding		3	0,540000	1,000000
GOTERM_MF_FAT	metal ion binding		4	0,660000	1,000000
GOTERM_MF_FAT	transition metal ion binding		3	0,670000	1,000000
GOTERM_MF_FAT	cation binding		4	0,670000	0,990000
GOTERM_MF_FAT	ion binding		4	0,680000	0,990000

Table 26. GO and functional terms for nuclear pull-down identified mCA repellors. Clustering was performed through the DAVID Functional Annotation Bioinformatics Microarray Analysis (<http://david.abcc.ncifcrf.gov/>)

<b>Repellers</b>				
<b>Annotation Cluster 1</b>		<b>Enrichment Score: 3.25</b>	<b>Count</b>	<b>P_Value</b>
	<b>Benjamini</b>			
KEGG_PATHWAY	Mismatch repair	3	0,000014	0,000056
KEGG_PATHWAY	Homologous recombination	3	0,000021	0,000043
KEGG_PATHWAY	DNA replication	3	0,000036	0,000048
KEGG_PATHWAY	Nucleotide excision repair	3	0,000055	0,000055
INTERPRO	Nucleic acid-binding, OB-fold	3	0,000370	0,010000
SP_PIR_KEYWORDS	dna replication	3	0,000790	0,007500
GOTERM_BP_FAT	DNA replication	3	0,003300	0,180000
GOTERM_BP_FAT	DNA metabolic process	3	0,024000	0,510000
SP_PIR_KEYWORDS	phosphoprotein	5	0,400000	0,590000
<b>Annotation Cluster 2</b>		<b>Enrichment Score: 1.71</b>	<b>Count</b>	<b>P_Value</b>
	<b>Benjamini</b>			
INTERPRO	Zinc finger, C2H2-type/integrase, DNA-binding	4	0,001800	0,024000
INTERPRO	BTB/POZ	3	0,001800	0,016000
INTERPRO	BTB/POZ-like	3	0,003500	0,023000
INTERPRO	BTB/POZ fold	3	0,003500	0,023000
INTERPRO	Zinc finger, C2H2-like	4	0,003900	0,021000
INTERPRO	Zinc finger, C2H2-type	4	0,004000	0,018000
SMART	BTB	3	0,005500	0,038000
COG_ONTOLOGY	Transcription / Cell division and chromosome partitioning	3	0,005900	0,012000
SMART	ZnF_C2H2	4	0,006800	0,024000
SP_PIR_KEYWORDS	zinc	5	0,010000	0,038000
GOTERM_MF_FAT	zinc ion binding	5	0,026000	0,350000
GOTERM_BP_FAT	regulation of transcription	5	0,029000	0,440000
SP_PIR_KEYWORDS	metal-binding	5	0,034000	0,090000
SP_PIR_KEYWORDS	transcription regulation	4	0,037000	0,085000
SP_PIR_KEYWORDS	Transcription	4	0,052000	0,110000
GOTERM_MF_FAT	transition metal ion binding	5	0,053000	0,450000
GOTERM_BP_FAT	transcription	4	0,075000	0,680000
SP_PIR_KEYWORDS	zinc-finger	3	0,120000	0,210000
GOTERM_MF_FAT	transcription regulator activity	3	0,160000	0,680000
GOTERM_MF_FAT	metal ion binding	5	0,180000	0,650000
GOTERM_MF_FAT	cation binding	5	0,180000	0,610000
GOTERM_MF_FAT	ion binding	5	0,190000	0,570000
<b>Annotation Cluster 3</b>		<b>Enrichment Score: 1.34</b>	<b>Count</b>	<b>P_Value</b>
	<b>Benjamini</b>			
GOTERM_CC_FAT	chromosome	3	0,002700	0,065000
GOTERM_CC_FAT	non-membrane-bounded organelle	3	0,063000	0,560000
GOTERM_CC_FAT	intracellular non-membrane-bounded organelle	3	0,063000	0,560000
SP_PIR_KEYWORDS	phosphoprotein			
<b>Unclassified</b>		<b>Count</b>	<b>P_Value</b>	<b>Benjamini</b>
GOTERM_MF_FAT	DNA binding	8	N/A	0,000180
SP_PIR_KEYWORDS	nucleus	8	N/A	0,009000
SP_PIR_KEYWORDS	acetylation	6	N/A	0,019000
SP_PIR_KEYWORDS	dna-binding	5	N/A	0,016000
SP_PIR_KEYWORDS	rna-binding	3	N/A	0,072000
GOTERM_MF_FAT	RNA binding	3	N/A	0,390000
GOTERM_BP_FAT	regulation of RNA metabolic process	3	N/A	0,940000

