Establishment and Maintenance of the DNA Methylation Pattern in the Human *α*-globin Cluster

Ricarda Gaentzsch

University College



A thesis submitted for the degree of

Doctor of Philosophy

of the

University of Oxford

MRC Molecular Haematology Unit

Weatherall Institute of Molecular Medicine

University of Oxford

Word number: ~43,300

Trinity Term 2013

Declaration

I, Ricarda Gaentzsch, declare that the work presented in this thesis is my own unless otherwise stated.

To my Family

Establishment and Maintenance of the DNA Methylation Pattern in the Human α-globin Cluster

Ricarda Gaentzsch
University College

Doctor of Philosophy
Hilary Term 2014

Abstract

DNA methylation is an epigenetic modification that plays an important role in development and differentiation. The patterns of DNA methylation are largely established in early embryogenesis and maintained during development. Abnormal DNA methylation patterns have been associated with many human diseases, including cancer. Despite its importance, little is currently known about the mechanisms that determine DNA methylation patterns throughout the genome. To shed light on the molecular mechanisms that regulate DNA methylation, this study investigates whether DNA methylation patterns are established and maintained normally when human DNA is placed into a heterologous murine environment as opposed to its natural, endogenous chromosomal environment.

Here, a previously generated transgenic mouse model, containing 117 kb of human DNA bearing the human *α*-globin cluster and all of its known regulatory elements, was analysed. The pattern of DNA methylation of the endogenous human *α*-globin cluster was compared with that of the transgenic cluster in the background of mouse embryonic stem cells (ESCs) and tissues. It was found that, although the normal human DNA methylation pattern was largely established and maintained in a mouse background, the region immediately around the human *α*-globin genes themselves is generally less methylated in mouse compared to human ESCs. It was found that regions adjacent and up to 2kb from the CpG islands (CGIs), so-called CGI shores, were unusually hypomethylated: this seems to be the result of an extension of CGIs in humanised mouse (hm) ESCs compared to human (h) ESCs. Furthermore, this hypomethylation appeared to increase during development in both erythroid and non-erythoid cells.

To identify any *cis*-regulatory sequences responsible for the hypomethylated state of human CGI shores in the mouse, 2-4 kb human test sequences containing the CGI associated with the human *α*-globin 2 (*α*2) gene and its adjacent hypomethylated shore were re-integrated into the mouse *α*-globin locus via recombination-mediated cassette exchange (RMCE). Human CGI shores became hypomethylated in the context of the re-integrated test sequences, indicating that the appearance of hypomethylation is determined by the underlying human DNA sequence in the test fragments.

In summary, the data presented here reveal that human CGIs become extended when placed in a mouse background leading to hypomethylation of human CGI shores in the mouse compared to the pattern of methylation at the normal endogenous human locus. These findings suggest that species-specific factors determine DNA methylation near CGIs. The transgenic mouse model provides an excellent system to dissect out species-specific regulation of CGI shore methylation. Furthermore, this study lays the foundation for future experiments addressing the role of DNA methylation in regulating human gene expression in the murine context, and examining the validity of transgenic mouse models for the study of human gene regulation.

Table of Contents

1 INTRODUCTION 1

1.1 EPIGENETICS 1
1.1 CHROMATIN 1
1.1.1 HISTONE MODIFICATIONS 3
1.2 DNA METHYLATION 6
1.2.1 THE MARK 6
1.2.2 DNA METHYLTRANSFERASES 7
1.3 ROLES OF DNA METHYLATION IN MAMMALS 15
1.3.1 X-INACTIVATION 16
1.3.2 GENOMIC IMPRINTING 16
1.3.3 TRANSPOSON SILENCING 17
1.3.4 GENE REGULATION 18
1.3.5 MAINTENANCE OF GENOME INTEGRITY 21
1.3.6 MOLECULAR EFFECTS OF DNA METHYLATION 21
1.4 GENOME-WIDE DNA METHYLATION PATTERNS 22
1.4.1 MAMMALIAN DNA METHYLATION PATTERNS 22
1.4.2 PHYLOGENY OF DNA METHYLOMES 30
1.4.3 NON-CPG METHYLATION 32
1.4.4 5HMC DISTRIBUTION 34
1.5 ESTABLISHMENT AND MAINTENANCE OF DNA METHYLATION PATTERNS 36
1.5.1 EPIGENETIC REPROGRAMMING 36
1.5.2 DNA DEMETHYLATION 37
1.5.3 ESTABLISHMENT OF DNA METHYLATION PATTERNS 41
1.5.4 MAINTENANCE OF DNA METHYLATION PATTERNS 41
1.5.5 REVISED MODEL OF ESTABLISHMENT AND MAINTENANCE OF METHYLATION 42
1.5.6 RECRUITMENT OF DNMTS TO DNA 43
1.5.7 KEEPING CGIS IN AN UNMETHYLATED STATE 44
1.6 THE A-GLOBIN GENES 48
1.6.1 GENOMIC ORGANIZATION 48
1.6.2 THE HUMAN A-GLOBIN GENES 49
1.6.3 REGULATION OF A-GLOBIN EXPRESSION 50
1.6.4 A TRANSGENIC MOUSE MODEL FOR THE HUMAN A-GLOBIN GENES 51
1.7 TRANSGENIC MOUSE MODELS TO STUDY HUMAN DISEASE 52
1.7.1 KEY OPEN QUESTIONS 54
1.7.2 OUTLINE OF THE THESIS 54

2 MATERIALS AND METHODS 59

2.1 MATERIALS 59
2.1.1 CHEMICALS, ENZYMES AND REAGENTS 59
2.1.2 KITS 65
2.1.3 EQUIPMENT 66
2.1.4 BUFFERS 67
2.1.5 PLASMIDS, CELL LINES AND TISSUES 69
2.2 DNA PROCEDURES 71
2.2.1 EXTRACTION OF GENOMIC DNA FROM CELL PELLETS (FROM EUKARYOTIC CELLS) 71
2.2.2 DNA EXTRACTION FROM BACTERIAL CULTURES (MINIPREPS) 72
2.2.3 BISULFITE CONVERSION 72
2.2.4 PCR AMPLIFICATION 73
2.2.5 RESTRICTION DIGESTION OF DNA 75
2.2.6 SOUTHERN BLOTTING 76

2.2.7 RADIOACTIVE PROBE LABELLING 77
2.3 CELL CULTURE PROCEDURE 78
2.3.1 ESC CULTURE CONDITIONS 78
2.3.2 ELECTROPORATION 78
2.4 BACTERIAL PROCEDURES 79
2.4.1 LIGATION INTO PGEM®-T EASY VECTOR 79
2.4.2 TRANSFORMATION OF COMPETENT CELLS 79
2.5 TARGETED BISULFITE SEQUENCING (TBS) 80
2.5.2 DATA ANALYSIS OF BISULFITE SEQUENCING DATA 84
2.6 TARGETED INTEGRATION OF TEST SEQUENCES INTO MES RMCE 85
2.6.1 MAPPING OF THE RMCE FRAGMENTS 87
2.7 BIOINFORMATIC ANALYSIS OF BIOCAP DATA 87

3 DNA METHYLATION PATTERNS OF THE HUMAN *A*-GLOBIN CLUSTER IN HUMAN AND MOUSE 89

3.1 INTRODUCTION 89
3.2 RESULTS 90
3.2.1 HUMAN CGIS MAINTAIN THEIR UNMETHYLATED STATE IN HUMANISED MOUSE COMPARED TO HUMAN DIFFERENTIATED CELLS 90
3.2.2 SEQUENCES SURROUNDING THE HUMAN A-GLOBIN GENES ARE HYPOMETHYLATED IN HUMANISED MOUSE VERSUS HUMAN TISSUES 92
3.2.3 HIGH RESOLUTION PROFILING OF DNA METHYLATION PATTERNS ACROSS THE HUMAN TRANSGENIC REGION IN ESCS 97
3.2.4 CHARACTERISATION OF METHYLATION CHANGES IN ESCS 101
3.2.5 HYPOMETHYLATION BECOMES EXACERBATED IN BRAIN 116
3.2.6 HYPOMETHYLATION OF CGI SHORES IN DIFFERENTIATED CELLS. 121
3.2.7 DNA METHYLATION PATTERN AROUND THE HUMAN A-GLOBIN GENES AT SINGLE NUCLEOTIDE RESOLUTION IN DEVELOPMENT 123
3.2.8 OVERVIEW OF THE CHANGES OF METHYLATION AROUND CGIS 126
3.3 DISCUSSION 128
3.3.1 PUTATIVE SPECIES-SPECIFIC DIFFERENCES IN THE TRANS-ACTING ENVIRONMENT OF HUMAN AND MOUSE CELLS 128
3.3.2 ESTABLISHMENT AND MAINTENANCE OF METHYLATION PATTERNS IN THE HUMANISED MOUSE MODEL 130
3.3.3 DIFFERENTIAL CGI SHORE METHYLATION 130
3.3.4 DNA METHYLATION ANALYSIS BY BISULFITE SEQUENCING AND REA 131
3.3.5 GENE BODY METHYLATION IN HUMAN ESCS 133
3.3.6 METHYLATION OF REPETITIVE SEQUENCES 134

4 DISSECTING OUT THE ROLE OF CGI H IN DETERMINING THE METHYLATION STATE OF ITS ADJACENT SHORES 136

4.1 INTRODUCTION 136
4.2 RESULTS 137
4.2.1 CHOICE OF INTEGRATION SITE 137
4.2.2 THE MOUSE A-GLOBIN LOCUS DOES NOT INDUCE HYPOMETHYLATION OF A METHYLATED FRAGMENT 140
4.2.3 CHOICE OF TEST SEQUENCES 144
4.2.4 DNA METHYLATION LEVELS OF THE INSERTED DNA FRAGMENTS 147
4.3 SUMMARY AND CONCLUSIONS 151
4.3.1 SEQUENCE DETERMINANTS OF DNA METHYLATION AROUND CGIS 152
4.3.2 ROLE OF CGI H IN DETERMINING THE HYPOMETHYLATED STATE OF ITS FLANKING DNA 152
4.3.3 ROLE OF THE A2 PROMOTER IN INITIATING A HYPOMETHYLATED STATE 153

5 NON-METHYLATED DNA PROFILE AT HUMAN CGIS IN MOUSE VERSUS HUMAN 154

5.1 INTRODUCTION 154
5.2 RESULTS 156
5.2.1 CHOICE OF METHOD AND CELL TYPES 156
5.2.2 MAPPING OF HUMAN AND TC1 LIVER AND TESTES DATASETS 156
5.2.3 COMPARISON OF SEQUENCE READS ACROSS CGIS 160
5.3 DISCUSSION: 167

6 FINAL DISCUSSION 170

6.1 PREVIOUS DNA METHYLATION STUDIES OF HUMAN TRANSGENES IN THE MOUSE 170
6.2 TRANSGENIC MOUSE MODEL TO DISSECT OUT REGULATORY FACTORS OF CGI SHORE METHYLATION 171
6.3 SEQUENCE DETERMINANTS OF DNA METHYLATION 172
6.4 MOLECULAR MECHANISM OF HYPOMETHYLATION 174
6.5 ROLE OF SPREADING IN DETERMINING METHYLATION PATTERNS 175
6.6 IDENTIFICATION OF DNA BINDING FACTORS 176
6.7 CGI SHORE HYPOMETHYLATION AND GENE EXPRESSION 177
6.8 FUNCTIONAL RELEVANCE OF CGI SHORE HYPOMETHYLATION 177
6.9 GENERALITY OF THE PHENOMENON 178

7 APPENDIX 180

7.1 TARGETED BISULFITE SEQUENCING (TBS) 180
7.1.1 COVERAGE OF CPGS IN ESCS AND BRAIN USING TBS 192
7.1.2 TESTING TBS FOR BIASES IN THE EFFICIENCY TO CAPTURE DIFFERENT TYPES OF DNA SEQUENCES 195
7.2 GEL LOADING CONTROLS 198
7.3 DIFFERENTIATION STAGE OF HUMAN AND MOUSE ERYTHROBLASTS 199
7.4 LIST OF PRIMERS 200
7.5 BIBLIOGRAPHY 206

Table of Figures

Figure 1: Schematic of the α-globin gene cluster in human and mouse. 57
Figure 2: Overview of human and mouse α-globin regions. 58
Figure 3: Principle of the recombination-mediated cassette exchange (RMCE) system. 86
Figure 4: Bioinformatic pipeline used for analysis of BioCAP data. 88
Figure 5: CGI H and I (associated with α2 and α1, respectively) are unmethylated in human and humanised mouse tissues. 91
Figure 6: Hypomethylation near the human α-globin genes in the humanized mouse model. 95
Figure 7: Hypomethylation near the α-globin genes in human/mouse hybrid cells. 96
Figure 8: Overview of DNA methylation pattern across the human transgenic region in human and humanised mouse ESCs. 100
Figure 9: Boxplot of the DNA methylation difference at CpG sites within unique sequences, repetitive elements (repeat) and Alu repeats (Alu). 102
Figure 10: Comparison of the DNA methylation difference between human and humanised mouse ESC with the GC and CpG content of the DNA surrounding each CpG site. 106
Figure 11: Boxplots showing the DNA methylation levels at CpG sites in CGIs, CGI shores, and regions outside of CGI shores in ESCs. 108
Figure 12: Verification of DNA methylation changes in the vicinity of CGIs by conventional bisulfite sequencing (CBS). 110
Figure 13: Single base pair DNA methylation profile of the human α-globin genes in human and humanised mouse ESCs. 112
Figure 14: DNA methylation pattern around CGI E (associated with the NPRL3 promoter) in human and humanised mouse ESCs. 114
Figure 15: Hypomethylation of the CGI shore upstream of CGI J in the re-derived hmES clone A (hmES cl.A). 115
Figure 16: Overview of the DNA methylation patterns in h and hm brain in comparison to the pattern in h and hm ESC. 118
Figure 17: Analysis of the methylation difference in ESCs and brain. 120
Figure 18: Comparison of the methylation difference between human and humanised mouse in ESC and brain. 121
Figure 19: DNA methylation in human and humanised mouse brain at CGIs, CGI shores and outside of CGIs and their shores. 122
Figure 20: Methylation pattern at CGI J and its upstream shore in h and hm brain in comparison to ECSs. 123
Figure 21: DNA methylation pattern of the α-globin cluster during development. 125
Figure 22: Relative location of the methylation difference at individual CpG sites with respect to their nearest CGI. 127
Figure 23: Relative location of the methylation difference relative to CGI E in h and hm ESC. 128
Figure 24: Mouse α-globin locus showing the site of integration of the human DNA fragments. 139
Figure 25: DNA methylation levels at the mouse DNA flanking the insertion site in wildtype mouse ES (mES WT) and in the targeted mouse ES cell line (mES RMCE). 139
Figure 26: Integration of fragment A into the pML7 vector to create the pML7-A donor plasmid. 141
Figure 27: Confirmation of integration of control fragment A. 142
Figure 28: DNA methylation levels of a naturally methylated fragment retains its methylation levels when integrated into the α-globin locus. 143
Figure 29: Location of the human test fragments targeted to the mouse α-globin locus. 144
Figure 30: Schematic showing the generation of the pML7-B/C donor plasmid. 145
Figure 31: Verification of correct targeting of the selected DNA fragments containing human CGI H and surrounding sequence to the mouse α-globin locus. 146
Figure 32: DNA methylation levels in selected human DNA fragments targeted to the mouse genome. 147
Figure 33: Bisulfite sequencing of a methylated control sample. 148
Figure 34: DNA methylation analysis at CGI neighboring regions. 150
Figure 35: DNA methylation levels at the mouse DNA flanking the insertion site. 151
Figure 36: Example locus showing the mapped sequence reads in UCSC for human and Tc1 liver, testes and input DNA. 160
Figure 37: Average read profile at CGIs in human and Tc1 liver (a) and testes (b). 161

Figure 38: Read density plot of Tc1 liver (two replicates), and human liver (two replicates) as well as Tc1 and human input samples. _____ *162*
Figure 39: Read density plot of Tc1 testes (two replicates), and human testes (two replicates) as well as Tc1 and human input samples. _____ *163*
*Figure 40: Example locus showing a loss of sequence reads in Tc1 over human liver.*_____ *164*
*Figure 41: Example locus showing a gain of sequence reads in Tc1 over human liver.*_____ *165*
*Figure 42: Workflow of targeted bisulfite sequencing (TBS).*_____ *183*
*Figure 43: Map of the human α-globin gene cluster showing the capture area within the transgenic region.*_____ *184*
*Figure 44: Close-up view of the capture area showing a 5kb window around α2 (grey) illustrating the distance between the capture oligos (red bars).*_____ *184*
*Figure 45: Enrichment of captured over uncaptured material.*_____ *185*
Figure 46: Verification of conversion of DNA by treatment with sodium bisulfite. _____ *186*
*Figure 47: Bioinformatics pipeline.*_____ *188*
Figure 48: Close-up view of the sequence reads obtained by TBS and mapped to the human genome. _____ *191*
Figure 49: Example locus of read sequences before and after collapsing of data. _____ *192*
*Figure 50: Efficiency of the method to capture CpG sites of high GC content.*_____ *197*
Figure 51: Ethidium bromide-stained gel corresponding to Southern shown in Figure 5. _____ *198*
Figure 52: Ethidium bromide-stained gel corresponding to Southern shown in Figure 6. _____ *198*
Figure 53: Ethidium bromide-stained gel corresponding to Southern shown in Figure 7. _____ *198*
*Figure 54: Differentiation stage of human erythroblasts.*_____ *199*
*Figure 55: Differentiation stage of humanised mouse erythroblasts.*_____ *199*

Acknowledgments

My first thanks go to my supervisors Richard Gibbons and Doug Higgs who gave me the opportunity to carry out the work presented in this thesis. I am thankful for their mentorship, time, valuable comments and probing questions throughout this project. Their enthusiasm and passion for research was highly motivational for me and I appreciate to have had the chance to be their student and work in such a pleasant and stimulating environment as the Weatherall Institute of Molecular Medicine.

I wish to express my warm and sincere thanks to David Garrick for his helpful support and scientific input, as well as to my college advisor, Catherine Pears, for her helpful advice. I am greatly indebted to Tom Milne, who has been available for help and support during the most challenging stages of this project. His constructive criticism and motivational support was greatly encouraging to me.

My deepest appreciation goes to all past and present members of the Gibbons' and Higgs' labs for creating a friendly and welcoming working environment. Without their scientific expertise and inspiration this work would not have been possible and it has been a great pleasure to work with them. In particular, I owe my profound gratitude to Helena Ayyub for much help with the Southern blots as well as much technical advice, Jackie Sharpe, Jackie Sloane-Stanley and Sue Butler for invaluable assistance with tissue culture and animal work. I am grateful to Danuta Jeziorska and David Clynes for their consistent help and advice as well as proofreading part of the manuscript, Noemi Roy, who introduced me to DNA methylation analysis and taught me many of the methods used in this thesis, Caroline Scott for the development of the targeted bisulfite sequencing technique, and Hannah Long for sharing her Tc1 data with me. I am thankful to Andrew Smith for providing me with the transgenic mouse model as well as human ESC lines and Magnus Lynch for the development of the RMCE system and providing me with some of the cell lines used in this study.

I would also like to give my sincere thanks to the computational bioinformatics research group (CBRG), who engineered the bioinformatics tools used in this study, helped me with the bioinformatic analyses carried out in this thesis, and

provided trainings and guidance during the analysis. In particular, I am greatly indebted to Jim Hughes for helping me with the pile-up analysis, generating heat maps and for great professional advice, as well as Nicki Gray for her invaluable assistance in analysing bisulfite sequencing data and for introducing me to bioinformatics.

My thanks go to Lorna Gregory, and Helen Butler from the Wellcome Trust Center for Human Genetics (WTCHG) for carrying out high-throughput sequencing as part of this work, as well as to the Sanger sequencing lab, specifically Tim Rostron and John Frankland, for their sequencing assistance, Joern Toedling of Bayer Schering Pharma AG for statistical advice, and Christina Rode for her help in the re-derivation of mouse ESC lines.

Last but not least, I would like to express my greatest gratitude to my family and friends – in particular to my parents and my brother Peer, as well as to my friend Bet. None of this work would have been possible without their incessant support and encouragement.

Finally, I wish to acknowledge the Medical Research Council, who funded this project.

Abbreviations

5caC	5-carboxylcytosine
5fC	5-formylcytosine
5hmC	5-hydroxymethylcytosine
5hmU	5-hydroxymethyluracil
5mC	5-methylcytosine
6TG	6-thioguanine
Ac	acetylation
AID	activation-induced deaminase
APOBEC	apolipoprotein B mRNA editing enzyme catalytic polypeptite
BACs	bacterial artificial chromosomes
BER	base excision repair
BSA	bovine serum albumin
C	cytosine
cat.no.	catalogue number
CBRG	Computational Biology Research Group
CBS	conventional bisulfite sequencing
CGI	CpG Island
chr	chromosome
cl.	clone
cpm	counts per minute
cRNA	complementary RNA
dd	double desalted
DMSO	Dimethyl sulphoxide
DNA	deoxyribonucleic acid
DNMT	DNA methyltransferase
dNTP	deoxyribonucleotide triphosphate
DTT	Dithiothreitol
EBV	Epstein-Barr virus
EDTA	ethylenediaminetetraacetic acid disodium
ESC	embryonic stem cell
FACS	fluorescence-activated cell sorting

FBS	fetal bovine serum
G	guanosine
A	adenosine
T	thymidine
GMEM	Glasgow minimum essential medium
h	human
H3K27me3	trimethylation of lysine residue 27 in histone H3
H3K4me(1/2/3)	(mono-/ di-/ tri-)methylation at lysine residue 4 in histone
H3ac	acetlyation of histone H3
H3K9	trimethylation of lysine residue 9 in histone H3
H4ac	acetlyation of histone H4
HAT	histone acetyltransferases
HCP	high CpG density promoter
HDAC	histone deacetylases
hm	humanised mouse
hs21	human chromosome 21
ICF	immunodeficiency, centromere instability and facial anomalies
ICP	intermediate CpG density promoter
K	lysine
kb	kilobases
L	liter
LB	lysogeny broth
LCP	low CpG density promoter
LIF	leukemia inhibitory factor
LINE	long interspersed nuclear element
m	mouse
M	Molar
m.ery	mouse erythroid cells
max	maximum
MBD	methyl-CpG binding domain
MCS-R	multispecies conserved sequence (regulatory element)
MEL	mouse erythroleukemia

metdiff	methylation difference
min	minimum
ml	milliliter
MLL	mixed-lineage leukaemia 1
MQ	MilliQ
MRC	Medical Research Council
n	number (total number of datapoints analysed)
N-EBV	human B-lymphocytes (from normal individuals transformed with EBV)
NEB	New England Biolabs
NLS	nuclear localization signal
ns	not significant
PACs	P1-derived artificial chromosomes
PBS	phosphate buffered saline
PCNA	proliferating cell nuclear antigen
PCR	polymerase chain reaction
PenStrep	Penicillin-Streptmycin
PEV	position-effect variegation
PGCs	primordial germ cells
PolII	RNA polymerase II
PRC1	Polycomb repressive complex 1
PRC2	Polycomb repressive complex 2
qPCR	quantitative real-time PCR
R	arginine
REA	methylation-restriction enzyme analysis
RMCE	recombination-mediated cassette exchange
RNA	ribonucleic acid
rpm	revolutions per minute
SDS	sodium dodecyl sulphate
SINE	short interspersed nuclear element
SSC	saline sodium citrate
TAE	Tris-acetate-EDTA
TBE	Tris-Borate-EDTA
TBS	targeted bisulfite sequencing

TDGs	thymine-DNA glycosylases
TE	Tris-EDTA
TF	transcription factor
TRD	target recognition domain
$tRNA^{Asp}$	transfer RNA Aspartate
TSS	transcription start site
U	unit(s)
UCSC	University of California, Santa Cruz
UREs	upstream regulatory elements
UV	ultraviolet
V	Volt
WT	wildtype
WTCHG	Wellcome Trust Centre for Human Genetics
X-gal	(5-bromo-4-chloro-3-indolyl-ß-D-galactopyranoside)
Xa	active X chromosome
Xi	inactive X
Xist	X inactive specific transcript
YACs	yeast artificial chromosomes
ZF-CXXC	zinc finger CXXC
µl	microliter

1 Introduction

1.1 Epigenetics

Since the human genome was sequenced more than 10 years ago, much research has been focused on trying to understand how the human genetic code is read and expressed by the cell, in other words, how it is regulated. The relevance of genome regulation becomes apparent when one considers the fact that most cells in the human body share the same genetic information. Naturally, the question arises as to how cells with identical genomes can carry out different functions. A link between genotype and phenotype that cannot be explained by genetic principles is historically described as 'epigenetics', a term that is often ascribed to Conrad Waddington (Waddington 1942). Literally translated 'epigenetics' refers to the code that sits on top (greek "epi" means above or around) of the genetic code. In a broad molecular sense, it was defined as "the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence" (Riggs, 1975). Certainly, the list of definitions for the term epigenetics is vastly expanding, reflecting the fast progress of the field and our increasing understanding of the molecular basis of epigenetics. In the following, I will give a brief introduction to the molecular principles that form the basis of epigenetics.

1.1 Chromatin

In eukaryotes, DNA is efficiently packaged into the nucleus of a cell by chromatin. Chromatin refers to a highly organised structure consisting of DNA and its associated protein scaffold. The fundamental unit of chromatin is the

nucleosome, which consists of a histone octamer and 146 bp of DNA wrapped around it (Luger et al, 1997). A histone octamer consists of two copies of each of the core histones H2A, H2B, H3, and H4. The core histones have structured globular domains and flexible amino-terminal tails that protrude from the nucleosome and can be modified in various ways, including acetylation, methylation, phosphorylation and ubiqitylation (for review see (Zentner & Henikoff, 2013)). Nucleosomes are themselves assembled into higher order structures by coiling around one another and associating with linker histones, such as histone H1 (for review see (Maeshima et al, 2010)). In this way, approximately 2 meters of genomic DNA is folded into the nucleus. Chromatin compaction does not only package DNA into the nucleus, but it also plays an important role in the regulation of gene expression by restricting access of the transcriptional machinery to the DNA. High levels of chromatin condensation, called heterochromatin, are hence associated with inactive genomic sequences, whereas the uncondensed accessible form of chromatin, called euchromatin, is a characteristic feature of regions of active genes (Kouzarides, 2007). Heterochromatin can be further subdivided into facultative heterochromatin, which is a temporarily condensed form characteristic of genomic regions containing developmentally regulated genes and constitutive heterochromatin, a permanently condensed form found in centromeric and telomeric regions (Bannister & Kouzarides, 2011). Regulation of chromatin structure is achieved by various mechanisms, including modifications of histone tails, which will be discussed below.

1.1.1 Histone modifications

Overall, more than 100 different histone modifications have been discovered to date (Tan et al, 2011; Zentner & Henikoff, 2013). They influence chromatin structure by modulating the association of histones with the DNA. The effects can be direct or indirect via the recruitment of other proteins that recognise specific chromatin marks. For example, acetylation of histones H3 and H4 (H3Ac and H4Ac) by histone acetyltransferases (HATs) directly affects chromatin structure by adding a negative charge to the histone tails, leading to a more loose association of the histones with the negatively charged DNA backbone (Bannister & Kouzarides, 2011). Histone acetylation acts, however, also indirectly via recruitment of proteins that contain a bromodomain, including HATs and chromatin remodelling complexes (Bannister & Kouzarides, 2011). Histone acetylation is a mark of actively transcribed promoters, and also occurs at remote regulatory regions (Wang et al, 2008). Responsible for the removal of acetyl groups from H3 and H4 are so-called histone deacetylases (HDACs).

In contrast to acetylation, the addition of methyl groups does not alter the charge of a histone tail, and hence, histone methylation is thought have a less direct effect than acetylation. Histones can be methylated single or multiple times: mono- and dimethylation occurs at lysine (K) and arginine (R) residues, while trimethylation is found only at lysine residues. Trimethylation of lysine 9 in histone H3 (H3K9me3), for example, a hallmark of heterochromatin, is bound by the chromodomain of heterochromatin protein 1 (HP1), and this is thought to be important for the maintenance of heterochromatin in pericentric regions (Bannister & Kouzarides, 2011; Bernstein et al, 2006b; Jenuwein &

Allis, 2001; Kouzarides, 2007). Although H3K9me3 is generally regarded as a repressive mark and a hallmark of constitutive heterochromatin, it is also found in the bodies of transcriptionally active genes, where it co-localizes with HP1γ in a transcription-dependent manner (Vakoc et al, 2005). Enzymes responsible for mediating H3K9me2/3 in human are G9A, SETDB1 and SUV39H1 and SUV39H2 (Peters et al, 2001; Schultz et al, 2002; Tachibana et al, 2001; Tachibana et al, 2002).

Trimethylation of lysine residue 27 in histone H3 (H3K27me3) is a repressive mark mediated by the Polycomb complex 2 (PRC2), which has been proposed to recruit another repressive Polycomb complex, polycomb complex 1 (PRC1) (Schwartz & Pirrotta, 2007). PRC1 was shown to cause chromatin compaction *in vitro* (Francis et al, 2004). However, whether or not such a mechanism operates *in vivo* is still under debate.

In contrast to the repressive marks mentioned above, methylation at lysine residue 4 of histone H3 (H3K4me) is generally considered as an active chromatin signature. High levels of H3K4me3 and low levels of H3K4me1 are found at promoters, while the opposite is observed at active regulatory elements that are distally located (enhancers) (Heintzman & Ren, 2009; Heintzman et al, 2007). H3K4me3 appears to prohibit the binding of DNMT3 enzymes via their ADD domain, which specifically binds unmethylated H3K4, and, in line with this, a strong inverse correlation is observed for H3K4me3 and DNA methylation, suggesting that H3K4 methylation may play a role in determining DNA methylation patterns (see section 1.5.7, (Meissner et al, 2008; Ooi et al, 2007).

Although H3K4me3 and H3K27me3 are generally regarded as active and repressive marks, respectively, they were shown to coexist in ESCs on the same histone H3 (Bernstein et al, 2006a). This so-called bivalent chromatin state has been proposed to maintain developmental genes in a silent state in embryonic stem cells (ESCs) while keeping them 'poised' for activation upon differentiation (Azuara et al, 2006; Bernstein et al, 2006a; Mikkelsen et al, 2007).

Due to its enrichment in the bodies of active genes, H3K36me2/3 is considered as an active mark (Barski et al, 2007; Mikkelsen et al, 2007). Its differential enrichment within gene bodies, with higher enrichment in exons versus introns may indicate a potential role in splicing (Zhou et al, 2011). Responsible for the removal of the H3K36 methylation mark are members of the KDM2 and KDM4 families, such as KDM2A and KDM2B. KDM2A and B are specifically recruited to unmethylated CpG Islands via their zinc finger CXXC (ZF-CXXC) domain[1]. Depletion of KDM2A results in increased levels of H3K36me2 in CGIs, which normally lack this modification (Blackledge et al, 2010). KDM2B can recruit PRC1 to unmethylated CpG islands, which leads to silencing of the associated gene (Farcas et al, 2012).

The combination of various histone modifications gives rise to a number of different chromatin configurations, also referred to as the "histone code" (Jenuwein & Allis, 2001), to regulate gene expression by altering chromatin structure. In comparison, DNA methylation appears rather simple as it is only

[1] ZF-CXXC domains are binding modules for the specific recognition of unmethylated CpGs (Allen et al, 2006, Lee et al, 2001)

involves the modification of a single base in the genome, yet it proves to be no less complex in its regulatory capacity.

1.2 DNA Methylation

1.2.1 The mark

First postulated as an epigenetic regulator in 1975 (Holliday & Pugh, 1975; Riggs, 1975), it is now the best-studied epigenetic modification we know and the only one which involves the chemical modification of the DNA itself. In eukaryotes, DNA methylation is exclusively found at cytosine residues in DNA, where a methyl group is covalently attached to the carbon-5 position of the pyrimidine ring of cytosine to form 5-methylcytosine (5mC). The enzymatic reaction is catalysed by DNA methyltransferases (see section 1.2.2).

In mammals, DNA methylation of cytosines is predominantly found in the context of CpG dinucelotides (when the cytosine is followed by a guanosine), although a growing body of evidence is challenging that view demonstrating that methylcytosine outside of the CpG context (CHG, CHH, where H=A,C,T) is also prevalent in mammals (see section 1.4.3). Due to its symmetrical nature (CpG dinucleotide on the top strand is mirrored by a CpG dinucleotide on the bottom strand), CpG methylation occurs normally on both forward and reverse strands (symmetrical methylation). This has important implications for the mechanisms of inheritance of methylation patterns during replication because it means that, during semi-conservative replication, the DNA methylation pattern at CpG sites is still preserved on each of the parental strands.

CpG methylation occurs at the vast majority CpG sites (60-80%) and has a bimodal distribution with CpG sites being either highly methylated (>70% of all CpG sites) or displaying low methylation levels (<30% of all CpG sites). CpG sites with low methylation levels are predominantly found in so-called CpG Islands, which are stretches of DNA that are particularly high in their content of Gs and Cs (GC content) as well as in their content of CpG dinucleotides (for further information see section 1.4.1.1).

Despite the fact that 5-methylcytosine is the most prevalent and best-studied modification, exciting recent studies have revealed other forms of cytosine modifications in DNA, including 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC) (see sections 1.4.2 and 1.5.2). In this thesis, if not otherwise stated, the term DNA methylation generally refers to 5mC methylation in the CpG context.

The chemical reaction that generates 5mC is catalyzed by DNA methyltransferases (DNMTs), which will be discussed in more detail below.

1.2.2 DNA methyltransferases

Responsible for the addition of a methyl group to carbon at position 5 in the pyrimidine ring of cytosine residues to generate 5-methylcytosine (5mC) are DNMTs. In eukaryotes, five DNA methyltransferases are known to date: DNMT1, DNMT2, DNMT3A, DNMT3B and DNMT3L. Almost all of them have been identified based on conserved sequence motifs that they share with the bacterial cytosine-5 methyltransferase (Goll & Bestor, 2005). While DNMT1, DNMT3A, and DNMT3B play pivotal roles in establishing and maintaining global DNA methylation patterns, DNMT2 appears to exert minimal or no

methyltransferase activity on DNA. DNMT3L has no catalytic activity, but serves as an important co-factor for DNMT3A and B.

1.2.2.1 DNMT1

DNMT1 is classically known as the maintenance methyltransferase, because of its preference for hemimethylated DNA (Goyal et al, 2006). Specificity to hemimethylated DNA is conferred by its N-terminal domain, which mediates interactions with other proteins to target DNMT1 to the replication fork. This is achieved via a peptide region that allows DNMT1 to interact with the replication factor proliferating cell nuclear antigen (PCNA), thereby targeting the enzyme to the replication machinery (Chuang et al, 1997). It furthermore interacts with ubiquitin-like with PHD and ring finger domains 1 (UHRF1, also known as NP95), which recruits DNMT1 to hemimethylated DNA at the replication foci leading to a base flipping mechanism exposing the unmethylated CpG to DNMT1 (Arita et al, 2008; Avvakumov et al, 2008; Hashimoto et al, 2008; Rountree et al, 2000; Sharif et al, 2007). The N-terminal domain also comprises a nuclear localization sequence (NLS) and a pair of bromo-adjacent homology (BAH) domains proposed to mediate protein-protein interactions (Callebaut et al, 1999; Goll & Bestor, 2005), as well as a ZF-CXXC domain. As mentioned, above, ZF-CXXC domains are binding modules for the specific recognition of unmethylated CpGs (Allen et al, 2006; Lee et al, 2001). In line with this, the ZF-CXXC domain of DNMT1 was shown to target the protein specifically to unmethylated CpG sites (Pradhan et al, 2008). In addition to its proposed role in recruiting DNMT1 to unmethylated DNA, the ZF-CXXC domain of DNMT1 appears to furthermore determine the enzymes preference for methylating hemimethylated over

unmethylated DNA (Song et al, 2011). The crystal structure of DNMT1 bound to fully unmethylated DNA revealed that the ZF-CXXC domain occludes the unmethylated target CpG from the active site, preventing *de novo* methylation in an autoinhibitory manner (Song et al, 2011)[2]. Song et al support this idea by presenting kinetic data of truncated versions of DNMT1 with or without the ZF-CXXC domain to show that the enzyme's preference for unmethylated DNA increases in the absence of the ZF-CXXC domain. Remarkably, a year later, the same group published the crystal structure of DNMT1 in complex with hemimethylated DNA (Song et al, 2012b). In this study, they show that binding of DNMT1 to hemimethylated DNA distorts the unmethylated target cytosine in such a way that it is exposed to the catalytic pocket of DNMT1. Taken together, they propose a model by which DNMT1 faithfully maintains methylation patterns of DNA as it emerges from the replication complex, which is attained by two mutually supportive mechanisms: (1) an autoinhibitory mechanism of DNMT1 that involves the ZF-CXXC domain to protect fully unmethylated CpG sites from *de novo* methylation and (2) specific recognition of hemimethylated sites by the target recognition domain (TRD) of the methyltransferase domain of DNMT1 due to conformational changes of the target cytosine. Interestingly, two independent studies characterised the crystal structure of DNMT1 fragments without DNA (Syeda et al, 2011; Takeshita et al, 2011). In the crystal structure of unbound DNMT1 the N-terminal replication focus targeting sequence (RFTS) domain of DNMT1 occupied the protein's DNA binding pocket, thereby preventing binding and methylation of DNA (Syeda et al, 2011; Takeshita et al, 2011). Based on their

[2] In Song et al (2011) the crystal structure of both human and mouse DNMT1 were determined. It was shown that the orthologous proteins were highly similar on both sequence level and structural level.

structural analysis Syeda et al propose a model in which unbound DNMT1 might be regulated by binding of UHRF1, which would remove the inhibitory domain from the active site to allow DNMT1 binding to DNA.

The N-terminal domain of DNMT1 is connected to the C-terminal domain via a flexible conserved Gly-Lys repeat. The C-terminal methyltransferase domain comprises the conserved sequence motifs that led to the identification of DNMT1. This domain is highly conserved throughout evolution and closely related to many of the bacterial enzymes (Bestor, 2000). The C-terminal domain is thought to confer sequence specificity of DNMT1 for CpG dinucleotides. This was demonstrated by replacing the C-terminal domain of mouse DNMT1 with that of prokaryotic DNMTs that do not have a preference for CpG dinucleotides (Pradhan & Roberts, 2000). The generated fusion proteins lost target specificity to CpG dinucleotides. The specificity for hemimethylated DNA, which is specific to eukaryotic N-termini, was retained in these fusion proteins. Hence, it was concluded that the eukaryotic N-termini confer specificity to hemimethylated DNA.

DNMT1 activity is regulated through protein modifications and interactions with other proteins. Furthermore, different isoforms of DNMT1 exist. These are generated by expression from alternative promoters or alternative splicing resulting in different isoforms with truncated N-termini altering the enzymes properties. Several isoforms are expressed at certain stages during development to ensure faithful propagation of DNA methylation patterns during development (for review see (Latham et al, 2008)). An alternative splice form of DNMT1, which is specific to the oocyte (DNMT1o), retains its

nuclear localization signal (NLS) to be localized to the nucleus at the early stages of embryogenesis (oocyte growth phase and implantation embryos at 8 cell stage), where it becomes fully active only after implantation (Mertineit et al, 1998). An alternative promoter regulates DNMT1 expression by producing a full length protein active in ESCs and somatic cells, as well as an untranslated sperm-specific version that is thought to play a role in regulating DNMT1 activity during spermatogenesis (for review (Latham et al, 2008)).

1.2.2.2 DNMT2

First identified in 1998 based on its sequence homology to other DNA methyltransferases, DNMT2 was soon revealed to be highly conserved over numerous phylogenetic branches, ranging from protists to higher eukaryotes, like *Drosophila melanogaster* and human (Bestor, 2000; Fisher et al, 2004; Kunert et al, 2003; Ponger & Li, 2005; Yoder & Bestor, 1998). In some eukaryotes, for instance, *Schizosaccharomyces pombe*, and *Drosophila melanogaster*, DNMT2 constitutes the sole putative DNA methyltransferase (Goll & Bestor, 2005; Hung et al, 1999). The relatively small protein of 415 amino acids is characterised by its lack of the N-terminal regulatory domain, which is commonly found in eukaryotic DNA methyltransferases. The function of the highly conserved catalytic C-terminal domain is currently still unclear. DNA methyltransferase activity has been documented in a number of species and in line with this, the *Drosophila* genome, which bears DNMT2 as the only methyltransferase, appears methylated, albeit at low levels (primarily at CpT and CpA) (Hermann et al, 2003; Kunert et al, 2003; Lyko et al, 2000). Yet, in all systems studied, DNMT2 was found to exert only weak DNA methyltransferase activity compared to other DNMTs and depletion of DNMT2

results only in minor or no morphological defects in most of the cases studied (Goll et al, 2006; Kuhlmann et al, 2005; Kunert et al, 2003; Okano et al, 1998b). An exception is zebrafish, where DNMT2 depletion results in severe differentiation defects in several organs, including retina, liver, and brain leading to considerable morphological abnormalities in embryos (Rai et al, 2007). Surprisingly, it was demonstrated more recently that human DNMT2 exhibits strong methyltransferase activity on RNA (Goll et al, 2006). More specifically, it was found to methylate $tRNA^{Asp}$ at cytosine on position 38 in the anticodon loop. Its role in mammalian cells is, however, still obscure.

1.2.2.3 DNMT3A and B

Homologous to DNMT1 in their C-terminal domain, DNMT3A and B are distinct in their N-terminal regulatory domains. The N-termini mediate protein-protein interactions and target the protein to chromatin. Chromatin binding as well as recognition of specific histone modifications is conferred by a PWWP domain and an ADD domain (ATRX-DNMT3-DNMT3L) in the N-terminus. The PWWP domain localizes DNMT3A and B to pericentric heterochromatin where it is required for methylation of satellite repeats (Chen et al, 2004; Ge et al, 2004). Mutations in the PWWP domain of *DNMT3B* disrupts chromatin binding and were found in patients with a rare autosomal-recessive human disorder called ICF (immunodeficiency, centromere instability and facial anomalies) syndrome (Ge et al, 2004; Xu et al, 1999). Patients with ICF syndrome are characterised by immunodeficiency, mild facial anomalies and, most interestingly, cytogenetic abnormalities in lymphocytes that largely affect pericentric regions on chromosome 1, 9, 16, which is associated with a loss of methylation at pericentromeric satellite repeats on these chromosomes. The

ADD domain of DNMT3A and B was shown to interact with unmodified H3K4 (Zhang et al, 2010), which is in line with the observation that the presence of H3K4me3 and DNA methylation show a strong anti-correlation (Meissner et al, 2008). Tight association of DNMT3A and B with chromatin has been documented (Jeong et al, 2009), and this interaction has been proposed to regulate cellular levels of DNMT3 by stabilizing the protein, while free DNMT3A and B may be selectively degraded (Sharma et al, 2011).

The crystal structure of DNMT3A revealed that it can form a tetrameric complex consisting of two DNMT3A-3L heterodimers, which binds DNA in a manner that the two active sites of DNMT3A are positioned at a distance of one helical DNA turn (Jia et al, 2007). Interestingly, it was proposed that this complex might methylate two CpG sites separated by one helical turn (~10bp) indicating that they might be able to methylate DNA at 10bp apart. In line with this notion, 9bp periodicity of 5mC in the CpG context has been observed in highly methylated regions of human chromosome 21 (hs21), at imprinted loci as well as repetitive elements, such as Alu (Glass et al, 2009; Jia et al, 2007; Zhang et al, 2009).

As neither DNMT3A nor DNMT3B show a preference for hemimethylated over unmethylated DNA, they are considered as *de novo* methyltransferases (Okano et al, 1998a). In line with this, DNMT3A and B are highly expressed in ES and germ cells where they are required for remethylation of DNA after genome-wide erasure of DNA methylation patterns as part of an epigenetic reprogramming event (Chen et al, 2002; Li et al, 1992; Okano et al, 1999; Okano et al, 1998a). In somatic cells, expression of *de novo*

methyltransferases is downregulated, yet still expressed. While DNMT3A is ubiquitously expressed, DNMT3B shows very low expression in somatic cells with the exception of testes, thyroid and bone marrow, where methylation levels are markedly elevated (Chen et al, 2002; Xie et al, 1999). DNMT3A2, a major isoform of DNMT3A, is expressed in tissues, including testis, ovary, spleen, and thymus (Chen et al, 2002).

DNMT3A and B play crucial roles in mammalian development, which is reflected by the observation that mice deficient in either of the enzymes die in early embryogenesis (DNMT3B) or early adulthood (DNMT3A) (Li et al, 1992; Okano et al, 1999). The different stages of lethality and the resulting phenotypes suggest that they have non-overlapping functions (Okano et al, 1999). Their distinct roles are furthermore demonstrated at the molecular level: DNMT3B was shown to be required for methylation at pericentromeric satellites, while DNMT3A is necessary for establishing imprints (Kaneda et al, 2004; Okano et al, 1999).

1.2.2.3.1 DNMT3L

DNMT3L is a mammalian-specific short paralogue of DNMT3A and B, which retains a cysteine-rich domain comprising an ADD domain in its N-terminus, but lacks the PWWP domain (Aapola et al, 2000; Goll & Bestor, 2005). Furthermore, key residues within the catalytic domain have been subject to substitutions, rendering the protein enzymatically inactive (Hata et al, 2002).

Although DNMT3L itself has no methyltransferase activity, it is required for full activity of DNMT3A and B by physically interacting with the *de novo* DNMTs (Chedin et al, 2002; Chen et al, 2005; Chen & Riggs, 2005; Gowher et al,

2005). Required for normal gametogenesis, it is expressed in germ cells and in the early embryo (at times when widespread *de novo* methylation occurs) where it is required for the establishment of maternal and paternal imprints (Aapola et al, 2000; Arima et al, 2006; Bourc'his et al, 2001; Hata et al, 2002; Webster et al, 2005). In oocytes, DNMT3L depletion causes severely reduced methylation at imprinting control regions and biallelic expression of normally imprinted genes (Bourc'his et al, 2001; Hata et al, 2002), while DNMT3L-deficient males display impaired spermatogenesis (Bourc'his & Bestor, 2004; Webster et al, 2005). During spermatogenesis, it was shown to not only be required for the establishment of paternal imprints, but also for the methylation and silencing of retrotransposable elements, specifically of dispersed repeat sequences, while pericentric tandem repeats were methylated normally (Bourc'his & Bestor, 2004; Hata et al, 2006; Webster et al, 2005). It is now established that DNMT3L stimulates and targets DNMT3 during waves of *de novo* methylation to chromatin in early embryogenesis and germ cells. Crystallographic studies indicate that DNMT3L specifically targets DNMT3A and B to chromatin that lack methylation at H3K4. Specific targeting is achieved via its ADD domain, an interaction that is inhibited by H3K4 methylation (Ooi et al, 2007; Otani et al, 2009)

1.3 Roles of DNA methylation in mammals

DNA methylation plays a fundamental role in many cellular processes. In particular, due to its nature as a direct modification of DNA itself, it is thought to be passed on during replication as a relatively stable mark and therefore has been considered as a major component of epigenetic memory. Here, I will

give a brief overview of some of the roles of DNA methylation in cellular processes.

1.3.1 X-inactivation

X-chromosome inactivation is a dosage compensation mechanism in mammals by which the majority of genes on one X chromosome are silenced in female cells (reviewed in (Augui et al, 2011). X-inactivation in the embryo is a random process[3] that is associated with hypermethylation of the majority of CGIs on the inactive X (Xi) compared to the active X (Xa). Genes that escape inactivation display low promoter methylation (Sharp et al, 2011). Despite promoter hypermethylation on the Xi, the inactive X is overall less methylated than its active counterpart and this can be attributed to lower levels of gene body methylation (Hellman & Chess, 2007). Efficient establishment of X-inactivation in the absence of DNMT3A and B in embryos suggests that DNA methylation is not required for the initiation of X-inactivation (Sado et al, 2004). DNA methylation was, however, found to play a crucial role in the maintenance of gene silencing on Xi as shown by reactivation of initially silenced genes on the Xi upon conditional knockout of DNMT1 in somatic tissues (Sado et al, 2000). In line with this, DNA methylation is a late modification in the process of X-inactivation (reviewed in (Heard, 2004).

1.3.2 Genomic Imprinting

Imprinting is the differential expression of paternally and maternally inherited alleles (reviewed in (Ferguson-Smith, 2011). Differential DNA methylation at imprinted loci plays a key part in the mono-allelic expression of imprinted

[3] While this is a random process in the cells of the inner cell mass that give rise to the embryo proper, in the cells that give rise to the extra-embryonic tissue it is the paternal X-chromosome that becomes inactivated.

genes. One of the best-studied imprinted loci is the *IGF2R/H19* locus. At this locus, a CTCF binding site separates the *IGF2R* gene from the *H19* gene. On the maternal allele, hypomethylation of that site allows CTCF to bind. Binding prevents the interaction of the *IGF2R* promoter with its distal enhancer elements that it shares with *H19* resulting in silencing of *IGF2R* and activation of *H19* on the maternal allele (Bell & Felsenfeld, 2000; Hark et al, 2000). On the paternal allele, methylation blocks CTCF from binding allowing interaction of *IGF2R* from its distal enhancers resulting in activation of *IGF2R* and silencing of *H19*. Although it is not clear how differential methylation at the *H19* locus is established, conditional knockout of *DNMT3A* in germ cells (but not *DNMT3B*) leads to the loss of methylation and biallelic expression, corroborating the notion that DNA methylation plays a major role in establishing maternal and paternal imprints (Kaneda et al, 2004). Once established, DNMT1 is required to maintain genomic imprints (Li et al, 1993).

1.3.3 Transposon silencing

Approximately half of mammalian genomes are composed of repetitive sequences, the majority (about 45%) of which are transposon-derived repeats, these include members of the small and long interspersed nuclear elements (SINEs and LINEs, respectively), such as L1 and Alu elements, as well as endogenous retroviruses (ERVs) (Lander et al, 2001). Although most of these retroelements have lost their transcriptional potential, a proportion of them has remained intact and these are capable of being actively transcribed (Coufal et al, 2009; Muotri et al, 2005). When active, they are believed to be able to integrate elsewhere in the genome via reverse transcription and integration of the produced cDNA (Havecker et al, 2004). This may pose a

threat on the host by causing mutations in functionally important regions in the genome or it may modulate the expression of nearby genes (Karimi et al, 2011; Rebollo et al, 2011). Hence, to keep retroelements inactive, silencing mechanisms have evolved, among which DNA methylation is thought to be important. In mammalian genomes, transposons are heavily methylated and this appears to be a widespread phenomenon observed in a wide variety of species (Zemach et al, 2010). In line with the proposed role of DNA methylation in silencing transposable elements (TEs), DNMT1-deficiency results in the upregulation of RNA arising from endogenous retroviruses (Karimi et al, 2011; Walsh et al, 1998). Furthermore, in support of a role for DNA methylation in keeping endogenous retroviruses inactive it was found that loss of Methyl-CpG-binding protein 2 (MeCP2), a protein that recognises and binds to methylated DNA, results in an increase of L1 transcripts and retrotransposition in mouse and human brain[4] (Muotri et al, 2010).

1.3.4 Gene regulation

An indication that DNA methylation plays a role in the regulation of gene expression stems from observations where methylation of promoter regions is associated with gene silencing. As mentioned above, allele-specific promoter methylation of genes on Xi as well as of imprinted genes is associated with mono-allelic gene silencing. Furthermore, aberrant promoter hypermethylation is frequently found in cancer and associated with inappropriate gene silencing (Jones & Baylin, 2002). An anticorrelation of DNA methylation at promoter regions and gene expression in healthy tissues was initially only demonstrated for a few genes (Song et al, 2005). A specific example of cell-

[4] MeCP2 mutation causes a neurodevelopmental disorder called Rett syndrome (Amir et al, 1999).

type specific gene regulation by DNA methylation is the tumor suppressor gene *SERPINB5*: the gene is expressed when the promoter is unmethylated and is silenced when the promoter is methylated (Futscher et al, 2002). More recently, however, this anticorrelation could be confirmed on a larger scale. In a DNA methylation study of 873 genes it was found that about one third of the analysed genes showed an inverse correlation of methylation at the transcription start site (TSS) with transcription (Eckhardt et al, 2006).

About half of all protein-coding genes are associated with CGIs at their promoter. CGI-associated promoters are usually unmethylated even when the gene is not expressed, suggesting that promoter hypomethylation *per se* is not sufficient for gene activation. Of note, however, some promoter CGIs do become methylated and this correlates with gene silencing (Shen et al, 2007). To better understand the role of DNA methylation on gene expression, genes have been classified according to the CpG-richness of their promoters (Meissner et al, 2008; Weber et al, 2007). High CpG density promoters (HCPs, >7% CpGs over 300bp), also called 'strong CGIs', were found to be nearly always unmethylated regardless of expression levels and are frequently associated with housekeeping genes or developmental regulators (Meissner et al, 2008), while low CpG density promoters (LCPs, <5% CpGs), also called 'non-CGI promoters' were overall methylated (Weber et al, 2007). By contrast, promoters with an intermediate CpG density (ICPs), also termed 'weak CGIs' did show a correlation between DNA methylation and gene expression. Many of them were linked to germ-line specific genes, which tended to become methylated during differentiation in somatic cells (Weber et

al, 2007). Similar observations were made in other studies, for example differential methylation at CGIs during reprogramming (Deng et al, 2009).

Promoter methylation seems to play a role in regulating pluripotency genes during lineage commitment. Many pluripotency genes acquire methylation at their promoters during differentiation of ESCs into neuronal progenitor cells when they are silenced (Mohn et al, 2008). In further support of a role for DNA methylation in gene expression, methylation changes occur dynamically during differentiation, suggesting that methylation changes might be involved in tissue-specific gene regulation (Meissner et al, 2008; Ziller et al, 2013).

Evidence suggests that, at least in some cases, DNA methylation may be a secondary event to gene repression important for maintaining a stable silent state rather than initiating gene silencing. For example, at the promoter of the *OCT4* gene, a master regulator of pluripotency, DNA methylation is only required at a later stage in differentiation after the gene has already been silenced, while the initiation of gene silencing is independent of DNA methylation (Feldman et al, 2006). Similarly, DNA methylation is required for the maintenance of gene silencing on the inactive X-chromosome, but is dispensable for its establishment (see section 1.3.1).

Apart from the observed anti-correlation of DNA methylation at promoter elements and gene expression, DNA methylation in gene bodies positively correlates with gene expression (Ball et al, 2009; Hellman & Chess, 2007; Laurent et al, 2010; Rauch et al, 2009). This observation has led to the proposal that gene body methylation may increase transcriptional efficiency by inhibiting spurious transcription initiation within gene bodies (Ball et al,

2009; Rauch et al, 2009). In four specific examples it could be shown that gene body hypomethylation coincided with transcription initiation at an intragenic alternative promoter, implying that methylation may regulate alternative promoter usage (Rauch et al, 2009). However, whether or not gene body methylation is a cause or consequence of transcription remains to be determined.

1.3.5 Maintenance of genome integrity

A role for DNA methylation in the maintenance of genome integrity is suggested by the observation that *DNMT3B* depletion leads to chromosomal abnormalities, poly- and aneuploidy (Dodge et al, 2005). Most compellingly, in patients with ICF syndrome *DNMT3B* depletion is associated with hypomethylation of pericentric repeats and cytogenic abnormalities at the chromosomes affected by hypomethylation (Xu et al, 1999). Furthermore, *DNMT1* hypomorphic mice display chromosomal instabilities with ploidy changes and are more prone to cancer (Eden et al, 2003; Gaudet et al, 2003). However, it is currently not known how DNA methylation may relate to the observed chromosomal aberrations, although it has been proposed that DNA methylation may be important to mask sites of homologous recombination, such as repetitive DNA (Robertson & Wolffe, 2000).

1.3.6 Molecular effects of DNA methylation

Despite the long-standing efforts to understand the role of DNA methylation and how it may affect gene expression, the molecular mechanisms of action are still not fully understood. The effects of DNA methylation on gene regulation can be explained on the molecular level in two different ways. 5-

methylcytosine can directly inhibit DNA-protein interactions by blocking the binding sites for sequence-specific binding proteins or, on the contrary, it can increase protein-DNA binding leading to the recruitment of proteins that specifically recognise this epigenetic mark. In support of the first (blocking of DNA binding sites) the most well known example is CTCF, which was shown to be excluded from the imprinting control region by DNA methylation at the H19 locus on the maternal allele (see section 1.3.2). Furthermore, proteins with a ZF-CXXC domain have been shown to bind unmethylated DNA. These include CFP1 (which is required for normal H3K4me3 levels at CGIs) (Thomson et al, 2010), KDM2A, and KDM2B (see section 1.1.1), and TET1 (see sections 1.4.2 and 1.5.2, for review on ZF-CXXC containing binding proteins see (Long et al, 2013a). In addition, there are several examples of proteins that specifically recognise methylated DNA, for instance the methyl-CpG-binding domain (MBD) proteins MBD1, MBD2, MBD4, and MeCP2 to name but a few (for a specific example see section 1.3.3).

1.4 Genome-wide DNA methylation patterns

1.4.1 Mammalian DNA methylation patterns

In recent years, the rapid development of elaborate techniques to study DNA methylation genome-wide has enabled the generation of quantitative genome-wide DNA methylation maps that greatly increases our knowledge of the distribution and specific pattern of DNA methylation in mammalian genomes. Here, I will give an outline of our current knowledge of the distribution of DNA methylation genome-wide.

Early work on DNA methylation made use of methylation-sensitive restriction enzymes to differentially cleave the DNA depending on its methylation state. This enabled the identification of a small fraction (<2%) of the genome, later termed CpG Islands, which is sensitive to cleavage by the methylation-sensitive endonuclease HpaII (Bird et al, 1985; Cooper et al, 1983). Genome-wide assessment of DNA methylation levels by HPLC showed that methylation levels vary between human tissues (Ehrlich et al, 1982). Repetitive elements were consistently found highly methylated across different types of tissues (Gama-Sosa et al, 1983).

Given the recent discovery of abundant levels of 5-hydroxymethylation (5hmC, see section 1.4.2) in the genome it is important to note that most of the studies discussed in this chapter were generated by techniques that do not distinguish between 5-methylcytosine and 5-hydroxymethylcytosine[5].

1.4.1.1 CGIs

As mentioned above, CGIs were initially identified due to their increased sensitivity to the methylation-sensitive restriction enzyme HpaII, which cleaves a small fraction of the genome (<2%) into small fragments ("HpaII-tiny-fragments") (Cooper et al, 1983). Subsequent sequencing of these fragments revealed that they are particularly rich in their content of Gs and Cs, as well as CpG dinucleotides, which gave rise to the term CpG Island (Bird et al, 1985)[6]. Their relative enrichment in CpG dinucleotides compared with the rest of the genome may be explained by an overall depletion of CpGs in vertebrate genomes. This was proposed to arise from spontaneous deamination of

[5] Most of the techniques, including bisulfite treatment or methylation-dependent cleavage of DNA, identify 5hmC as 5mC.
[6] Of note, the recognition site of HpaII is CCGG, hence, this method is biased towards interrogating GC-rich DNA.

methylcytosine to uracil, which can (if not correctly repaired) eventually lead to conversion to thymidine residues in place of cytosines (C to T conversion). Global methylation in vertebrate genomes was therefore proposed to lead to the natural depletion of cytosine residues genome-wide (Bird, 1980; Coulondre et al, 1978). CGIs, on the contrary, which largely lack methylation, are less prone to C to T conversions, and, hence, retain their Cs resulting in the relative CpG-richness compared to the rest of the genome (Deaton & Bird, 2011).

Based on the knowledge that CGIs display elevated levels of CpG dinucleotides as well as overall GC content compared to the rest of the genome, algorithms were soon created to predict the presence of unmethylated CGIs throughout the genome. The earliest algorithm was established by Gardiner-Garden and Frommer (1987), which define CGIs as regions of at least 200 bp in length, which display a GC content of 50% or higher, and a CpG content that is higher than the expected CpG content (as judged on the basis of the number of Gs and Cs in the segment) (Gardiner-Garden & Frommer, 1987). In contrast to the sequence-based approach for the identification of CGIs, functional approaches have been developed more recently, which take into account the methylation state of a CGI (Blackledge et al, 2012; Illingworth et al, 2008). This is achieved by capturing unmethylated DNA in the genome using a protein domain, the ZF-CXXC domain, which specifically recognises unmethylated DNA. Illingworth and colleagues use the ZF-CXXC domain of MBD1 to compare the number of functionally derived

CGIs in human and mouse[7] (Illingworth et al, 2010). They identified a similar number of CGIs in human (25,495) and mouse (23,021). In comparison, a bioinformatics approach predicts ~27,000 CGIs for human, but far fewer CGIs (~15,500) for mouse (Antequera & Bird, 1993; Lander et al, 2001; Waterston et al, 2002). The authors reason that this discrepancy may be explained by the fact that the algorithms do not take into account the smaller content of CpGs in the mouse genome compared to the human genome. In support of this notion, they find that despite being relatively CpG-rich (compared to the bulk genomic DNA) the newly identified mouse CGIs do not meet the criteria of the bioinformatic prediction algorithms. Apart from verifying the conservation of the number of CGIs in human and mouse genomes, it was furthermore found that the genomic location is highly preserved in the two species with an equal distribution of promoter, inter- and intragenic CGIs.

Using this method, it was found that about half of the CGIs coincide with TSSs, while the remaining half is located distal from annotated promoters, so-called “orphan” CGIs. Yet many of the “orphan” CGIs do share characteristics of functional transcription start sites (as judged by H3K4me3 levels and RNA polymerase II occupancy), which led the authors to propose that they might represent as yet unidentified functional promoter elements (Illingworth et al, 2010). In support of a functional role of these newly identified “orphan” CGIs is their tendency to be developmentally regulated with the appearance of

[7] This technique has been refined more recently by labelling the CXXC domain of human KDM2B with biotin, which can be immobilised on avidin beads rather than using a chromatography column-based approach as it is done in the previous approach (Blackledge, Klose 2011). This technique is more applicable for conventional laboratory use as it requires lower amounts of DNA, smaller elution volumes and is more time-efficient. Furthermore, BioCAP requires only standard laboratory equipment and facilitates parallel sampling. I will be using the data generated using the BioCAP approach in chapter 6.

methylation in tissues and a coincident loss of active H3K4me3 mark (Illingworth et al, 2010).

In summary, CGIs are GC- and CpG-rich stretches of DNA, on average 1kb in size, generally unmethylated, although some CGIs, particularly remote CGIs (distal of TSS) are tissue-specifically methylated (Deaton & Bird, 2011). Human and mouse genomes harbour roughly 25,000 CGIs, which are on average 1kb in size, constituting roughly 1% of the genome (Cooper et al, 1983). About half of them coincide with promoter elements (the majority of which are house keeping genes as well as some developmental genes), the other half are located in inter- or intragenic regions.

On the basis that the bioinformatic approach is an accurate predictor of CGIs in the human genome, I will employ the University of California, Santa Cruz (UCSC) browser CGI annotation as the definition for CGIs throughout my thesis. The UCSC genome browser defines CGIs based on the most routinely used definition established by Gardiner-Garden and Frommer (Gardiner-Garden & Frommer, 1987) and refines prediction by taking out repetitive elements, which constitute a primary confounding factor as they share many of the features of CGIs (Illingworth & Bird, 2009).

1.4.1.2 Gene promoters

Although a large number of genes are associated with a CGI at their TSS, an equal number of gene promoters do not have a CGI at their 5' end and these will be discussed in this section. Similar to CGI promoters the majority of non-CGI promoters show low levels of methylation. In B-cells, for example, only 10% of all analysed promoters are methylated (Rauch et al, 2009). As

mentioned in section 1.3.4, promoters were classified according to their CpG density into high, intermediate and low CpG-density promoters (HCPs, ICPs and LCPs) and it was found that methylation levels were partly determined by the CpG density of promoters (Meissner et al, 2008; Weber et al, 2007). While HCPs were found to be nearly always unmethylated regardless of expression levels, and LCPs nearly always methylated, ICPs have a higher tendency of methylation and were frequently methylated in ESCs as well as in somatic cells (Meissner et al, 2008). However, promoter sequence and gene function was also proposed to play a role in determining DNA methylation states of ICPs (Weber et al, 2007). Interestingly, the majority of HCPs are associated with housekeeping genes or developmental genes (Meissner et al, 2008). Those associated with housekeeping genes tend to be marked by H3K4me3 in ESCs ("univalent mark") and those associated with developmental genes were enriched for the bivalent mark (H3K4me3 and H3K27me3, see section 1.1.1, (Meissner et al, 2008). Despite the observed trend regarding the methylation levels of promoters and their CpG content, it was noted that for low CpG density regions, the histone methylation patterns were better predictors of DNA methylation than CpG density (Meissner et al, 2008).

1.4.1.3 CGI shores

In a DNA methylation study carried out by Irizarry et al it was found that tissue-specific differences in methylation (T-DMRs) predominantly localize to regions near CGIs, in so-called CpG Island shores (Irizarry et al, 2009). CGI shores were defined as up to 2kb away from CGIs and also found to harbor the majority of DNA methylation variation in normal colon versus colon cancer DNA, between human induced pluripotency stem cells and paternal

fibroblasts, and during lineage commitment (Doi et al, 2009; Irizarry et al, 2009; Ji et al, 2010). In accordance, a recent in-depth DNA methylation survey, which includes 42 whole genome sequencing data sets undertaken on 20 different types of human cells and tissues, identifies CGI shores amongst the most dynamically methylated regions during normal development (Ziller et al, 2013). Interestingly, it was found that differential methylation, particularly at CGI shores, correlated with the expression of the associated gene (Irizarry et al, 2009; Ji et al, 2010). Although it is functionally still not clear how CGI shore methylation may affect gene expression, the majority of differential methylation (~ 70%) was located within 500 bp of an alternative TSS, and not near (within 500bp of) an annotated TSSs. Therefore, tissue-specific differential methylation was proposed to influence transcription by regulating promoter choice (Irizarry et al, 2009). In support of a functional role of differential methylation of CGI shores it was found that about half of the identified DMRs showed sequence conservation as well as corresponding tissue-specific methylation between human and mouse (Irizarry et al, 2009). Mechanistically, cancer-specific differential methylation has been associated with the loss of boundaries at CGIs resulting in a methylation boundary shift into the adjacent region or into the interior of the island (Hansen et al, 2011). This was proposed to lead to hypo- and hypermethylation of CGI shores depending on the direction of the shift. In an independent study, a similar observation was made with boundaries of hypomethylated regions (HMRs) fluctuating resulting in a cell-type-dependent widening of HMRs corroborating the observation of “tidal methylation changes at CGI shores” (Hodges et al, 2011). The widest HMRs were found in germ cells and it was proposed that

they might “presage the extent of shores” in other cell types (Molaro et al, 2011). Further studies are required to establish a putative role for CGI shore methylation in transcriptional gene regulation.

1.4.1.4 Gene bodies

Gene body methylation is an interesting feature in that it is evolutionarily ancient. A possible role for gene body methylation in the regulation of transcription has been proposed due to its enrichment in the bodies of active genes, in *Arabidopsis* (Cokus et al, 2008; Zhang et al, 2006; Zilberman & Henikoff, 2007), as well as in mammals (Ball et al, 2009; Laurent et al, 2010; Lister et al, 2009; Rauch et al, 2009). In the *Arabidopsis*, gene body methylation is highest at moderately transcribed genes, whereas expression at either extreme correlates with lower methylation levels (Zilberman et al, 2007) and a similar observation was made in mammalian cells (Maunakea et al, 2010). Furthermore, gene-body methylation is a feature of the active human X chromosome compared to its inactive counterpart (Hellman & Chess, 2007) and hypomethylation of gene bodies has been associated with silent genes in cancer cell lines (Shann et al, 2008). The role of gene body methylation is still unclear. An interesting genome-wide study indicated that tissue-specific gene-body methylation might suppress alternative transcripts originating within intragenic CGIs *in vivo* and *in vitro* (Maunakea et al, 2010).

In contrast, in another study gene body methylation was shown to decrease transgene expression, which was accompanied by depletion of RNA polymerase II (PolII), H3K4me2, H3K4me3, H3K9ac and H3K14ac in the methylated region, representing a more closed chromatin conformation

(Lorincz et al, 2004). It was proposed that dense intragenic DNA methylation could induce the formation of a closed chromatin structure leading to the reduced efficiency of PolII elongation (Lorincz et al, 2004).

As mentioned previously (see section 1.1.1) H3K36me3 also marks gene bodies of active genes. Interestingly, the PWWP domain of DNMT3A was shown to bind to H3K36me3 *in vitro* and this link may explain the observed positive correlation between gene body methylation and transcription (Dhayalan et al, 2010).

1.4.1.5 Exon/Introns

The concept of gene body methylation has been refined by a close examination of single base resolution methylation maps revealing differential methylation patterns across gene bodies. Methylation levels are usually low at TSSs and slowly rise again within the gene body, the first exon usually being less methylated than the rest of the gene body (Lister et al, 2009). Methylation in exons is generally higher than in introns. Furthermore, the splice junctions (exon-intron boundaries) display characteristic features of methylation with a methylation spike at the 5' splice junction and a sharp drop at the 3' boundary (Laurent et al, 2010). The specific methylation pattern at exon-intron boundaries has been proposed to work in concert with histone marks to influence the recognition of splice junctions and thereby regulate the generation of alternative splice forms (Laurent et al, 2010).

1.4.2 Phylogeny of DNA methylomes

To gain insight into the roles of DNA methylation, phylogenetic studies on DNA methylomes have been undertaken and revealed fascinating insights.

Most organisms have DNA methyltransferases, and in line with this, methylcytosine is present in most species analysed. Yet, some exceptions do exist, for instance the model organisms *Saccharomyces cerevisiae, Schizosaccharomyces pombe, Caenorhabditis elegans* and *Drosophila melanogaster,* which have little or no 5mC in their genomes (Jeltsch, 2010)[8], suggesting that methylation can be replaced by other mechanisms. Methylated genomes can be broadly classified according to their pattern of methylation, which can be global or mosaic (Suzuki & Bird, 2008). Global DNA methylation is typical of vertebrate genomes, where 5mC occurs throughout the genome with the exception of CpG Islands (Lister et al, 2009; Suzuki & Bird, 2008). Mosaic methylation patterns are frequently found in invertebrate genomes, which are characterised by large domains of highly methylated DNA interrupted by long stretches of methylation-free regions. Mosaic methylation in certain fungi, including *Neurospora crassa*, is due to methylation targeted mainly to transposable elements (TEs). TE Methylation is also commonly found in the plant genome, but plants are furthermore characterised by features that they share with mammalian genomes, including methylation of transposons, gene body methylation and lack of methylation at promoters of expressed genes (Jeltsch, 2010). Gene body methylation appears to be highly conserved and was proposed to predate the divergence from plants and animals, while methylation of TEs is more variable and seems to have evolved independently in plants and vertebrates and has not evolved in invertebrates (Feng et al, 2010; Zemach et al, 2010). Gene body

[8] Interestingly, in line with the above-mentioned hypothesis that methylated genomes are depleted in cytosine residues, the genomes of *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae* bear cytosine residues at the expected frequency and CpG-rich stretches similar to CpG islands are not seen (Deaton, Bird, 2011).

methylation typically shows a parabolic relationship with gene expression, i.e. modestly expressed genes being highly methylated, while genes with expression levels at either extreme display lower methylation levels. As discussed in section 1.4.1.4 gene body methylation has been linked to the transcriptional activity of the associated gene, although its significance is currently still under debate.

Another feature characteristic of plant genomes is a high abundance of 5-methylcytosine outside of the CpG context, a feature that – as the most recent genome-wide studies have revealed – is not exclusive to plant genomes..

1.4.3 Non-CpG methylation

5-methylcytosine outside of the CpG context (CHG, CHH, where H=A,C,T) has long been known to exist in plant genomes, where it is thought to play a role in gene silencing and genomic imprinting. Although significant levels of methylation at CpA and to a lesser extent CpT dinucleotides have been previously detected in mouse ESCs (but not in tissues), its significant contribution to the overall levels of methylation has not been widely appreciated until recently and its distribution in the genome has only been revealed with the more recent genome-wide bisulfite sequencing surveys (Ramsahoye et al, 2000). The first high-throughput whole genome bisulfite sequencing study in mammalian cells uncovered the remarkable abundance of non-CpG methylation: it constitutes 25% of all methylation in human ESCs (17% in CHG and 7% in CHH, where H=T,C,A) (Lister et al, 2009). Interestingly, non-CpG methylation was found to be strand-specific, with 98% occurring on one strand only, preferentially on the antisense strand of a gene

body (Lister et al, 2009). Although the relevance of this strand-bias has not been addressed it raises important questions as to its link to transcription. For example, it is conceivable that the antisense strand of a transcribed gene may be better accessible to DNA methylation resulting in the observed strand-biased.

Subsequent genome-wide surveys confirmed the high abundance of non-CpG methylation in ESCs and found mCpA to be the predominant form of non-CpG methylation in human ESCs (Laurent et al, 2010). Furthermore, it was shown that, similar to the distribution of mCpG in the genome, CpA methylation is absent from promoters (CGI and non-CGI promoters) and enriched in gene bodies of active genes.

A recent genome-wide DNA methylation survey in human and mouse brain revealed new fascinating insights into non-CpG methylation in somatic tissues (Lister et al, 2013). Firstly, non-CpG methylation was found to be the dominant form of all methylation constituting 53% and 38% of all cytosine methylation in human and mouse brain, respectively. Secondly, it was highly conserved showing inter-individual correlation of mCH levels genome-wide between independent samples of the same cell type and also between human and mice, indicating that it might be tightly regulated. Interestingly, DNMT3A expression and binding pattern correlate with mCH patterns in brain, indicating a potential role of DNMT3A in mediating mCH methylation. Finally, in support of a putative functional role intragenic mCH is correlated with low expression of the associated gene, (although this contrasts the with its distribution in ESCs where a positive correlation between gene expression

and intragenic non-CpG methylation is observed). Intriguingly, while mCH levels are high, this study demonstrated that mCH levels are negligible in fetal brain, but rapidly accumulates at birth suggesting that this signature is only added later in development and may play an important functional role.

As mentioned above, the occurrence of non-CpG methylation correlates with the expression of DNMT3 enzymes, suggesting that they may be important to establish this mark (Laurent et al, 2010; Lister et al, 2013; Ziller et al, 2011). In support of this notion, knockdown of *DNMT3A* and *B* reduces global levels of non-CpG methylation in human ESCs (Ziller et al, 2011), indicating that non-CpG methylation might be mediated by *de novo* methyltransferases. This is in line with the fact that non-CpG methylation is asymmetric and methylation symmetry is required for replication-based maintenance by DNMT1. It has, therefore, been speculated that non-CpG methylation might reflect hyperactivity of *de novo* methyltransferase activity in ESCs, where DNMT3 enzymes are most highly expressed (Ramsahoye et al, 2000; Smith & Meissner, 2013).

1.4.4 5hmC distribution

The relative contribution of 5hmC to the total amount of methylcytosine was only revealed in 2009, but since then an enormous number of studies have focused on this new base in the genome (Kriaucionis & Heintz, 2009; Tahiliani et al, 2009). It is highly abundant in Purkinje cells, where it constitutes ~40% of all methylation (Kriaucionis & Heintz, 2009). In a simultaneous report, it was shown to make up ~5% of all cytosine bases in MspI cleavage sites in ESCs (Tahiliani et al, 2009). Abundant levels of 5hmC in ESCs and brain have since

been confirmed in a large number of subsequent studies (Ficz et al, 2011; Lister et al, 2013). With the development of new techniques genome-wide maps of 5hmC have shed light on the distribution and potential functions of 5hmC. (For review on methods to assess 5hmC levels see (Shen & Zhang, 2013). 5hmC is mostly associated with euchromatin, where it is particularly enriched at promoters, CGIs, and gene bodies (particularly exons).

It is now well established that 5hmC is generated from 5mC via TET-mediated oxidation (for further details see section 1.5.2). It is therefore proposed to constitute an intermediate of an active DNA demethylation pathway (see section 1.5.2). In line with this notion, 5hmC was found particularly enriched at CGIs of low to medium GC content and in regions of intermediate levels of CpG density (weak promoters), which is reminiscent of the binding preference of TET1 to CpG-rich regions (Xu et al, 2011). Importantly, its enrichment at promoters is dependent on TET proteins, as shown by knockdown of *TET1* and *TET2* leading to an increase in promoter methylation. This was found for a number of promoters of pluripotency genes, which were silenced upon accumulation of methylation at their promoters and ESCs showed an increased propensity for differentiation (Ficz et al, 2011).

In addition to its role in DNA demethylation, it has furthermore been proposed to serve as a stable epigenetic mark with a unique regulatory role. For example, 5hmC was found particularly enriched at gene promoters with a bivalent chromatin signature, which led the authors to propose that it may contribute to the 'poised' state (Pastor et al, 2011). In another study it was shown to recruit MBD3, which preferentially binds to 5hmC relative to 5mC

(Yildirim et al, 2011). Finally, a specific mutation found in some RETT syndrome patients preferentially affects 5hmC binding, suggesting that it plays an important role in human disease (Mellen et al, 2012).

1.5 Establishment and maintenance of DNA methylation patterns

With the rapid advances in technologies that allow the characterisation of genome-wide DNA methylation patterns it has become apparent that DNA methylation patterns are much more dynamic than initially anticipated. The most dramatic changes in methylation are seen during two phases in the life cycle, where global methylation patterns are erased and re-established as part of an epigenetic reprogramming event (see section 1.5.1). Demethylation can occur passively via replication-dependent dilution of 5mC or actively through the activity of catalytic enzymes that remove 5mC from DNA (see section 1.5.2). Re-establishment and subsequent maintenance of DNA methylation patterns requires the activity of DNA methyltransferases (see section 1.5.1 and 1.5.4). Here, I will first give an overview of the dynamics of DNA methylation in early development to then go on to describe our current mechanistic insights as to how these patterns may be determined.

1.5.1 Epigenetic reprogramming

During two phases in development, an epigenetic reprogramming event takes place, in which DNA methylation patterns are erased genome-wide (with few exceptions) to reset the cell's identity and regain pluripotency. The first wave of global demethylation occurs in early development upon fertilization of the oocyte and is important for the establishment of totipotency; parental imprints

are exempt. After fertilization, the paternal genome becomes actively demethylated, while the maternal genome is passively demethylated. Active demethylation occurs via TET3-mediated oxidation (Gu et al, 2011; Iqbal et al, 2011; Wossidlo et al, 2011). Protection from active demethylation of the maternal genome (and a few parental imprinted loci) is achieved via protection from TET3 activity by PGC7 (also known as Stella) (Nakamura et al, 2012; Wossidlo et al, 2011). Methylation patterns are re-established in the inner cell mass of the blastocyst from which ESCs are derived (Li, 2002; Reik et al, 2001). A second wave of global demethylation and remethylation resets parental imprints in primordial germ cells (PGCs) during gametogenesis. Methylation patterns in PGCs are erased genome-wide via a combination of active and passive mechanisms with the exception of a few novel repeat sequences, which escape genome-wide demethylation (Popp et al, 2010). Re-establishment of methylation of the paternal genome begins pre-natally with the establishment of imprints and is complete at the onset of meiosis in the adult, while the maternal genome acquires imprints in the postnatal growth phase of the oocyte. The establishment and maintenance of methylation patterns requires the activity of DNA methyltransferases (see section 1.2.2). How this may be achieved will be outlined in the following sections 1.5.3 and 1.5.4.

1.5.2 DNA demethylation

Due to its rather stable nature, DNA methylation has long been viewed as unidirectional and researchers have been puzzled by the question of how genome-wide DNA methylation patterns are erased during epigenetic

reprogramming. We now know that both active and passive DNA demethylation pathways exist.

Both active and passive DNA demethylation pathways play an important role in epigenetic reprogramming. While passive DNA demethylation occurs simply by replication-dependent dilution of 5mC (or modifications thereof, see below), active DNA demethylation involves the catalytic activity of enzymes to remove this modification from DNA (for review see (Pastor et al, 2013)). From a chemical perspective, the simplest ways to achieve this would be the enzymatic removal of the methyl group of 5mC itself. However, breaking the strong carbon-carbon bond that connects the methyl group and the cytosine base would require enormous catalytic power and is therefore thermodynamically unfavourable. Whether or not such a pathway exists in mammalian cells is still under debate.

Another possible mechanism may be deamination of 5mC to produce T, followed by base excision repair (BER) to replace the mismatched T with unmethylated C. Activation-induced deaminase (AID) and proteins of the apolipoprotein B mRNA editing enzyme catalytic polypeptite (APOBEC) family are enzymes that can carry out this functions[9]. In zebrafish, AID, MBD4 and the DNA repair protein GADD45a can cooperate in demethylating a methylated DNA duplex (Rai et al, 2008). In mammals, AID deficiency results in a partial deficit in DNA methylation erasure (Popp et al, 2010). However, because CpG methylation is usually symmetrical deamination of both strands would give rise to a TG-GT double mismatch, which would generate DNA

[9] AID and APOBEC proteins are normally required for the generation of antibody diversity, RNA editing and retroviral defence. AID was discovered to be essential for somatic hypermutation and class switch recombination of immunoglobulin genes in B-cells.

double strand breaks if processed via BER. It is currently not clear if such a pathway plays a crucial role in demethylation in mammalian cells.

The best-studied mechanism of DNA demethylation involves oxidative demethylation of 5mC to create 5hmC. Based on homology searches of trypanosome thymidine hydroxylases JBP1 and JBP2 the TET family of proteins were identified as candidate enzymes that mediate the conversion of 5mC to 5hmC (more info see section 1.4.2). The TET proteins received their name from the ten-eleven translocation (t(10;11)(q22;q23)) (a chromosomal translocation of chromosome 10 and 11), which fuses the *mixed-lineage leukaemia 1* (*MLL1*) gene located on human chromosome 10 with the *TET1* gene on human chromosome 11 (Lorsbach et al, 2003; Ono et al, 2002).

Since the simultaneous discovery of 5hmC and the TET proteins as candidate enzymes that oxidize 5mC to generate 5hmC in 2009 an enormous amount of evidence has been produced to support their roles in active demethylation pathways. First, TET1 was shown to convert 5mC to 5hmC *in vitro* as well as in cultured cells (Tahiliani et al, 2009). Furthermore, it could be shown that knockdown of *TET1* led to a decrease in the levels of 5hmC, supporting a role of TET1 in active demethylation (Ficz et al, 2011; Tahiliani et al, 2009). It is now well-established that 5hmC is generated from 5mC by TET-mediated oxidation, an assertion supported by the requirement of both pre-existing 5mC and the presence of TET enzymes for the generation of 5hmC in the genome: 5hmC is lost in 5mC-depleted DNMT triple knockout ESCs (Ficz et al, 2011; Williams et al, 2011). 5hmC levels are reduced in TET1-depleted ESCs (Ficz et al, 2011). Additionally, TET3 deficiency leads to a failure of 5mC to 5hmC

to conversion in the paternal genome of zygotes (Gu et al, 2011). In further support of their role in active DNA demethylation, ectopic expression of TET proteins in cell lines leads to a reduction in 5mC levels and an increase in 5hmC levels (Ito et al, 2010).

The question remains how 5hmC as an intermediate of an active DNA demethylation pathway is removed. Removal of 5hmC may occur via passive dilution of 5hmC during DNA replication as reported for pre-implantation embryos or via the activity of TET proteins, which can further oxidize 5hmC to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) (Inoue & Zhang, 2011; Song & He, 2013). These two additional modifications of cytosine residues have been recently detected in the mammalian genome (for review see (Song & He, 2013). 5fC and 5caC occur, however, at extremely low levels (10 to 1000 lower levels compared to 5mC) (Song & He, 2013). A unique role for 5fC and 5caC has so far not been identified and they are instead merely believed to be demethylation intermediates that are actively (via removal by thymine-DNA glycosylases, TDGs) or passively (via dilution) removed. 5hmC can also be deaminated to 5-hyrdoxymethyluracil (5hmU) (Seisenberger et al, 2012). Active removal of 5hmU, 5fC and 5caC is thought to be mediated by TDGs resulting in an abasic site that will be repaired by BER to replace modified C with unmethylated cytosine (He et al, 2011). Passive demethylation of 5fC and 5caC may also occur (Inoue et al, 2011).

Hence, to our current knowledge, cytosine can exist in 6 different forms in mammalian genomes: unmethylated, 5mC, 5hmC, 5fC, 5caC, and 5hmU. For

mapping techniques of these new 5mC derivatives see (Shen & Zhang, 2013; Song & He, 2013; Song et al, 2012a).

1.5.3 Establishment of DNA methylation patterns

Re-establishment of DNA methylation patterns requires the activity of the *de novo* methyltransferases DNMT3A and DNMT3B. This view is based on the finding that DNMT3 enzymes are highly expressed in ESCs and germ cells during the periods of *de novo* methylation and downregulated in somatic cells (Okano et al, 1998a). It is furthermore supported by the finding that DNMT3 enzymes exert *de novo* activity towards fully unmethylated DNA *in vitro* (Okano et al, 1998a). Once DNA methylation patterns are established early in ESCs (and in primordial germ cells), DNA methylation patterns are maintained largely by the activity of the maintenance methyltransferase DNMT1.

1.5.4 Maintenance of DNA methylation patterns

Maintenance of DNA methylation patterns is thought to be mainly achieved by the activity of DNMT1. DNMT1 is mostly transcribed during S-phase (Robertson et al, 2000), where it localizes to the replication fork (Leonhardt et al, 1992; Liu et al, 1998) via interaction with PCNA and/or UHRF1 to transfer methyl groups onto hemimethylated DNA, which is in line with the marked preference of DNMT1 for hemimethylated DNA (Bestor & Ingram, 1983; Hermann et al, 2004). However, this strict separation of the functions of DNMT1 and DNMT3 enzymes into maintenance and *de novo* methyltransferases seems oversimplified when considering several observations (as outlined below). A revised model has been proposed instead.

1.5.5 Revised model of establishment and maintenance of methylation

Although DNMT1 and DNMT3A/B enzymes are classically referred to as maintenance and *de novo* methyltransferases, respectively, this long-held view seems rather obsolete and has been challenged by several observations (Jones & Liang, 2009). There is evidence that DNMT3A/B contribute to DNA methylation maintenance: double knock out of DNMT3 A and B in ESCs results in a gradual loss of methylation with increased levels of hemimethylation of CpG-rich sequences (including L1 elements) indicative of an impairment in maintenance methylation (Chen et al, 2003; Liang et al, 2002). In support of this, DNMT3A and B display the same propensity for methylating hemimethylated DNA as for unmethylated duplex DNA (Latham et al, 2008), and DNMT3A and B are still expressed in somatic tissues, albeit at reduced levels (Jones & Liang, 2009; Latham et al, 2008; Okano et al, 1998a). Conversely, DNMT1 does, in fact, exert *de novo* methyltransferase activity on unmethylated DNA (Pradhan et al, 2008; Pradhan et al, 1999).

In summary, the observations suggest that maintenance methylation might be more complex than previously described. In particular, it has been noted that the simplified model does not take into account any mechanism of error correction, a feature common to most, if not all, biological systems (Jones & Liang, 2009). Hence, a revised model has been put forward instead (Jones & Liang, 2009). In this model, the bulk of DNA methylation in dividing cells is maintained by DNMT1 following the replication fork to transfer methyl group to hemimethylated DNA. After DNMT1 has left the replication fork error correction may still be performed, which may be achieved via additional targeting of DNA methyltransferases to DNA, including DNMT3 enzymes.

Having outlined how DNA methylation patterns are erased and re-established, a key open question is what determines DNA methylation patterns. In particular, a key open question is how DNMTs may be recruited to or excluded from specific regions in the genome.

1.5.6 Recruitment of DNMTs to DNA

Given the observation that most of the CpG content of the genome is highly methylated and because DNMTs lack intrinsic sequence specificity, it has been proposed that DNA methylation may be an indiscriminate process that affects all CpGs irrespective of the underlying sequence (Bird, 2002). The exceptions are unmethylated regions, such as CGIs, which appear to be protected from methylation. The reasoning behind this model is that any type of DNA sequence could potentially be methylated, i.e. there is nothing refractory about a DNA sequence and even CpG-rich DNA stretches, such as CpG Islands, can be methylated under certain circumstances, for example intra- or intergenic CGIs which become methylated during development or in abnormal cells (Esteller, 2002; Mohn et al, 2008). Yet, depletion of each of the different DNMTs results in the loss of methylation in specific regions in the genome (see section 1.2.2), suggesting that DNMTs do discriminate between DNA sequences and targeting may be specific. Furthermore, cofactors have evolved to target DNMTs to specific regions in the genome, including DNMT3L (see section 1.2.2.3.1). Other co-factors, for instance ATRX and LSH2, members of the SWI/SNF chromatin remodelers, are required for normal methylation of repetitive DNA, suggesting the need for modulation of the chromatin structure by remodelers to enable access of DNMTs to DNA

(Dennis et al, 2001; Gibbons et al, 2000). Given the evidence for selectivity of DNA methylation targeting renders the description “default” rather obsolete.