## 10.2 Additional figures

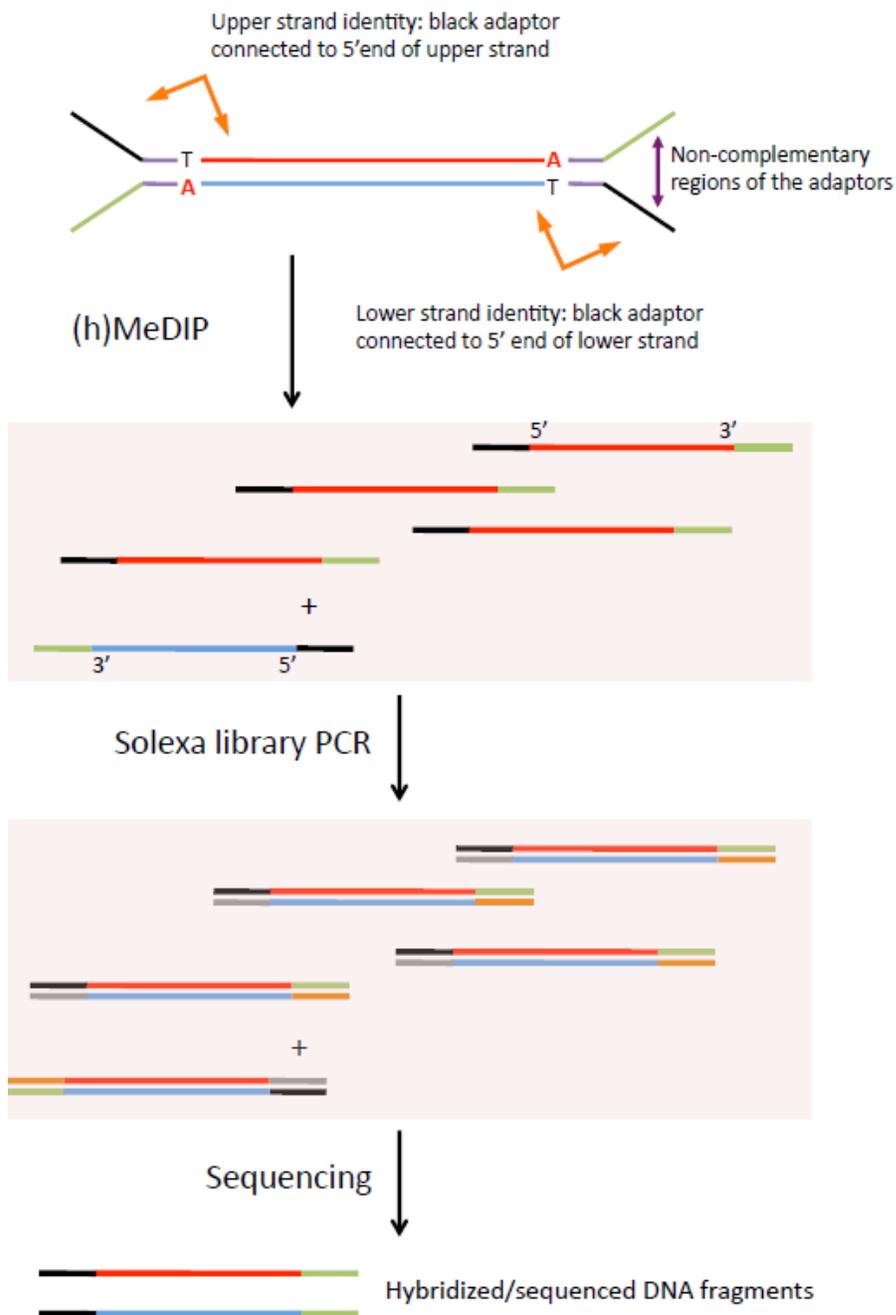


Figure 71. A schematic representation of the conservation of strand information in paired end NGS datasets. Adapter sequence from the original strands differs from the adapter sequences of their PCR products, making it possible to map original forward and reverse strands. Figure borrowed from Ficiz, Branco et al. 2011.