In summary, several lines of evidence suggest that DNMTs are specifically recruited to DNA, a process, which is currently not well understood.

1.5.7 Keeping CGIs in an unmethylated state

A key question regarding the establishment and maintenance of genome-wide DNA methylation patterns in mammalian cells is how CGIs remain unmethylated. It has been put forward that unmethylated regions in the genome, such as CGIs, are protected from DNA methylation (Bird, 2002). How this may be achieved is currently not well understood, but a few possible factors involved will be discussed below.

A putative role of H3K4me3 in inhibiting methylation has been mentioned above (see section 1.1.1). Binding of ADD-containing DNMT3 is allosterically blocked by H3K4 methylation (Ooi et al, 2007; Otani et al, 2009). CXXC finger protein 1 (CFP1), which is found at CGIs, can recruit H3K4methyltransferases to CGIs via its ZF-CXXC domain. It has been proposed that the CG-rich DNA is sufficient for CFP1 recruitment as CG-rich transgenes containing no promoter sequences are sufficient to induce a hypomethylated state of the CG-rich DNA (Thomson et al, 2010). Yet, the role of H3K4me3 in preventing *de novo* methylation is not clear, as loss of H3K4me3 by CFP1 depletion does not result in promoter hypermethylation (Clouaire et al, 2012). A possible explanation may be redundant protective mechanisms involved (Clouaire et al, 2012). Members of the H3K4 methyltransferases (MLL) were shown to inhibit DNA methylation at developmental genes, which likely involves its ZF-

CXXC domain (Erfurth et al, 2008). Of note, another class of ZF-CXXC domain proteins that may be important for protecting CGIs from methylation are the TET enzymes, which have been implicated in DNA demethylation (see section 1.5.2). Interestingly, TET1 is enriched at CGIs and this has been proposed to remove aberrant DNA methylation and increase DNA methylation fidelity at CGIs (Williams et al, 2011). In agreement with this idea, TET1 knockdown leads to increased methylation levels at CGIs (Xu et al, 2011).

Several lines of evidence suggest that transcription factors (TFs) play a key role in determining a hypomethylated state of a sequence. Firstly, mutation of SP1 binding sites in the CGI at the mouse *APRT* gene promoter leads to hypermethylation of the CGI (Brandeis et al, 1994; Macleod et al, 1994). Secondly, 1kb promoter fragments retain their hypomethylated state at an ectopic site and this is critically dependent on binding motifs within the fragments. Mutating DNA binding motifs (including SP1, CTCF) induces *de novo* methylation of the fragments (Lienert et al, 2011). Thirdly, the significance of TFs is supported by genome-wide surveys, which show that cell type-specific TFs are enriched in hypomethylated regions in a cell type-specific manner (Stadler et al, 2011). Finally, depletion of the TF REST in mouse ESCs results in hypermethylation of the binding sites of REST, which are normally unmethylated (Stadler et al, 2011).

In support of this, CpG Islands often coincide with promoter elements. Even non-promoter CGIs often share characteristics of promoter elements and carry active marks (Illingworth et al, 2010), see section 1.4.1.1). In fact, in several instances CGIs identified in unusual locations prove to be promoters

(Lyle et al, 2000; Macleod et al, 1998; Wutz et al, 1997). How promoter elements may protect from DNA methylation is not known. In particular, the question arises whether protein binding *per se* can prohibit methylation or whether transcription initiation is required. In support of the first, promoter hypomethylation is observed in the absence of measureable transcription (Lienert et al, 2011). In support of the latter is an elegant study that gave mechanistic insights into how transcription may influence DNA methylation. It was observed that at many housekeeping promoters, an unusual structure, called R-loop, could be formed through RNA-DNA hybrid formation between the nascent RNA and the template DNA strand. This structure is formed when the gene is transcribed and prohibits the recruitment of *de novo* methyltransferases (Ginno et al, 2012).

In conclusion, it appears that several mechanisms exist which protect CpG Islands from methylation through binding of transcription factors and/or recruitment of antagonistic histone marks.

1.5.7.1 Recognition of silent DNA

Several lines of evidence suggest that DNA methylation may be targeted to regions in the genome that are already silent constituting a secondary effect of silencing rather than being an initiating effect. A few examples were mentioned previously, such as the dispensability of *de novo* methylation for the initiation of X-inactivation, where the genes become methylated when they have already been silenced (see section 1.3.1). Indeed, X-inactivaton, which is initiated by expression of the non-coding Xist (X inactive specific transcript) RNA, precedes methylation of its CGIs by several days (Lock et al, 1987).

Another example mentioned above is *OCT4*, which becomes methylated only after it has already been transcriptionally silenced (see section 1.3.4).

Common to all of the above-mentioned examples is that methylation is established at genes that are already transcriptionally silent. Hence, it has been proposed that transcriptional repression might trigger *de novo* methylation during embryogenesis (Bird, 2002). The question arises what then may trigger methylation of silent regions? A potential trigger for methylation of silent regions may be histone modifications, such as H3K9me3, which is associated with transcriptionally inactive heterochromatin (see section 1.1.1). Depletion of the H3K9me3 methyltransferases SUV39H1 and 2 in mESCs results in the delocalization of DNMT3B from heterochromatic foci and alterations of DNA methylation at pericentric heterochromatin (Lehnertz et al, 2003). Vice versa, DNMT1 or DNMT3A/B-deficient mESCs do not show impaired H3K9me3 patterns suggesting that DNA methylation is not important for H3K9 methylation of pericentric heterochromatin (Lehnertz et al, 2003). The histone methyltransferase G9A, responsible for H3K9me2 in euchromatic regions (Tachibana et al, 2002), is required for recruitment of DNMT3A to retrotransposons and methylation thereof (as well as a subset of non-repetitive DNA, for instance CpG-rich promoters) (Dong et al, 2008). Whether or not the histone mark itself can be recognized by DNMTs is not known, but physical interactions between DNMTs and the histone methyltransferases G9A and SUV39H1 indicate that targeting of DNA methylation to sites of H3K9 methylation may be indirect (Esteve et al, 2006; Fuks et al, 2003; Geiman et al, 2004). DNMT3B may also be recruited to sites of H3K9 methylation via interaction with proteins that recognize H3K9me2/3, such as

HP1 (Geiman et al, 2004; Lehnertz et al, 2003). Furthermore, UHRF1, which interacts with DNMT1 (see 1.5.4), was found to also bind to H3K9me3 (and unmethylated H3K4), providing a further link between H3K9 methylation and DNA methylation (Nady et al, 2011; Rottach et al, 2010). In summary, recruitment of DNMTs to silent DNA may be triggered directly or indirectly via H3K9 methylation.

1.6 The α-globin genes

The *α*-globin genes are amongst the most well characterised genes in terms of gene regulation. In human and mouse they are expressed in late-stage erythroid cells where their gene product forms part of a large protein complex, hemoglobin, which transports oxygen from the lung to the rest of the body. Downregulation of the *α*-globin genes in human results in a certain form of anemia, called *α*-thalassemia.

1.6.1 Genomic organization

The genomic organization of the α-globin genes and their *cis*-regulatory elements in different species were revealed by comparative sequence analysis of the *α*-globin genes in 22 species covering 500 million years of evolution (Hughes et al, 2005). They are located within a gene-dense region of constitutively expressed genes, where they are arranged in the same order in the different species analysed (conserved synteny). In all of the 22 species studied, there is an embryonic globin gene (*ζ* in human and mouse), at least one *α*-globin gene and in most species there are two minor globin genes (*αD* and *θ*, pronounced theta). A 135kb region of conserved synteny is considered

to bear all of the *cis*-regulatory elements necessary for regulation of the *α*-globin genes (Hughes et al, 2005).

Comparative sequence analysis of the *α*-globin genes revealed four highly conserved elements; the multispecies conserved elements 1 - 4 (MCS-R1-4) (Hughes et al, 2005). Their relevance has been supported by functional studies on patient-derived mutants as well as experimental mutational analyses. Natural mutants, in which the *α*-globin genes are downregulated, often show large deletions that either remove the *α*-globin genes themselves or their upstream regulatory elements (UREs) (Horsley et al, 2001; Viprakasit et al, 2006; Viprakasit et al, 2003). Mutational analyses in experimental systems confirmed the relevance of the UREs MCS-R1-4 for *α*-globin expression and identified MCS-R2 as the major URE, because it appears to be the only element capable of driving high levels of *α*-globin gene expression on its own (Higgs et al, 1990). The remote regulatory elements are associated with erythroid-specific DNase I hypersensitive sites (DHSs) reflecting their cell type-specific activity (Flint et al, 1997; Vyas et al, 1992) and harbour erythroid-specific TF binding motifs, such as GATA1, GATA2, SCL, NF-E2 and EKLF.

1.6.2 The human α-globin genes

The human *α*-globin genes are located in the subtelomeric region of the short arm of chromosome 16 (16p13.3) (Figure 1**,** Figure 2). This region comprises a cluster of genes, including the *α*-like globin genes (*ζ* (zeta), *α*D, *α*2, *α*1, and Θ (theta)), which are regulated by four multispecies conserved sequences (MCS-R1-4). The *α*-globin genes are associated with large CGIs, which have been predicted bioinformatically as well as functionally (Figure 1, Figure 3).

Their hypomethylated state has been verified at a few loci in ESCs, erythroid and non-erythroid tissues (Bird et al, 1987). CGI hypermethylation of the *α*-globin genes was found in a patient with *α*-thalassemia (Tufarelli et al, 2003). Other than that, little is currently known about the pattern of DNA methylation in the human *α*-globin cluster. The role of DNA methylation in the human *α*-globin cluster is not well understood.

1.6.3 Regulation of α-globin expression

The *α*-globin genes are located within a region of generally open chromatin, in which the *α*-globin genes are tightly regulated in a tissue-specific manner. Cell type-specific expression is achieved via the assembly of general transcription factors (TFs) as well as erythroid-specific TFs (reviewed in (Higgs et al, 1990)). Erythroid-specific TFs, including GATA2, SCL, NF-E2, first bind to the UREs in early progenitors (GATA2 is subsequently replaced by GATA1) and then to the *α*-globin gene promoter elements. Further erythroid-specific TFs are recruited to the UREs, which is associated with the appearance of DHSs (Anguita et al, 2004). At this stage expression is readily detectable by *in situ* RNA hybridization, but expression levels are still low (Brown et al, 2006). Active transcription involves conformational changes, which brings promoter and UREs in close proximity via chromosomal looping (Vernimmen et al, 2007). Expression levels are highest levels in intermediate erythroblasts (Anguita et al, 2004; Brown et al, 2006; De Gobbi et al, 2007).

The *α*-globin genes are transcriptionally silent in hematopoietic stem cells, early progenitors and non-erythroid cells. In ES and progenitor cells they are marked by a bivalent chromatin state (H3K4me3 and H3K27me3), which

renders them silent (although low levels of transcription can be detected in ESCs) (De Gobbi et al, 2011). Polycomb complex remains bound in non-erythroid differentiation, while it is cleared specifically during erythropoiesis resolving the bivalent state during erythroid differentiation (Garrick et al, 2008). Gene activation is associated with activating histone marks, including H3Ac, H4Ac, H3K4 methylation (H3K4me1/2/3) (De Gobbi et al, 2007). During the final stages of maturation the entire *α*-globin domain becomes acetylated with peaks at the MCS-Rs as well as promoter elements.

1.6.4 A transgenic mouse model for the human α-globin genes

In order to study gene regulation of the human *α*-globin locus, a transgenic mouse was created in which a 117 kb DNA fragment containing the human *α*-globin cluster was placed in the mouse replacing the mouse syntenic region on chromosome 11 (Figure 1) (Wallace et al, 2007). It was previously shown that this 117 kb DNA fragment containing the *α*-globin genes and their UREs (MCS-R1-4), which lie up to 70 kb away, is sufficient to drive normal expression of the human *α*-globin genes in the mouse. In this mouse model the human *α*-globin genes are expressed in the correct tissue- and developmental stage-specific manner, albeit at reduced levels (50% of the endogenous mouse *α*-globin expression level), and viable mouse strains homozygous for the humanised allele could be created (Wallace et al, 2007). Correct tissue- and developmental stage-specific regulation of the genes was also verified on the molecular level: TF binding (GATA1), histone modifications (H3K4me1 and H3 acetylation) and interaction of promoter with enhancer elements was found to be normal in this mouse model (Vernimmen

et al, 2009). The DNA methylation patterns of the human *α*-globin genes in the mouse model have so far not been characterised.

1.7 Transgenic mouse models to study human disease

To gain insight into the molecular mechanisms that underlie human diseases it is pivotal to understand how human genes are normally regulated. However, studies on human gene regulation can be difficult due to limitations in the accessibility of human cells, ethical constraints and the involvement of high costs. Therefore, mouse models are frequently used as model organisms to gain insight into the principles of mammalian gene regulation. Although mouse model systems have proven to be valuable tools to further our understanding of the principles underlying mammalian gene regulation, several studies have previously reported that engineering human disease-causing mutations in the mouse do not phenocopy the same disease patterns in the mouse (Anguita et al, 2002; Engle et al, 1996; Garrick et al, 2006). For example, deletion of the major URE in the mouse *α*-globin cluster causes a much milder form of *α*-thalaessemia than in a similar deletion in human, indicating possible differences in the regulation of the *α*-globin genes in human and mouse (Anguita et al, 2002).

To overcome this problem, transgenic mouse models have been developed over the past 30 years to place human genes in the mouse to have an easily accessible system for the study of human gene regulation in the mouse. This was made possible with the development of techniques that enable the introduction of foreign DNA into the germ line of mice to create transgenic mouse models in the early 1980s and has led to the generation of mouse

models carrying small transgenes of human genes (for review see (Bedell et al, 1997). It was soon realized, however, that human genes might require sequences located far away from the gene so that a transgene construct often does not contain all of the necessary regulatory sequences. Further limitations of small transgenes include variable copy number and random integration in the genome, both of which are confounding factors in the analysis of gene expression of the transgene. However, one of the major problems with small transgenes is that they often undergo silencing (Garrick et al, 1996). Their activity can be influenced by their flanking DNA by chromatin-induced effects (position-effect variegation, PEV) or as a result of repetitive integration of multiple copies of the transgene in concatameres in the same position, which can lead to repeat-induced gene silencing (Alami et al, 2000; Garrick et al, 1998; Kaufman et al, 1999; Wilson et al, 1990). Even with large transgenic DNA fragments, which were made possible with the development of P1-derived, bacterial or yeast artificial chromosomes (PACs, BACs, YACs), these problems persist. Although large transgenes are more likely to contain the distal regulatory elements required for normal expression, are less likely to be affected by genomic position and transgene copy number is usually lower, a problem with large transgenes is that they are often subject to rearrangements, and due to their large size, it is often difficult to analyse their structural integrity and copy number determination is difficult, which may confound the analysis and interpretation of studies on their gene regulation.

As a next step was the development of models in which large chromosomal segments of human DNA were placed into another species (O'Doherty et al, 2005). In an extreme example an aneuploidy mouse strain carrying hs21 was

developed to recapitulate Down's syndrome (Tc1 mouse). This mouse model was later used to dissect out *cis*- and *trans*-regulatory mechanisms of gene regulation by comparing TF binding and chromatin modifications in this mouse model compared to normal human and mouse cells (Wilson et al, 2008). Similarly, a transgenic mouse model carrying a large human DNA segment in place of the mouse syntenic region has been engineered to study human *α*-globin gene regulation in the mouse (Wallace et al, 2007) (see section 1.6.4). Our current knowledge of the epigenetic patterns of such large DNA segments in a heterologous environment is rather limited.

1.7.1 Key open questions

Some of the key open questions in the field of DNA methylation are how genome-wide DNA methylation patterns are established and maintained. More specifically, it is not well understood how DNMTs are recruited to specific regions in the DNA, how certain regions, such as CGIs, may be protected from methylation by DNMTs, and what role DNA demethylation plays in determining genome-wide DNA methylation patterns. In particular, a key open question is how do cellular factors recognise DNA to confer DNA methylation, protection from methylation or, given the recent discovery of abundant levels of 5hmC in mammalian genomes, potentially lead to DNA demethylation.

1.7.2 Outline of the thesis

The aim of this project is to gain insight into the molecular mechanisms that are involved in establishing and maintaining DNA methylation patterns in mammalian cells. To do this, a humanised mouse model, which carries the

human *α*-globin genes in place of the mouse syntenic region (described in section 1.6.4) was employed to find out how well DNA methylation patterns are established and maintained in a heterologous murine environment compared to its natural state in human cells. DNA methylation patterns of the human *α*-globin genes were analysed in the humanised mouse model and compared to the pattern in normal human cells. First, DNA methylation patterns were screened using a low-resolution approach and it was found that the region around the *α*-globin genes is hypomethylated in various humanised mouse tissues compared to human tissues (chapter 3). Next, DNA methylation patterns were characterised in detail in human and humanised mouse ESCs to find out how well DNA methylation patterns are established in humanised mouse ESCs compared to its natural environment in human ESCs. It was found that the human *α*-globin gene region becomes hypomethylated in humanised mouse versus human ESCs, particularly in regions near CpG islands, in so-called CpG island shores. Next, to investigate whether the pattern changes in differentiated cells, DNA methylation patterns were characterised in erythroid and non-erythroid tissues and it was found that hypomethylation near the *α*-globin genes becomes exacerbated in somatic cells due to an extension of the hypomethylated region in both erythroid and non-erythroid cells in the humanised mouse model compared to human.

To identify any cis-regulatory sequences responsible for the hypomethylated state of human CGI shores in the mouse, 2-4 kb human test sequences containing CGI H (associated with human α2) and its adjacent hypomethylated shore were re-integrated into the mouse α-globin locus via

recombination mediated cassette exchange (chapter 4). Human CGI shores became hypomethylated in the context of the re-integrated test sequences, indicating that the appearance of hypomethylation is determined by the underlying human DNA sequence in the test fragments. Finally, it was sought to investigate whether the unmethylated region around CGIs generally becomes extended when human DNA is placed into the mouse. To do this, a previously generated mouse model carrying a copy of hs21 was employed to compare the extent of non-methylated regions around CGIs of hs21 in somatic cells of mouse versus human (chapter 5). The significance of these findings is discussed in chapter 6.

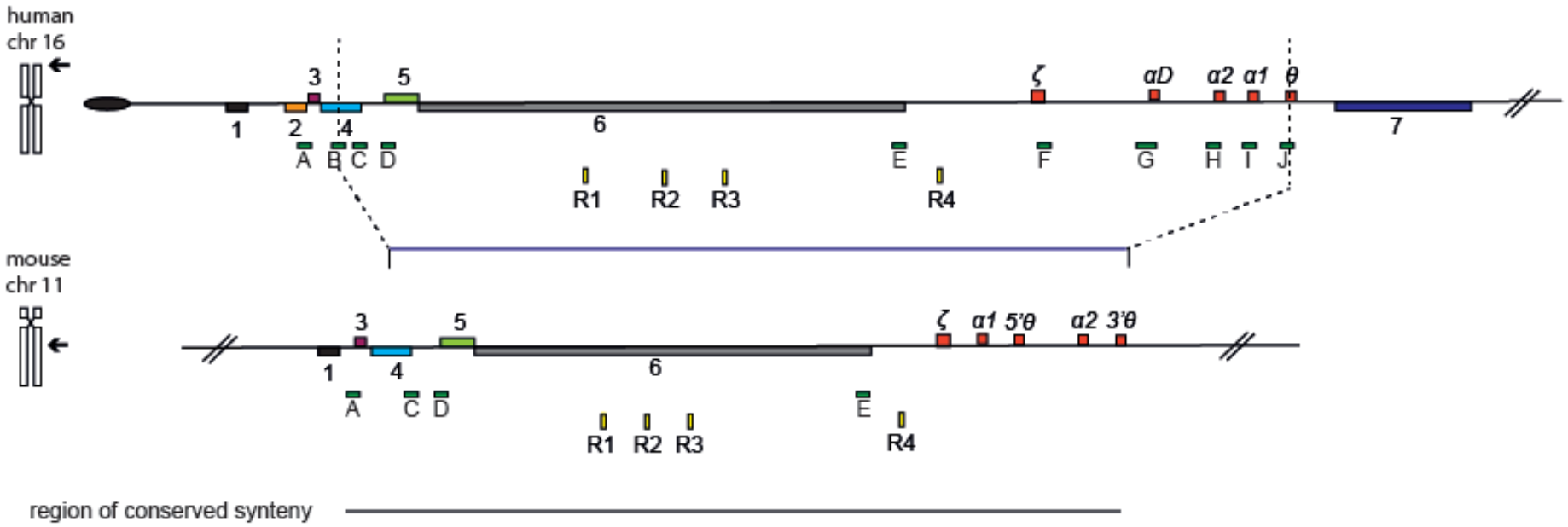

Figure 1: Schematic of the *α*-globin gene cluster in human and mouse.

The region of conserved synteny (as indicated) contains the same sets of genes in the same order in both species. The *α*-globin genes are shown in red. They lie in a relatively gene-dense area. Other genes in the region (indicated as colored boxes) are 1: *IL9RP3*, 2: *POLR3K*, 3: *C16ORF33*, 4: *RHBDF1*, 5: *MPG*, 6: *NPRL3*, 7: *LUC7L*. Genes depicted above the line are transcribed from left to right and those below are transcribed from right to left. CGIs (A-J) are depicted as green boxes. MCS-R1-4 (R1-4) are indicated as yellow boxes. In the humanised mouse model the human transgenic region (chr16: 53417-170901, hg18, demarcated by dashed lines) replaces the mouse syntenic region on chromosome 11 (indicated by a blue line). (Figure adapted from (Wallace et al, 2007)).

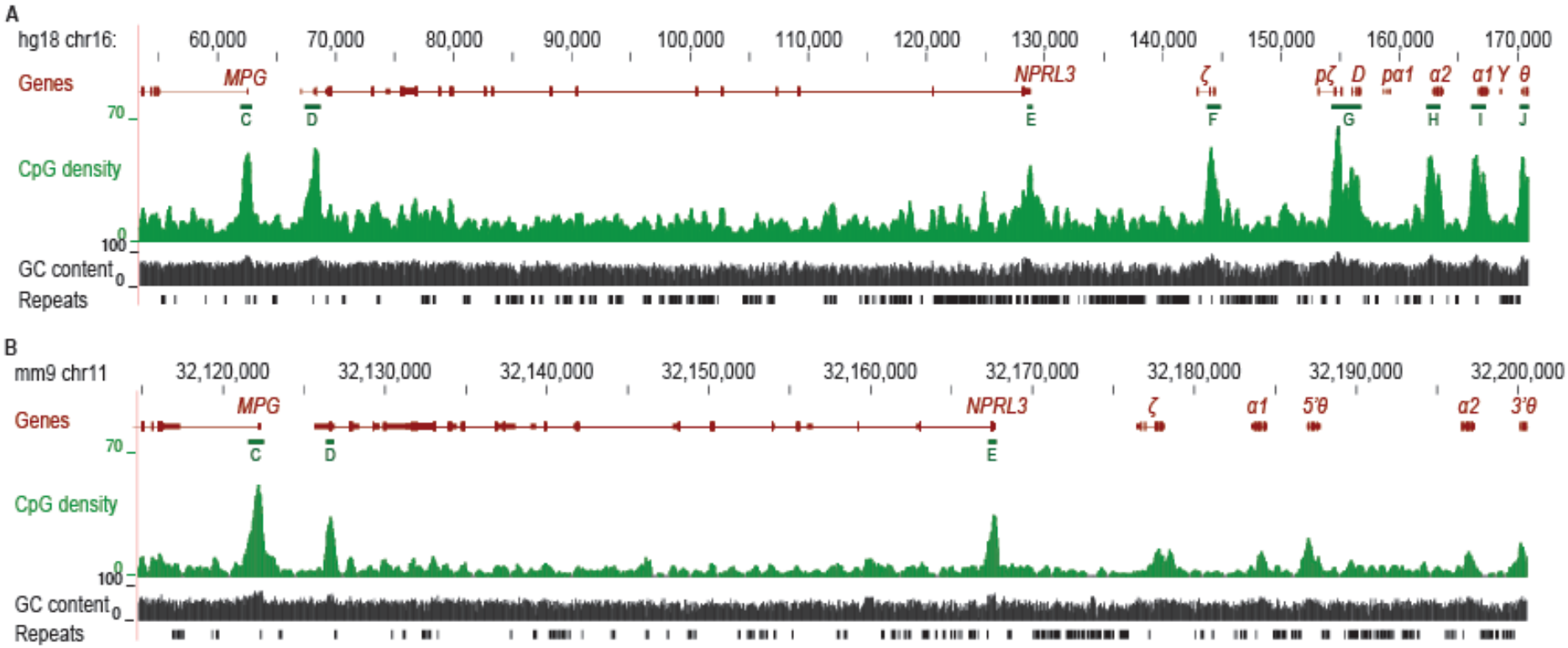

Figure 2: Overview of human and mouse *α*-globin regions.

A) The human *α*-globin gene cluster is located on chromosome 16 (53,417-170,901, NCBI36/hg18). B) The mouse syntenic region is located on chromosome 11 (chr11:32,114,759-32,200,537, NCBI36/mm9). Genes are shown in red (Ensembl Gene Predictions), CGIs C-J and the CpG density are shown in green. UCSC tracks (from top to bottom) show the genomic location, genes, CpG density per 300bp, GC content (in 5bp sliding windows), repeats (RepeatMasker).

2 Materials and Methods

2.1 Materials

2.1.1 Chemicals, enzymes and reagents

1-thioglycerol, 1-TG (Sigma-Aldrich, cat.no. M6145-25ml)

2-Mercaptoethanol (Sigma-Aldrich, cat.no. M7522-100ml)

AcuI (New England Biolabs, NEB, cat.no. R6041L)

Agarose, UltraPure™ (Invitrogen, cat.no. 16500500)

Amersham Autoradiography Hypercassette (Amersham, cat.no. RPN11643)

Amersham Hyperfilm MP (Amersham, cat.no. 28-9068-46)

Amersham Intensifying Hyperscreen (Amersham, cat.no. RPN1663)

Ampicillin (Invitrogen, cat.no. 11593027)

Amplitaq Gold® 360 DNA polymerase (Roche, cat.no. 4398823)

AmpliTaq Gold® PCR Buffer (10x) (Roche, cat.no. 4398823 or 4398853)

BD Falcon™ centrifuge tube, 15ml (VWR, cat.no. 430790)

BD Falcon™ centrifuge tube, 50ml (VWR, cat.no. 352070)

BglII (NEB, cat.no. R0144L)

Boric acid (Sigma-Aldrich, cat.no. 0394)

Bovine Serum Albumin (BSA) (NEB, cat.no. B9000S/ Sigma-Aldrich, cat.no. A2153)

Bromophenolblue (Sigma-Aldrich, cat.no. B8026)

Chloroform (Sigma-Aldrich, cat.no. C2432-500ml)

Corning® cell culture flasks, 175cm^2 (Sigma-Aldrich, cat.no. CLS431080)

Corning® cell culture flasks, 25cm^2 (Sigma-Aldrich, cat.no. CLS431639)

Corning® cell culture flasks, 75cm^2 (Sigma-Aldrich, cat.no. CLS431641)

Corning® tissue culture treated cell culture dish, 100mmx20ml (Sigma-Aldrich, cat.no. CLS431067)

CpGenome Universal Methylated DNA (Millipore, cat.no. S7821)

Complementary RNA (cRNA) [1µg/µl] (Qiagen)

Cryogenic vials, sterile, 1.5ml (Evergreen Scientific, cat.no. 222-3902-085)

dATP (NEB, cat.no. N0440S)

Dextran sulphate (Sigma-Aldrich, cat.no. D6001-100g)

DMSO, Dimethyl sulphoxide dehydrated AnalR NORMAPUR® (VWR, cat.no. 23500.260)

DNA LoBind Tube (Eppendorf, cat.no. 022431021)

dNTP (Invitrogen, cat.no. 10297018)

DTT, DL-Dithiothreitol (Sigma-Aldrich, cat.no. D9779-10g)

Dynabeads® M-280 Streptavidin, 2ml (Invitrogen, cat.no. 11205D)

EagI (NEB, cat.no. R0505S)

EDTA, Ethylenediaminetetraacetic acid disodium salt dehydrate (Sigma-Aldrich, cat.no. 5134-500g)

Ethanol (VWR, cat.no. 10107-2L)

Ethidium Bromide, UltraPure™ [10mg/ml] (Invitrogen, cat. no. 15585011-10ml)

FBS, fetal bovine serum Gibco®, each batch tested in-house for ESCs (Life Technologies, cat.no. 10270106-500ml)

Ficoll® PM 400 (Sigma-Aldrich, F4375 – 100g)

Formamide (Sigma-Aldrich, cat.no. 47671)

GelRed Nucleic Acid Gel Stain, 10 000x in water (cat.no. 41003)

gelatin Sigma 2% G1393-100ml

Gelatin, 2%, 100ml (Sigma-Aldrich, cat.no. G1393)

Gene Pulser/MicroPulser Electroporation cuvette (BioRad, cat.no. 165-2088)

GMEM, Glasgow's Minimal Essential Medium (Gibco, cat.no. 21710-025)

Hand towels (Kimberly-Clark)

HCl, hydrochloric acid, concentrated, 37% (VWR, cat.no. 20252.335)

HotStarTAQ™ DNA polymerase (Qiagen, cat.no. 203205)

HpaII (NEB)

HT (Gibco, cat.no. 41065-012-100ml)

hypoxanthine-aminopterine-thymidine (HAT) 21060-017

illustra ProbeQuant G-50 Micro Columns (GE Healthcare, 28-9034-08).

Isopropanol, 2-propanol AnalR NORMAPUR® (VWR, cat.no. 20842.323)

Klenow DNA polymerase (3' > 5' exo^-) (NEB M0212S)

L-Glutamine (Gibco, cat.no. 25030024-100ml)

Library Efficiency® DH5α^{TM} Competent Cells (Invitrogen 18263-012)

Leukemia inhibitory factor (LIF)-conditioned medium (prepared by Sue Butler)

low melting point agarose gel (SeaPlaque^{TM}, Lonza)

MethylEasy Xceed Kit (Human Genetic signatures, ME002)

$MgCl_2$ (Roche)

microTUBE AFA Fiber Pre-Slit Snap-Cap 6x16mm (Covaris, cat.no. 520045)

NaOH (VWR, cat.no. 28244.295)

Na_2HPO_4 (Sigma-Aldrich, cat.no. 255793)

NaH_2PO_4, sodium dihydrogen orthophosphate dehydrate (BDH, cat.no. 301324Q)

Non-essential Amino Acids Solution, 100x (Gibco, cat.no. 11140-035-100ml)

NotI (NEB, cat.no. R0189S)

Orange G (Sigma-Aldrich, cat.no. O3756-25g)

phosphate-buffered saline (PBS) pH7.4 (Gibco, cat.no. 0010-015 – 500ml or 10010-056-10x500ml)

PenStrep, Penicillin-Streptomycin, 100x (Gibco, cat.no. 15140122)

Pfu Hotstart Cx DNA Polymerase (2.5U/ul Stratagene cat.no. 600410)

pGEM®-T Easy Vector (Promega, cat.no. A1360)

phenol solution, equilibrated with 10mM Tris HCl (Sigma-Aldrich, cat.no. P4557)

Polyvinylpyrrolidone (Sigma-Aldrich, cat.no. PVP-360 – 100g)

Proteinase K (Sigma-Aldrich, cat.no. 25530-015)

PstI (NEB, cat.no. R0140S)

Pwo 5U/ul (Roche, cat.no. 11408835)

Radioactive [α^{32}P]-dCTP (3000Ci/mmol, 10mCi/ml, Perkin Elmer, cat.no. NEG513H250UC)

Reservoir (Sigma-Aldrich, cat.no. CSL4871)

Sodium dodecyl sulphate, SDS, 20% (Amresco, cat.no. 0837)

Sequenase version 2.0 [13U/µl] (Affymetrix, cat.no. 70775Y/Z)

SOC medium (supplied with Library Efficiency® DH5α^{TM} Competent cells)

Sodium Pyruvate solution 100mM (Gibco, cat.no. 11360-039-100ml)

MAX Efficiency® Stbl2™ Competent Cells (Invitrogen, cat. no. 10268-019)

SSC, saline sodium citrate (Severn Biotech Ltd, cat.no. 20-6500-50)

Sterile Water (Sigma-Aldrich, cat.no. 77321-18-5)

Streptavidin-coated magnetic Dynabeads (Life Technologies™, cat.no. M-280)

Sucrose (Sigma-Aldrich, cat.no. S0389)

Syringe filters, 0.2µM (Appleton Woods, cat.no. FC121)

Syringe filters, 0.45µM (Appleton Woods, cat.no. FC122)

T4 DNA Ligase (Roche, cat.no. 799 009)

T4 DNA polymerase [3U/ul] (NEB, cat.no. M0203S)

T4 Polynucleotide kinase [10U/ul] (NEB, cat.no. M0201S)

Tris-EDTA (TE) buffer solution E (Sigma-Aldrich, cat.no. 93283)

Tris-Acetate (Sigma-Aldrich, cat.no. T1258)

Tris-Base (Sigma-Aldrich, cat.no. T1503)

Trypsin-EDTA (1x), 0.25%, phenol red (Gibco, cat.no. 252000-056)

Tryptone (Sigma-Aldrich, cat.no. T7293)

Whatman™ 3MM chromatography paper (GE healthcare, cat.no. 3030917)

X-gal, 5-bromo-4-chloro-3-indolyl-ß-D-galactopyranoside (Qiagen, cat.no. 129931)

Zeta-Probe Membrane (Bio-Rad, 162-0159)

2.1.2 Kits

Qiaprep Spin Miniprep Kit (Qiagen, cat.no. 27106)

QIAfilter Plasmid Maxi Kit (Qiagen, cat.no. 12263)

MiniElute PCR Purification Kit (Qiagen, cat.no. 28006)

PCR purification kit (Qiagen, cat.no. 28106)

EZ DNA Methylation-Gold™ Kit (Zymo Research, cat.no. D5005)

Amersham Megaprime DNA Labelling System (GE Healthcare, cat.no. RPN1607)

Rapid DNA Ligation Kit (Roche, cat.no. 11635379001)

MethylEasy™ *XCeed* Rapid DNA Bisulfite Modification Kit (Human Genetic Signatures, cat.no. ME002)

2.1.3 Equipment

Qubit® 2.0 Fluorometer (Life Technologies)

video gel documentation system (Gel-Doc2000; BioRad Laboratories, Hercules CA, USA)

UV light (Gel-Doc2000; BioRad Laboratories, Hercules CA, USA).

ABI Prism 7000 sequence detection thermocycler (Applied Biosystems)

UV Stratalinker 2400 (Stratagene)

Hybridization incubators (Techne)

BioRad Gene Pulser^{TM} & Capitance Extender (BioRad)

Ecotron incubation shaker (Infors HT)

Covaris Adaptive Focus Acoustics (K bioscience)

SpeedVac lyophiliser, Christ Vacuum Concentrator 2-25 (SciQuip)

ThermoMixer Comfort (Eppendorf)

Centrifuge (Eppendorf)

NanoDrop spectrophotometer (Thermo Scientific)

Qubit fluorometer (Life Technologies)

Incubator (LTE Scientific, IP-60)

PCR machine (BioRad)

video gel documentation system (Gel-Doc2000; BioRad Laboratories, Hercules CA, USA).

storage phosphor imaging system (GE healthcare, Typhoon™)

2.1.4 Buffers

Lysis buffer

10mM	Tris pH8
10mM	NaCl
10mM	EDTA
0.5%	SDS
50μg/ml	proteinase K

LB medium

10g	Tryptone
5g	yeast extract
10g	NaCl/L
100μg/ml	ampicillin

TBE (Tris-Borate-EDTA) buffer

890mM Tris

890mM Boric Acid

20mM EDTA

Neutralization Buffer

1M Tris

3M NaCl

80ml/L concentrated HCl

Hybridisation buffer (for Southern blot)

50% formamide

3x SSC

1x Denhardt's

20µg/ml salmon sperm DNA

6% Dextran sulphate

0.5% SDS

Wash buffer (for Southern blot)

0.1% SDS

0.1x SSC

100x Denhardt's

2% Ficoll 400

2% Polyvinylpyrolidone

2% BSA (Sigma-Aldrich)

ddwater (MQ)

Bromophenolblue loading buffer (10x)

0.25% bromophenolblue (or Xylene cyanol)

40% sucrose

Orange loading dye (6x)

0.2% orange G

40% sucrose

Hybridisation buffer (for TBS)

1.5 M NaCl

40 mM Sodium Phosphate buffer (pH 7.2)

10 mM EDTA (pH 8.0)

10x Denhardt's

0.2% SDS

Sodium Phosphate Buffer

1M NaH_2PO_4

1M Na_2HPO_4

Bead Binding Buffer

10mM Tris-HCl (pH7.5)

1mM EDTA

1M NaCl

Wash buffer (for TBS)

1x SSC

0.1% SDS

10x Annealing Buffer

100mM Tris-HCl (pH7.5)

1M NaCl

10mM EDTA

Bead Binding Buffer

10mM Tris-HCl (pH7.5)

1mM EDTA

1M NaCl

2.1.5 Plasmids, cell lines and tissues

mES RMCE	kindly provided by Magnus Lynch
hES cl.1	kindly provided by Andrew Smith (“Edi2”)
hES cl.2	kindly provided by Andrew Smith (“Shef4”)
hmES cl.1	kindly provided by Andrew Smith (“6TGR clone 4”)
hmES cl.2	kindly provided by Andrew Smith (“6TGR clone 6”)
hmES cl.A	re-derived ESC clone (Christina Rode)

pML7	kindly provided by Magnus Lynch
G1/2	kindly provided by Magnus Lynch
M5	kindly provided by Magnus Lynch
N11	kindly provided by Magnus Lynch
O12	kindly provided by Magnus Lynch
hbrain (1)	kindly provided by Richard Gibbons (MRC13171, collection date 10/08/1993, suction termination, menstrual age not determined, date supplied: 26/09/1994)
hbrain (2)	kindly provided by Richard Gibbons (MRC8685)
hmbrain (1)	homozygous HWT324 E9.5 B1-4, kindly provided by Jackie Sharpe (HWT324 = cross humanised WT male, F1 (C57BL6male x CBA female) females, embryonic day E9.5, B1-4= brains from 4 litters pooled)
hmbrain (2)	homozygous HWT324 E13.5 brain 1, kindly provided by Jackie Sharpe (HWT324 = cross humanised WT male, F1 (C57BL6 male x CBA female) females, embryonic day E13.5, brain 1 = from one mouse)
hFibach	kindly provided by Jackie Sloane-Stanley, late Fibach 11/18, collected on 31.10.12
hmTer119	kindly provided by Helena Ayuub, homozygous X1009

2.2 DNA procedures

2.2.1 Extraction of genomic DNA from cell pellets (from eukaryotic cells)

Genomic DNA was extracted from cell lines and homogenized tissues. Cells in culture were spun down at 1000 revolutions per minute (rpm) for 5 minutes and pellets were washed once in PBS. Tissues were finely minced using a scalpel. Cells were lysed by incubation in 0.2-1.5 ml lysis buffer (10mM Tris pH8, 10mM NaCl, 10mM EDTA, 0.5% SDS, proteinase K 50μg/ml lysis buffer) for approximately 5 hours at 37°C, with occasional gentle pipetting during incubation.

Following cell lysis DNA was extracted according to the standard phenol/chloroform procedure. Briefly, an equal volume of phenol (Fluka, phenol saturated with Tris pH8) and an equal volume of chloroform were added to the DNA. Each step was repeated once and each time the mix was spun down at 10,000 rpm for 3 minutes and the aqueous upper layer was retained after the spin. The DNA was precipitated with 1/10 volume 4M NaCl and double volume 100% ethanol, mixed and centrifuged at 13,000 rpm for 5 minutes. The DNA was washed twice in 70% ethanol and resuspended in TE (Fluka). The concentration was determined by spectrophotometry (measuring absorbance at 260nm) (NanoDrop spectrophotometer, Thermo Scientific) or by Qubit fluorometer (Life Technologies). DNA was stored at -20ºC for long-term storage, otherwise at 4ºC for continuous use.

2.2.2 DNA extraction from bacterial cultures (MiniPreps)

Bacterial cultures were obtained by picking single colonies from agar plates and growing bacterial cultures overnight in 5ml LB medium (10g Tryptone, 5g yeast extract, 10g NaCl/L) supplemented with 100μg/ml ampicillin at 37°C, vigorously shaking (220rpm). Cell pellets were obtained by centrifugation of 1-3 ml of cell culture at 13000 rpm for 1 minute in an Eppendorf tube. Plasmid DNA was extracted using the Qiagen MiniPrep kit, an alkaline lysis method. Cell lysis was performed according to the manufacturer's instructions. DNA was resuspended in water.

2.2.3 Bisulfite conversion

Bisulfite sequencing is considered the gold standard technique for quantitative measurement of methylcytosine in DNA. In this approach, cytosine is converted to uracil by a deamination step that requires the presence of sodium bisulfite. Methylcytosine is protected from this step due to a slower reaction of this deamination step. Subsequent PCR amplification of selected regions of the bisulfite modified genomic DNA results in thymines at positions, which previously contained cytosines and cytosines at previous methylcytosine positions.

Bisulfite conversion of DNA was carried out using the EZ DNA Methylation-Gold™ Kit (Zymo Research, D5005) according to the manufacturers instructions. In brief, 20μl of DNA [12.5 ng/μl] was mixed with 130μl CT conversion reagent and incubated at 98ºC for 10 minutes to denature the DNA followed by 2.5 hours at 64ºC for conversion of DNA. After a 20-minute

desulphonation step on the column, samples were washed and eluted in 10µl elution buffer.

2.2.4 PCR amplification

PCRs were carried out using HotStarTaq from Qiagen. 2µl of bisulfite modified DNA was used for a 25µl PCR reaction (1x PCR buffer, 200mM each dNTP, 0.1-1mM each primer, 2.5 units (U) polymerase). PCR conditions were: 95ºC 15 minutes to activate the enzyme, followed by 35 cycles of 95ºC 1 minute denaturation, 2 minutes annealing, 72ºC 3 minutes extension, with a final extension at 72ºC for 10 minutes. The efficiency and specificity of the PCR amplification was tested by gel electrophoresis of 5µl of the PCR products on a 2% agarose gel containing 0.5x GelRed or ethidium bromide [0.5µg/ml] in TBE buffer at 130 volt (V) for 1 hour. The DNA was visualised using a video gel documentation system (Gel-Doc2000; BioRad Laboratories, Hercules CA, USA). PCR products were purified with the PCR purification kit (Qiagen) as per manufacturer's instructions.

For nested PCR, two rounds of PCRs were carried out. In the first round of PCR amplification was carried out using outer primers (F1 + R1). In a total volume of 25µl the reaction mixture contained 2µl of bisulfite-treated DNA, 0.2mM dNTPs, 0.5µM of each primer, 0.5U of HotStarTAQ^{TM} DNA polymerase (Qiagen, Germany) in the provided buffer (final concentration 1x) supplemented with Q solution (1x final concentration). The second round of PCR was performed in 50µl volumes of reaction mixtures using 5µl of the first round of PCR products and 1U of HotStarTAQ^{TM} DNA polymerase. PCR was performed in a thermal cycler under the following conditions: 95ºC, 15

minutes, followed by 26 cycles of 1 minute denaturation at 95ºC, 2 minutes at the primer pair-specific annealing temperature (50-60ºC), and 3 minutes at 72ºC for elongation, followed by a final 10 minutes elongation at 72ºC.

2.2.4.1 Primer design

Primer pairs for bisulfite-converted DNA were designed to yield on average 250bp products. The primers were designed by eye and designed to match sequences that do not contain CpG sites. All Cs in the sequence of the primer were assumed to be non-methylated (hence converted to Ts after bisulfite modification). If possible the primers were designed to contain a GC clamp at the 3' end (2-3 Gs in the last 5 bp of the primer) and complementarity between the primer pairs, particularly at their 3' end was avoided to minimize amplification of primer dimers. Each primer pair was aimed to have a melting temperature (Tm) in the range of 50-60ºC.

Nested bisulfite-specific primers were designed by Supachai Ekwattanakit.

For a list of primers see Appendix section 7.2.

2.2.4.2 Colony PCR

Colony PCRs were carried out on single colonies grown on agar plates using HotStarTaq Polymerase Kit (Qiagen). Each picked colony was transferred to a 25µl PCR reaction (1x PCR buffer, 200mM each dNTP, 0.4µM colony PCR primer, 2.5U HotStarTaq polymerase). The PCR program was as follows: 95ºC, 15 min, followed by 30 cycles of 30 seconds denaturation at 95ºC, 30 seconds primer annealing at 66ºC, and 1-1.5 minute at 72ºC for elongation, followed by a final 10 minutes elongation step. The quality of the PCR product

was tested by gel electrophoresis as described in section 2.2.4. PCR products were purified using Qiagen PCR purification kit as per manufacturer's instructions.

2.2.4.3 Quantitative real-time PCR (qPCR)

q-PCR was carried out in 25µl reaction contained 6.5µl water, 12.5µl 2×PCR mastermix, 3µl primer/probe mix and 5µl template DNA. The reactions were performed in triplicates with an ABI Prism 7000 sequence detection thermocycler (Applied Biosystems). Data analysis was carried out using the Applied Biosystems software. The $2^{\Delta\Delta CT}$ method described in 7.1.1 was used to calculate enrichment of DNA with primers located inside and outside the capture area.

2.2.5 Restriction digestion of DNA

Restriction digests were carried out with enzymes (Roche or NEB) as per manufacturers instructions. Briefly, 5-20 µg of DNA was digested in a 50 µl reaction containing 1x concentration of the supplied reaction buffers, and 1/10 volume of the enzymes. The reaction was incubated at the recommended temperature for each of the enzymes, typically 37°C, for 3-4 hour or over night. Digests were size-separated by agarose gel electrophoresis (1.5% agarose in 1x TBE) and visualised by staining with ethidium bromide or GelRed and exposure to ultraviolet (UV) light (Gel-Doc2000; BioRad Laboratories, Hercules CA, USA).

2.2.6 Southern blotting

Genomic DNA was cleaved as described in above. In a 50µl reaction volume, 10-50µg of genomic DNA was cleaved by 30-100U of the methylation-sensitive restriction enzyme HpaII (NEB, recognition site C'CGG), or EagI (C'GGCCG, NEB) in the presence of 5µl of NEB4 buffer at 37ºC for 3 hours.

Similar reaction conditions were used for the methylation-insensitive restriction enzymes PstI (CTGCA'G) and BglII (A'GATCT). The recommended buffers were used accordingly. For double digests (HpaII and BglII) were incubated for 3 hours at 37°C with HpaII followed by the addition of NaCl to 50mM and BglII and incubation for a further 3 hours. The resulting fragments were separated according to size by gel electrophoresis (1% agarose gel in 1xTAE over night at 50-60V constant). Typically, 8µg of DNA was loaded, but for MEL hybrid versus human Epstein-Barr virus (EBV)-transformed B lymphocytes the amount of DNA loaded was adjusted to get similar signal intensities from all samples (typically 3µg MEL Hybrid and 0.1µg human wildtype). The gel was stained with ethidium bromide and the DNA was visualised using a video gel documentation system (Gel-Doc2000; BioRad Laboratories, Hercules CA, USA).

For Southern blotting, the gel was soaked in 1M NaOH (VWR) for 40-50 minutes to denature the DNA, rinsed in water, and neutralized in neutralization buffer for at least the double amount of time (1.5 - 8 hours).

The gel was assembled as follows to enable transfer of DNA to a membrane: a sponge in a deep dish was soaked in 6xSSC and Whatman^{TM} 3MM chromatography paper was placed on top covering the sponge. Gel and

membrane were assembled as follows (from bottom to top): gel (rinsed in water), Zeta-Probe Membrane (rinsed in water), two layers of Whatman™ 3MM chromatography paper (soaked in 6xSSC), a stack of towels, and approximately 0.5kg of weight using a glass sheet and a bottle. The hybridization was carried out over night. The membrane was then removed, air-dried for 10-15 minutes. The DNA was cross-linked to the membrane by UV light and baking in an oven for 80ºC for 2 hours.

The membrane was rehydrated in 2x SSC, equilibrated in 5 ml hybridisation buffer (formamide 50%, 3xSSC, 1xDenhardt's, 20µg/ml salmon sperm DNA, 6% Dextran sulphate, 0.5% SDS) for 15 minutes to 2-3 hours and incubated in 10ml hybridization buffer containing the labelled probe (see section 2.2.7 for the probe labelling protocol) over night at 42ºC to allow hybridisation of the probe to the DNA. The membrane was rinsed in 2xSSC, washed once in 2xSSC for 10 minutes to 1 hour. Washes were carried out by rinsing the membrane in 2xSSC, washing in 2xSSC for 10 minutes to 1 hour at room temperature, followed by 1-2 hot washes at 65ºC in wash buffer (0.1% SDS, 0.1x SSC) for 5 minutes (until the radioactive signal was about 2-5 counts per minute (cpm) as measured by Geiger counter). The DNA was visualised using a storage phosphor imaging system (GE healthcare, Typhoon™) or by exposure to an Amersham Hyperfilm MP (Amersham, 28-9068-46) at -70ºC in cassette for 1-21 days.

2.2.7 Radioactive probe labelling

Radioactive probe labelling was performed using the Amersham Megaprime DNA Labelling System (GE Healthcare, RPN1607) as per manufacturer's

instructions. Briefly, 50ng of the probe and 5µl random primers in a final volume of 35µl were denatured at 95ºC for 5 minutes. 10µl of labelling buffer, 1µl of radioactive [α^{32}P]-dCTP (3000Ci/mmol, 10mCi/ml, Perkin Elmer, NEG513H250UC), and 2µl Klenow fragment was added and the reaction mixture was incubated at 37ºC for 10 minutes, followed by purification using illustra ProbeQuant G-50 Micro Columns (GE Healthcare, 28-9034-08). The labelled probe was denatured at 95ºC for 5 minutes and added to 10 ml hybridization buffer (50% formamide, 3xSSC, 1xDenhardt's, 20µg/ml salmon sperm DNA, 6% Dextran sulphate, 0.5% SDS). The probes used for the Southern blots were provided by Helena Ayyub.

2.3 Cell culture procedure

2.3.1 ESC culture conditions

mES RMCE culture were grown at 37°C, 5% CO_2 feeder-free on gelatinized plates (0.1% Gelatin diluted in PBS, Sigma) in Glasgow minimal essential medium (GMEM; Invitrogen), 10% fetal calf serum (PAA laboratories), 1% LIF conditioned medium (prepared by Sue Butler), 1% sodium pyruvate (Invitrogen), 1% non-essential amino acids (Invitrogen), 0.1mM β-mercaptoethanol.

Human ESCs were kindly provided by Andrew Smith. These were grown at 37°C with 5% CO_2.

2.3.2 Electroporation

Plasmid DNA obtained by Maxiprep was precipitated with 100% ethanol on dry ice for 5 minutes and washed three times in 70% ethanol, resuspended

over night at 4°C in 100µl sterile PBS. Single cell suspensions of ESCs in a total volume of 600µl PBS were electroporated in a 0.4cm gap cuvette (BioRad) at 0.8k, 3F capacitance, time constant 0.1 seconds followed by recovery of cells for 10 minutes at room temperature and plating onto gelatinized 10cm tissue culture plates.

2.4 Bacterial procedures

2.4.1 Ligation into pGEM®-T Easy Vector

Ligation of a PCR product into pGEM®-T Easy Vector (Promega, A137A) using the Rapid DNA Ligation Kit (Roche, 11635379001) according to the manufacturer's instructions. Briefly, 50ng of vector was mixed with 25ng of insert with 1x DNA dilution buffer (Roche) to a final volume of 10µl. 10µl of DNA Ligation buffer and 5U of T4 DNA Ligase was added, mixed, and incubated at room temperature for 5-10 minutes. The ligation reaction was used immediately for transformation of *E. coli* cells or stored at -20ºC.

2.4.2 Transformation of competent cells

60 µl of Library Efficiency® DH5α^{TM} Competent Cells (Invitrogen 18263-012) were mixed gently with 6µl of ligation reaction (see section 2.4.1), and kept on ice for 30 minutes in a 5ml Polypropylene Round-Bottom Tubes (BD Falcon). Chemical transformation was carried out by heat shocking the cells at 42ºC for 45 seconds after which the cells were immediately put on ice for 2 minutes, 540 µl SOC medium (supplied with Library Efficiency® DH5α^{TM} Competent cells) was added at room temperature and the cells were then

incubated at 37ºC for 1 hour, shaking at 220rpm in an Ecotron incubation shaker (Infors HT).

125µl of the transformation mixture was plated onto agar plates supplemented with 40µl Ampicilin [100mg/ml] and 40µl X-gal [50mg/ml] prior to addition of the transformation reaction. The plates were incubated at 37ºC over night. The next day colonies were picked for colony PCR (see section 2.2.4.2) or for DNA extraction (see section 2.2.2).

2.5 Targeted bisulfite sequencing (TBS)

2.5.1.1 Sonication

DNA was sonicated to an approximate size of 500bp. Briefly, 5µg genomic DNA in a total volume of 100µl EB (Qiagen) was sonicated in a microTUBE AFA Fiber Pre-slit Snap-Cap using the Covaris Adaptive Focus Acoustics (K bioscience) for 90 seconds (duty cycle 5%, intensity 3, cycles/burst 200 and frequency sweeping). Sonication was verified on a 1.5% agarose gel. The volume of the sample was reduced in a SpeedVac lyophiliser in a DNA LoBind Tube (Eppendorf) to a final volume of about 14µl (60°C for approximately 30 minutes for one tube). The concentration of the sample was measured using NanoDrop.

2.5.1.2 Capture

5µg sonicated DNA was mixed with 1ng biotinylated oligo library (designed by Jim Hughes). In a PCR tube, the 15µl reaction was overlaid with mineral oil. DNA was denatured at 95°C for 5 minutes and slowly annealed at 65°C 15min. 15µl of prewarmed (65°C) 2x hybridisation buffer B (1.5 M NaCl, 40

mM Sodium Phosphate buffer (pH 7.2), 10 mM EDTA (pH 8.0), 10x Denhardt's, 0.2% SDS,) was added. The reaction was incubated overnight at 65°C shaking (rpm) on a ThermoMixer Comfort (Eppendorf). Next day, 50µl of streptavidin-coated magnetic Dynabeads® (M-280) was washed twice with 500µl Bead Binding Buffer (10mM Tris-HCl (pH7.5), 1mM EDTA, 1M NaCl). Beads were resuspended in 150µl Bead Binding Buffer and 25µl of the overnight hybridization reaction (at 65ºC) was added to the beads, mixed and incubated rotating for 30 minutes, at room temperature. The beads were washed in 1ml wash buffer (1x SSC, 0.1% SDS) for 15 minutes rotating at room temperature. This step was repeated an additional three times using pre-warmed wash buffer (65ºC) and incubating for 15 minutes rotating at 65ºC. To elute the DNA, 50µl 0.1M NaOH was added to the beads and incubated at room temperature for 10 minutes, gently mixing occasionally during the incubation time. The supernatant was transferred to 50µl Tris-HCl (pH 7.5) and DNA was purified using illustra ProbeQuant G-50 Micro Columns. 20µl of the purified sample were removed to test for the capture efficiency by q-PCR (see section 7.1.1). This was done using primers outside and inside the area of capture to measure the amount of DNA from the captured region to a non-captured region. If the capture efficiency was 3000 fold or more, the captured DNA was used for downstream processing.

2.5.1.3 Bisulfite conversion

Captured DNA was modified by bisulfite treatment using the MethylEasy™ *XCeed* Rapid DNA Bisulfite Modification Kit (Human Genetic Signatures, ME002) according to the manufacturer's instructions. In the heat

denaturation/desulphonation step temperature spikes were included to allow efficient denaturation of the DNA fragments: 3 cycles of 80ºC for 10 minutes, followed by 99ºC for 30 seconds, and a final incubation at 80ºC for 15 minutes. 0.5µl of cRNA (1µg/µl, Qiagen) were added to pooled samples at step 7. DNA was eluted twice in each 20µl of pre-warmed reagent 5 at 95°C for 20 minutes.

2.5.1.4 Random Primer Amplification (RPA)

DNA was first amplified for two cycles using random primers (Round A) followed by 30 cycles of specific amplification using primers that match sequences of the first primer (Round B).

Round A: Two cycles of synthesis were performed using random primers (RG3A). To allow denaturation, 7µl of DNA and 4µM random primers in a 10µl reaction of 1x sequenase reaction buffer (GE Heathcare) were incubated at 94ºC for 10 minutes. After denaturation, the samples were placed on ice. Sequenase was added in a 5µl volume (4U Sequenase, 3mM dNTPs, 0.1M DTT (NEB) and 0.75µg BSA (NEB) and incubated at 37°C 10 minutes, 94°C 2 minutes, 8°C 2 minutes (at which point a further 4U Sequenase (Sequenase version 2.0, 13U/µl, GE healthcare) in 1µl of 3.5x Sequenase buffer was added to the reaction), followed by ramping to 37ºC in 8 minutes (12°C, 15°C, 19°C, 22°C, 25°C, 29°C, 33°C, 36°C, each for 1 minute) and a final 37°C incubation for 10 minutes. The volume was increase to 50µl by adding 35µl of 1x TE (Fluka) to each PCR tube.

Round B: 5µl of the diluted template from Round A was used in the second round of amplification. PCR was carried out in a 100µl reaction of 1x

AmpliTaq Gold® 360 PCR Buffer without Mg^2+, 1µl dNTPs [25mM], 1µl RG4 primer [50µM], 0.5µl AmpliTaq Gold® 360 DNA polymerase[5U/µl] (Roche), 0.2µl PWO [5U/µl] (Roche), 8µl $MgCl_2$ [25mM] (Roche), 5µl template DNA. PCR conditions were 5 minutes at 94°C followed by 30 cycles of 30 seconds at 92°C, 30 seconds at 40°C, 30 seconds at 50°C, and 1 minute at 72°C. 10µl of PCR product was visualised on a 1.5% agarose gel stained with ethidium bromide. The remainder 90µl of the reaction was purified using Qiagen MiniElute PCR Purification Kit.

2.5.1.5 AcuI digestion

The adapter sequences that were added during the PCR reactions were cleaved off using AcuI. 10µl of (50U AcuI (NEB), 10µl Buffer 4, 8µl 40µM SAM) for 90 minutes at 37°C. 10µl of this reaction was removed for later analysis by gel electrophoresis (1.5% agarose) along with the uncut DNA.

Successful bisulfite conversion of the DNA was verified by sequencing of a subcloned fraction of the captured, bisulfite treated and amplified DNA. For this, 32µl of AcuI-digested material was A-tailed using Klenow-DNA polymerase (3'→5' exo-, [5U/µl], NEB M0212S), 1mM dATP, 1x NEB buffer, 50µl total volume, DNA: 1/3 of the 100µl AcuI digestion mixture) and ligated into pGEM®-T Easy Vector as described in section 2.4.1.

2.5.1.6 Gel purification

AcuI cleaved DNA was gel-purified from a 1% low melting point agarose gel (SeaPlaque™, Lonza) run in 1x TAE at 50V, 1 hour. Fragments above 300bp were excised and purified using the MiniElute Kit as per manufacturers

protocol, eluting in 10µl of Qiagen elution buffer. The DNA concentration was measured by fluorometric quantitation (Qubit). The Qubit® 2.0 Fluorometer uses fluorescent dyes specifically target the biomolecules of interest and emit signals only when bound to their specific target molecules. This allows accurate measurement of DNA even at low concentrations.

10ng of DNA was used for high throughput sequencing using the Illumina sequencing platform carried out by the Wellcome Trust Centre for Human Genetics (Oxford). Each sample was sequenced in one lane by paired-end sequencing of 100bp.

2.5.2 Data analysis of bisulfite sequencing data

Bioinformatic analysis of high throughput bisulfite sequencing data was carried out with the help of Nicki Gray (MRC Molecular Haematology Unit, Weatherall Institute for Molecular Medicine). Bismark was used to map reads and call methylation states of cytosines. Bismark is a set of tools for the time-efficient analysis of bisulfite sequencing (BS-Seq) data. It performs alignments of bisulfite-treated reads to a reference genome and cytosine methylation calls at the same time. Bisulfite-treated reads are mapped using the short read aligner Bowtie (Langmead, 2010). Sequence reads are first transformed into fully bisulfite-converted forward (C→T) and reverse read (G→A conversion of the forward strand) versions, before they are aligned to similarly converted versions of the genome (also C→T and G→A converted). Sequence reads that produce a unique best alignment from the four alignment processes against the bisulfite genomes (which are running in parallel) are then compared to the normal genomic sequence and the methylation state of

all cytosine positions in the read is inferred. Due to low quality of reverse reads in the brain dataset[10], all sequence reads (ESCs and brain) were mapped using single-end approach allowing one mismatch in the first 50bp of the sequence reads. Sequences of identical length and methylation status were regarded as sequence clones arising from the previous PCR amplification step and only a single of the identical sequences were allowed for analysis. CpG sites covered by a minimum of 10 sequence reads were allowed for further analysis. Seqmonk was used to view and quantitate the data and produce reports, such as percentage of methylation and quantitation of read numbers at each CpG position. Further analysis was performed using Excel.

2.6 Targeted integration of test sequences into mES RMCE

In a previous study, the *HPRT*-deficient mES E14Tg2a cell line has been previously targeted with an RMCE cassette that replaces the mouse *α*-globin genes on chromosome 11. The cell line containing the RMCE cassette will be referred to as mES RMCE hereafter. The RMCE cassette contains the 5' portion of a *HPRT* minigene and a *NEO* resistance marker flanked LoxP and Lox511 sites. The targeted cell line, here referred to as mES RMCE, was kindly provided by Magnus Lynch.

To integrate a sequence of interest into the mouse *α*-globin locus the test sequence is first cloned into the NotI site of the donor plasmid pML7 (kindly provided by Magnus Lynch). Once integrated into the plasmid the test sequence lies next to the 3' portion of the *HPRT* minigene and LoxP and

[10] Sequence reads with low sequence quality were excluded from further analysis. For more information see section 7.1.

Lox511 sites are located on either site of the *HPRT* gene and the test sequence (Figure 3). The correct sequence of the insert was verified by Sanger sequencing of the donor plasmid. Co-transfection of the donor plasmid with a Cre expression construct into the mES RMCE cell line, results in the Cre recombinase-mediated replacement of the *NEO* gene in the mouse genome with the test sequence in the donor plasmid. In this way, a functional *HPRT* minigene is created by placing the 3' portion of the *HPRT* minigene from the donor plasmid next to the 5' portion in the mouse genome. Positive selection of cells containing a functional *HPRT* gene was carried out in medium supplemented with hypoxanthine-aminopterine-thymidine (HAT).

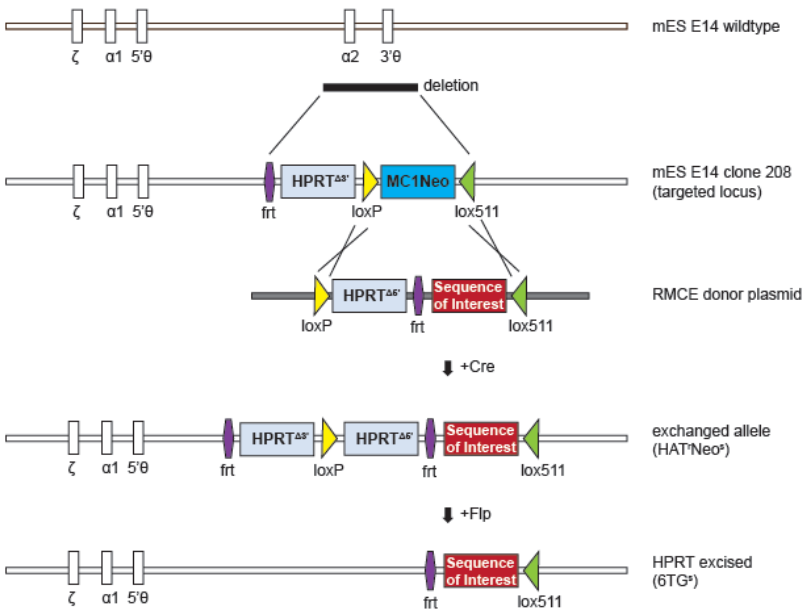

Figure 3: Principle of the recombination-mediated cassette exchange (RMCE) system.
The mES RMCE cell line bears a RMCE cassette in place of the mouse α2 and 3'θ genes. The RMCE cassette contains the 5' half of an *HPRT* resistance minigene and a MC1Neo cassette. Test sequences flanked by the 3' portion of the *HPRT* minigene and loxP sites in a RMCE donor plasmid are integrated into the mouse *α*-globin locus using the Cre/lox system. Correctly exchanged clones are identified by the reconstituted *HPRT* minigene. In the final step the *HPRT* minigene is excised using the Flp/frt system to leave only the test sequence in the defined genomic locus with minimal exogenous flanking sequences. (Figure adapted from (Lynch et al, 2012)).

The reconstituted *HPRT* minigene is flanked by FRT sites, recognition sites for Flp recombinases. The *HPRT* marker was excised in a second step by

transfecting *HPRT*-positive cells with a vector containing the Flp recombinase. *HPRT*-negative cells were selected by the addition of the nucleoside analog 6-thioguanine (6TG), which is toxic to *HPRT*-positive cells. This is done to avoid any effects on DNA methylation that might arise from transcription of the *HPRT* minigene.

2.6.1 Mapping of the RMCE fragments

The read quality of high-throughput sequencing data was tested using FastQ program and the reads were mapped using Bismark allowing one mismatch per read. Mapping of the reads in each sample was performed against the human DNA sequence that had been placed in the mouse. This corresponds to the region 161,060 – 164,998 in G1, G2, and methylated control, 161,060-163,300 for M5, 161060-162911 for N11, 162,911-164,998 for O12 and 163,728-165,529 for fragment 163-165k, on chromosome 16 (hg18).

2.7 Bioinformatic analysis of BioCAP data

Reads were mapped single-end using bowtie allowing 1 mismatch per read. The bioinformatics pipeline used is briefly outlined in Figure 4. Subsampling of sequence reads uses an algorithm that randomly discards sequence reads of a given dataset to bring it down to a pre-defined number of datapoints. The algorithm was created and kindly provided by Steve Taylor in the Computational Biology Research Group (CBRG) at the WIMM. Heat maps were generated by aligning all CGIs at their midpoint (formula: CGI startpoint + (CGI endpoint – CGI startpoint)/2) and plotting the read number across each CGI. CGIs were sorted by their enrichment in Tc1 liver1 over human liver1 or Tc1 testes1 over human testes1. Enrichment was calculated by

subtracting the sequence read number in human from that in Tc1 (Tc1-hs) for each given CGI. Average read profiles (pile-up analysis) was carried out using an algorithm designed and kindly provided by Jim Hughes, CBRG, WIMM. Briefly, sequence reads were aligned at the midpoint of each CGI as described above, sequence reads at each position along a CGI were added up and normalised to the total number of sequence reads in the dataset (covering hs21). To compare the distribution of reads across CGIs, reads were further normalised to the number of reads across all CGIs (6kb around the midpoint of CGIs).

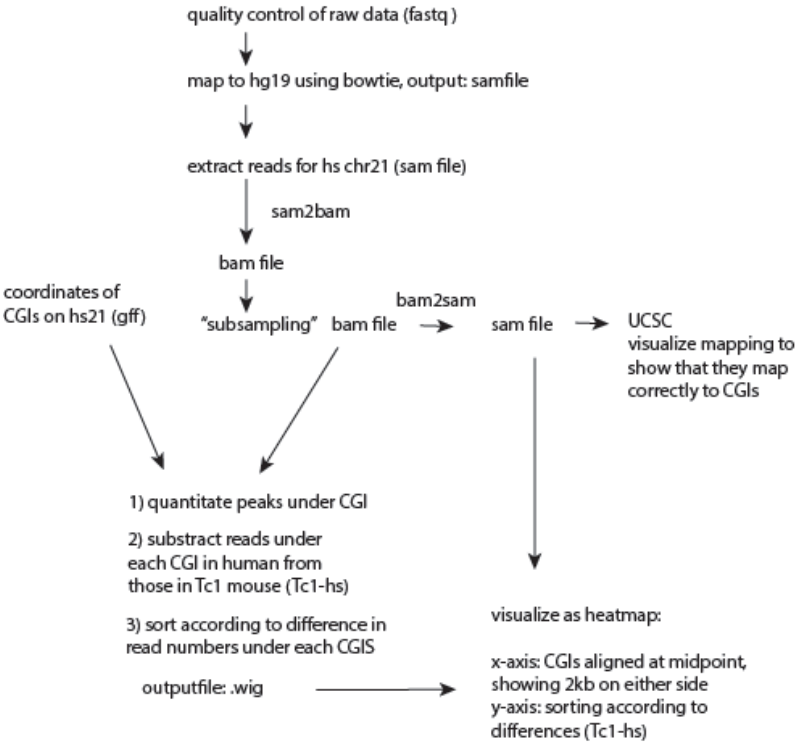

Figure 4: Bioinformatic pipeline used for analysis of BioCAP data.

3 DNA methylation patterns of the human *α*-globin cluster in human and mouse

3.1 Introduction

Despite the fact that DNA methylation was the first epigenetic mark to be identified several decades ago, it is still not well understood how DNA methylation patterns are determined. A key open question is the role of the DNA sequence in establishing and maintaining DNA methylation patterns. To shed light on the role of the DNA sequence in determining DNA methylation patterns, this chapter investigates whether DNA methylation patterns are established and maintained normally when human DNA is placed into a heterologous murine environment as opposed to its natural, endogenous chromosomal environment. To do this, DNA methylation was analysed in a previously created humanised mouse model, which carries 117kb of human DNA comprising the human *α*-globin cluster in place of the mouse syntenic region. DNA methylation patterns of the human *α*-globin genes was extensively characterised in humanised mouse ESCs and compared to their normal pattern in human ESCs. Furthermore, to find out how well the established patterns are maintained, DNA methylation wasanalysed in differentiated cell type, including erythroid and non-erythroid cells of the humanised mouse in comparison to human. A detailed investigation of DNA methylation patterns of the human *α*-globin cluster in mouse versus human might help to shed light on molecular mechanisms involved in establishing and maintaining DNA methylation patterns in a normal cell.

3.2 Results

3.2.1 Human CGIs maintain their unmethylated state in humanised mouse compared to human differentiated cells

The human *α*-globin genes are associated with CGIs of about 1kb in size and are *bona fide* unmethylated (Bird et al, 1987; Tufarelli et al, 2003). To investigate whether the human *α*-globin CGIs are correctly unmethylated in humanised mouse tissues, DNA methylation levels were interrogated at CGI H and I (associated with α2 and α1, respectively) by DNA methylation-sensitive restriction enzyme analysis (REA) (Figure 5a). A limiting digest was performed with PstI to cleave the DNA at sites flanking CGIs H and I (at positions 162'303 and 163'794 as well as 166'107 and 167'613, respectively) (Figure 5b). Using a probe (Pstα) that recognises both globin genes α2 and α1[11], the limiting digest results in a 1.5kb band (Figure 5a, b, and last lane in Figure 5c). Further cleavage by the methylation-sensitive restriction enzyme EagI cleaves the DNA at positions 162,619 and 166,623 within CGI H and CGI I, respectively, if the DNA is unmethylated resulting in fragments of 1kb and 0.5kb in size.

[11] Human *α1* and *α2* are of almost identical sequence (due to gene duplication), wherefore the probe Pstα recognises both genes.

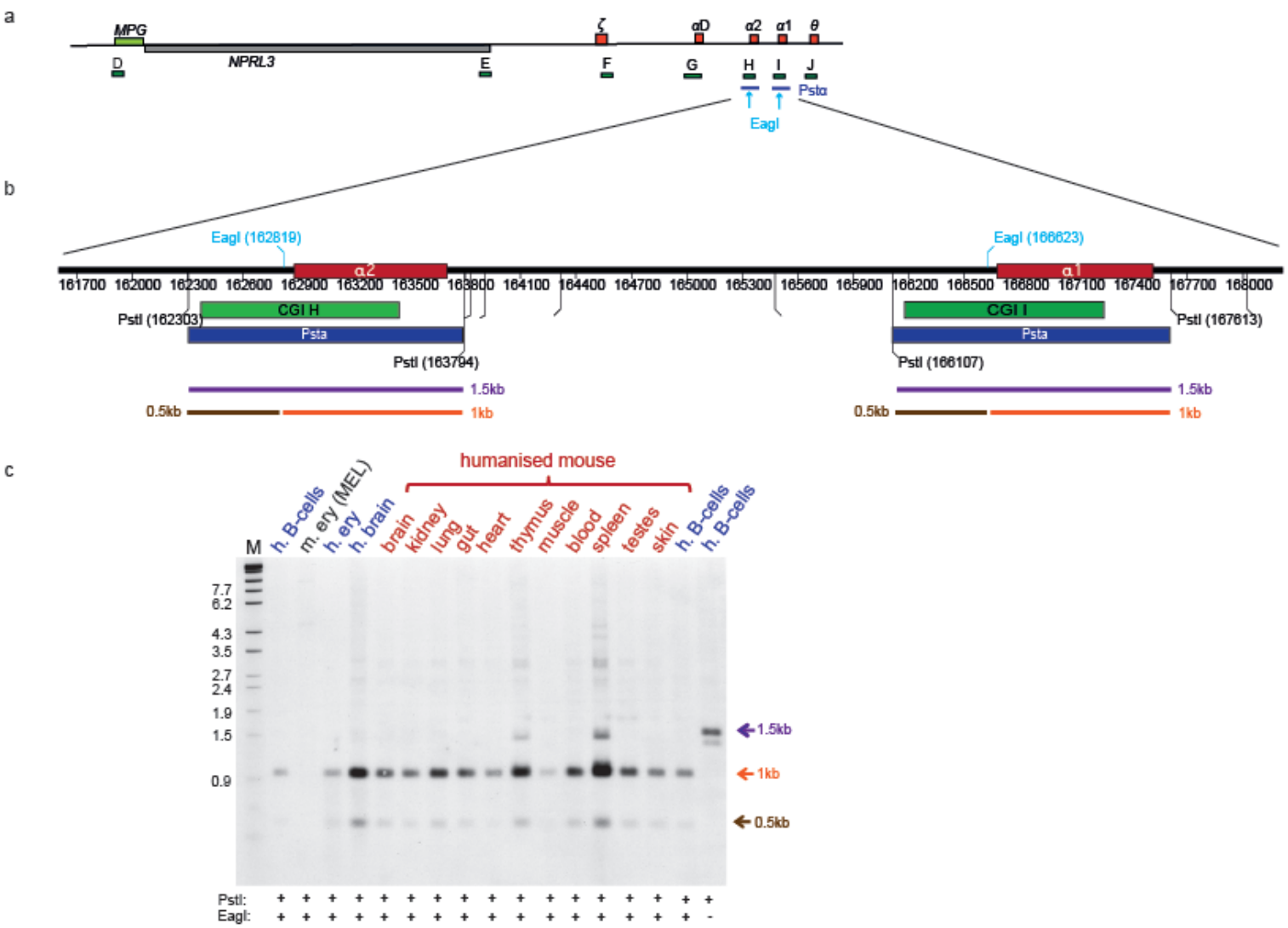

Figure 5: CGI H and I (associated with *α2* and *α1*, respectively) are unmethylated in human and humanised mouse tissues.

(a) Schematic representation of the human *α*-globin cluster showing the location of the probe Pstα (dark blue) and the restriction sites for EagI (light blue) relative to the *α*-globin genes. (b) Restriction map of the *α*-globin locus with the exact location of Pstα probe and EagI and PstI restriction sites (black vertical lines). The limiting digest with PstI results in 1.5kb fragments (last lane in c). Further cleavage by EagI yields 0.5 kb and 1kb fragments if the DNA is unmethylated. CGIs H and I are shown in green and *α*-globin genes in red. (c) Southern Blot of human and humanised mouse genomic DNA from various tissues cut with PstI and EagI and probed with Pstα. This Southern Blot was carried out by Helena Ayyub. (ery = erythroblast)

As can be seen in Figure 5, DNA from human erythroblasts, brain and B-lymphocytes (N-EBV) is cleaved into fragments of 0.5kb and 1kb in size suggesting low methylation levels at the restriction sites 162,619 and 166,623 within CGIs H and I. Likewise, the DNA from all of the analysed humanised mouse tissues (brain, kidney, lung, gut, heart, thymus, muscle, blood, spleen, testes, skin) is cleaved to 0.5kb and 1kb fragments, although a weaker 1.5kb band is seen in humanised mouse thymus and spleen, which may be explained by loading differences and cleavage efficiencies between the samples (see Appendix section 7.2). The specificity of the probe Pstα towards human DNA was verified by the lack of cross-hybridisation of the probe with DNA from mouse erythroid cells (m.ery, Figure 5c).

Similarly, methylation levels were measured in other CGIs within the transgenic region. More specifically, this was done for the following CGIs: CGI D (associated with *MPG*) at the positions 68,385 and 68,403, CGI E (*NPRL3*) at 128,791, CGI F (*α*D) at 144,758, and CGI G (ζ) at 154,725 and 154,742. In all of the cases the interrogated CpG sites were unmethylated in all of the analysed human and humanised mouse tissues (data not shown). Hence, this data suggests that human CGIs are, by and large, correctly unmethylated in humanised mouse compared to human tissues.

3.2.2 Sequences surrounding the human α-globin genes are hypomethylated in humanised mouse versus human tissues

Having demonstrated with a few selected CpG sites that human CGIs are correctly unmethylated in the humanised mouse model, the next aim was to ascertain whether faithful methylation patterns were maintained at other

regions across the transgenic region. The sequences surrounding the *α*-globin genes were analysed using REA (Figure 6). A limiting digest was carried out with the methylation-insensitive restriction enzyme BglII leading to fragments of 12.1kb and 7.4kb in size when probed with Pstα (Figure 6a, b, 3rd and last lane in Figure 6c, 1st lane in Figure 6d). Further cleavage of the DNA with the methylation-sensitive restriction enzyme HpaII leads to a 3kb band in human B-lymphocytes, erythroblasts and brain, and an additional 1.2kb band in human erythroblasts and brain (Figure 6). In all of the analysed humanised mouse tissues (erythroid (Ter119+) and non-erythroid (Ter119-) blood cells, whole blood, liver, kidney, gut, testis, thymus, brain, lung) the 3kb band is absent after HpaII digestion, indicating that the region between CGI I and J is less methylated in the humanised mouse tissues than in human. Hypomethylation of this region is observed in all of the analysed humanised mouse tissues compared to human erythroblasts and human brain.

To rule out that the observed methylation differences are due to tissue-specific differences in methylation, the same restriction digest was performed on equivalent cell types: three human liver and four human brain samples were compared to two humanised mouse liver and five humanised mouse brain samples confirming that the region between CGI I and J (approximate boundaries of 167'145 to 170'171) is hypomethylated in matching human and humanised mouse tissues (Figure 6).

To eliminate the possibility that the mouse chromosomal (and presumably nuclear) environment affects the methylation state of the human *α*-globin genes, DNA methylation was analysed at the human *α*-globin genes in

human/mouse hybrid cells derived from fusing human B-lymphoblasts and mouse erythroleukemia (MEL) cells, which carry human chromosome 16. REA was performed on seven hybrid cell lines using BglII/HpaII restriction digest and probed with Pstα. Although some of the cell lines have lost human chromosome 16 it becomes apparent that the human *α*-globin genes remain hypomethylated even in the context of a whole copy of human chromosome 16 in a mouse cell, indicating that the methylation changes are not due to the influence of the surrounding mouse chromosomal environment.

Further regions across the transgenic region were also screened by REA. By and large, it was found that most of the transgenic region displays comparable methylation patterns in humanised mouse compared to human tissues (data not shown). In conclusion, while methylation patterns of most of the transgenic region appeared normal in humanised mouse versus human tissues the region around the *α*-globin genes becomes hypomethylated. Hypomethylation was found in all of the analysed humanised mouse tissues, including erythroid (liver) and non-erythroid tissues (brain) and this effect was not due to a change in the chromosomal environment.

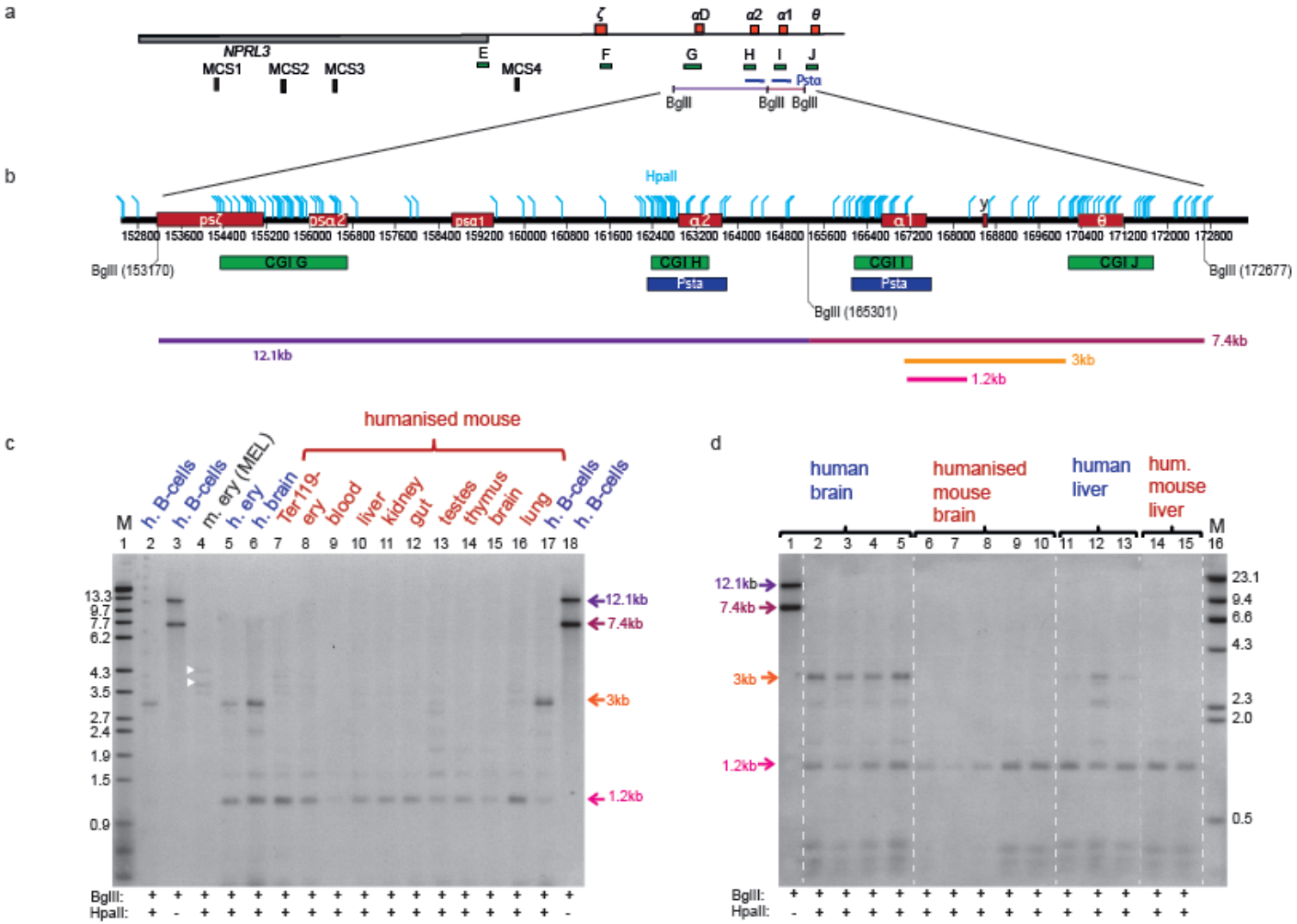

Figure 6: Hypomethylation near the human *α*-globin genes in the humanized mouse model.

a) Schematic representation of the human α-globin cluster showing the location of the probe Pstα and the limiting digest with BglII. b) Restriction map of the α-globin locus indicating the location of the probe Pstα (dark blue), and BglII and HpaII (light blue) restriction sites. BglII cleavage results in 12.1kb and 7.4kb fragments. Further cleavage by HpaII (light blue) results in 3kb and 1.2kb fragments. CpG islands H and I are shown in green, α-globin genes in red. c) and d) Southern Blots showing human and humanised mouse DNA from various tissues cut with BglII and HpaII and probed with Pstα. White triangles indicate bands resulting from cross-hybridization of Pstα with mouse DNA. ery = erythroblast. This Southern Blot was carried out by Helena Ayyub.

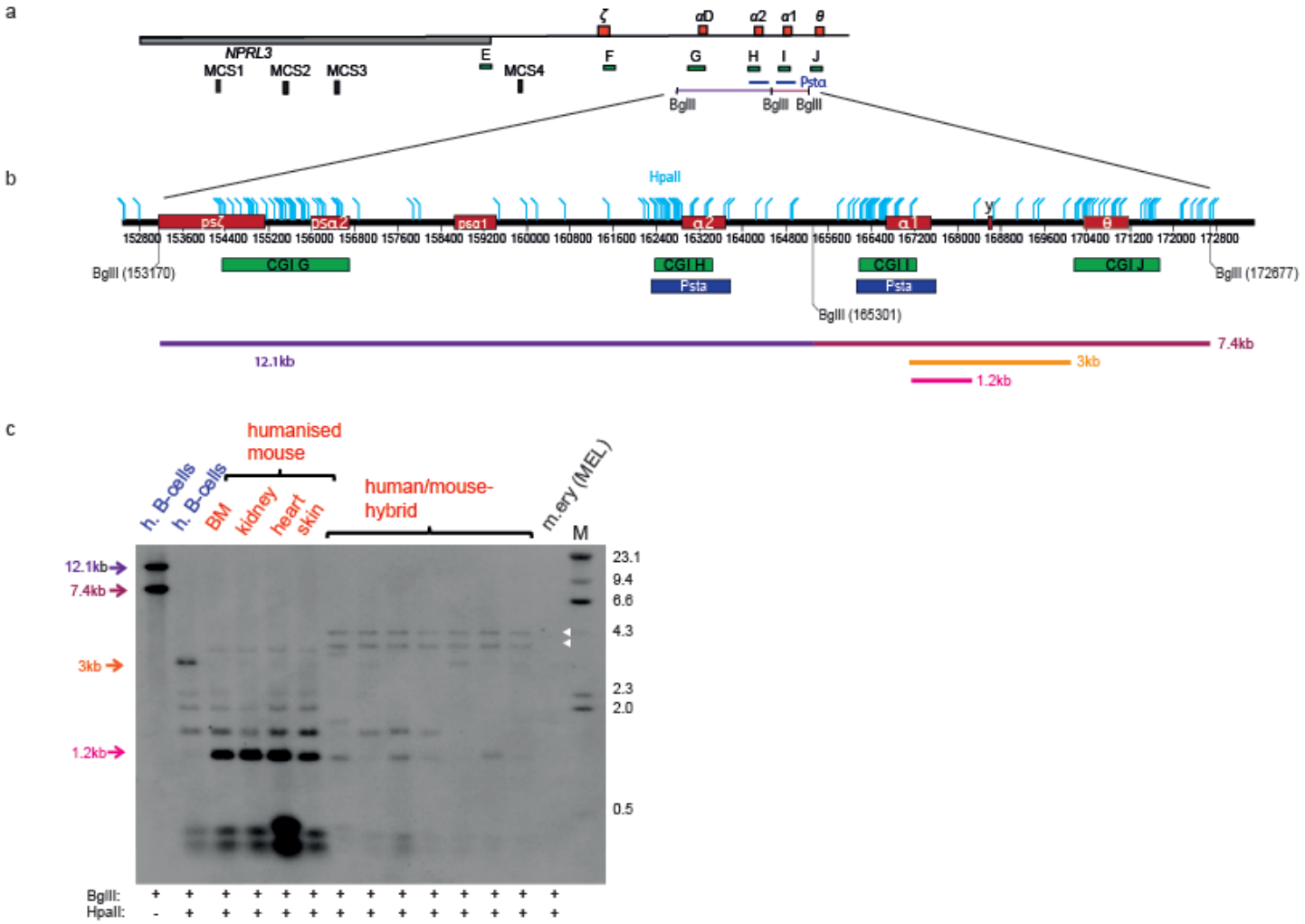

Figure 7: Hypomethylation near the *α*-globin genes in human/mouse hybrid cells.

a) Schematic representation of the human *α*-globin cluster showing the location of the probe Pstα (dark blue) and limiting digest with BglII. b) Restriction map of the *α*-globin locus the location of the probe Pstα, and BglII and HpaII (light blue) restriction sites. BglII cleavage results in a 12.1 kb fragment and a 7.4 kb fragment, which is further cut to 3kb and 1.2kb fragments by HpaII. CpG islands H and I are shown in green and *α*-globin genes in red. c) Southern Blot of genomic DNA from human and humanised mouse tissues as well as human/mouse-hybrids cut with BglII and HpaII and probed with Pstα. White triangles indicate bands resulting from cross-hybridization of Pstα with mouse DNA. ery = erythroblast, BM = bone marrow. This Southern Blot was carried out by Helena Ayyub.