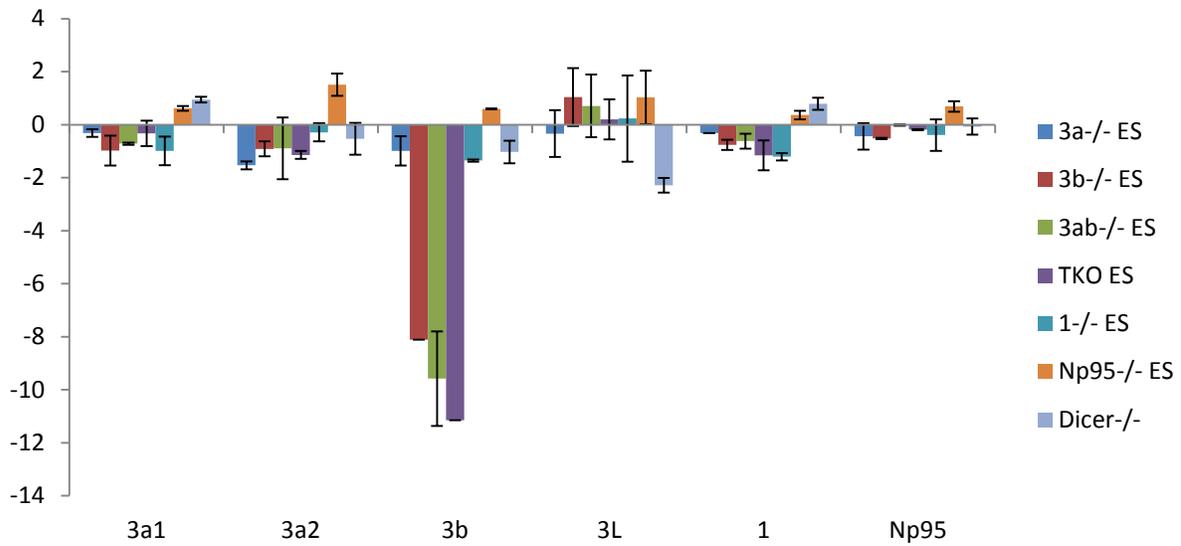


Figure 72. Log2 expression of Dnmts and Np95 in a panel of Dnmt-KO and Np95-KO mES cell lines used for this project. The Dnm3a, Dnm3b, Dnm3ab and Dnmt1 KO were normalised to J1 WT, Np95-KO to E14 WT, and the Dicer-KO clones have their own WT controls.