3.2.3 High resolution profiling of DNA methylation patterns across the human transgenic region in ESCs

Having demonstrated that the human *α*-globin cluster becomes hypomethylated at a subset of sites near the *α*-globin genes in all of the analysed humanised mouse tissues compared to its endogenous state, the next aim was to find out whether hypomethylation is a problem of establishment or maintenance of DNA methylation patterns in mouse versus human. To address this question, DNA methylation patterns were analysed in human and humanised mouse ESCs with the premise that methylation defects in ESCs likely reflect a problem of establishment. To enable mapping of the aberrantly methylated sequences across the transgenic region a quantitative high-resolution DNA methylation profile across the region was created. Bisulfite sequencing was used to measure DNA methylation levels across the human *α*-globin cluster in humanised mouse versus human as it generates quantitative single nucleotide resolution data. Bisulfite sequencing depends on the propensity of sodium bisulfite to specifically convert unmethylated Cs into Us, while methylation protects from conversion. Subsequent PCR amplification and DNA sequencing enables inferral of the methylation state of individual CpG sites from the sequence.

Because conventional sequencing of individual fragments across a region as large as the transgenic region (117kb) is a labour-intensive endeavour, bisulfite conversion was coupled to high-throughput sequencing to generate a high resolution profile of DNA methylation levels across the human *α*-globin gene cluster. To keep the costs for high-throughput sequencing in the scope

of our laboratory, a targeted approach (targeted bisulfite sequencing, TBS) was used to specifically enrich the *α*-globin region prior to bisulfite conversion. Enrichment was achieved using DNA oligonucleotides complementary to a 78kb region comprising the human *α*-globin gene cluster. For more details on design and capture efficiency of the oligo capture technique TBS see the supplementary chapter 7.1.

An overview of the DNA methylation pattern of the human *α*-globin cluster in human and humanised mouse ESCs is shown in Figure 8. DNA methylation levels of single CpG sites ranging from 0% to 100% methylation are plotted over the genomic region of the transgenic region. Most of the analysed CpG sites are highly methylated in human ESCs. Unmethylated CpG sites are concentrated in CGIs around the *α*-globin genes. The pattern in humanised mouse ESCs is largely similar: Methylation levels are high at most CpG sites across the transgenic region, while CGIs are correctly unmethylated in humanised mouse ESCs. In contrast to human ESCs, the *α*-globin region appears less methylated in humanised mouse ESCs (highlighted in yellow in Figure 8). To better visualise the differences in DNA methylation levels at individual CpG dinucleotides, differences in methylation states were calculated by substracting methylation levels in human ESCs from those obtained for humanised mouse ESCs. Negative values therefore represent hypomethylation in the humanised mouse compared to human, while positive values indicate hypermethylated CpG sites. A plot of the methylation differences across the transgenic region shows that in humanised mouse ESCs some CpG sites are hypomethylated while others are hypermethylated

compared to human ESCs (examples are indicated by red arrowheads in Figure 8).

Using the sensitive TBS analysis it is evident that methylation changes appear across the whole transgenic region in humanised mouse versus human ESCs with regions of both hypo- and hypermethylation (Figure 8). The most obvious changes appear around the *α*-globin genes confirming the REA analysis. In conclusion, this data suggests that the human *α*-globin cluster is already hypomethylated in humanised mouse compared to human ESCs, with the most obvious changes around the *α*-globin genes. This infers a defect in the establishment of normal human methylation patterns in the mouse context.

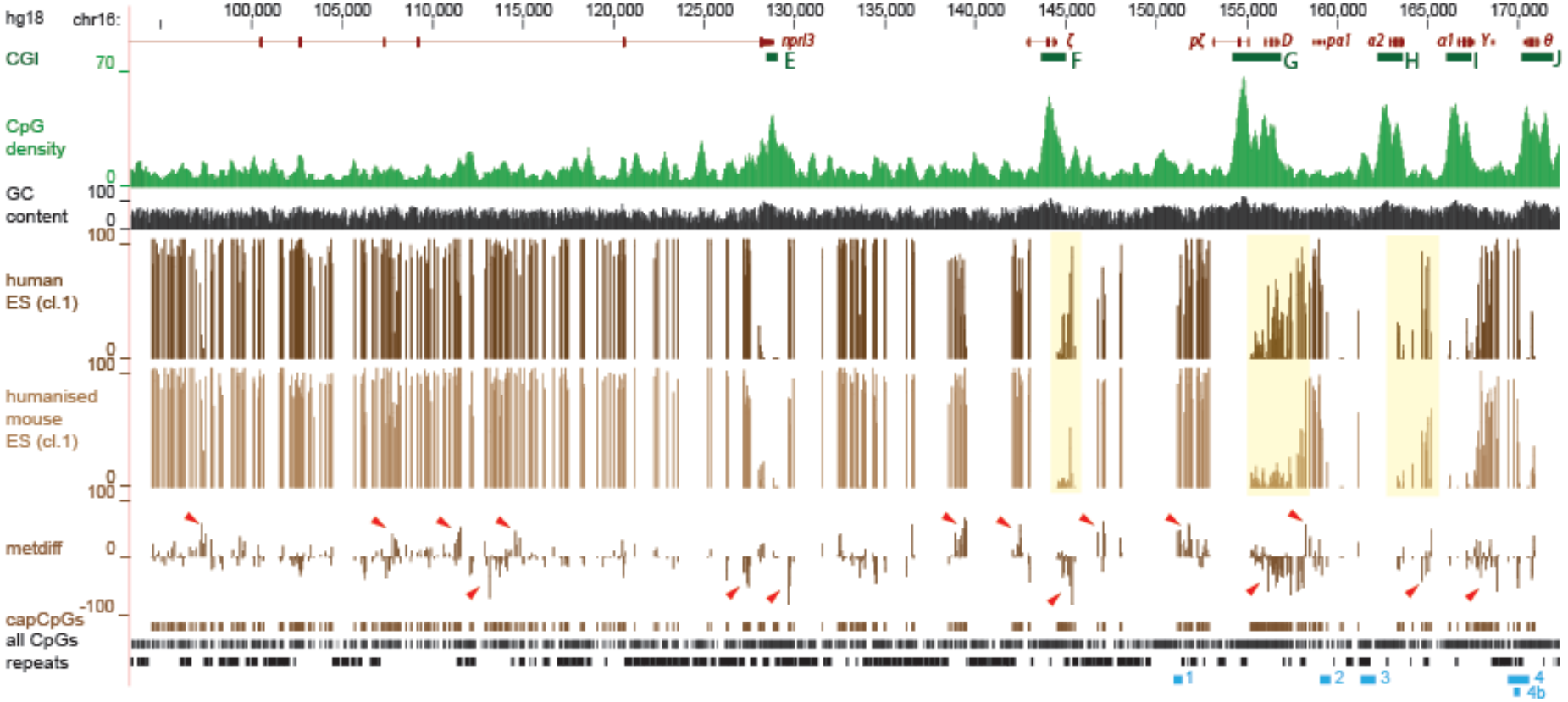

Figure 8: Overview of DNA methylation pattern across the human transgenic region in human and humanised mouse ESCs.

The tracks show the genomic location on chromosome 16 (chr16) in human genome build hg18, genes (Ensembl gene prediction) are shown in red, CGIs in green, CpG density, GC content, DNA methylation level [%} in human ESC and humanised mouse ESC, methylation difference (metdiff) between human and humanised mouse ESCs, captured CpG sites in both human (h) ES and and humanised mouse (hm) ESCs (cap CpGs; n= 785), all CpGs in the area, repetitive elements (repeatmasker). Highlighted in yellow are regions of obvious methylation changes between human and humanised mouse ESCs. Red arrowheads indicate examples of hypo- and hypermethylated sites in the plot showing the methylation difference between h and hmESCs. The location of amplimers 1-4 and 4b later used for bisulfite sequencing of individual subclones are indicated.

3.2.4 Characterisation of methylation changes in ESCs

Having established that methylation changes already appear in humanised mouse ESCs relative to the endogenous human cellular environment, the second aim was to characterise the DNA sequences that become differentially methylated in the mouse versus human. Gathering knowledge about which DNA sequences are not recognised and methylated by the mouse might help to dissect out possible mechanisms involved in normally methylating these sequences in a human cell.

3.2.4.1 DNA methylation of repetitive elements in humanised mouse versus human ESCs

Approximately half of mammalian genomes is made up of repetitive elements, including transposable elements (TEs), which can be harmful to the organism when active. Hence, mechanisms have evolved to silence TEs in mammalian genomes, including DNA methylation. These silencing mechanisms are thought to have co-evolved with TEs at a rapid rate owing to an incessant arms race between repeats and their silencing mechanisms. This is reflected by the appearance of lineage-specific TEs, such as human-specific Alu repeats. Hence, the question arises whether the mouse DNA methylation machinery fails to recognise human-specific repeats in the human *α*-globin cluster.

To address this question, the methylation data generated by TBS was stratified according to CpG sites located within or outside of repeats. Of the 785 CpG sites analysed 580 are located in unique sequences and 205 are in

repetitive elements, of which 148 coincide with Alu repeats (Table 1, Figure 9).

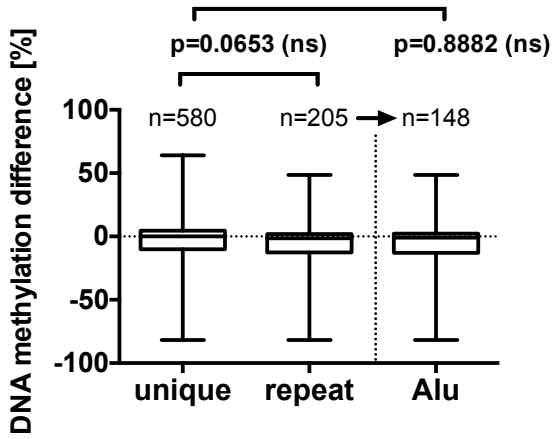

Figure 9: Boxplot of the DNA methylation difference at CpG sites within unique sequences, repetitive elements (repeat) and Alu repeats (Alu).

The y-axis shows the methylation difference, which was calculated by substracting DNA methylation levels at individual CpG sites in human ESCs from that in humanised mouse ESCs. The data was stratified according to the location of individual CpG sites within repeats (all repeats including Alu repeats) and within unique sequences of the human α-globin cluster. A subpopulation of the repeat fraction, the Alu sequences, was separately analysed. P-values were determined using the Mann-Whitney-test. A p-value >0.05 is deemed not significant (ns).

Table 1: Descriptive statistics of the methylation changes in the unique content versus the repeat content of the human *α*-globin cluster.

Of the 785 CpG sites, 580 coincide with unique sequences and 205 with repetitive elements. Of the latter 205, 148 fall into the category of Alu repeats. The mean and the median of the methylation difference between human and humanised mouse ESCs in each group is indicated.

	unique	repeat	Alu
n	580	205	148
median	0.0	-1.44	-0.96
mean	-2.1	-4.6	-4.5

Figure 9 shows a boxplot of the methylation difference, which, as mentioned above, was calculated by subtracting the level in human ESC from the level in humanised mouse ESC so that negative values represent hypomethylation in humanised mouse ESC versus human ESC and positive values indicate hypermethylated CpG sites. The average methylation difference at CpG sites located within repeat sequences is -4.6%, which is slightly, but not significantly (ns, p = 0.0653, Figure 9) lower than the average methylation difference in unique sequences (-2.1%) (Table 1). Similarly, further stratification of the repeat content into Alu repeats does not reveal a significant change in methylation compared to unique sequences with an

average methylation difference of Alu repeats of -4.5% compared to -2.1% in unique sequences (p=0.8882, Figure 9, Table 1). Hence, this data suggests that, although the repeat content of the human *α*-globin cluster is on average hypomethylated in humanised mouse versus human ESCs, the changes at repeats appear to largely reflect the changes observed across the whole transgenic region and do not seem to be specific to repeats.

3.2.4.2 Methylation changes at CpG sites within regions of differing GC- and CpG-content

It is hypothesized that human-specific sequences features foreign to the mouse genome might be improperly recognised by the mouse DNA methylation machinery which may account for the methylation differences observed when human DNA is placed into the mouse genome. With the availability of complete mammalian genome sequences, features specific to the human genome, which are not found in the mouse genome, have been identified. This includes the GC content, as well as the CpG content[12] of the human genome, both of which reach more extreme levels in the human genome compared with that of the mouse genome (Lander et al, 2001)[13]. As both GC- and CpG-content have previously been implicated in playing a role in determining the methylation state of a given sequence (Lienert et al, 2011), it was tested here whether the GC- or the CpG content of the DNA influences the extent of the change in methylation observed in humanised mouse versus human ESCs.

[12] The p in CpG stand for the phosphodiester bond between the two bases highlighting that C is followed by a G. The G+C content, in contrast, is the content of Gs and Cs in a given sequence irrespective of the order of the two bases.
[13] Overall, the mouse genome contains a higher number of G+Cs, but these seem to be more evenly distributed across the genome, because at smaller window sizes, the G+C content in the human genome exceeds that of the mouse reaching more extreme levels (low and high) (Lander et al, 2001).

3.2.4.2.1 GC content

To examine methylation changes of CpG sites in regions of differing GC content, for each given CpG site the methylation difference was plotted against the GC content of the DNA surrounding that CpG site. The GC content was calculated for each CpG site as the percentage of Gs and Cs in a 200 bp window surrounding that CpG site. Figure 10a shows a scatterplot of the methylation difference at single CpG sites in human versus humanised mouse ESCs over the GC content in a 200 bp window. Each dot represents a single CpG site. As can be seen in Figure 10a, hypo- and hypermethylated CpG sites cluster in regions with intermediate levels of Gs and Cs (40-65% GC content), while methylation of CpG sites within GC-rich (above 65% GC content) or GC-poor regions (30-40% GC content)[14] shows less extreme differences in mouse compared to human. Figure 10b shows boxplots of the methylation data grouped according to the GC content around each CpG site to better visualise the methylation variations in humanised mouse versus human ESCs. CpG sites in regions of intermediate to high levels of GC content (50-70%) show a significant change in the DNA methylation level (Figure 10b). Although CpG sites of low GC content (30-40%) tend to be more methylated in humanised mouse versus human ESCs, the methylation change is not significant due to the low coverage of CpG sites in GC-poor regions (Figure 10d). Note that the GC content anti-correlates with the overall methylation level of CpG sites in human and humanised mouse ESCs: GC-poor regions (30-40%) are densely methylated, regions of intermediate GC content show intermediate methylation levels, and GC-rich regions (70-80%)

[14] Note, however, that the coverage in GC-poor regions (30-40% GC content) is very low.

(mostly reflecting CpG sites in CGIs) are unmethylated (Figure 10c). In conclusion, hypomethylation does not directly correlate with the GC content of the sequence surrounding each CpG site. Yet there seems to be a tendency for hypomethylated CpG sites to cluster in regions of intermediate GC content. These regions coincide with sequences of intermediate methylation levels.

3.2.4.2.2 CpG content

To find out whether the methylation difference at a given CpG site between humanised mouse and human ESCs correlates with the CpG density of the surrounding DNA sequence the methylation difference was plotted over the CpG density (Figure 10e). The CpG density was calculated by counting the number of CpGs in a 200bp window surrounding each CpG site. CpG sites in CpG-poor regions (1-5 CpGs per 200bp), where most CpG sites lie (Figure 10h), are highly methylated in human ESCs and show on average a similar methylation level in humanised mouse ESCs, although hypo- and hypermethylated sites are seen in these regions (Figure 10f,h). In contrast, CpG-rich regions (16-30 CpGs/200bp) are largely unmethylated in human ESCs and show similarly low methylation levels in humanised mouse ESCs, suggesting that they are relatively well templated in humanised mouse versus human ESCs (Figure 10f,g). Significant changes in the level of DNA methylation occur in regions of intermediate levels of CpG density (6-15 CpGs per 200bp). In conclusion, it appears that unmethylated CpG-rich regions (above 16 CpGs per 200 bp) are well templated in humanised mouse versus human ESCs, largely reflecting the proper recognition of human CGIs as such by the mouse DNA methylation machinery, while most methylation

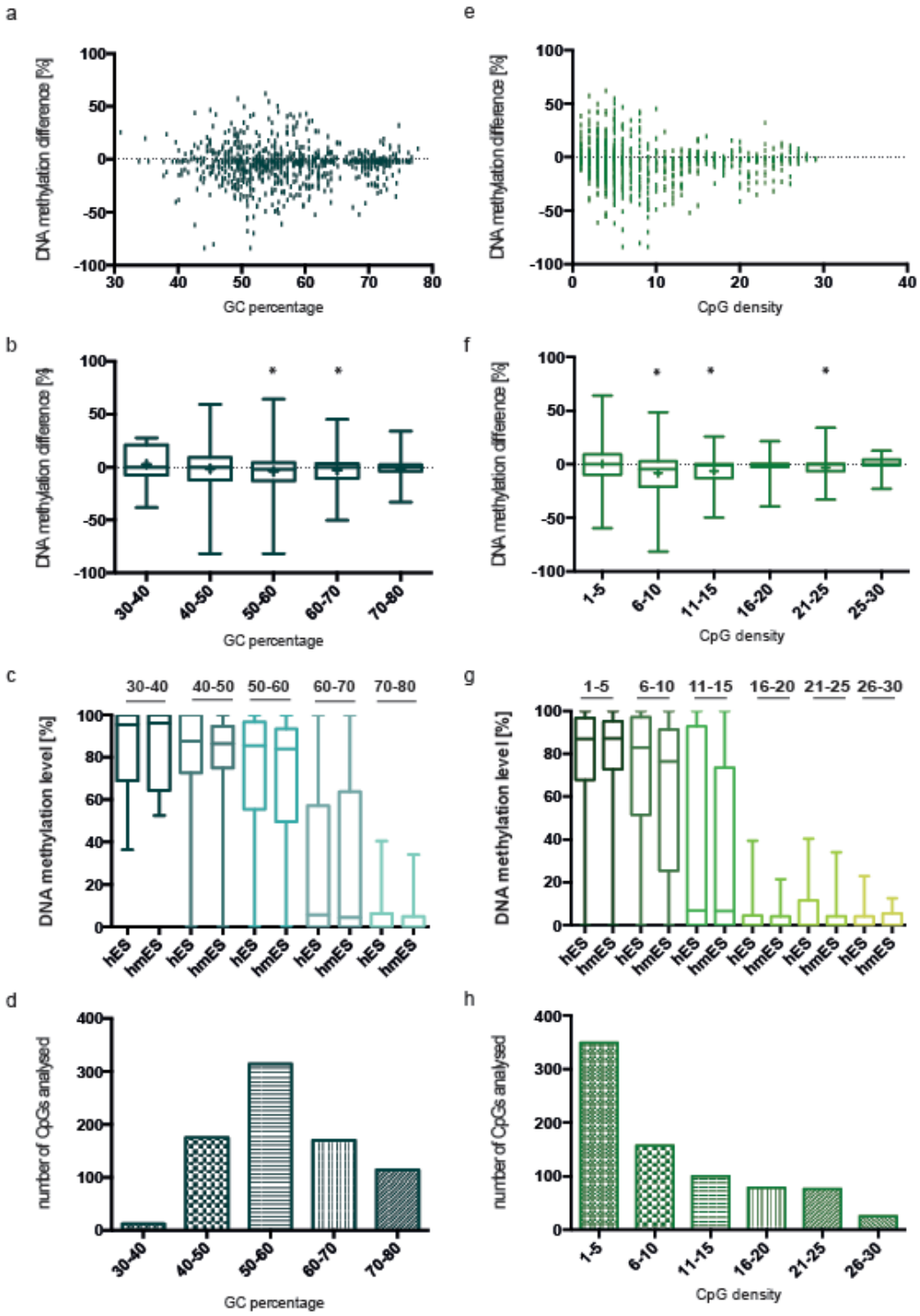

Figure 10: Comparison of the DNA methylation difference between human and humanised mouse ESC with the GC and CpG content of the DNA surrounding each CpG site.

(a) Scatterplot correlating the DNA methylation difference with the GC content (percentage of Gs and Cs in a 200bp window surrounding each CpG site). Each dot represents a single CpG sites. (b) Boxplots showing the methylation difference stratified according to GC content. (c) Boxplot showing the absolute methylation levels in human and humanised mouse ESCs according to GC content. (d) Number (n) of analysed CpG sites in each group of differing GC content. (e) Scatterplot of the DNA methylation difference against the CpG density (number of CpGs within a 200bp window surrounding each analysed CpG site). Each dot represents a single CpG sites. (f) Boxplots showing the methylation difference according to CpG density. (f) Boxplot of the absolute methylation levels according to the CpG density. (h) Number (n) of analysed CpG sites in each group. P-values between human and humanised mouse samples were calculated using the Mann-Whitney-Test.

aberrations occur in regions of low to intermediate CpG density (1-15 CpGs/200bp), with hypomethylation predominantly concentrating in regions of intermediate CpG density (6-15 CpGs/200bp). Significant hypomethylation in regions of 21-25 CpGs/200bp most likely reflect hypomethylation of the alpha globin gene bodies, which coincide with CGIs.

3.2.4.3 Hypomethylation of CGI shores

In the previous section genomic sequence features that differ between human and mouse genomes were analysed to find out whether the mouse DNA methylation machinery does not recognise human-specific sequence features. The data, however, did not support the notion that human-specific sequence features, such as the content of repetitive elements, GC- or CpG-rich sequences, become specifically hypomethylated due to the failure of the mouse methylation machinery to recognise these sequences. In the following, I will focus on the most obvious differences in methylation between human and humanised mouse and characterise the underlying genomic features.

The most obvious changes in methylation across the human *α*-globin cluster in humanised mouse versus human ESCs occur at sequences near the *α*-globin CGIs (Figure 8), which is consistent with the initial observation made by REA (Figure 6). It appears that the unmethylated regions of the CGIs associated with the *α*-globin genes are extended in humanised mouse compared to human ESCs spreading further into neighboring sequences. This is most evident at the region adjacent to CGI G (Figure 8). A similar pattern, however, can be seen at other CGIs, such as CpG Island F, H, and I.

As mentioned in section 1.4.1.3 studies have shown that regions up to 2kb away from CGIs, so called CGI shores, are subject to differential methylation in different tissues as well as in cancer (Irizarry et al, 2009). To investigate whether hypomethylation of the human transgenic region in the mouse was specific to the CGI shores the methylation data was stratified according to CpG sites that lie within CGI shores, CGIs, and those that lie outside of CGIs and their shores (Figure 11).

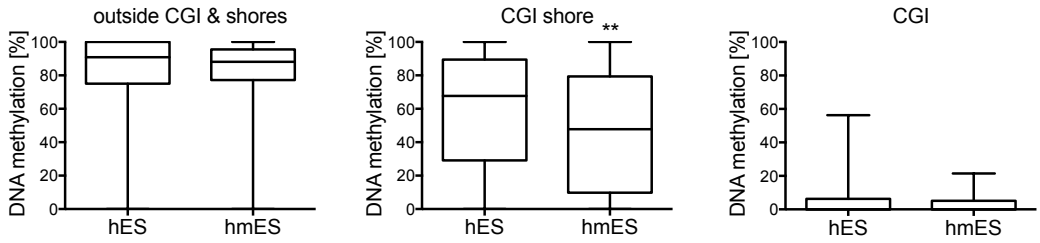

Figure 11: Boxplots showing the DNA methylation levels at CpG sites in CGIs, CGI shores, and regions outside of CGI shores in ESCs.

CGI shores were defined as 2kb regions flanking CGIs (Irizarry et al, 2009). P-values were calculated using the Mann-Whitney-Test (see Table 2).

Table 2: Statsitics of methylation differeces at CGIs, their shores and regions outside of CGIs and their shores in ESCs.

The number (n) analysed is indicated. Medians and means of the methylation difference between human (h) and humanised mouse (hm)ESCs are shown. P-values for the methylation difference between h and hmESCs were calculated using the Mann-Whitney-test (significant p-values < 0.05).

	Outside CGI&shores	CGI shore	CGI
n	409	147	229
Median	0.0%	-9.9%	-4.1%
Mean	0.1%	-11.1%	-2.5%
P value	0.0554 (ns)	0.0052 (**)	0.6281 (ns)

As can be seen in Figure 11, human CGIs are largely unmethylated in humanised mouse and human ESCs, although the average methylation difference (-2.5%) indicates that CGIs are slightly less methylated in humanised mouse versus human ESCs (Table 2). On the contrary, CGI shores within the human transgenic region are significantly less methylated in mouse ESCs compared to human ESCs. Hypomethylation seems to be a particular feature of CGI shores, as regions outside of CGIs and their shores are methylated at similarly high levels in humanised mouse compared to

human ESCs. Hence, this data suggests that CGI shores are significantly hypomethylated in humanised mouse versus human ESCs.

3.2.4.4 Validation of CGI shore hypomethylation

To ensure these changes were not attributable to a clonal effect, the finding was validated in another ESC clone from human and humanised mouse. DNA methylation patterns were measured by conventional bisulfite sequencing (CBS) of selected amplicons located within CGI shores, CGIs or outside CGIs and their shores. The location of the amplimers 1-4 is shown in Figure 8. The analysed CpG sites that lie outside of CGIs and their shores (amplicons 1 and 2) are highly methylated in both human and humanised mouse ESCs. In contrast, CpG sites next to CGI H (amplicon 3) and CGI J (amplicon 4) are hypomethylated in humanised mouse versus human ESCs, confirming the previous finding that regions adjacent to CGIs are hypomethylated in mouse ESCs (Figure 12). In summary, hypomethylation of regions could be verified in another set of ES cell lines of human and humanised mouse. The data supports the previous finding that hypomethylation is specifically found at regions directly adjacent to CGIs and this seems to be due to an extension of the unmethylated region of CGIs in humanised mouse ESC compared to human ESC.

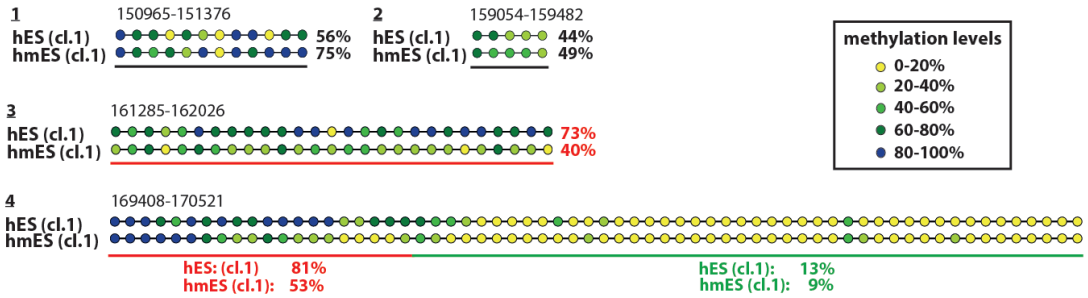

Figure 12: Verification of DNA methylation changes in the vicinity of CGIs by conventional bisulfite sequencing (CBS).

The location of the amplimers 1-4 is indicated in Figure 8. Each dot represents a CpG site, the methylation levels are color coded. Amplimers 1 and 2 are located outside of CGIs and shores (further than 2kb away from CGIs, indicated by a black line), amplimer 3 is located within a CGI shore (indicated by a red line), amplimer 4 covers a CGI (underlined in green) and its shore (red). The chromosomal location of the amplimers is indicated. Average methylation levels of all CGIs across each analysed region are given for human ES (hES) and humanised mouse ES (hmES).

3.2.4.5 Single base pair resolution profile of the DNA methylation pattern around the human α-globin genes in human and humanised mouse ESCs

The data shown above indicates that hypomethylated CpG sites are concentrated in CGI shores in humanised mouse versus human ESCs, suggesting an extension of the hypomethylated sequences around CGIs. To investigate changes in DNA methylation at CGIs and their shores in greater detail, a higher resolution DNA methylation profile of representative CGIs and their surrounding sequences were desired. (The limitations of the TBS approach are described in 7.1.2.1). It was chosen to carry out bisulfite sequencing of amplimers spanning a 20kb region around the *α*-globin genes in human and humanised mouse ESCs. The region was divided into 82 fractions of an average size 250 bp and primers were designed for PCR amplification of bisulfite converted DNA. For 74 of these it was possible to get specific PCR products. The location of the amplimers is shown in Figure 13. Bisulfite conversion of DNA from h and humanised mouse (hm) ESC was carried out in parallel to ensure similar conversion efficiencies of the samples and eliminate biases resulting from over-conversion of one sample over another. In addition, an *in vitro* methylated control sample was converted in

parallel. This was done to verify that methylated CpGs were protected from conversion. For reasons of cost- and time-efficiency it was chosen to sequence all PCR products by high-throughput sequencing on the Illumina platform (MiSeq) once successful conversion has been validated in a number of subclones. (For more details on the design of the method see section 7.1).

Figure 13 shows the DNA methylation pattern in human and humanised mouse ESCs across the 20kb region amplified. The region covers four CGIs (CGI G, H, I J) associated with the *α*-globin genes (*α*D, *α*2, *α*1, and θ, respectively). Note, that due to the small size of the *α*-globin genes, the CGIs span most of the gene. All four CGIs are unmethylated in human ESC, with the exception of the *α*-globin gene bodies, which show a slightly elevated DNA methylation level. Methylation levels increase just outside of the borders of the annotated CGIs. Most CpG sites outside of CGIs are highly methylated, with the exception of two dips in methylation at the promoter region of pseudo *α* and just upstream of it. In line with the data presented above (see chapter 3.2) the CGIs are also unmethylated in humanised mouse ESCs, although the increase in methylation at the gene bodies is not observed in humanised mouse ESCs. Importantly, in confirmation of the previous data, the unmethylated region at CGI G, H and J are extended in humanised mouse versus human ESCs.

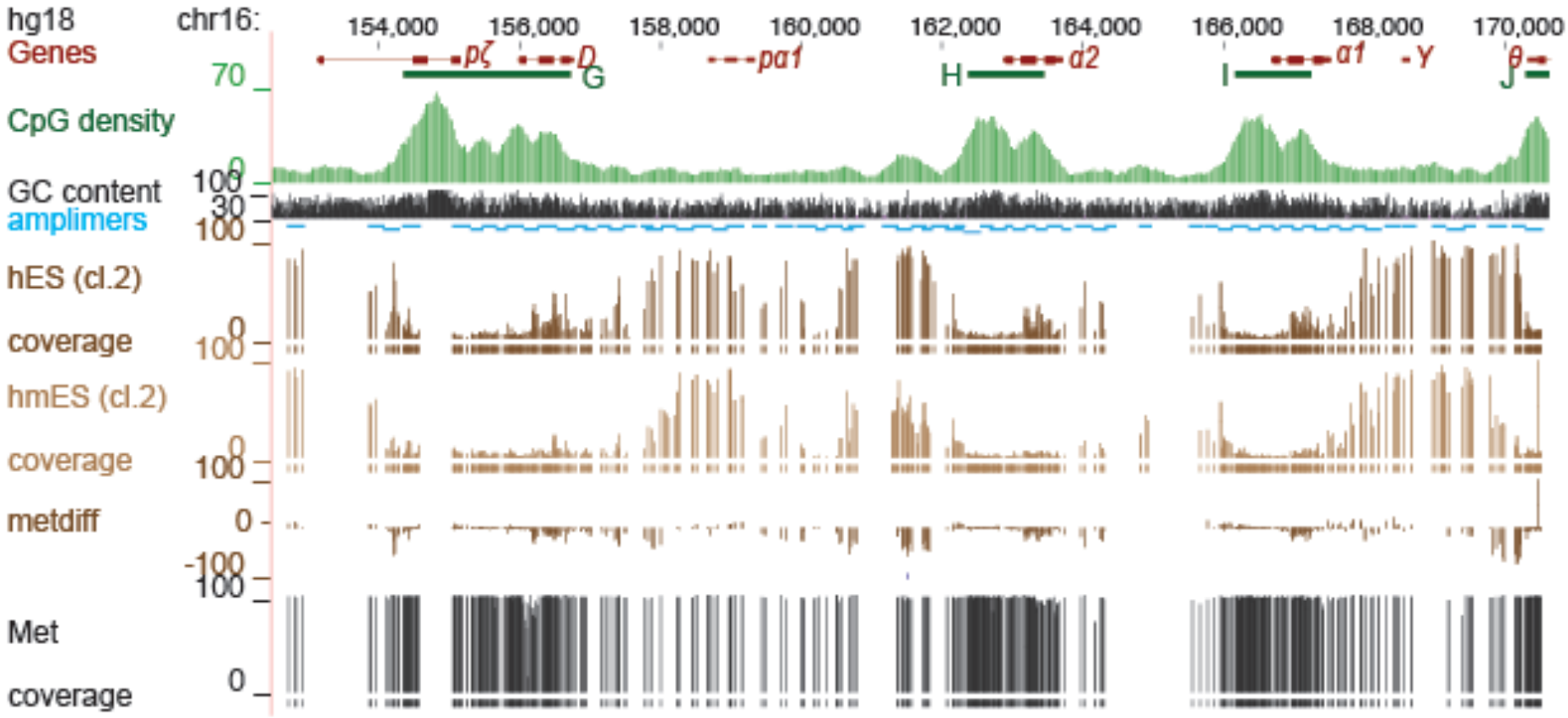

Figure 13: Single base pair DNA methylation profile of the human *α*-globin genes in human and humanised mouse ESCs.

DNA methylation levels were assessed by bisulfite sequencing of a 20kb region surrounding the *α*-globin genes and their associated CGIs. DNA methylation [%] in human ES (hES cl.2), humanised mouse ES (hmES cl.2) and in an *in vitro* methylated control sample (Met) are shown for individual CpG sites (vertical bars). The coverage of CpG sites in each dataset is indicated below the given methylation track. The methylation difference between human and humanised mouse ESCs (metdiff) shows hypomethylated CpG sites as negative values and hypermethylation as positive values. Genes are shown in red, CGIs in green.

In conclusion, the data presented here suggests that human CGIs are extended in humanised mouse versus human ESCs leading to hypomethylation of CGI shores, confirming the previous findings. Furthermore, mild methylation of the *α*-globin gene bodies, which are located within their associated CGIs, was observed in human ESC and this methylation is lost in humanised mouse ESC.

3.2.4.6 Methylation pattern at the promoter CGI of the NPRL3 gene

Above it has been shown that the CGIs associated with the human *α*-globin genes are extended in humanised mouse versus human ESC. The *α*-globin gene, however, constitute a class of unusual gene in which the CGIs span most of the gene. A better representative of a typical gene structure (a large gene body with a CGI that spans the promoter and first eon region) is the *NPRL3* gene located upstream of the human *α*-globin genes (Figure 8). Bisulfite sequencing was carried out to investigate the methylation pattern around the *NPRL3* CGI. Primers were designed to span the CGI and about 2.5kb of DNA on either side of it. As can be seen in Figure 14 CGI E is unmethylated in human and humanised mouse ESCs confirming the initial REA results (mentioned in section 3.2.1). Furthermore, the hypomethylated region at CGI E is extended in humanised mouse ESCs compared to human ESC, confirming previous results for another CGI.

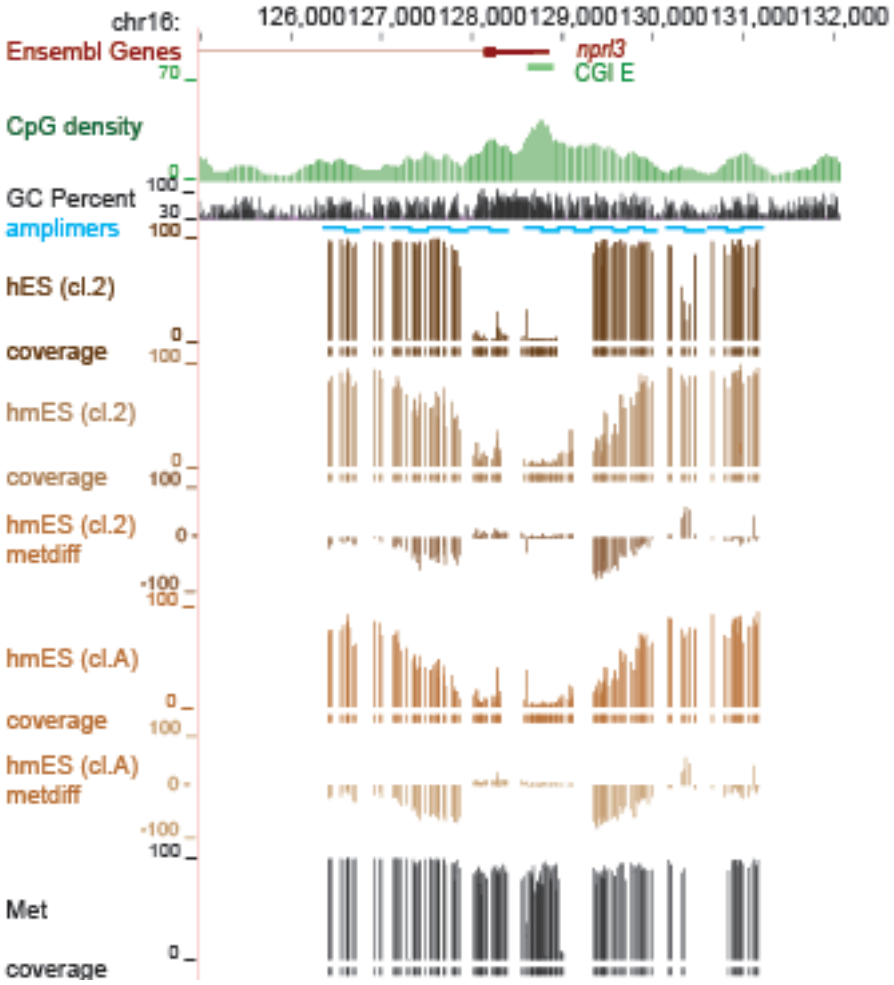

Figure 14: DNA methylation pattern around CGI E (associated with the *NPRL3* promoter) in human and humanised mouse ESCs.
DNA methylation levels [%] analysed by bisulfite sequencing across CGI E (green) associated with *NPRL3* (red) in human ES**C** clone 2 (hES cl.2), humanised mouse ES**C** clone 2 (hmES cl.2) as well as in a re-derived humanised mouse ES**C** clone A (hmES cl.A). PCR amplimers used for PCR amplification are shown in light blue.

In development, global DNA methylation levels are erased and re-established during two waves of epigenetic reprogramming (see section 1.5.1). In contrast to human ESCs, the human sequence in humanised mouse ESCs analysed so far have never gone through the germ line, hence, it has never undergone the epigenetic reprogramming event *in vivo*. Therefore, it was asked whether the methylation pattern observed in the hmESC lines observed so far truly reflect the pattern in humanised mouse ESC *in vivo*. To address this question, fresh ESC clones were derived from a humanised mouse and methylation patterns were analysed at the *NPRL3* promoter CGI as a representative

example of an extended CGI. Methylation patterns across the CGI and adjacent regions were analysed by bisulfite high-throughput sequencing in the re-derived ESC line (hmES cl.A) (Figure 14). It was found that the pattern of methylation around CGI E at the *NPRL3* gene is highly similar to the pattern in the humanised mouse ESC clone that has been generated *in vitro* and never been in the mouse, suggesting that the genetically engineered ESC clones reflect the pattern *in vivo*.

To confirm this finding methylation levels were also measured at the region 5' to CGI J in the re-derived ESC line by conventional bisulfite sequencing (Figure 15). The average methylation level of 12 CpG sites in amplimer 4b located in the shore upstream of CGI J (Figure 8) is 80% in human ESCs, while in the re-derived hmESC clone it is 46%, which is comparable to the level in the previously analysed hmESC clone 2 (35%). In conclusion, these findings suggest that the documented methylation patterns around human CGIs in humanised mouse ESCs reflect the pattern in humanised mouse ESCs *in vivo*.

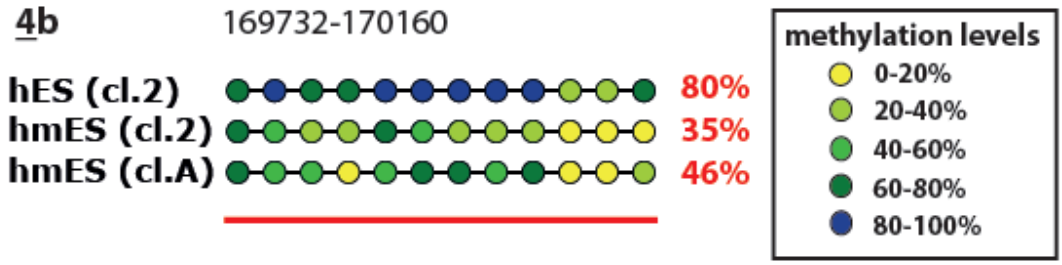

Figure 15: Hypomethylation of the CGI shore upstream of CGI J in the re-derived hmES clone A (hmES cl.A). DNA methylation levels were measured by CBS. Amplimer 4b covers the region 169732-170160 on chromosome 16 (indicated in Figure 8). The average DNA methylation level of all CpG sites in the amplified region (red line) is indicated (red).

3.2.5 Hypomethylation becomes exacerbated in brain

Having established that humanised mouse ESCs are normally methylated across most of the transgenic region with the exception of CGI shores, the next aim was to find out whether the methylation patterns established in ESCs are properly maintained in differentiated cells. Two differentiated cells types from human and humanised mouse were examined, including erythoid and non-erythroid cell types. For reasons of accessibility and ability of choosing matched developmental stages, human and humanised mouse erythroid cells and brain tissues were chosen for methylation analysis (see section 3.2.7).

To get a quantitative profile across most of the transgenic region TBS was carried out on DNA from human and humanised mouse brain. Due to poor sequence qualities of the reverse strands in this dataset, only the forward reads were allowed for analysis, resulting in a coverage of 540 and 327 CpG sites in human and humanised mouse ESCs, respectively. 247 of these are covered in both datasets. The methylation levels of these 247 CpG sites were visualised in UCSC genome browser and compared to the methylation pattern of human and humanised mouse ESCs described in the previous section (section 3.2.3, Figure 16).

Despite the low coverage it is apparent that the overall methylation pattern looks largely similar in humanised mouse brain compared to human brain. Methylation differences are, however, visible in these differentiated cells. An interesting pattern is observed at CGI G, where the region that is less methylated in humanised mouse compared to human ESCs is unmethylated in both human and humanised mouse brain leading to the disappearance of

the hypomethylated sites (underlined with a black line in Figure 16). Furthermore, hypomethylation seems to spread further away from CGI G in humanised mouse versus human brain. Despite reservations because of the low coverage, it seems that the unmethylated region at CGI G becomes further extended in humanised mouse versus human brain leading to hypomethylated sites located further away from the CGI (Figure 16). Furthermore, the data suggests that hypo- and hypermethylated CpG sites display more extreme values, indicating that the aberrations become exacerbated during development.