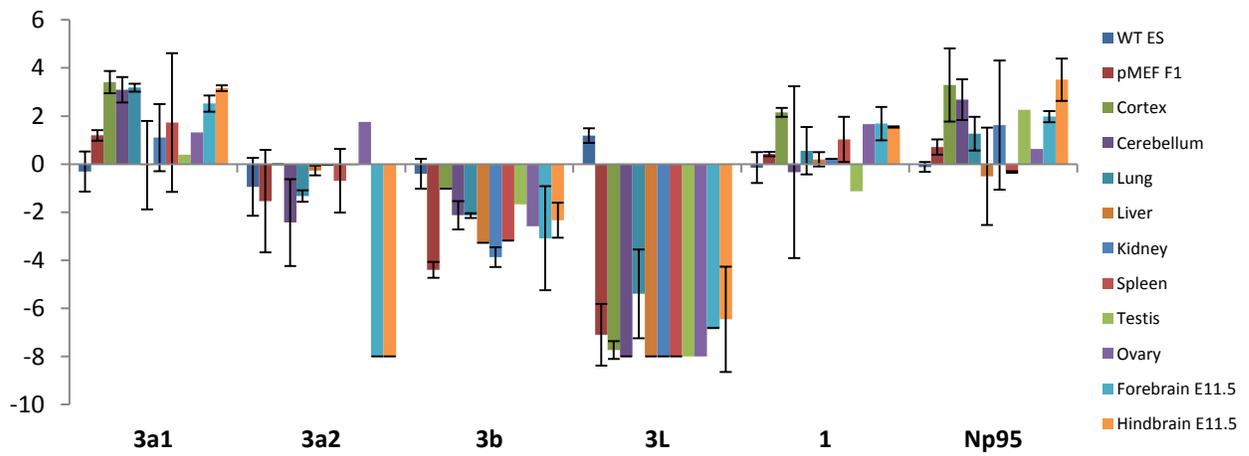


Figure 73. Log2 expression of Dnmts and Np95 in a panel of tissues. All were normalised to J1 and E14 WT mESCs.

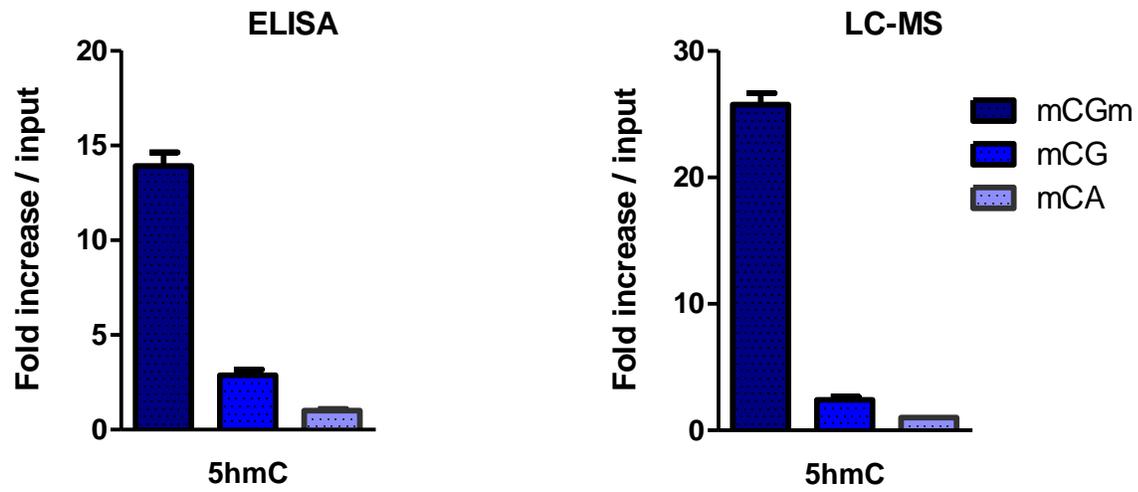


Figure 74. A comparison between ELISA and LC-MS measurements of absolute gained levels of 5hmC after Tet1 oxidation in three different methylation contexts.