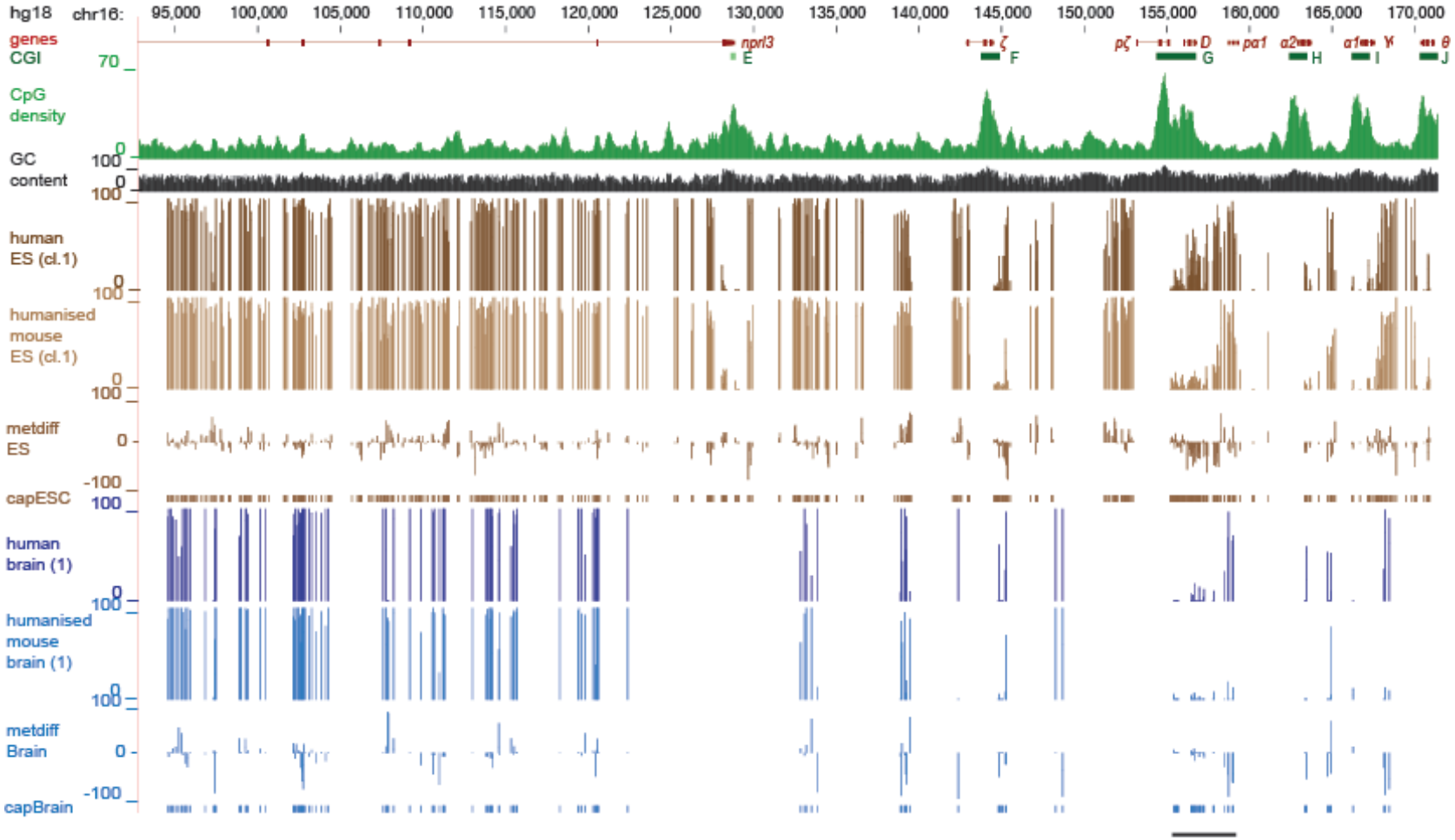

Figure 16: Overview of the DNA methylation patterns in h and hm brain in comparison to the pattern in h and hm ESC.

DNA methylation levels were analysed by targeted bisulfite sequencing (TBS). Track show genes (red), CGIs (green), CpG density (green), GC content, methylation levels [%] in h and hm ESCs and brain. Methylation levels are shown for CpG sites covered in both datasets of human and humanised mouse ESC (n=785) or brain (n=247). The methylation difference (metdiff) was calculated for ESC and brain so that negative values represent hypomethylation and positive values hypermethylation. CpG sites covered in ESCs (capESC) and brain (capBrain) are indicated.

To better visualise the distribution of the methylation difference in ESCs and brain the data was divided into hypo-, hyper- and normally methylated CpG sites. Methylation differences between -15% and +15% were deemed normal, whereas differences below -15% were considered as hypomethylated and above +15% hypermethylated (Figure 17). In humanised mouse versus human ESCs 70% of CpG sites are within a range of -15% and 15% of methylation difference, indicating that most CpG sites are normally methylated (Figure 17a). Similarly, 75% of CpG sites are methylated to a similar degree (in the range of -15% to 15% methylation difference) in humanised mouse versus human brain suggesting that most CpG sites are normally methylated in humanised mouse compared to human brain (Figure 17c). In humanised mouse ESCs 19% of CpG sites are hypomethylated. In brain a slightly smaller percentage (17% of CpG sites) is hypomethylated. At the other end of the spectrum, 11% and 8% are hypermethylated in humanised mouse and human ESCs versus brain, respectively (Figure 17a,c). Interestingly, however, most of the methylation changes between humanised mouse and human ESCs are relatively mild (1% of CpG sites are hypo- and hypermethylated by more than 50% Figure 17b). In brain the methylation changes are more extreme (7% and 3% of the CpG sites are hypo- and hypermethylated, respectively, by more than 50% in humanised mouse versus human brain) (Figure 17d). In conclusion, in line with the initial observation the human *α*-globin cluster is hypomethylated in humanised mouse versus human. Furthermore, the methylation changes become exacerbated in

humanised mouse versus human brain in comparison to ESCs as shown by the greater deviation of the methylation differences from 0% (Figure 18 and Table 3).

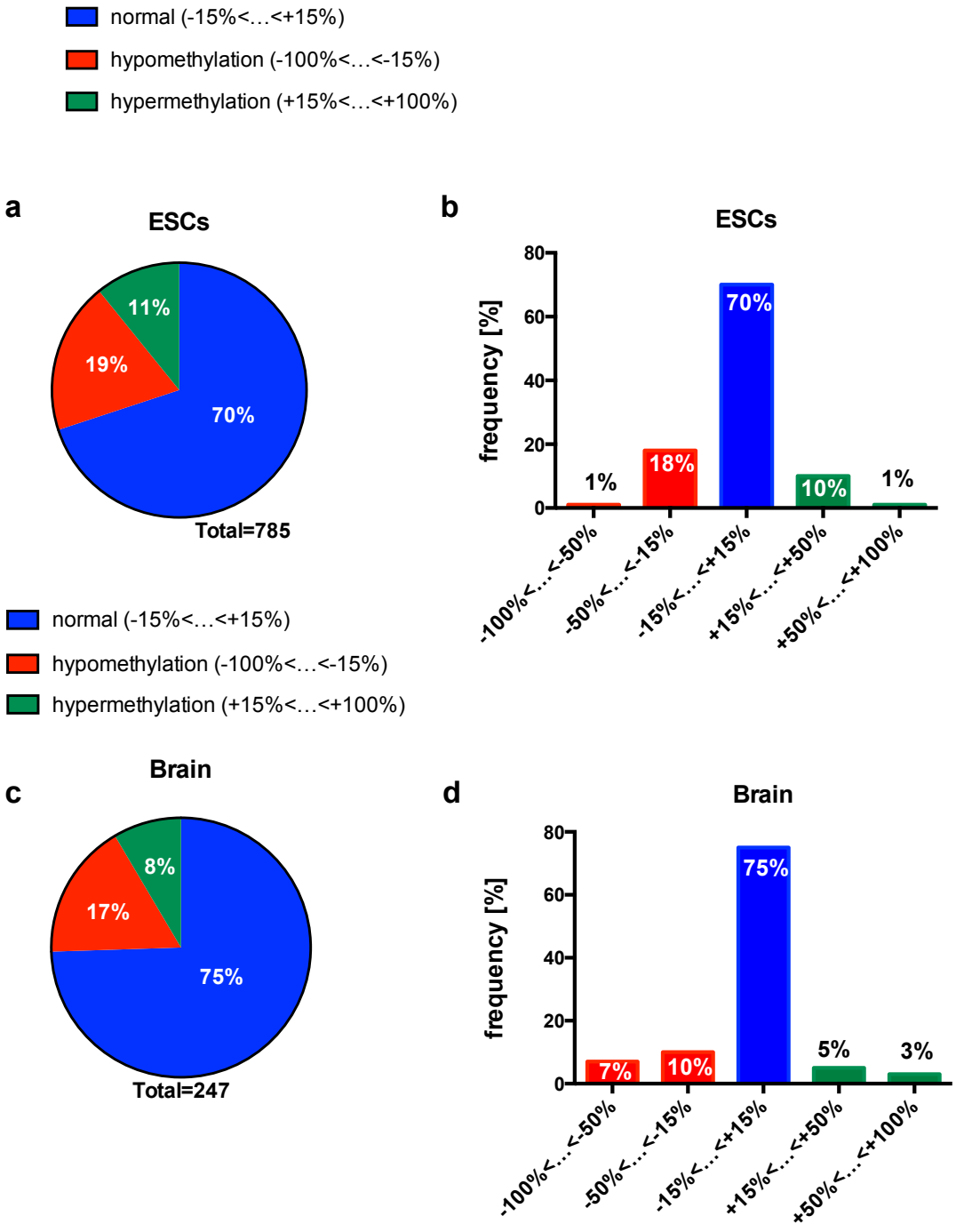

Figure 17: Analysis of the methylation difference in ESCs and brain.

Hypo and hypermethylation of humanised mouse versus human in ES (a, b) and brain (c, d) analysed by TBS. The data was stratified according to the degree of hypomethylation in humanised mouse versus human ESC and humanised mouse versus human brain. A methylation difference between -15% and 15% was deemed normal. CpG sites with a methylation difference that exceeds +15% or -15% in humanised mouse versus human was called hypermethylated or hypomethylated, respectively.

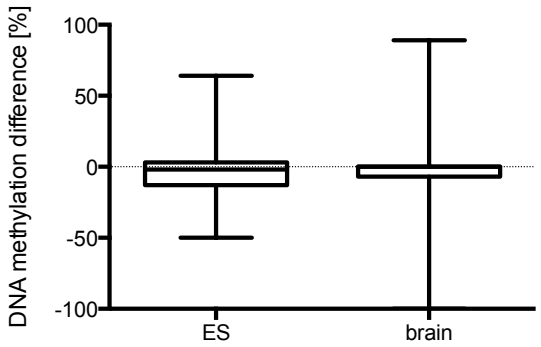

Figure 18: Comparison of the methylation difference between human and humanised mouse in ESC and brain.
Methylation changes become exacerbated reaching more extreme levels in brain versus ESC (p=0.001978, Ansari-Bradley-test).

Table 3: Statistics of the methylation difference in human and humanised mouse ESC versus brain.
The total number (n) of CpG sites analysed in ESCs and brain is indicated, as well as the median of the methylation difference. The standard deviation (stdev) of the methylation difference between human and humanised mouse ESCs versus brain, reflecting the spread of the data, is given.

	ES	brain
n	236	236
median	-2,0	0.0
stdev	16	26.23

3.2.6 Hypomethylation of CGI shores in differentiated cells.

To this point, the evidence suggests that hypomethylation at the human α-globin genes in humanised mouse versus human becomes exacerbated in brain. To find out whether hypomethylation in humanised mouse versus human brain affects CGI shores, the data obtained by TBS (s. chapter 3.2.5) were stratified according to the location of CpG sites in CGI shores, CGIs or outside of CGIs and their shores. It was found that also in brain cells, hypomethylation predominantly occurs at CGI shores, which are significantly less methylated in humanised mouse versus human brain. In contrast, human CGIs and regions outside of CGIs and their shores are not significantly less methylated in humanised mouse compared to human brain, suggesting that hypomethylation is a specific feature attributed to CGI shores.

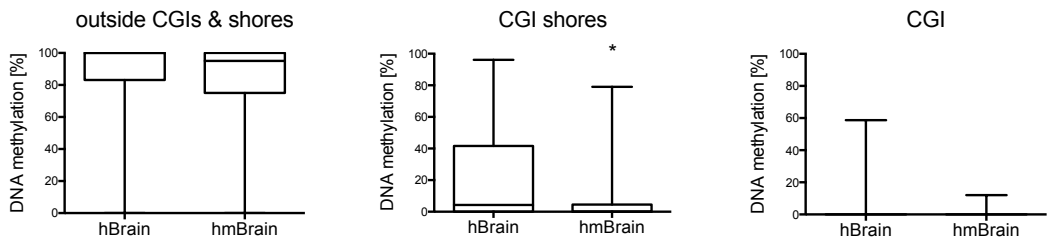

Figure 19: DNA methylation in human and humanised mouse brain at CGIs, CGI shores and outside of CGIs and their shores.

Stratification of the data according to the location of CpG sites within CGIs, CGI shores (up to 2kb away from CGIs) or outside of CGIs and their shores. Significant hypomethylation in humanised mouse brain versus human brain is observed in CGI shores (p=0.0114, Mann-Whitney-test; n=237)

Table 4: Statsitics of methylation differeces at CGIs, their shores and regions outside of CGIs and their shores in h and hm brain.

The number **of CpG sites** analysed (n) is indicated. Medians and means of the methylation difference between h and hmBrain are shown. P-values for the methylation difference between h and hmBrain were calculated using the Mann-Whitney-test (significant p-values < 0.05).

	Outside CGI&shores	CGI shore	CGI
n	145	32	59
Median	0.00	-3.064	0.0
Mean	-3.69	- 9.67	-2.04
P value	0.0965 (ns)	0.0144 (*)	0.3589 (ns)

3.2.6.1 Validation of CGI shore hypomethylation in brain

To verify that CGI shores become hypomethylated in humanised mouse versus human brain, DNA methylation levels of the region spanning CGI J and its adjacent shore were analysed by CBS. The location of the chosen amplimer (amplimer 4) is depicted in Figure 8. The average methylation level of CGI J is 2% in human brain and 4% in humanised mouse brain. Methylation levels increase at the border of the CGI with an average methylation level of 25% in humanised mouse brain in fragment 4b (located directly adjacent to the annotated CGI J) versus 38% methylation of fragment 4b in human brain. Hence, the region directly adjacent to CGI J is hypomethylated in humanised mouse versus human ESCs. Judging from the pattern of methylation (Figure 20) the unmethylated region at CGI J seems to spread further into the 5' adjacent region leading to hypomethylation of the shore of CGI J in humanised mouse versus

human brain. Note that the hypomethylated region at CGI J seems to extend in normal human development (human brain versus human ESCs). This is in line with the previous observation of an extension of CGI G in human brain versus human ESC (see chapter 3.2.5), indicating that changes at CGIs occur during normal development.

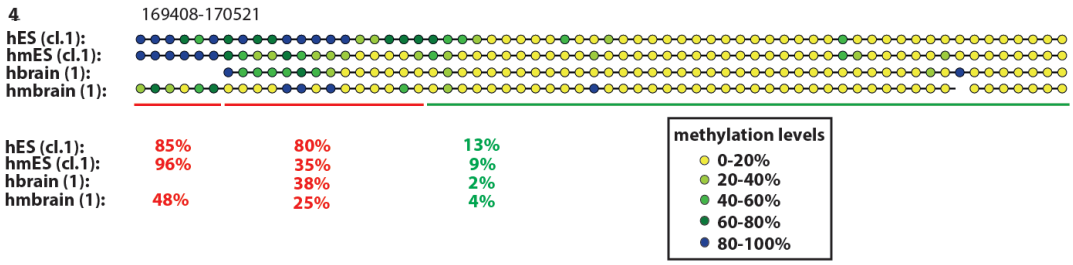

Figure 20: Methylation pattern at CGI J and its upstream shore in h and hm brain in comparison to ECSs.
The hypomethylated region of CGI J is extended in humanised mouse brain compared to human brain. The genomic coordinates (chr16, hg18) of amplimer 4 is indicated, its location relative to CGI J is shown in Figure 8. CpG sites located within CGIs or their shores are indicated by green and red lines, respectively, and their average methylation levels are indicated.

3.2.7 DNA methylation pattern around the human α-globin genes at single nucleotide resolution in development

To further validate the proposed hypothesis that CGIs become extended in differentiated cells, a single nucleotide resolution profile of DNA methylation levels across the human *α*-globin genes in human and humanised mouse erythroid cells and brain were generated. (For a comparison of differentiation stages see section 7.3). This was done by bisulfite conversion coupled to high-throughput sequencing as described in section 3.2.4.5.

The DNA methylation pattern of the *α*-globin gene region in human and humanised mouse erythroid cells and brain is visualised in UCSC genome browser in Figure 21. As expected human CGIs at the *α*-globin genes are

unmethylated in both human and humanised mouse tissues, in both erythroid and non-erythroid cells, confirming the previous data suggesting that human CGIs are correctly recognised and unmethylated in humanised mouse tissues as observed in human tissues (see chapter 3.2). Interestingly, in normal development, gene body methylation of the *α*-globin genes seems to disappear in both human erythroid and non-erythroid cells compared to human ESCs. In contrast, in humanised mouse, human *α*-globin gene bodies, which as mentioned above are already unmethylated in ESCs, stay unmethylated in erythroid and non-erythroid tissues. Overall, the analysed region across the *α*-globin genes displays an overall decrease in methylation levels in humanised mouse versus human erythroblasts. In a similar fashion, the human *α*-globin genes are associated with severe hypomethylation across the whole 20kb region analysed in humanised versus human brain. Hence, these findings support the above described observation that hypomethylation becomes exacerbated during differentiation in humanised mouse versus human brain. Furthermore, it could be shown that hypomethylation increases in humanised mouse versus human erythroblasts. In conclusion, it was shown that the mouse DNA methylation machinery fails to properly maintain DNA methylation patterns of the human *α*-globin cluster in erythroid and non-erythroid cells leading to an exacerbation of hypomethylation during development.

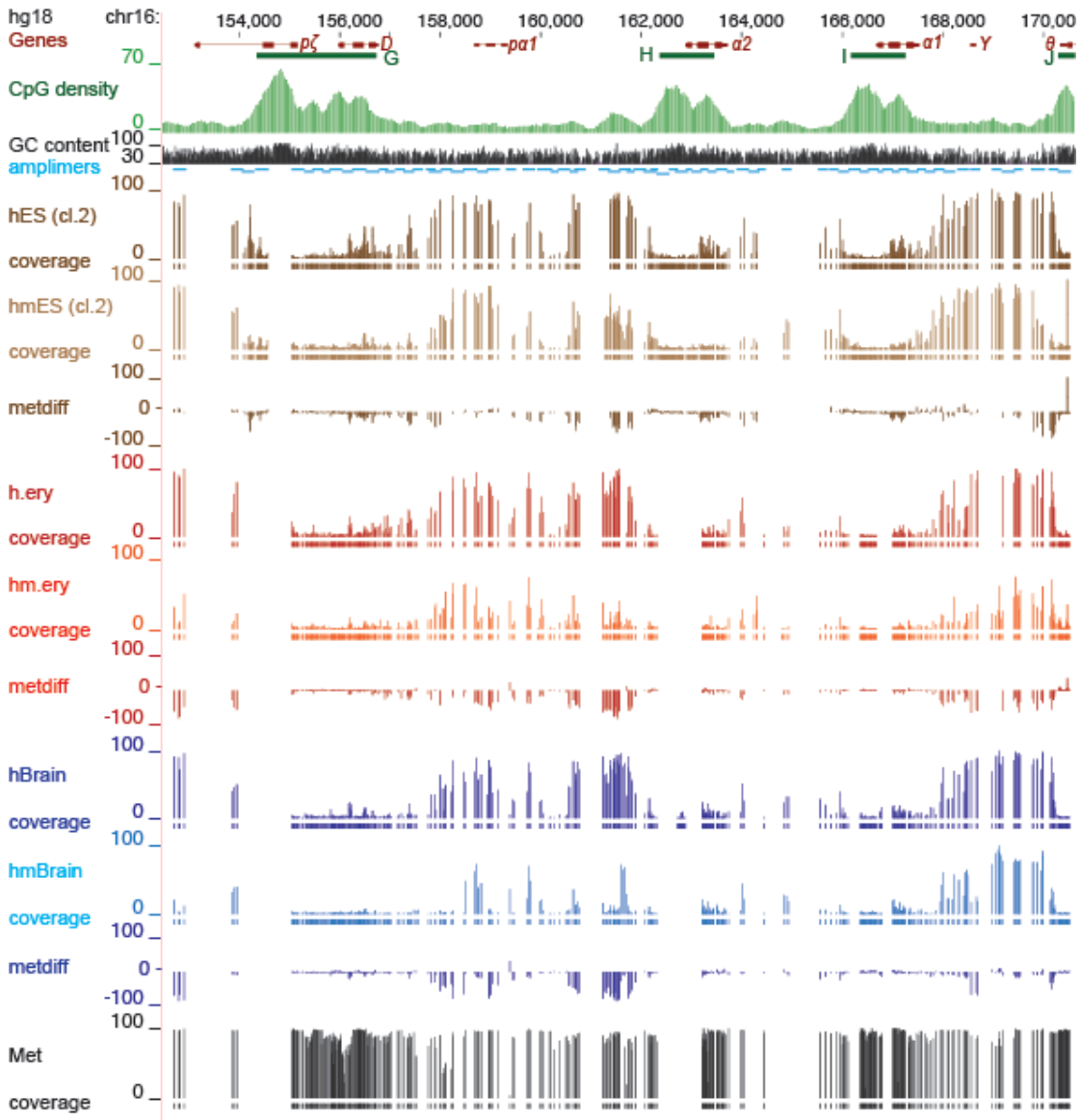

Figure 21: DNA methylation pattern of the *α*-globin cluster during development.

DNA methylation levels were analysed by bisulfite sequencing of a 20kb region surrounding the α-globin cluster in human and humanised mouse erythroblasts (h.ery and hm.ery, shown in red) as well as in human and humanised mouse brain (hbrain and hmbrain, shown in blue). Methylation differences were calculated for individual CpG sites in humanised mouse versus human erythroblasts and brain (negative values = hypomethylation, positive values = hypermethylation). An *in vitro* methylated control was converted in parallel to verify successful conversion of the DNA. For comparison to the pattern in pluripotent cells, the methylation data for h and hm ESCs from Figure 13 is also shown.

3.2.8 Overview of the changes of methylation around CGIs

To better visualise the changes in methylation according to their location relative to CGIs, the methylation data obtained by profiling the 20kb region covering *α*-globin CGIs G, H, I, and J in humanised mouse versus human ESC, erythroid cells and brain was replotted relative to their location to the nearest CGI and all CGIs analysed (G, H, I, and J) were overlayed (after normalising for their width) (Figure 22). In ESCs hypomethylated sites cluster largely directly upstream (0 to -1000bp) of CGIs although a few hypomethylated sites are also seen further away (for example around -1700bp), while at the downstream site of CGIs hypomethylated cluster mainly around +500 to +1500bp. Furthermore, about 25% at the 3' end of the CGIs is mildy hypomethylated, which most likely represents the observed lack of gene body methylation in humanised mouse ESCs. A few hypermethylated CpG sites are found at around -500bp and -1300bp as well as at +500bp to +1500. In erythroid cells and brain, hypomethylated sites are located further away from the CGIs reflecting the documented extension of the unmethylated regions in humanised mouse versus human differentiated cells. In both erythroid and brain the most obvious hypomethylation appears directly upstream of CGIs and up to -2000bp away, while at the other end hypomethylation is most prominent at about +700bp and +2500bp downstream of CGIs in erythoid cells and +1200bp to 2500 bp in brain (Figure 22).

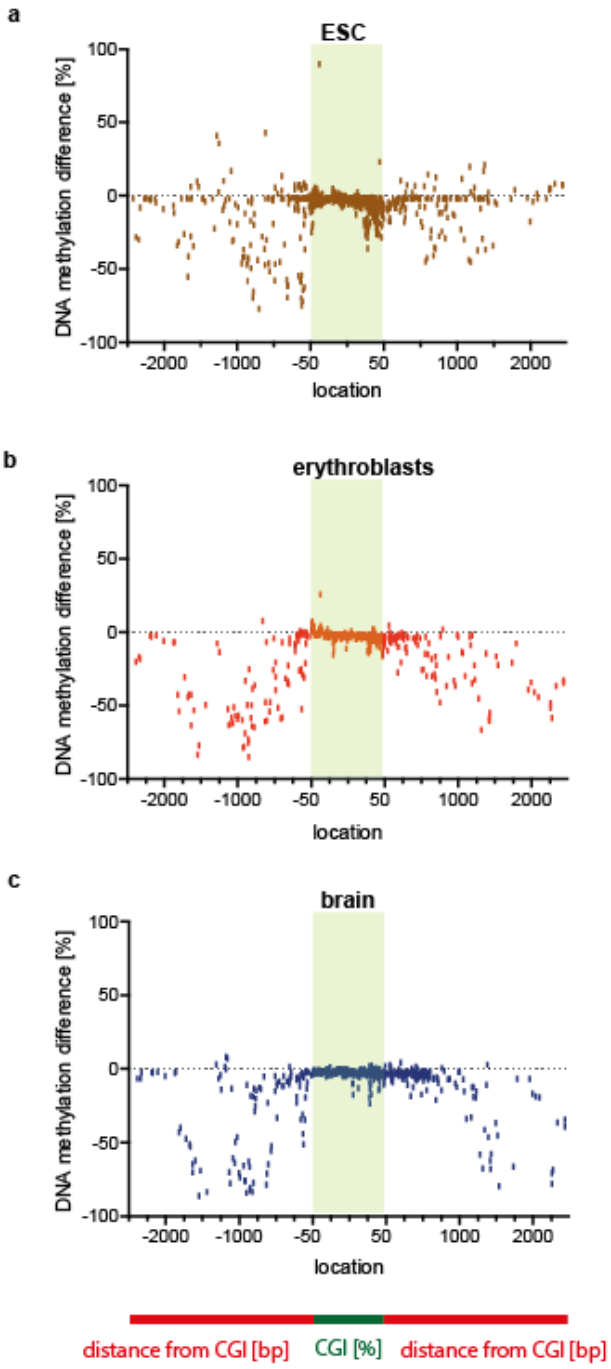

Figure 22: Relative location of the methylation difference at individual CpG sites with respect to their nearest CGI.
CGIs G, H, I, and J were overlayed after normalizing for their width [%]. The position of CpG sites adjacent to CGIs is expressed as the distance to its nearest CGI. The methylation data for humanised mouse versus human ESC (a), erythroblasts (b) and brain (c) was obtained by bisulfite sequencing of the 20kb region covering CGIs G, H, I, and J; the pattern in UCSC is shown in Figure 21.

The same was done for the data obtained for *NPRL3* in human and humanised mouse ESCs, where hypmethylation is most extreme at approximately -800bp to -1500bp and +500bp to +1100 away from CGI E (Figure 23). In summary, this data suggests that hypomethylated CpG sites largely fall in the regions previously defined as CGI shores (2kb regions adjacent to CGIs), although they can extent even a bit further than 2kb away from CGIs in erythroid cells and brain.

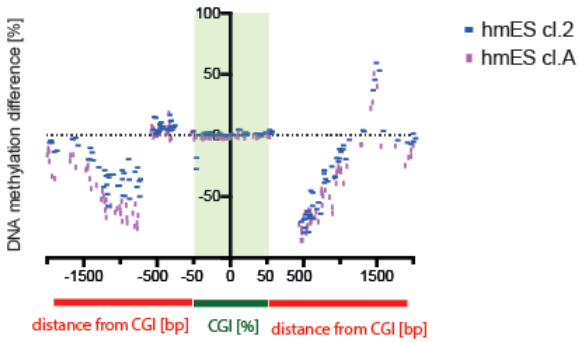

Figure 23: Relative location of the methylation difference relative to CGI E in h and hm ESC.
The methylation data obtained by bisulfite sequencing of a ~5kb region surrounding CGI E (associated with *NPRL3*) was shown in UCSC in Figure 14. The width of CGI E (from UCSC) is expressed as the percentage. The location of each analysed CpG site outside of CGIs is given as the relative distance to its nearest CGI.

3.3 Discussion

3.3.1 Putative species-specific differences in the trans-acting environment of human and mouse cells

In this chapter, a humanised mouse model, in which the human *α*-globin cluster replaces the mouse syntenic region on chromosome 11, was employed to study the establishment and maintenance of DNA methylation in a heterologous mouse environment compared to its endogenous chromatin state. It was found that human DNA methylation patterns across most of the transgenic region are

relatively well established and maintained in the humanised mouse model compared to normal human cells, suggesting that the human transgenic region is sufficient to determine its methylation pattern in a heterologous murine environment. An exception is sequences near the *α*-globin genes, which become hypomethylated in all of the analysed humanised mouse tissues, including erythroid and non-erythroid tissues. Importantly, hypomethylation of the human *α*-globin genes is unlikely to be caused by a change in the chromosomal location, as demonstrated by the appearance of hypomethylation at the human *α*-globin genes when a whole copy of human chromosome 16 is placed in a mouse cell (Figure 7).

With the most recent genome-wide DNA methylation surveys it is becoming increasingly apparent that methylation changes are much more dynamic than initially anticipated. DNA methylation changes occur during development in brain, erythropoiesis, as well as under different culturing conditions (Blaschke et al, 2013; Ficz et al, 2013; Lister et al, 2013; Shearstone et al, 2011). Here, hypomethylation of human DNA in the mouse environment was, however, consistently found in every tissue analysed (Figure 6) and highly similar patterns were described in different ESC clones rendering it unlikely that the observed changes are due to developmental stage-specific differences. In conclusion, the methylation changes observed here likely stem from changes in the *trans*-acting milieu. In line with this assertion, differences in the expression of *trans*-acting factors in human and mouse are documented (Sankaran et al, 2009).

3.3.2 Establishment and maintenance of methylation patterns in the humanised mouse model

Interestingly, hypomethylation was already seen in humanised mouse versus human ESCs, but becomes exacerbated in somatic cells, implying that the mouse DNA methylation machinery does not properly establish methylation patterns in ESCs and furthermore fails to maintain the established patterns in somatic cells. In ESCs, the majority of hypomethylated CpG sites cluster in CGI shores. A similar observation was made in differentiated cells, although hypomethylation spreads further away from CGIs, thereby, affecting the whole region around the *α*-globin genes. Interestingly, the spread of hypomethylation at the *α*-globin genes in humanised mouse versus human somatic cells is reversed in humanised mouse ESCs *in vivo* after an epigenetic reprogramming, suggesting not only that the ESC clones truly reflect the *in vivo* DNA methylation pattern, but also that methylation at the human *α*-globin genes is dynamically regulated during development in the humanised mouse model.

3.3.3 Differential CGI shore methylation

Previous studies report changes in the width of hypomethylated regions in different cell types suggesting cell type-specific regulation of the extent of hypomethylated regions (Hodges et al, 2011; Molaro et al, 2011). Dynamic regulation of the extent of hypomethylated regions during development may explain cell type-dependent differential methylation at CGI shores (Hodges et al, 2011). Here, differential methylation at CGI shores was observed at the *α*-globin-associated CGIs when the same DNA was analysed in a heterologous murine

environment and compared to its natural human environment. This finding was confirmed for the *NPRL3*-associated CGI in human versus humanised mouse ESCs, which is structurally closer to other genes genome-wide. Whether or not this is a general phenomenon observed when any type of human DNA is placed in the mouse remains will be addressed in chapter 5.

3.3.4 DNA methylation analysis by bisulfite sequencing and REA

Here, two different approaches to study DNA methylation were employed, bisulfite sequencing and restriction enzyme analysis. Both methods yield reproducible results underpinning that the findings are not the result of technical biases. Hypomethylation of human *α*-globin cluster in the humanised mouse model was furthermore confirmed using a third approach, matrix-assisted laser desorption/ionization time of flight mass spectrometry (MALDI-TOF MS), carried out previously by Supachai Ekwattanakit (Ekwattanakit). Of note, however, none of these methods is able to distinguish between hydroxymethylation and 5-methylcytosine (Huang et al, 2010) (add reference for restriction enzymes). Given the most recent reports of a relatively high abundance of 5hmC in ESCs and brain (Ficz et al, 2011; Kriaucionis & Heintz, 2009; Lister et al, 2013), a caveat must be added to the data, in that it is not possible to make out whether the observed loss in methylation is due to a loss of 5hmC or 5mC in the humanised mouse model versus human cells. In light of its proposed role as a stable mark with distinct binding properties to proteins on one hand, and with regard of its potential role as an intermediate in an active demethylation pathway on the other, it would be interesting to test whether the changes observed at CGI

shores stem from differences in 5hmC or 5mC. Distinguishing between a difference in 5hmC and 5mC might help shed light on possible mechanisms by which CGI shore methylation might be regulated and how this might influence transcription. Methods to discriminate between the two modifications have been developed (for review see (Song et al, 2012a)).

Another interesting feature that genome-wide methylation studies recently unveiled is a high prevalence of non-CpG methylation in ESCs and brain (Lister et al, 2013; Lister et al, 2009). Although the data generated by TBS allows to analyse methylation in the non-CpG context, this study has focussed on investigating methylation differences between human and humanised mouse only in the CpG context. Nevertheless, it would be interesting to test whether the observed differences coincide with changes in methylation in the non-CpG context. As non-CpG methylation is thought to mainly mediated by the activity the *de novo* methyltransferases (see section 1.4.3), investigating non-CpG methylation in human and humanised mouse may reveal important mechanistic insights into the molecular principles involved in methylating CGI shores.

Of note, given the high prevalence of non-CpG methylation it is noteworthy that data generated by bisulfite sequencing using specific primers that match bisulfite-converted DNA is biased towards amplifying DNA that contains low amounts of non-CpG methylation. This has to do with the design of the primers match *in silico* converted genomic DNA (i.e. all Cs in the non-CpG context are converted *in silico* to Ts). Hence, DNA with high levels of non-CpG methylation will by large not be amplified. Importantly, however, this bias does not seem to affect the

obtained result of CpG methylation pattern as the TBS approach, in which random primers are used for PCR amplification (hence no bias against non-CpG methylation) results in a very similar CpG methylation pattern.

Finally, on a technical note, it is worth mentioning that the bisulfite sequencing data is strand-biased. In TBS, capture was performed on the top strand, so that the methylation data reflects the pattern on the top strand, whereas the specific PCR amplification of the 20kb region spanning the human α-globin genes and the 5kb region across the *NPRL3* gene interrogate methylation levels of the bottom strand. Both datasets generate similar patterns, suggesting that there is no difference in the methylation pattern of top and bottom strand.

3.3.5 Gene body methylation in human ESCs

A number of studies have reported elevated levels of methylation at the bodies of active genes (Ball et al, 2009; Laurent et al, 2010; Lister et al, 2009; Rauch et al, 2009). In line with this, the *NPRL3* gene, a ubiquitously expressed gene, is highly methylated in ESCs and brain of human and humanised mouse as measured by bisulfite sequencing (by TBS) (Figure 16). The *α*-globin genes, in contrast, are largely unmethylated in most cell types irrespective of expression and this is likely to be due to the fact that constitute a special class of genes in which the associated CGIs spans most of the gene. Interestingly, however, in human ESCs, the gene bodies display slightly elevated levels of methylation, which disappears during normal development. Both expressing and non-expressing human cells (erythroid and brain) lack gene body methylation. In *Arabidposis*, the

methylation levels were highest at genes with intermediate levels of transcription (Zilberman et al, 2007). A similar observation was made in mammals (Maunakea et al, 2010). This may provide a possible explanation for the observed pattern of methylation at gene bodies during normal human development: in non-expressing and highly expressing cells *α*-globin gene bodies are unmethylated, while gene body methylation is elevated levels human ESCs, where basal transcription levels have been documented (De Gobbi et al, 2011). In this respect, it would be interesting to test whether basal transcription levels are found at the human *α*-globin genes in ESCs of the humanised mouse model, which lack human *α*-globin body methylation. Whether or not this correlation may any biological meaning is very hypothetical and remains to be verified.

3.3.6 Methylation of repetitive sequences

To my knowledge, at the start of this project, hypomethylation of a human transgene in the mouse has not been documented before. A recent study, however, report that in a mouse model, which carries a complete copy of hs21 (Tc1 mouse model), a number of human loci become activated in the mouse, and these activated regions are enriched in repetitive elements, specifically in LTR and LINE elements (Ward et al, 2013). Activation of these elements is concomitant with depletion in DNA methylation at repeat sequences. Activated DNA methylation-depleted regions were shown to be enriched in TE-derived repetitive elements, specifically in LTR and LINE elements. Although the data presented in this study suggests that hypomethylation is not specifically found at repetitive elements when the human *α*-globin cluster is placed in the mouse, this

seeming disparity may be due to differences in the cell types analysed, the size of the region analysed, or the repeat content analysed.

4 Dissecting out the role of CGI H in determining the methylation state of its adjacent shores

4.1 Introduction

In the previous chapter it was shown that methylation of a human transgenic DNA fragment of 117kb placed in the mouse is largely preserved. Some regions, however, do become hypomethylated in the transgenic mouse model, particularly around the *α*-globin genes, and this seems to be associated with a wider spread of the unmethylated region of CGIs resulting in CGI shore hypomethylation. A key outstanding question concerns the locality of the sequence cues that determine hypomethylation of CGI shores. Are they encoded locally within the hypomethylated DNA sequence or are they distally located? It has already been shown that the human *α*-globin genes become hypomethylated even when the whole of human chromosome 16 is placed in the mouse (Figure 7), showing that the propensity for hypomethylation of the human sequences is intrinsic to the human sequence and not dependent on mouse sequences in *cis*. This chapter dissects out the human *cis*-regulatory sequences that may be involved in determining the methylation state of CGI shores.

Two possible scenarios are likely: (1) the DNA sequence of a CGI shore may autonomously determine its hypomethylated state or (2) sequences distal to the shores may be responsible for hypomethylation of the given shore. In support of the former, promoter sequences that determine their hypomethylated state independent of their surrounding sequence were

identified (Lienert et al, 2011). In support of the latter, hypomethylation of DNA can be induced via spreading of hypomethylation from an unmethylated promoter sequence into a methylated DNA fragment (Lienert et al, 2011). This raises the question whether hypomethylation of CGI shores results from spreading of hypomethylation from the unmethylated promoter CGI into the shore, or whether hypomethylation is, in fact, independent of the surrounding DNA and intrinsic to the DNA sequence of the CGI shore itself.

To test these two possibilities, an exemplary CGI shore, namely the shore adjacent to CGI H (associated with *α*2), was re-integrated into the mouse genome in the presence or absence of its adjacent CGI H. This was done by employing an RMCE system as a reproducible system for the integration of DNA fragments into the same genomic locus. Furthermore, by analysing fragments with or without the promoter the relative contribution of the *α*2 promoter in regulating CGI shore hypomethylation will be evaluated. Narrowing down the locality of the sequence cues responsible for the hypomethylated state of CGI shores in mouse versus human may shed light into the cellular factors that are involved in normally regulating CGI shore methylation.

4.2 Results

4.2.1 Choice of integration site

To exclude biases due to the surrounding chromatin state and a difference in the copy number of transgenes that can affect the DNA methylation state of transgenes (Garrick et al, 1998; Robertson et al, 1995), it was

desirable to integrate single copies of the selected test sequences into the same genomic locus in the mouse. To achieve this, an RMCE system was chosen as a tool, because it enables the rapid and reproducible integration of test sequences into the same target locus[15]. An RMCE system employs recombination sites for bacterial site-specific recombinases, which are targeted to a specific genomic locus of interest via conventional homologous recombination. A test sequence is first integrated into a donor plasmid in such a way that identical recombination sites will flank it. Co-transfection of the donor plasmid and an appropriate recombinase expression construct enables integration of the test sequence into the targeted genomic locus.

A previously described RMCE system was chosen to integrate all test fragments into the same mouse genomic locus, the mouse *α*-globin locus on chromosome 11 (32194007-32203113, mm9) (Lynch et al, 2012). For more details see section 2.6 and Figure 3). Importantly, the RMCE cassette is located in a region of open chromatin (as judged by the absence of H3K27 trimethylation marks near the integration site (Lynch)) and does not contain CpG Islands nearby (Figure 24).

To verify that there are no unmethylated CGIs nearby, which might confer a hypomethylated state of the test sequences DNA methylation levels were measured at the mouse DNA flanking the RMCE cassette by bisulfite sequencing in wildtype (WT) mouse ESCs and targeted mouse ESCs (mouse ES RMCE, kindly provided by Magnus Lynch). A 1kb region

[15]Alternatively, genomic integration of transgenes can be achieved by random genomic integration, which is a fast and relatively easy approach, but unspecific regarding integration site and transgene copy number, or, by conventional homologous recombination, which permits integration of single copies of transgenes into a defined genomic locus, but is confounded by a great variability in targeting frequency.

comprising 6 CpG sites located at the 5' site of the integration site, as well as a 500 bp region containing 3 CpG sites on the 3' site of the integration locus were analysed. The location of the amplimers used to analyse the methylation state at the mouse DNA is indicated in Figure 24.

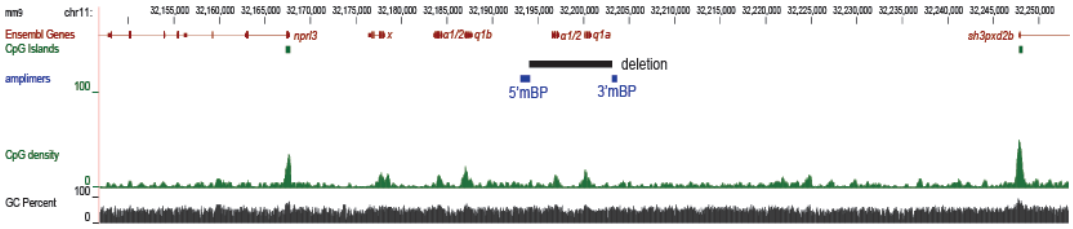

Figure 24: Mouse *α*-globin locus showing the site of integration of the human DNA fragments.
Test sequences were integrated into the same genomic locus in the mouse *α*-globin locus to replace the deletion indicated by a black bar (chr11: 32,194,007-32,203,113, mm9). The location of the amplimers 5'mBP and 3'mBP used to test the methylation level of the mouse DNA flanking the RMCE cassette is shown in blue. Genes are shown in red, CGIs in green. (The snapshot shown here comprises the region chr11: 32,146,881-32,253,252, mm9).

High methylation levels were found on both the 5' and the 3' site of the mouse DNA flanking the RMCE cassette (Figure 25) in the mES RMCE cell line similar to the levels in mouse wildtype ES E14 cells (mES WT) implying that the genetic manipulation of the cell line does not influence DNA methylation levels near the integration site.



Figure 25: DNA methylation levels at the mouse DNA flanking the insertion site in wildtype mouse ES (mES WT) and in the targeted mouse ES cell line (mES RMCE).
DNA methylation levels of amplicons were analysed by bisulfite sequencing on either site of the RMCE cassette. DNA methylation levels at individual CpG sites (represented by circles) are color-coded. The location of the primers 5'mBP (5' site of the mouse breakpoint) and 3'mBP (3' site of the mouse breakpoint) are indicated in Figure 24.

4.2.2 The mouse α-globin locus does not induce hypomethylation of a methylated fragment

Next, it was important to verify that the mouse *α*-globin locus does not induce aberrant hypomethylation of the integrated test sequences. Hence, a methylated DNA fragment was inserted into the mouse *α*-globin locus via the RMCE system. The chosen fragment is 2kb in size, spans the region 101676-103710 on chromosome 16 (hg18) and was shown stay highly methylated in the humanised mouse model (Figure 8). The fragment, hereafter referred to as fragment A, was generated by PCR amplification using primers that contain NotI restriction sites. The PCR product was gel purified and integrated into the pML7 donor plasmid (kindly provided by Magnus Lynch) via the NotI site to generate the pML7-A vector (Figure 26). Successful generation of the pML7-A donor plasmid was verified by gel electrophoresis of NotI digested pML7-A as well as sequencing of the insert to verify correct sequence and orientation of fragment A in the vector.

The pML7-A donor plasmid was transfected into mES RMCE cells along with a vector containing the Cre recombinase (Figure 27). Positive clones were selected in HAT medium. Single clones were picked and transfected with a vector containing Flp recombinase to allow excision of the *HPRT* marker. *HPRT* negative clones in which the *HPRT* were selected in 6TG containing medium, which is toxic to cells expressing a functional *HPRT* gene (Figure 27).

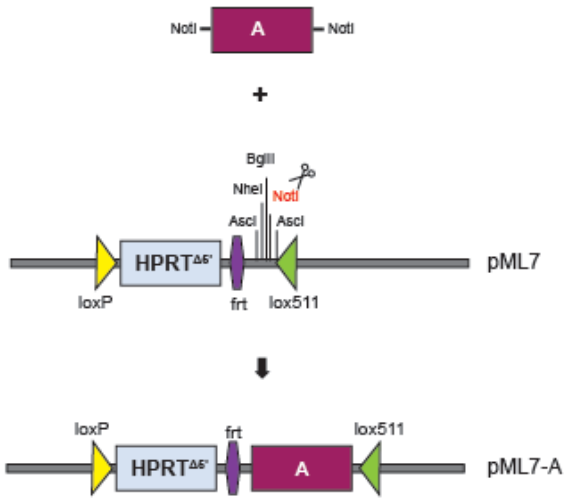

Figure 26: Integration of fragment A into the pML7 vector to create the pML7-A donor plasmid.
Fragment A was generated by PCR amplification using primers containing NotI restriction sites. Fragment A and pML7 were cleaved with NotI and ligated to create pML7-A. Adapted from (Lynch et al, 2012).