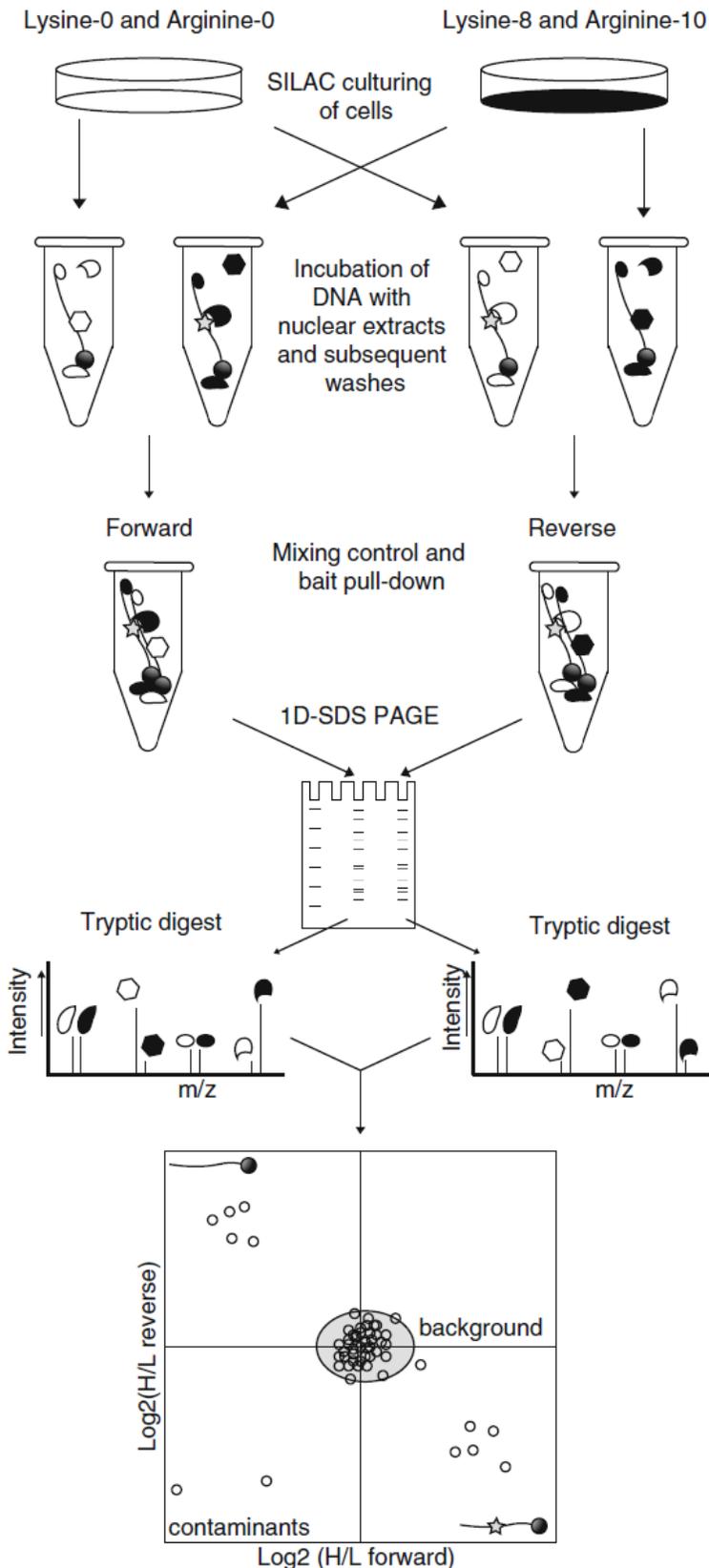


Figure 75. A schematic representation of the SILAC-based technique workflow for identification of nuclear protein binders.

Bait and control DNA are separately incubated with light and heavy nuclear extracts (NE) from cells grown in light or heavy SILAC media. Bait DNA incubated with heavy NE is combined with control DNA incubated with light NE (forward experiment) and bait DNA incubated with light NE is combined with control DNA incubated with heavy NE (reverse experiment). The two experiments are fractionated using 1D SDS-PAGE, followed by in-gel digestion and mass spectrometry. The results can be visualised in a scatterplot. Specific interactors of the bait DNA are located in the lower right quadrant (high forward ratio, low reverse ratio) whereas proteins that are repelled by the bait DNA end up in the upper left quadrant (low forward ratio, high reverse ratio). High-abundant background proteins and nonspecific DNA binders cluster together around the origin of the graph.

Adapted from (Spruijt et al. 2013).