Three clones of mES RMCE cells containing fragment A were chosen for analysis of methylation levels by bisulfite conversion (A5, A7, A11). Six amplimers (on average 250 bp in length) to span the fragment (shown in blue in Figure 28) were designed to interrogate methylation levels by high-throughput sequencing (Illumina). As can be seen in Figure 28 the DNA methylation levels of this fragment stays high even in the context of this smaller fragment in the new locus and are comparable to the level observed in humanised mouse ESCs, suggesting that the mouse genomic locus does not reduce methylation levels of inserted exogenous DNA. Taken together, these data suggest that the mouse *α*-globin locus does not significantly alter DNA methylation levels of an integrated human test sequence and hence, constitutes a suitable site to test whether methylation of integrated test fragments is sequence- or context-dependent.

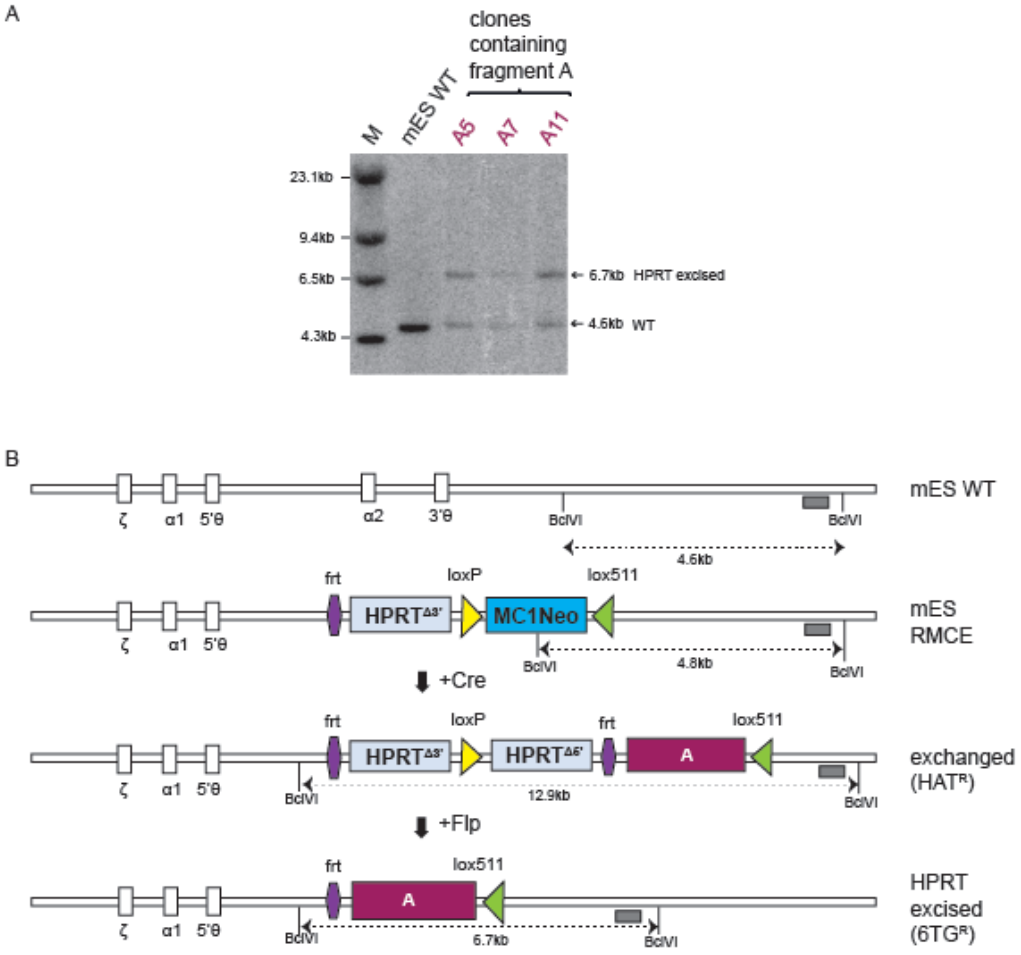

Figure 27: Confirmation of integration of control fragment A.

A) Southern blot showing integration of fragment A and excision of *HPRT* marker gene. B) Schematic diagram showing the expected band sizes in A.

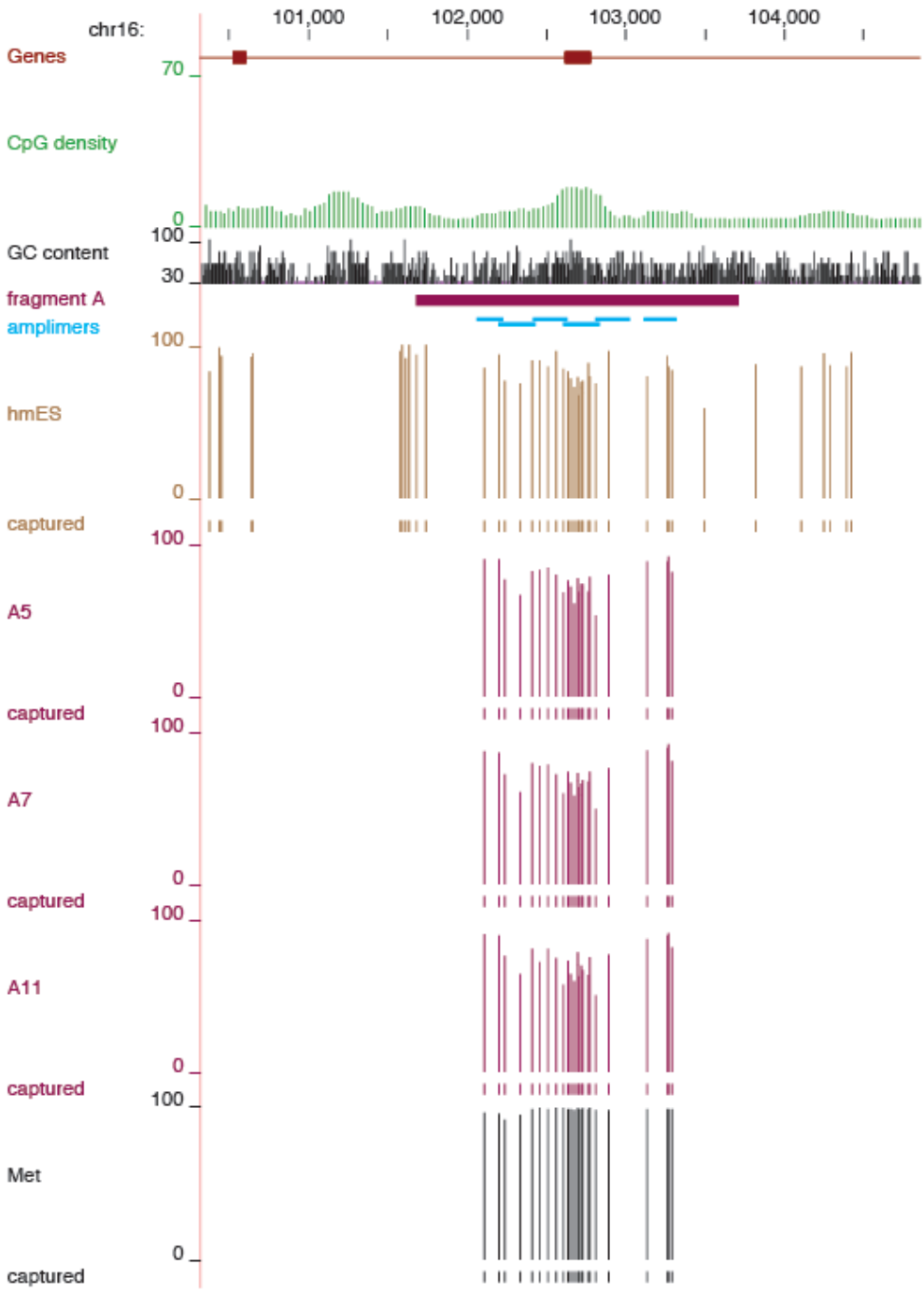

Figure 28: DNA methylation levels of a naturally methylated fragment retains its methylation levels when integrated into the *α*-globin locus.

Methylation levels were analysed by bisulfite sequencing. The methylated control sample (Met) was used to verify that methylcytosine is protected from conversion after bisulfite treatment.

4.2.3 Choice of test sequences

Having established that the mouse *α*-globin locus is a suitable site to test whether methylation of DNA sequences is sequence- or context-dependent, the next aim was to dissect out the *cis*-regulatory sequences determining CGI shore hypomethylation. CGI H (associated with *α*2) was chosen as a good representative of an extended CGI (Figure 13). Several fragments containing this CGI (or parts thereof) and its adjacent CGI shore were selected for analysis. The location of the fragments are shown in Figure 29.

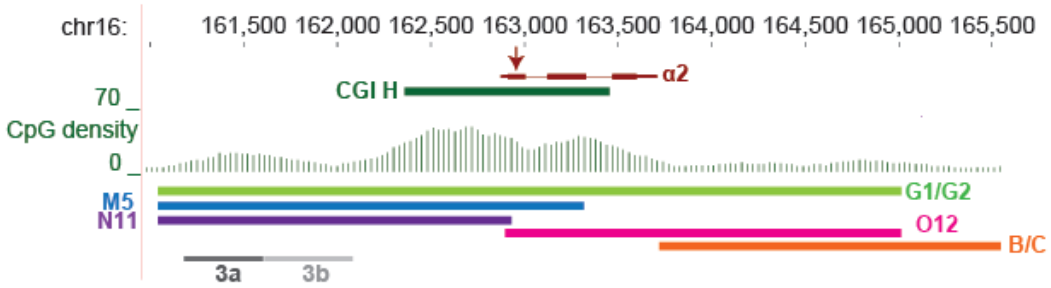

Figure 29: Location of the human test fragments targeted to the mouse *α*-globin locus.
The DNA fragments comprise CGI H associated with the human *α*-globin gene *α*2 and the surrounding sequence. PCR amplimers 3a and b used to interrogate DNA methylation levels of these fragments in Figure 32 are indicated as grey bars.

Fragment G1/G2 comprises all of CGI H and ~1kb and ~1.5 kb on either side (upstream and downstream, respectively) of the CGI. Fragment G1/G2 is divided into two fragments, one of which bears the 5' part of CGI H and 1kb upstream of it (N11), the other one contains the 3' part of CGI H and the 1.5kb downstream region (O12). Fragment M5 is slightly longer than N11 and contains part of *α*2 including its promoter element. Fragment B/C comprises a 2kb region downstream of CGI H and is completely missing CGI H and *α*2. Cell lines bearing the fragments G1/G2, M5, N11

and O12 were a kind gift of Magnus Lynch (Lynch; Lynch et al, 2012). The cell line containing fragment B/C was generated in this study.

To generate fragment B/C, the region of interest was amplified by PCR using primers that contain NotI restriction sites. The PCR product was gel purified, digested with NotI and integrated into the NotI-digested pML7 donor plasmid (kindly provided by Magnus Lynch) to generate the pML7-B/C vector (Figure 30).

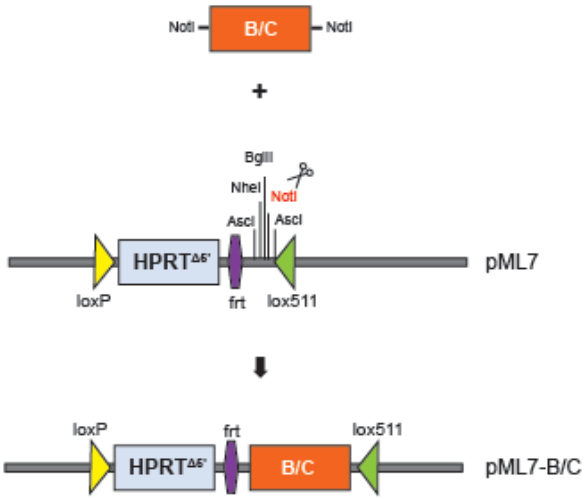

Figure 30: Schematic showing the generation of the pML7-B/C donor plasmid.
The test sequence B/C, generated by PCR amplification using primers containing NotI restriction sites, was digested with NotI and integrated into the NotI-digested pML7 donor plasmid to generate the pML-B/C donor plasmid.

Successful integration into the vector was tested in single clones by restriction digest with NotI and the insert was sequenced to verify correct orientation of the fragments in the vector and to verify that no mutations were introduced during PCR amplification of the fragment.

The pML7-B/C plasmid bearing the correct sequence and orientation of the insert were transfected into mES RMCE cells along with a vector containing the Cre recombinase and positive clones were selected in HAT

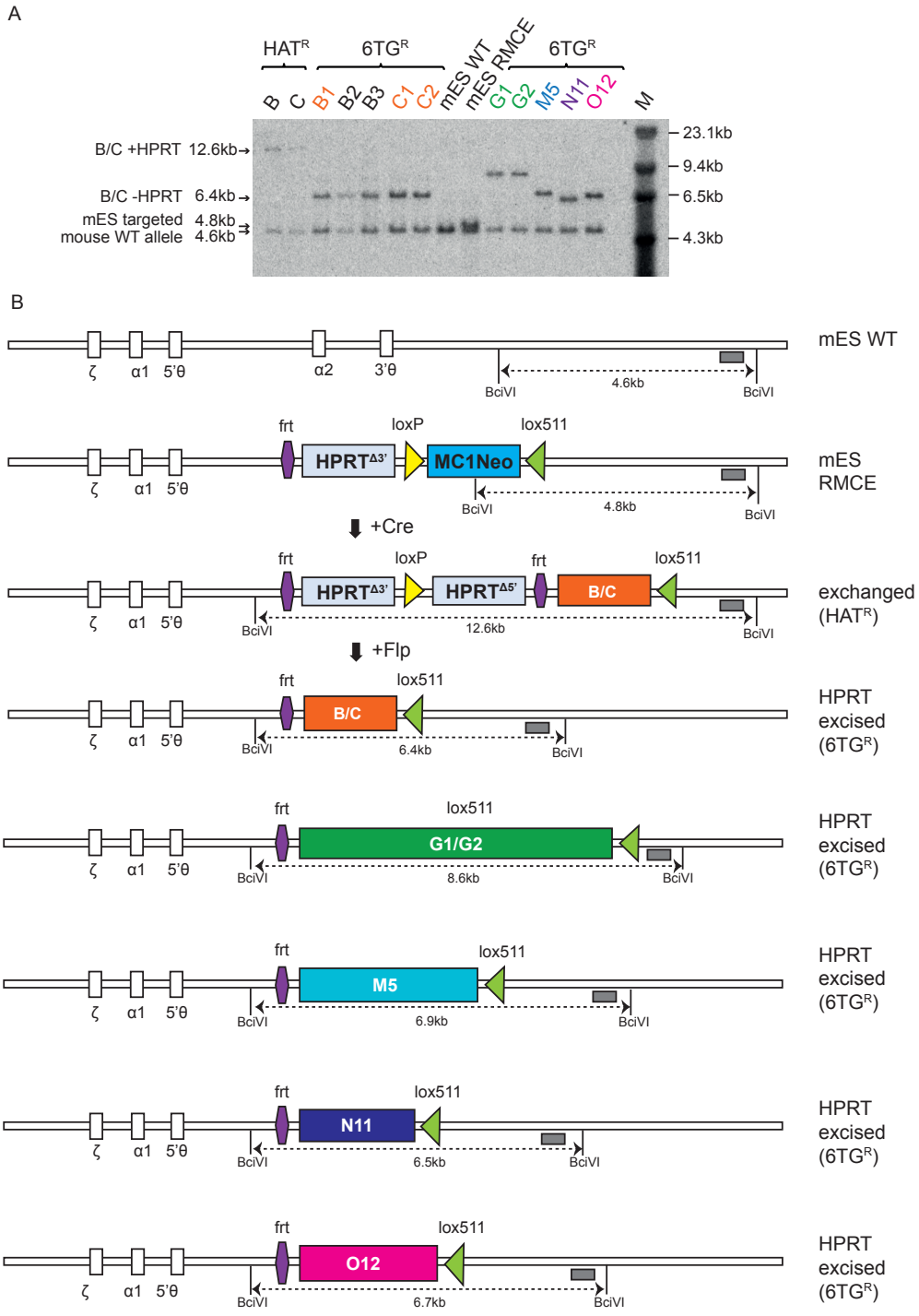

Figure 31: Verification of correct targeting of the selected DNA fragments containing human CGI H and surrounding sequence to the mouse *α*-globin locus.

A) Southern blot of BciVI digested DNA probed with probe 11+12 (indicated in b as grey box) confirming the predicted band sizes. B) Schematic diagram of the predicted band sizes upon digestion with BciVI and probe 11+ 12 (indicated as a grey box) in mES WT, mES RMCE and in the cell lines targeted with the selected human test fragments B/C, G1/G2, M5, N11, and O12.

medium. Single positive clones were picked and the *HPRT* marker was excised. Clones in which the *HPRT* gene was successfully excised were selected in 6TG containing medium, which is toxic to cells expressing a

functional *HPRT* gene. Correct integration of the cell lines containing the selected human test sequences was validated by Southern blot (Figure 31).

4.2.4 DNA methylation levels of the inserted DNA fragments

To get a first glance of DNA methylation of the integrated fragments, methylation levels were first measured at two selected amplimers located in the shore upstream of CGI H (Figure 29). Two clones containing the 4kb region including CGI H and the shores on both sides (G1/G2) both become hypomethylated in the mouse *α*-globin locus (Figure 32). Similarly, a truncated version of this fragment (N11) comprising only the 5' half of CGI H and its adjacent sequence also becomes hypomethylated in the mouse locus. Hence, this data suggests that hypomethylation of shores may be intrinsic to the sequence of the shore and/or the 5' half of CGI H (fragment N11). Of note, fragment N11 does not comprise the promoter element of *α*-globin 2, indicating that hypomethylation is independent of transcription from the *α*-globin gene promoter.

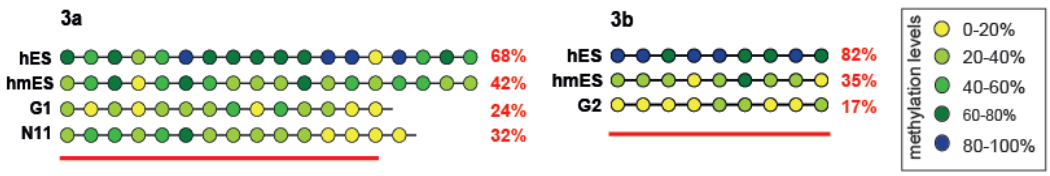

Figure 32: DNA methylation levels in selected human DNA fragments targeted to the mouse genome.
DNA methylation levels were measured by bisulfite sequencing at amplimers 3a and 3b. The average methylation level of the amplimers (analysed CpG sites are underlined in red) are indicated. The location of amplimers 3a and 3b is indicated in Figure 29.

To get an overview of DNA methylation levels across all fragments, DNA methylation was next analysed by bisulfite high-throughput sequencing using the amplimers designed in chapter 3.2.4.5. *In vitro* methylated

genomic DNA was processed in parallel to demonstrate that methylated Cs are, indeed, protected from conversion after treatment of genomic DNA with sodium bisulfite (Figure 33). Figure 34 shows an overview of the DNA methylation pattern of the analysed fragments in UCSC genome browser in comparison to the methylation patterns in human and humanised mouse ESCs. For each sample, a track showing the methylation difference in comparison to that of human ESCs is shown (negative values represent hypomethylated sites in the inserted fragment compared to human ESCs). The high resolution profile indicates that the 4kb fragment (G1 and G2, both green in Figure 34) is largely sufficient to establish normal human patterns, except for the upstream CGI shore and the *α*2 gene body, which are hypomethylated in this 4kb fragment mimicking the scenario observed in humanised mouse ESCs. From this data it can be concluded that a 4kb fragment comprising CGI H and its adjacent shores contains the signal for hypomethylation of the upstream CGI shore as well as the *α*-globin gene body.

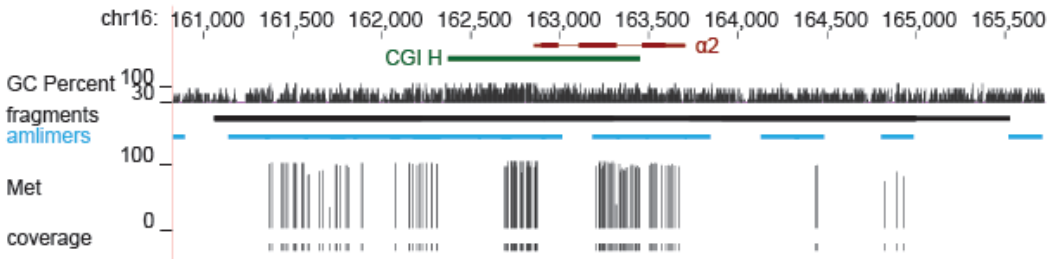

Figure 33: Bisulfite sequencing of a methylated control sample.
In vitro methylated human genomic DNA was converted in parallel with the fragments shown in Figure 34 as a control.

To evaluate the relative contribution of CGI H on the methylation state of the surrounding sequences, truncated versions of the 4kb fragment containing parts or none of CGI H and the adjacent region were analysed

(fragments M5, N11, O12, and B1, C1, C2). Both 2kb fragments containing the upstream shore and the 5' parts of CGI H stay hypomethylated (fragments M5 in purple and N11 in blue in Figure 34), suggesting that the 3' half of CGI H, including the promoter element of *α*2, is not involved in creating a hypomethylated state of the upstream CGI shore. Fragment O12 comprising the other (3') end of CGI H, including the *α*2 gene body, is hypomethylated (at the body of the gene), indicating that hypomethylation of the gene body is determined by 3' end of CGI H and/or its shore. Hypomethylation is unlikely to be due to overconversion of the samples (i.e. conversion of all cytosines, including methylated cytosines into uracil) as shown by a protection of methylated Cs in the *in vitro* methylated control sample (Figure 33). For fragments B1, C1, and C2, which lack all of CGI H, firm conclusions cannot be made due to gaps in the dataset. The gaps are likely to be due to loss of PCR amplimers during library generation.

Taken together, the data presented here suggests that the 2-4kb human DNA fragments comprising all or parts of CGI H and its adjacent shore become hypomethylated when re-integrated into the mouse *α*-globin locus indicating that the appearance of hypomethylation of human DNA in the mouse is determined by the underlying DNA sequence of the inserted DNA fragments. Hypomethylation is unlikely to result from genetic manipulation of the mouse locus as methylation of the mouse DNA flanking the inserted fragments remains unaltered after integration (Figure 35).

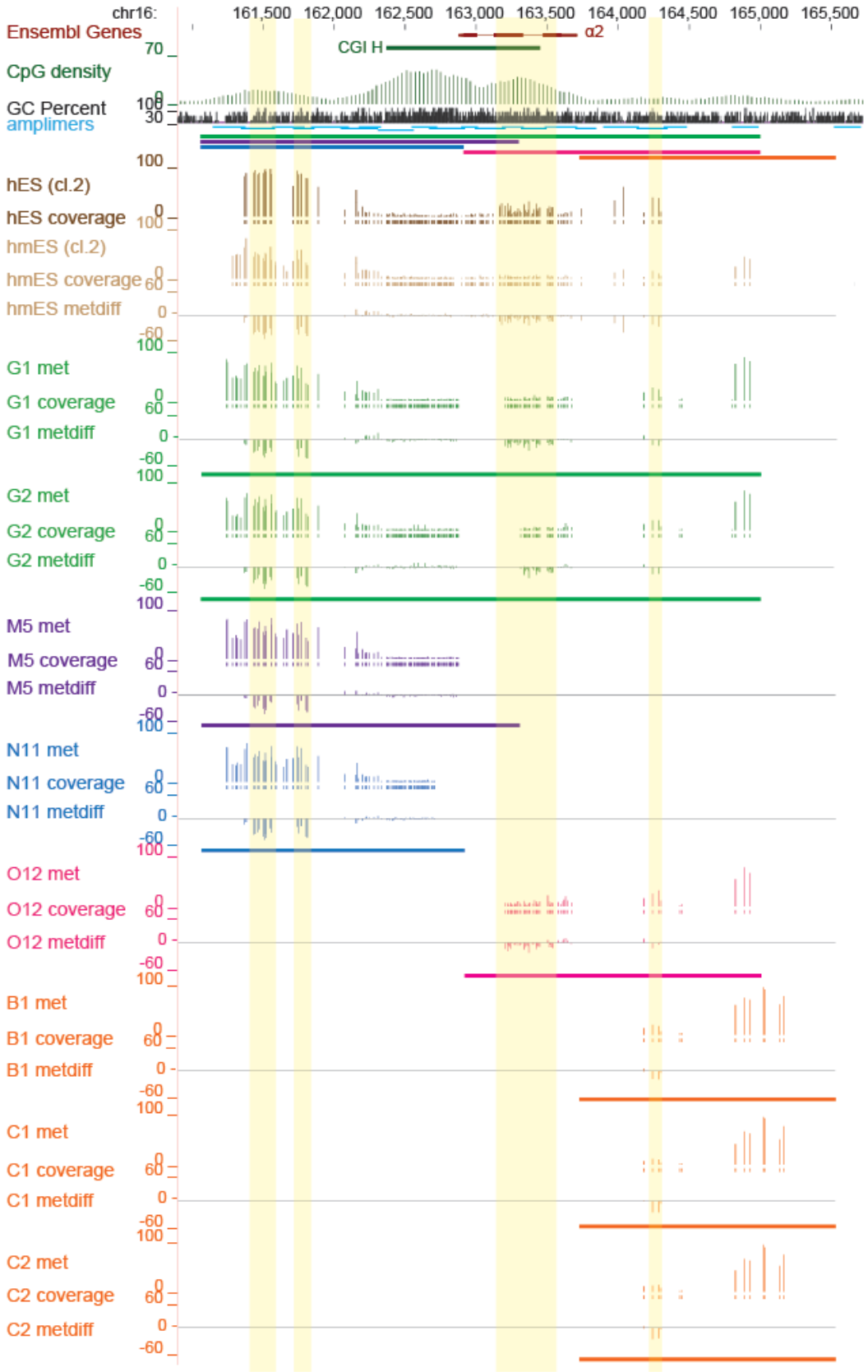

Figure 34: DNA methylation analysis at CGI neighboring regions.

The *α2* gene is shown at the top in red. The location of the different fragments are indicated as coloured bars and the PCR amplimers used are shown above in blue. DNA methylation levels ranging from 0 to 100% are shown for each fragment. DNA methylation levels were calculated for CpG sites that were covered by at least 100 reads. The position of each analysed CpG site is indicated (coverage).

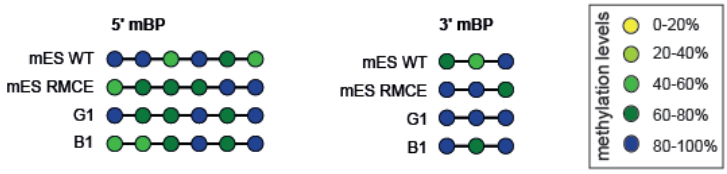

Figure 35: DNA methylation levels at the mouse DNA flanking the insertion site.
Methylation levels at individual CpG sites (circles) are color-coded. A 1kb region containing 6 CpG sites were analysed on the 5' site of the mouse breakpoint (5'mBP). On the 3' site of the mouse breakpoint (3'mBP) a 500bp region comprising 3 CpG sites screened for methylation levels. Methylation levels were analysed in the clones G1 and B1 and compared to the levels in wildtype mouse **ESCs** (mES WT). The methylation level in targeted m**ouse ESCs** (mES RMCE) are also shown.

4.3 Summary and Conclusions

In this chapter an RMCE-based approach was employed to narrow down the *cis*-regulatory sequence responsible for a hypomethylated state by re-integrating human test sequences into the mouse genome and evaluating their methylation state. To eliminate possible biases arising from the surrounding genomic locus all test sequences were integrated into the same genomic location in the mouse genome. The mouse *α*-globin locus was chosen as the target site because it is deprived of unmethylated CGIs nearby, which may induce a hypomethylated state of integrated test sequences. This was verified by bisulfite sequencing of a 2kb region surrounding the integration site confirming the expected high methylation levels in the mouse *α*-globin locus. Furthermore, it was demonstrated that a normally methylated human test fragment re-establishes high methylation levels at the mouse *α*-globin locus confirming that the mouse locus does not induce hypomethylation of the integrated test sequences.

4.3.1 Sequence determinants of DNA methylation around CGIs

To identify the sequences responsible for a hypomethylated state of human DNA in the mouse, 2-4kb test fragments were integrated into the mouse *α*-globin locus. These test sequences consist of human CGI H or parts thereof and its adjacent upstream and downstream shores. It was found that a 4kb fragment consisting of CGI H and its flanking shores (fragments G1, G2) became hypomethylated when integrated into the mouse *α*-globin locus, suggesting that the signal for hypomethylation of the test fragments is contained within the 4kb fragment. Likewise, 2kb truncated versions of the fragment consisting of parts of CGI H and its flanking DNA became hypomethylated upon re-integration into the mouse locus. These data suggest that the appearance of hypomethylation is determined by the underlying human DNA sequence in the test fragments.

4.3.2 Role of CGI H in determining the hypomethylated state of its flanking DNA

Differential methylation of CGI shores has been observed in different cell types as well as in cancer (Hansen et al, 2011; Irizarry et al, 2009). Another study described methylation changes at CGI shores as “tidal” methylation at CGI shores, which seem to result from a cell type-specific changes in the width of hypomethylated regions (HMRs), a characteristic feature of many promoter HMRs (Hodges et al, 2011). How the width of hypomethylated regions may be regulated is currently unknown.

Here, two possible models for the regulation of methylation around CGIs were tested. The first model proposes that differential shore methylation

may result from differential spreading of hypomethylation from a CGI into its shore. In an alternative scenario, hypomethylation may be intrinsic to the CGI shore. The data presented here shows that ~2kb DNA fragments are sufficient to reproduce hypomethylation of CGI shores, suggesting that they contain the necessary *cis*-regulatory sequence determinants to induce hypomethylation of the upstream shore of CGI H. Although it remains to be determined whether the complete lack of CGI H can restore methylation levels of a CGI shore[16], narrowing down possible sequence determinants to a region of 2kb may help to elucidate possible mechanisms that methylate human CGI shores in a normal human cell.

4.3.3 Role of the α2 promoter in initiating a hypomethylated state

Of note, whilst fragments G1, G2, M5 and O12 all contain the promoter element of *α*2, fragment N11 does not. Yet, despite the absence of the promoter sequence of *α*2 in fragment N11 the upstream CGI shore remained hypomethylated, suggesting that the human *α*-globin promoter and hence transcription from that promoter do not play a role in determining the methylation state of the surrounding DNA. This is in line with the study carried out by Lienert et al mentioned above, where it was shown that the establishment of hypomethylation at promoter MDRs was independent of measurable transcription of the associated gene (Lienert et al, 2011).

[16] Due to time constraints the analysis of fragment B/C (as well as another fragment covering only the upstream shore – not shown here) could not be finished, but is planned in the future. Although coupling bisulfite conversion with high-throughput sequencing is highly valuable in order to accelerate the generation of methylation data it has proven to be cause loss of data for some PCR products. Gaps will be filled by conventional bisulfite sequencing of subcloned PCR products.

5 Non-methylated DNA profile at human CGIs in mouse versus human

5.1 Introduction

Results presented in this thesis suggest that CGIs in the human *α*-globin cluster become extended in the humanised mouse model compared to its natural environment in a human cell. A key open question is whether this phenomenon is restricted to the human *α*-globin genes or whether it is a general phenomenon observed at other human CGIs when placed in the mouse. To address this question, in this chapter CGIs across a whole human chromosome placed in the mouse will be analysed and compared to their state in a human cell. To do this, a transchromosomal mouse model will be employed which carries a whole human chromosome in addition to the mouse genome.

A transchromosomal mouse model, which bears a whole copy of hs21 was generated in a previous study to recapitulate Down's syndrome, a genetic disorder which arises from aberrant gene dosage effects due to three copies of hs21 (trisomy 21) (O'Doherty et al, 2005). The generation of this mouse model (Tc1) was achieved by fusing microcells containing single chromosomes with ESCs. In this way, ES cell lines were generated with which it was possible to create viable mouse strains bearing an almost complete copy of freely segregating hs21 (42Mb of 46.9 Mb) in addition to the mouse genome. Retention of hs21 in somatic cells (brain, spinal chord and spleen) of this mouse strain was confirmed, although the chromosome tends to be lost and was only present in ~25-70% of nuclei (depending on the tissue)

(O'Doherty et al, 2005). Aspects of Down syndrome were recapitulated in this mouse strain (O'Doherty et al, 2005).

This mouse model was used in an ensuing study to dissect out the role of *cis*- and *trans*-acting factors in determining the binding of the mouse hepatocyte nuclear factors HNF1α, HNF4α, and HNF6 to hs21 as well as its mouse orthologous chromosome 16 in hepatocytes. It was found that most binding events on hs21 in Tc1 were similar to the patterns observed on hs21 in normal human hepatocytes and different to the pattern of its murine orthologue, wherefore the authors concluded that the DNA sequence is the primary determinant of TF binding events rather than the TFs themselves. Interestingly, in a follow-up study, the mouse model was used again to specifically interrogate the non-coding portion of hs21 in Tc1 compared to human liver and it was found that many repetitive elements become activated as judged by active H3K4me3 chromatin marks and PolII binding. A number of these activated repetitive elements showed decreased methylation levels, suggesting that methylation patterns are improperly templated in Tc1 versus human liver (Ward et al, 2013). Methylation patterns of CGIs were not investigated.

The aim of this chapter is to interrogate the methylation profile around CGIs across hs21 in Tc1 and human to find out whether there is a general tendency of human CGIs outside of the human *α*-globin genes to become extended in the context of a mouse versus a human cell. This will be done by analysing the non-methylated DNA profile obtained by pull-down of unmethylated DNA by ZF-CXXC-affinity purification followed by high-throughput sequencing of

the captured material. It is anticipated that the spread of the reads across a CGI may reflect the width of the unmethylated region around that CGI. Although firm conclusions cannot be made, the data presented in this chapter may provide a basis for future studies to analyse the methylation state of hs21 in Tc1 and human.

5.2 Results

5.2.1 Choice of method and cell types

Genome-wide non-methylated DNA profiles, which were generated using biotinylated CXXC affinity purification (Bio-CAP) were analysed. BioCAP employs the ZF-CXXC domain of human KDM2B to capture unmethylated DNA. Subsequent high-throughput sequencing of the captured material generates peaks of sequence reads representing the non-methylated DNA fraction genome-wide. Two biological replicates of each of human liver and testes, as well as an input sample were chosen for analysis. These were derived from publicly available data (Long et al, 2013b) and compared to two biological replicates of each of Tc1 liver, testes and input DNA, kindly provided by Hannah Long and Rob Klose.

5.2.2 Mapping of human and Tc1 liver and testes datasets

Sequence reads of each sample were mapped against the human genome (hg19, Table 5, Table 6). Of the ~54-68 million reads in human liver, testes and input, 37-52 million reads map to the human genome, which corresponds to a mapping efficiency of 69% to 81% in human samples (Table 5). The majority of non-matching sequence reads arise from sequencing the adapters

in the library, which was verified by FastQC (data not shown, for more details on FastQC see section 7.1). Only sequence reads that map to hs21 were considered for further analysis, resulting in 712'977, 614'040, 561'304, 560'828, 704'787 reads that map to hs21, equivalent to 1.39%, 1.40%, 1.51%, 1.49, and 1.34% of all reads that map respectively to hg19 in human liver (two replicates), human testes (two replicates) and human input sample (Table 5). This is in line with the expected recovery of reads from chromosome 21 compared to the whole genome as judged by the percentage of CGIs on chromosome 21 in comparison to the whole genome (1.27%, Table 7).

In the six samples from Tc1 (two replicates each of liver and testes, as well as an input sample), a total number of 43 to 57 million reads were analysed. Of these 805,098, 607'804, 675'183, 651'075, and 405'364 map to the human genome, corresponding to a mapping efficiency of 1.41%, 1.12%, 1.29%, 1.24%, and 0.83% in in Tc1 liver (two replicates), Tc1 testes (two replicates) and an input sample, respectively (Table 6). This is comparable to the percentage of CGIs on hs21 compared to the rest of the genome (1.27%, Table 7). Of the reads that map to the whole human genome, 32%, 21%, 32%, 35%, and 53% map to hs21, indicating that a proportion of reads map to the human genome outside of hs21 (Table 6). This may be due to sequence similarities in human and mouse genomes resulting in mouse sequence reads matching to human DNA or reflect the presence of human DNA fragments other than hs21 in the mouse. Hence, it is noteworthy that a small fraction of sequence reads possibly arising from the mouse genome may erroneously map to the human genome, possibly confounding the analysis.

Table 5: Mapping efficiency of human liver, testes, and input.

Sequence reads were mapped to hg19 using bowtie. The table shows the total number of reads analysed, the number of reads that map to hg19, the mapping efficiency, the number of reads that match to hs21 and the percentage of reads that map to hs21 all reads that map to hg19.

hg 19	h liver 1	h liver 2	h testes 1	h testes 2	h input
total no of reads	68'094'501	58'619'926	53'637'386	55'766'604	64'886'867
mapped reads (hg19)	51'262'250	43'761'110	37'132'939	37'560'437	52'465'504
mapping efficiency (hg19)	75%	75%	69%	67%	81%
mapped reads (hs21)	712'977	614'040	561'304	560'828	704'787
reads on hs21/reads in hg19 [%]	1.39%	1.40%	1.51%	1.49%	1.34%

Table 6: Mapping efficiency of Tc1 liver, testes, and input.

Sequence reads were mapped to hg19 using bowtie. The table shows the total number of reads analysed, the number of reads that map to hg19, the mapping efficiency, the number of reads that match to hs21 and the percentage of reads that map to hs21 all reads that map to hg19. Subsampled read numbers in human and Tc1 liver or testes are highlighted in bold.

hg 19	Tc1 liver 1	Tc1 liver 2	Tc1 testes 1	Tc1 testes 2	Tc1 input 1
total reads	56'998'703	54'110'232	52'349'410	52'497'335	48'948'153
mapped reads (hg19)	805'098	607'804	675'183	651'075	405'364
mapping efficiency (hg19)	1.41%	1.12%	1.29%	1.24%	0.83%
mapped reads (hs21)	258'085	**127'642**	215'004	225'918	**214'583**
reads on hs21/reads in hg19 [%]	32.06%	21.00%	31.84%	34.7%	52.94%

Table 7: Number and size of CGIs genome-wide (hg19) and on hs21.
The total number of CGIs (total no.) is given for hg19 and hs21, as well as the percentage of CGIs on hs21 of all CGIs genome-wide (in parenthesis). Furthermore shown are minimum (min), maximum (max), average and median sizes of CGIs (in basepair, bp).

	hg19	hs21
total no.	28'691	365 (1.27%)
min size [bp]	201	202
max size [bp]	45'712	5'698
average size [bp]	761	735
median size [bp]	559	487

Successful mapping was verified in UCSC genome browser. An example locus depicting the mapped sequence reads from human and Tc1 testes is shown in Figure 36, exemplifying that the sequence reads in both human and Tc1 mouse correctly align to annotated CGIs. Furthermore, it becomes apparent that non-methylated DNA profiles are highly similar between biological replicates, suggesting that the generated Bio-CAP data is highly reproducible.

As differences in the number of reads between the samples may confound the results, the read numbers in all samples were equalized prior to downstream analyses. This was done by randomly removing excessive sequence reads in samples with more reads to the number of reads in the sample with the lowest number of reads ("subsampling"). In human and Tc1 liver, this brings the total number of reads that map to hs21 down to 127'642, which is the number of reads in sample Tc1 liver 2 (highlighted in bold in Table 5). In human and Tc1 testes 214'583 number of sequence reads were considered for further analysis, which is the number of reads in the Tc1 input sample (highlighted in bold in Table 5).

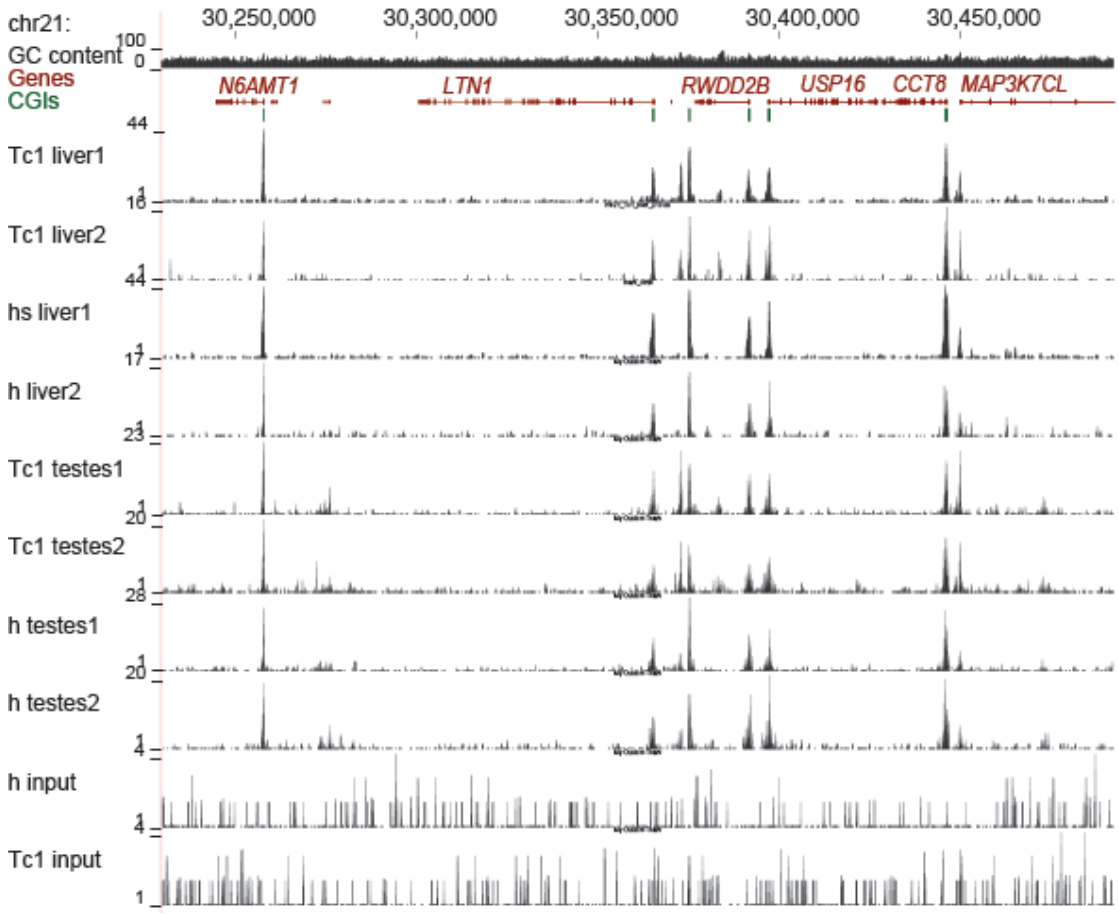

Figure 36: Example locus showing the mapped sequence reads in UCSC for human and Tc1 liver, testes and input DNA.

The region (chr21: 30'214'910-30'490'984, hg19) comprises the following genes (red): N-6 adenine-specific DNA methyltransferase 1 (N6AMT), LTN1 E3 ubiquitin protein ligase 1 (LTN1), RWD domain containing 2B (RWDD2B), ubiquitin specific peptidase 16 (USP16), chaperonin containing TCP1, subunit 8 (theta) (CCT8), MAP3K7 C-terminal like (MAP3K7CL). CGIs are shown in green.

5.2.3 Comparison of sequence reads across CGIs

To probe possible differences among non-methylated DNA profiles in Tc1 and human, the average read numbers around annotated CGIs was plotted chromosome-wide using a pile-up approach (Figure 37). This reveals a similar distribution of the sequence reads in a 6kb region around the midpoint of CGIs in Tc1 liver1 and liver2 compared to human liver 1 and liver2, as well as Tc1 testes1 and testes 2 compared to human testes1 and testes2. The similar distribution of the data is reflected by a similar standard deviation in all samples (Table 8, Table 9). Assuming that the distribution of the sequence

reads near CGIs reflects the extent of non-methylated DNA at CGIs on hs21, this data suggests that the width of unmethylated regions at CGIs may be similar in Tc1 compared to human. Yet, it is important to note that it is currently not known whether the method is sensitive enough to pick up possible small differences in the methylation profile around CGIs and this preliminary conclusion certainly requires further validation.

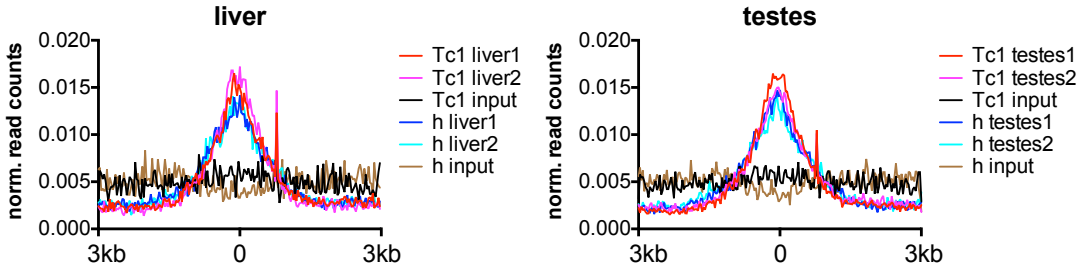

Figure 37: Average read profile at CGIs in human and Tc1 liver (a) and testes (b).
Average read densities are shown across CGIs, which were aligned at their midpoint (0). 3kb on either side of the midpoint are shown.

Table 8: Comparison of standard deviations between human and Tc1 liver. The total number of datapoints (n) analysed and the standard deviation (stdev) are given for each sample.

	h liver 1	h liver 2	h input	Tc1 liver1	Tc1 liver2	Tc1 input
n	200	200	200	200	200	200
stdev	0.003201	0.003232	0.001018	0.003685	0.004229	0.001022

Table 9: Comparison of standard deviations between human and Tc1 testes.
The total number of datapoints (n) analysed and the standard deviation (stdev) are given for each sample.

	h testes 1	h testes 2	h input	Tc1 testes1	Tc1 testes2	Tc1 input
n	200	200	200	200	200	200
stdev	0.003497	0.003142	0.0008716	0.004212	0.003652	0.0008254

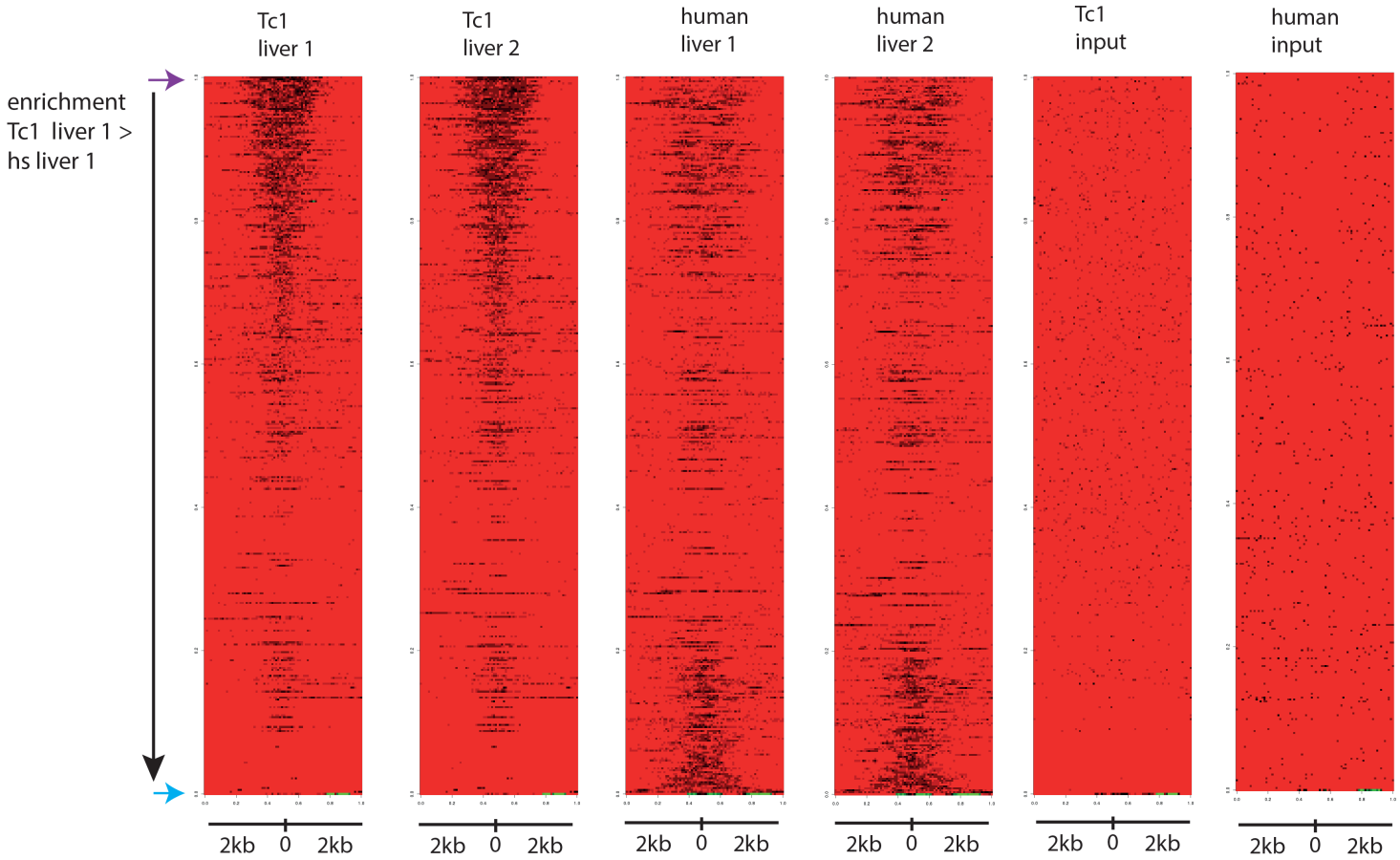

Figure 38: Read density plot of Tc1 liver (two replicates), and human liver (two replicates) as well as Tc1 and human input samples.

The x-axis shows the read density across annotated CGIs aligned at their midpoint (0) and 2kb on either side of the midpoint. CGIs were sorted according to the enrichment of reads in Tc1 liver 1 over human liver 1 (y-axis), so that CGIs for which there are more reads in Tc1 liver1 than in human liver 1 are shown at the top (indicated by a purple arrow), while CGIs with an enrichment of reads in human over Tc1 is shown at the bottom of the heat map (indicated by a blue arrow). Black and red depict high and low read densities, respectively.

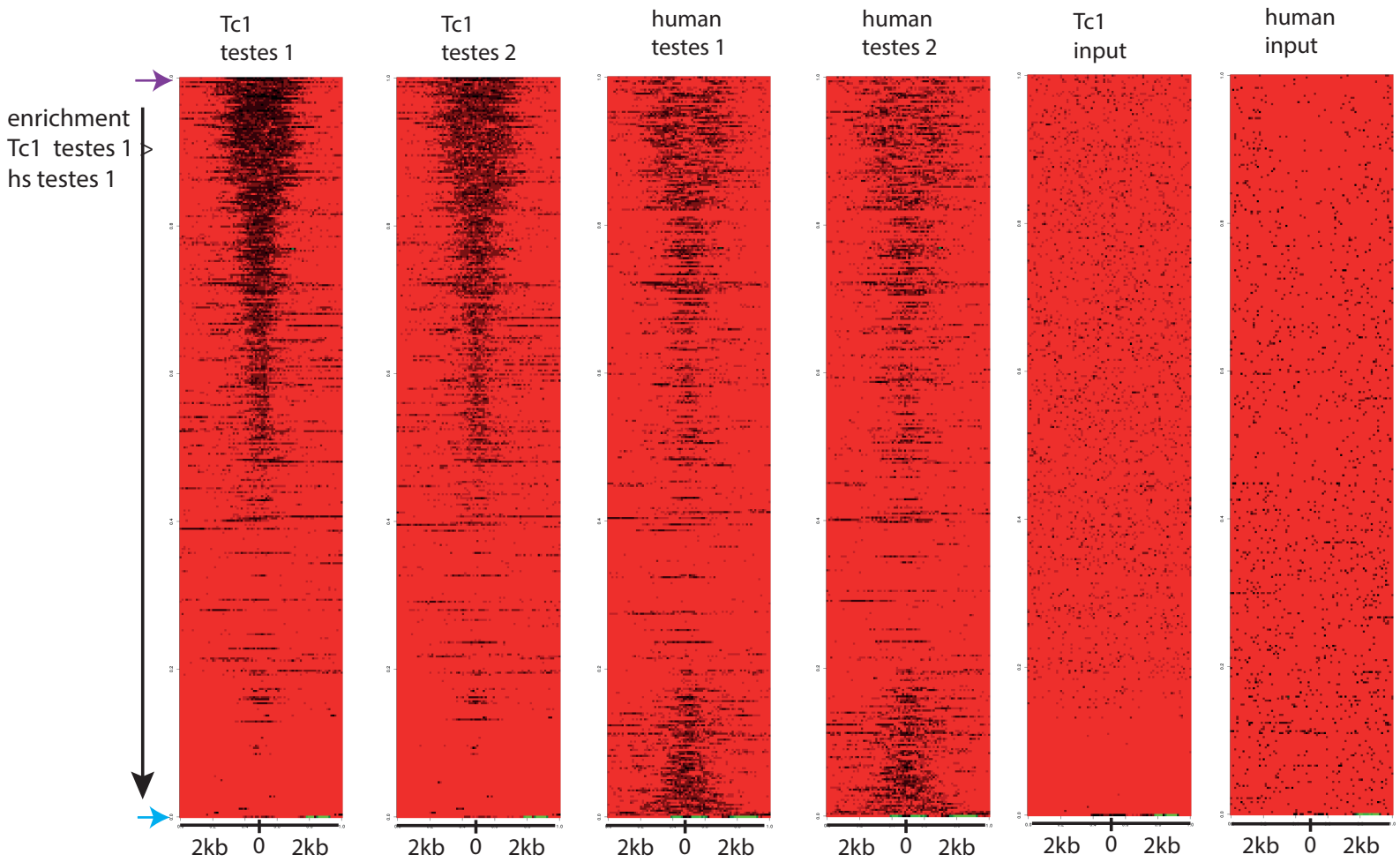

Figure 39: Read density plot of Tc1 testes (two replicates), and human testes (two replicates) as well as Tc1 and human input samples.

The x-axis shows the read density across annotated CGIs aligned at their midpoint (0) and 2kb on either side of the midpoint. The y-axis shows each of the 365 CGIs on hs21 sorted from top to bottom according to the enrichment of reads in Tc1 testes1 (at the top, purple arrow) over human testes 1 (at the bottom, blue arrow). Black and red depict high and low read densities, respectively.

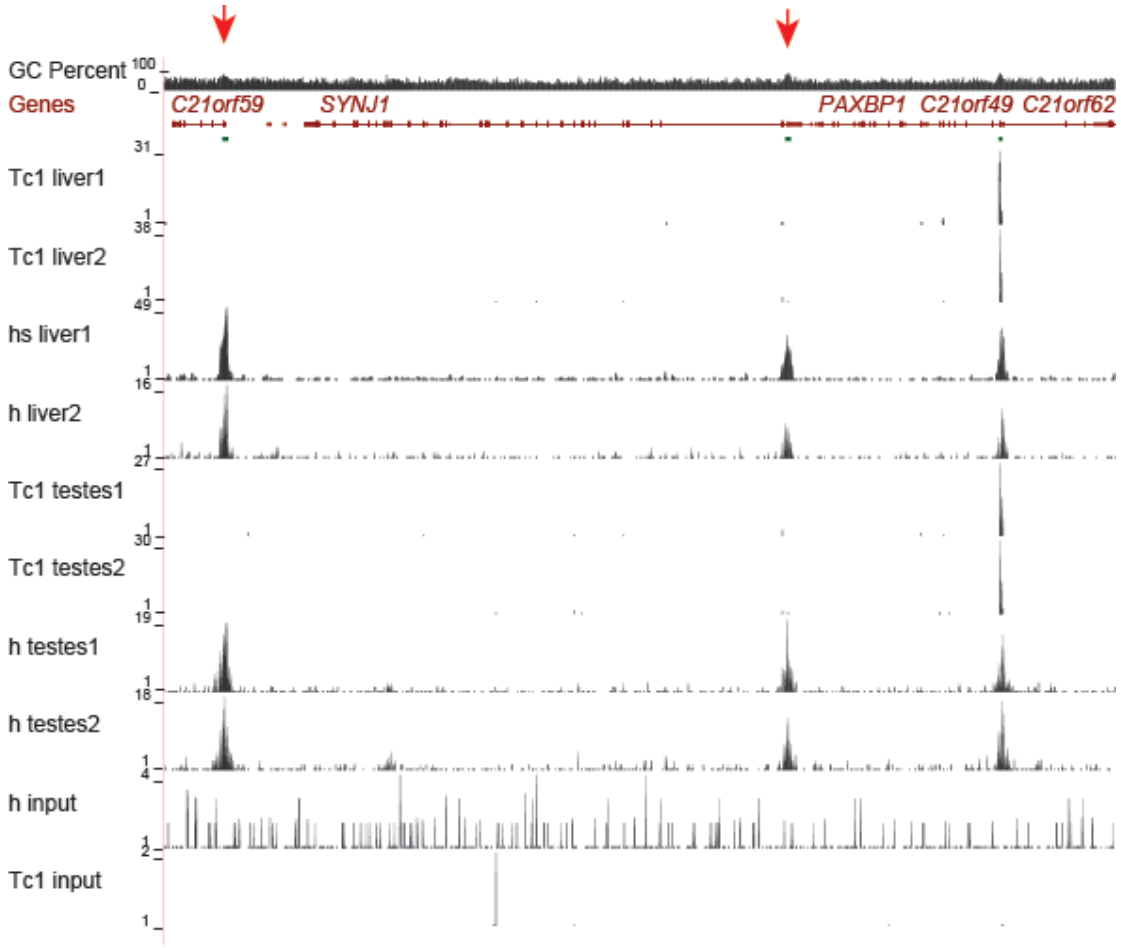

Figure 40: Example locus showing a loss of sequence reads in Tc1 over human liver.
This CGI was selected by lower enrichment of sequence reads in Tc1 compared to human liver1 and corresponds to the position indicated with a blue arrow in Figure 38. The CGIs that show a difference in read densities in Tc1 versus human are indicated with a red arrow.

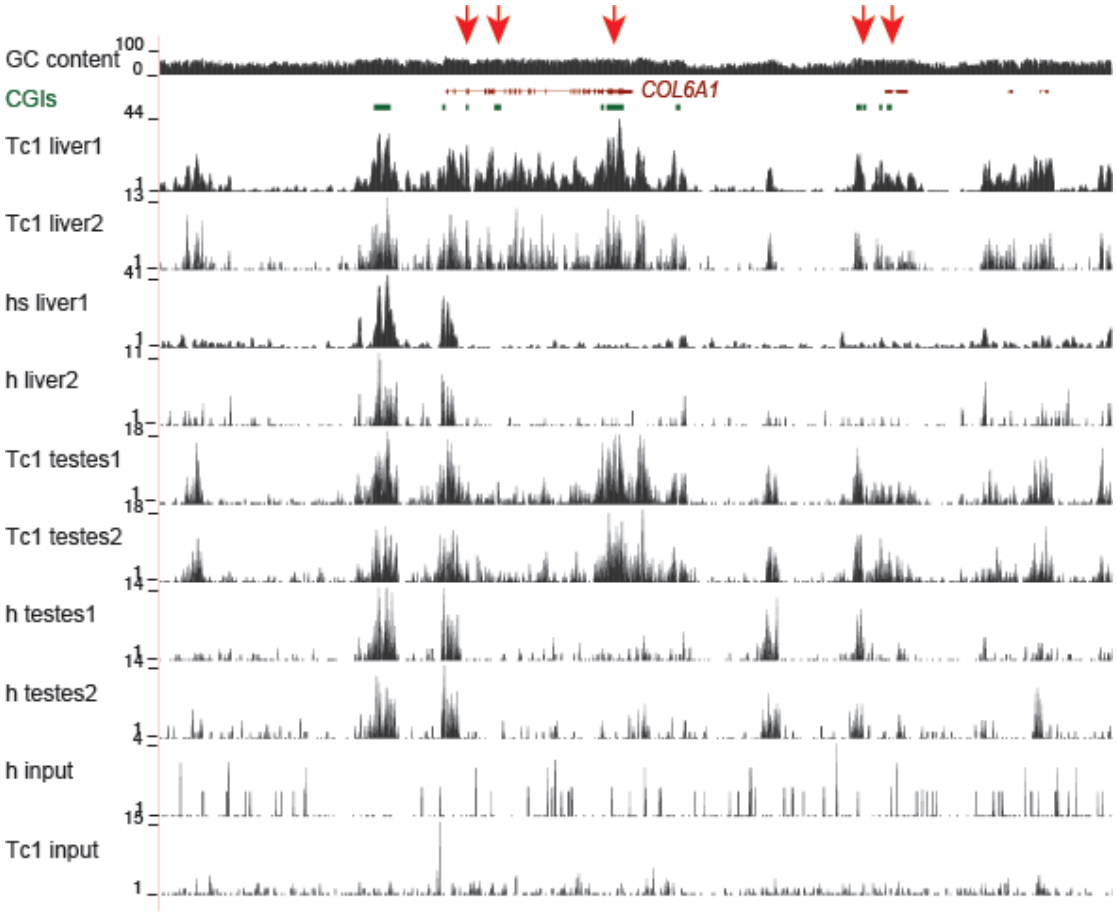

Figure 41: Example locus showing a gain of sequence reads in Tc1 over human liver.
This CGI was selected by its higher enrichment of sequence reads in Tc1 compared to human liver1 and corresponds to the position indicated with a purple arrow in Figure 38. The CGIs that show a difference in read densities in Tc1 versus human are indicated with a red arrow.

Although the average read density profiles at CGIs are similar in Tc1 and human, this analysis can only reveal an overall trend in non-methylated DNA profiles, but it does not take into account possible changes in the sequence read profile at individual CGIs. To address this notion, read densities were plotted across individual CGIs and CGIs were sorted according to the enrichment of reads in Tc1 over human liver as well as testes (Figure 38, Figure 39). This way, although no obvious changes in the width of CGIs seem to occur between human and Tc1 liver as well as testes, sorting of CGIs according to their enrichment in human or Tc1 shows that there is a considerable fraction of reads that bear more reads in human liver and testes compared to Tc1 liver and testes, respectively, which appear at the bottom of

the heat maps in Figure 38 and Figure 39 (indicated by a blue arrow). Conversely, the fraction of CGIs that bear more reads in Tc1 than in human appear at the top of the read density profiles in Figure 38, Figure 39 (highlighted by a purple arrow), indicating a possible difference in the methylation state around CGIs in Tc1 versus human. Example loci showing CGIs with the largest differences between Tc1 mouse and human selected from the very top and bottom of the read density profiles are shown in UCSC (Figure 40, Figure 41). In Figure 40 across this region peaks present in human and testes are missing in Tc1. This may be due to a lack of data for this region as suggested by a lack of sequence reads in Tc1 input. A possible explanation may be that this region is not present in Tc1 mice, as the mouse model only contains an incomplete copy of hs21 (90%, see introduction), although this remains to be verified. Conversely in Figure 41 it is apparent that sequence reads in Tc1 are completely absent from human liver and testes. In this case, the gain of peaks in Tc1 versus human may very well reflect the appearance of hypomethylated regions in Tc1 versus human as there is a consistent read coverage across the region in all samples.

In summary, it appears that the extent sequence read peaks is overall similar in Tc1 and human liver as well as testes, but a few examples of gains or losses of peaks are observed. Yet, whether or not the observed changes in the peaks reflect changes in the methylation profile requires further investigation.

5.3 Discussion:

In this chapter, non-methylated profiles at CGIs of a whole human chromosome transplanted into the mouse was analysed. The analysis revealed that CGIs are generally preserved on hs21 in Tc1 versus human liver and testes, suggesting that the majority of CGIs are properly recognised and remain unmethylated when hs21 is placed in the mouse, which is in line with the observation made in the human *α*-globin cluster, where human CGIs are properly recognised and unmethylated in the context of the humanised mouse model. Furthermore, the sequence read profiles around CGIs appear similar in human and Tc1 liver as well as human and Tc1 testes, which may suggest that no obvious changes in the methylation state of CGIs across hs21 occur, although it is currently not known whether the technique is sensitive enough to pick up small changes in the methylation state similar to those observed in the humanised mouse *α*-globin cluster when placed in the mouse compared to human cells. It is also currently not known whether differences in the initial read number in each sample (prior to subsampling) may result in differences in the distribution of the peaks. Of note, Tc1 mice tend to lose hs21, so that possible differences in the starting material of hs21 used in the BioCAP assay may confound the interpretation of these results; it is for example conceivable that a difference in starting human material may result in differences in the retention of human to the ZF-CXXC domain when pulling down unmethylated DNA. Hence, further studies are required to find out whether human CGIs become extended in regions in the genome other than the human *α*-globin cluster when placed in the mouse compared to human. This will be done by picking a sample of CGIs and comparing the methylation

profile in Tc1 liver and testes versus human liver and testes by bisulfite sequencing.

Although the present data is currently not sufficient to make conclusions about possible changes in the extent of non-methylated DNA at CGIs another interesting observation was made in the study presented here, which is a gain in peaks in Tc1 liver and testes versus human liver and testes (Figure 41). Whether or not this observation is made at CGIs in other regions that the one shown in Figure 41 will be tested in the future. Although this data requires further validation it indicates a possible appearance of hypomethylation in Tc1 at CGIs, which may be methylated in human. To confirm the methylation state at these CGIs, publicly available data will be screened and if available it will be used to validate their methylation state in normal human liver and testes. Furthermore, bisulfite sequencing of these CGIs in Tc1 and human liver is planned in the future. Finally, Figure 41 suggests that the CGIs, which show a gain in peaks in Tc1 liver and testes versus human liver and testes, represent non-promoter CGIs. However, it remains to be determined whether this observation can be made at CGIs in other regions than the one shown in Figure 41. In the future, is planned to quantitate the number of CGIs that show a gain in peaks in Tc1 versus human and determine the number of promoter, inter- and intragenic CGIs. Furthermore, bisulfite sequencing of a sample of CGIs that represent promoter and non-promoter CGIs will be carried out to validate the preliminary findings presented here.

Finally, the question remains whether there are also losses of peaks in Tc1 liver and testes versus human liver and testes. Figure 40 shows such an

example, but as mentioned this loss of peaks in Tc1 may be due to the absence of this region in Tc1 mice. To verify the absence of this region in Tc1, PCR of genomic DNA from Tc1 will be carried out and compared to human genomic DNA. Of note, Figure 39 and Figure 40 indicates that there is a substantial fraction of other CGIs in which there are more reads in human liver and testes than in Tc1 liver and testes. Whether or not this loss in peaks in Tc1 is due to the lack of data in Tc1 or whether these differences truly reflect methylation changes in Tc1 versus liver will be addressed in the future.

6 Final Discussion

Gene expression in normal development is tightly regulated and DNA methylation plays an important role in this process. Abnormal patterns of DNA methylation have been associated with many human diseases (Egger et al, 2004; Esteller, 2005; Feinberg & Tycko, 2004; Herman & Baylin, 2003; Robertson & Wolffe, 2000). To understand how DNA methylation regulates gene expression it is important to have a profound knowledge of the patterns of DNA methylation and an understanding of how these are established and maintained in mammalian cells. The study presented here furthers our knowledge of DNA methylation patterns of human DNA in a heterologous murine environment compared to its natural endogenous state and reveals species-specific regulation of DNA methylation between human and mouse. The presented results may provide a useful basis to identifying the *cis*- and *trans*-acting regulatory factors involved in determining DNA methylation patterns in human and mouse cells and furthermore call into question the validity of transgenic mouse models for the study of human gene regulation.

6.1 Previous DNA methylation studies of human transgenes in the mouse

Since it was possible to create lines of transgenic mice in the early 1980s, the use of these mouse models to study human gene regulation has dramatically increased. A number of previous studies, however, report abnormal gene silencing of transgenes when integrated into heterochromatic regions or when arranged in concatamers (Alami et al, 2000; Garrick et al, 1998; Kaufman et al, 1999). In one specific example it was shown that gene silencing is

associated with increased methylation of the transgene (Garrick et al, 1998). Other than that, studies investigating methylation of transgenes are, to my knowledge, very limited and most of the studies that observed transgene silencing focussed on small transgenes. In the study presented here we examined the DNA methylation of a large 117kb human DNA fragment carrying the human *α*-globin genes in the mouse syntenic region, in which the mouse *α*-globin genes are normally regulated. In this model tissue- and developmental-stage specific expression of the human *α*-globin genes in mouse cells is seen with expression reaching ~40% of the endogenous mouse genes (Wallace et al, 2007). The reason for suboptimal expression is not known. Here we show that specific regions (CGI shores) in the human *α*-globin cluster become hypomethylated when the human DNA was inserted into the mouse genome compared to its natural state, a human cell. Although a previous study has documented hypomethylation of repetitive DNA on hs21 in Tc1 mice in comparison to human cells (Ward et al, 2013), the data presented here is, to my knowledge, the first systematic in-depth analysis of DNA methylation patterns of a large human transgenic region integrated into the mouse genome and the first documentation that reveals hypomethylation of CGI shores of human DNA when placed in the mouse compared to human.

6.2 Transgenic mouse model to dissect out regulatory factors of CGI shore methylation

DNA methylation patterns are grossly perturbed in cancer and these perturbations include CGI shore hypomethylation (Irizarry et al, 2009). To understand what leads to the observed DNA methylation abnormalities, DNA methylation is frequently studied in these cancer cells. However, gross

abnormalities in the expression of many genes in cancer cells as well as selective pressure makes it difficult to dissect out the factors involved in normally determining DNA methylation patterns in a healthy cell. In contrast, in this transgenic model we are able to study the pattern of methylation of a segment of DNA, which is presumed to play a neutral role in the establishment and maintenance of DNA methylation, and to examine methylation dynamics *in vivo* during normal mouse development. Using this model hypomethylation of CGI shores was documented to occur when human DNA is placed into a mouse compared to human. Clearly, changes in the methylation profile of human DNA in mouse versus human cells must reflect differences in the mouse versus human nuclear environment. As these mouse cells represent normally functioning healthy cells, it is proposed that this model system provides a useful tool to dissect out the nuclear factors involved in determining CGI shore methylation.

6.3 Sequence determinants of DNA methylation

In this study it was found that when placing a 117kb region comprising the human *α*-globin genes into the mouse, methylation of most of the human transgenic region is properly established and maintained in mouse compared to human cells: most of the transgenic region displays high levels of methylation in mouse versus human, while unmethylated CGIs remain unmethylated in the mouse context. Finding that most of the human *α*-globin cluster is properly methylated in the humanised mouse model suggests that the mouse DNA methylation machinery recognises human DNA in the same way as the human nuclear components do. This is in line with our current knowledge that DNA methylation is high at most CpG sites in both species

and the enzymes responsible for this modification are highly conserved between the species (Bestor, 2000; Goll & Bestor, 2005). In agreement with this idea, large human DNA segments introduced into the mouse preserve the binding of proteins to human DNA compared to a normal human cell, suggesting that human DNA is properly recognised by mouse proteins in a heterologous murine environment (Lynch et al, 2012; Wilson et al, 2008).

In the present study, however, it was shown that in the context of the humanised mouse model CGIs become extended when placed in the mouse. Furthermore, preliminary data suggests that methylated CGIs become aberrantly hypomethylated in the context of a whole copy of human chromosome 16 in mouse versus human cells indicating that hypomethylation of the *α*-globin genes persists in mouse cells even when the *α*-globin cluster remains in its normal chromosomal context and is not the result of a relocation of the genes from human chromosome 16 to mouse chromosome 11 as in the humanised mouse model (see Figure 7). Rather, these results suggest that putative differences in the *trans*-acting milieu between human and mouse cells lead to in changes in the DNA methylation pattern of human DNA. Factors that are involved in binding to CGIs and potentially keeping CGIs in an unmethylated state have been previously proposed. These include TFs, DNA binding proteins with a ZF-CXXC domain, such as TET enzymes, as well as proteins that recruit H3K4me3, which is inhibitory to DNA methylation. The differences in the *trans*-acting milieu could result from differential expression of cellular proteins or from species-specific binding properties of proteins to the same sequence (human DNA). Interestingly, differential expression of orthologous human and mouse proteins has been previously documented for

a specific case in which the γ-globin transcriptional repressor BCL11A is differentially expressed in human and mouse (Sankaran et al, 2009). On the other hand, in support of possible differences in binding properties of human and mouse proteins, it was found that functionally conserved proteins in human and mouse show large differences in their binding patterns at orthologous genes (Odom et al, 2007). In conclusion, it is proposed that differences in the *trans*-acting milieu, either due to differences in protein expression or differences in the binding proteins of human and mouse proteins, may explain the differences in DNA methylation pattern observed when human DNA is placed in the mouse compared to normal human cells.

6.4 Molecular mechanism of hypomethylation

Clearly, differences in the methylation pattern of human DNA in mouse versus human cells reflect differences in the interaction of mouse versus human proteins to human DNA. With this in mind, the question remains how on a molecular level may hypomethylation be induced. It is, on the one hand, conceivable that mouse proteins bind human DNA less efficiently (due to differences in binding properties or expression) resulting in the loss of recruitment of DNA methyltransferases to human DNA. On the other hand, it is possible that mouse proteins bind more efficiently to human DNA resulting in protection of the DNA from methylation. (The latter may occur passively by restricting access of DNMTs to the DNA or via an active DNA demethylation pathway.) In the future, it would be interesting to test for species-specific differences in protein binding to the same human DNA (see section 6.6).

6.5 Role of spreading in determining methylation patterns

A study carried out by Lienert et al suggests that the unmethylated state of sequences can spread into normally methylated sequences (Lienert et al, 2011). On this basis, it was hypothesized in the study presented here an increased spreading of hypomethylation from CGIs into normally methylated CGI shores may lead to hypomethylation of CGI shores (see chapter 4). Of note, however, an alternative explanation of the observed extension of CGIs may the lack of spreading of methylation from regions outside of CGIs into CGIs.

How spreading of methylated or unmethylated states may be achieved is currently not known. For heterochromatin, however, the molecular mechanism of spreading is relatively well established. In essence, it involves a self-reinforcing crosstalk between DNA and histones, in which cytosine methylation, histone deacetylation, and H3K9 methylation act synergistically to create a positive epigenetic feedback loop that self-perpetuates and spreads across large regions of DNA. Although the exact sequence of events is still unclear to date, crosstalk between repressive histone marks and DNA methylation is well documented (see chapter 1.1.1).

Boundaries of heterochromatin spreading may be established by the presence of active epigenetic marks (e.g. H3K4me3) that counteract heterochromatic marks (see chapter 1.1.1) or via a disruption of the contiguous array of nucleosomes, which is thought to be necessary for heterochromatin spreading in *S. cerevisiae* (Bi & Broach, 2001). Importantly, analogous mechanisms in which histone methylation serves as a recruitment

platform for effector proteins also function at euchromatic loci, where methylation at H3K4 and H3K36 recruit chromatin remodelling and HDAC activities, respectively, to transcribed genes (Carrozza et al, 2005; Joshi & Struhl, 2005; Keogh et al, 2005; Pray-Grant et al, 2005; Wysocka et al, 2006). In this light, it is conceivable that in a similar fashion to the spread of heterochromatin methylated states may spread into CGIs, where they are counter-acted by active marks (or via a specific nucleosomal array) at CGIs. The counteractive forces of active and repressive marks may determine boundaries of CGIs and thereby the methylation state of CGI shores. Whether or not hypomethylation of CGI shores is the result of a spreading mechanism will be addressed in the future by analysing a fragment that solely consists of the sequences of the CGI shore upstream of CGI H.

6.6 Identification of DNA binding factors

Further studies are required to elucidate sequence features of CGI shores that might determine the methylation state of the underlying DNA and shed light on possible mechanisms involved in the establishment of DNA methylation levels in these regions. Previous reports have implicated DNA binding factors in determining methylation states of DNA. In the future, it would be interesting to test whether aberrant binding of mouse DNA binding proteins to these fragments might explain hypomethylation of the underlying DNA.

In this study, the sequence determinants for hypomethylation have been greatly narrowed down to 2kb sequences. In the future, these hypomethylated sequences could be screened for putative DNA binding motifs and tested for

their role in determining methylation states by mutating them. Furthermore, putative differences in the binding properties of human and mouse proteins to CGI shore fragments could be analysed *in vitro* by band shift assay using nuclear extracts from human and mouse cells.

6.7 CGI shore hypomethylation and gene expression

Previous studies suggest that CGIs are enlarged when the associated gene is active compared to their state when the gene is silent, although for specific examples the opposite was observed (Molaro et al, 2011). These observed correlations, however, do not reveal whether CGI shore hypomethylation is a cause or a consequence of gene expression.

In the transgenic mouse model, CGIs are extended irrespective of cell type (erythroid and non-erythroid), suggesting that hypomethylation of CGI shores is independent of transcription. To corroborate the idea that the methylation states of human CGIs in the mouse is independent of the expression level of the *α*-globin genes it would be interesting to measure methylation of the *α*-globin CGI in a fragment with a constitutive promoter, that induces high levels of expression. This fragment has already been generated in a previous study and could be analysed in the future. Of note, however, this experiment would only address a possible role for transcription in the regulation of methylation, while the vice versa scenario could be addressed as outline below (see section 6.8).

6.8 Functional relevance of CGI shore hypomethylation

Finally, the question arises as to what extent CGI extension is functionally relevant. To address this question, it would be interesting to investigate

whether the observed CGI extension in human DNA in a mouse cell is associated with a change in the expression level of the gene.

This could be done by a chromosome-wide correlation approach in the Tc1 mouse model correlating DNA methylation data with publicly available data on active marks (H3K4me3 levels and RNA polII binding). Bisulfite sequencing data could, for example, be generated for hs21 bisulfite-high throughput sequencing data of FACS-sorted hs21 in Tc1 versus human. Furthermore, as it is currently not known whether the hypomethylated region of CGIs is enlarged in Tc1 compared to human, it would also be very interesting to instead find out whether the appearance of hypomethylated CGIs in Tc1 versus human liver and testes correlates with gene expression.

6.9 Generality of the phenomenon

The observation that human CGIs become extended was documented for all of the tested CGIs in the human *α*-globin cluster, suggesting that this might be a general phenomenon. However, to draw a conclusion it is necessary to extent this analysis and test further CGIs. Here, a mouse model system carrying a whole copy of hs21 was used to study whether hypomethylation occurs at other human CGIs when placed in the mouse. Although the data presented here could not answer this question and it remains to be determined whether the method itself is sensitive enough to reveal species-specific changes in the width of CGIs, this experiment did reveal further regions of hypomethylation in the transgenic Tc1 mouse. It would be of interest to explore the functional consequence of this hypomethylation – for example by looking for changes in gene expression and possible activation of

cryptic promoters. Future studies will assess the methylation state of other CGIs on hs21 in mouse versus human by picking a number of selected CGIs on hs21 and measuring DNA methylation levels by bisulfite sequencing in human and Tc1 to probe for changes in CGI shore methylation.

7 Appendix

7.1 Targeted bisulfite sequencing (TBS)

Targeted bisulfite sequencing (TBS) is an approach that was developed previously in the lab to interrogate DNA methylation patterns of the human *α*-globin cluster. The technique was designed by Professor Richard Gibbons. Establishment of the protocol and initial optimisation was carried by Caroline Scott and Nicola Malik and further developed by Noemi Roy (Roy). High-throughput sequencing was carried out at the Wellcome Trust Centre for Human Genetics (WTCHG) under supervision of Lorna Gregory and Helen Butler. The downstream processing of the high throughput sequence data was carried out with the help of Nicki Gray (Molecular Haematology Unit).

In contrast to other targeting approaches, in our protocol DNA capture was carried out prior to bisulfite conversion. The advantage of carrying out the capture on native DNA is that the oligo library can be designed independent of the methylation state, while capture after bisulfite conversion requires taking into account the methylation state of the converted DNA when designing the oligos. Furthermore, a greater specificity of the capture can be achieved using oligos designed against the normal genome that contains all four bases. (Note that bisulfite treatment converts the vast majority of Cs in the genome reducing the complexity of the genome to essentially a 3 base genome.)

The workflow of the TBS is illustrated in Figure 42. Genomic DNA was sheared by sonication to fragments of around 500 bp (Figure 42). Sonicated

DNA fragments were denatured and hybridised in solution to a library of biotinylated oligonucleotides spanning the human *α*-globin genes.

The oligo library was designed by Richard Gibbons and Jim Hughes to capture the human *α*-globin genes and their UREs. The library consists of 456 oligos with an average length of 52bp. The oligos match to 23.7kb of human DNA distributed across a 78kb region spanning the human *α*-globin genes and their UREs. Figure 43 shows a map with the capture area in the human *α*-globin gene cluster. The captured area ranges from 93'800 to 172'600 kb[17]. The capture oligos match to the top strand so that only the bottom strand is captured. Repetitive regions were excluded from the library (i.e. the oligo library was designed to only match to unique regions of the *α*-globin cluster). A 5kb representative area zoomed into *α2* is depicted in Figure 44 to illustrate the spacing between individual probes and their location with respect to repetitive elements.

Successful enrichment of the captured region was verified by qPCR choosing primers located within and outside of the captured area (Figure 45). An adequate enrichment was considered to be at least 3000 fold. Enriched material was denatured and bisulfite converted. The bisulfite converted DNA was then amplified for two cycles using non-specific primers that contain an AcuI cleavage site followed by 30 cycles of specific PCR amplification using primers that match to the primer sequences used in the first random primer

[17] The location of the primers ranges from 94308 (start position of first oligo) to 172114 (end position of last oligo, in human genome build hg18) (covering 2262 CpG sites). The capture area is estimated to range from 93800-172600 (covering a total number of 2303 CpGs), which is equivalent to the location of the oligos extended by 500 bp (the average length of sonicated genomic DNA fragments) on either side. Note that the captured area in the transgenic mouse model is slightly smaller, because the transgenic region (coordinates chr16: 53417-170901, hg18) ends before the last capture oligo reducing the capture area in humanized mouse by 1.7kb (resulting in 120 fewer CpG sites = 2183 total CpGs in the captured area).

amplification (RPA) step. Efficient bisulfite conversion of the DNA was verified by subcloning and sequencing a small amount of DNA and analysing the conversion of Cs to Ts in the non-CpG context of 20 subclones. If successfully converted, the primer sequences attached to the DNA fragments during PCR amplification are trimmed off by AcuI digestion[18]. The cleaved DNA is size selected and gel purified. A minimum of 10 ng of recovered material was used for downstream high throughput sequencing using the Illumina platform (100bp paired-end sequencing). High-throughput-sequencing was performed at the Wellcome Trust Centre for Human Genetics (WTCHG).

[18] The use of RPA primers introduces a further 17 bases to each capture fragment. AcuI restriction enzyme cleaves 22 bp downstream of its recognition site within DNA fragment thereby ensuring that no primer sequences will be sequenced on the Illumina platform.

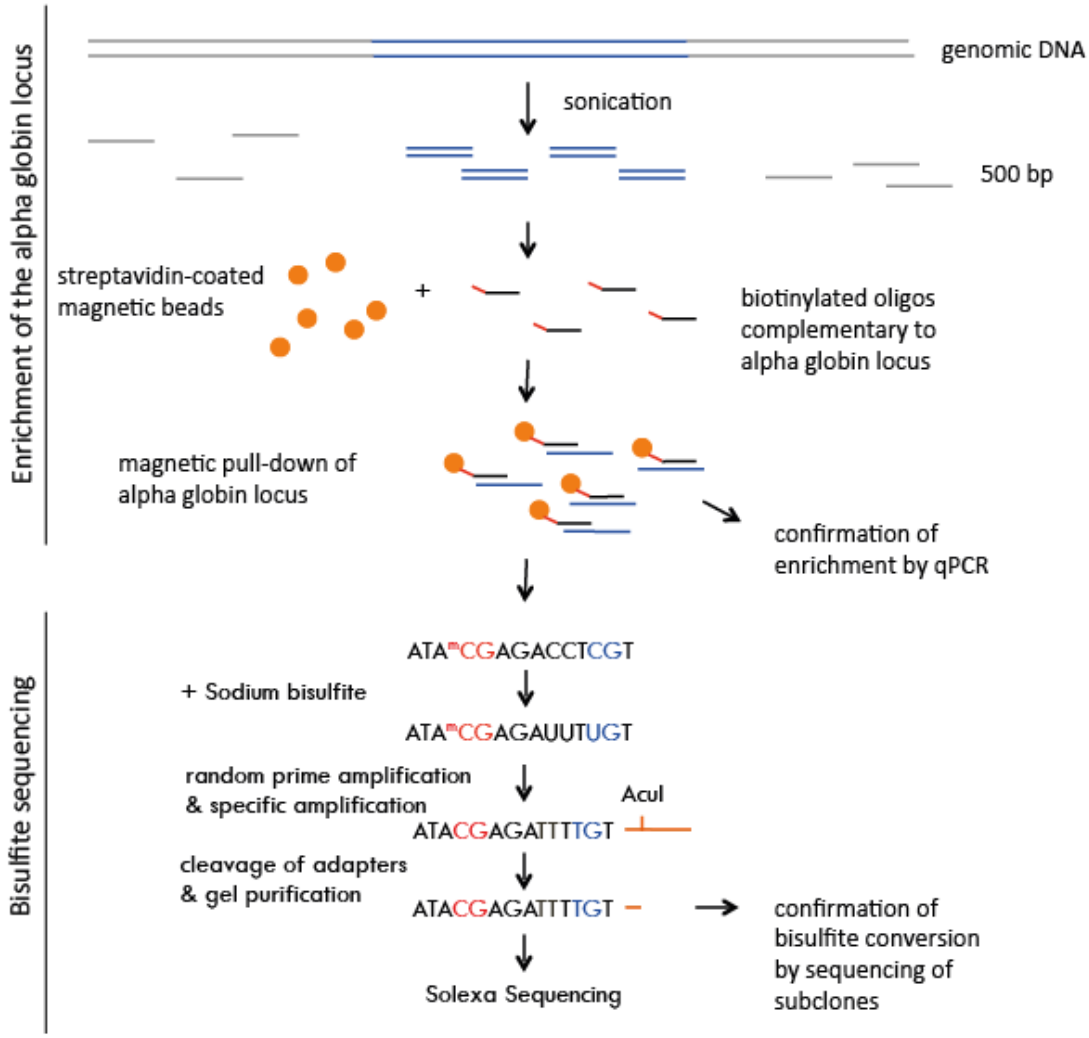

Figure 42: Workflow of targeted bisulfite sequencing (TBS).

Genomic DNA is fragmented by sonication to an average length of 500 bp. DNA fragments are denatured and hybridized to a library of biotinylated olignucleotides complementary to the α-globin gene region. Hybridizing DNA fragments are subsequently pulled down using streptavidin-coated magnetic beads. qPCR is performed on the recovered material to verify enrichment of the α-globin gene region over regions outside the captured area. Successfully captured DNA is modified by treatment with sodium bisulfite selectively turning unmethylated cytosines into uracil, while methylated cytosines are protected from this modification. Subsequent random prime amplification (RPA, two rounds) adds 17 bp of known DNA sequence comprising an AcuI restriction site to the captured DNA fragments. RPA is followed by 30 cycles of specific PCR amplification using primers that match the adapters added in the first rounds of random PCR amplification. The adapters are subsequently trimmed off by AcuI digestion and the cleaved DNA fragments are gel purified. A fraction of the converted material is subcloned and sequenced to confirm successful bisulfite conversion. Finally, the converted material is processed by high throughput sequencing using the Illumina platform to enable the direct measurement of DNA methylation levels by calculating the number of cytosines and thymidines at each CpG site.

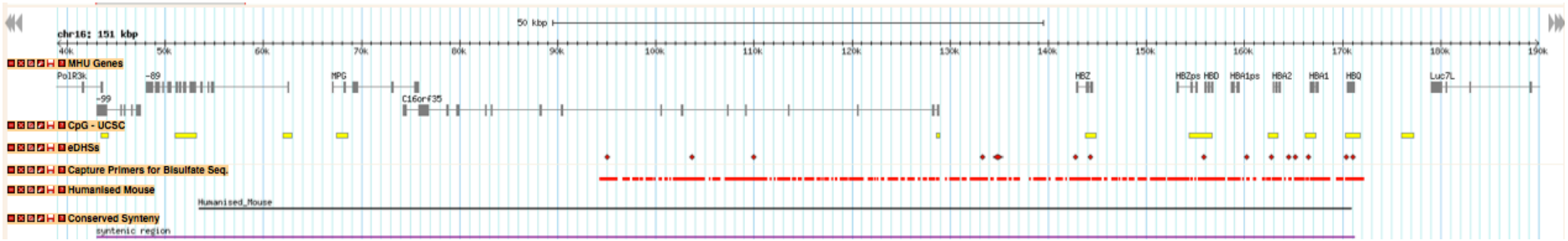

Figure 43: Map of the human *α*-globin gene cluster showing the capture area within the transgenic region.

Depicted are (from top to bottom): genomic location (in kb) on 16p, genes (grey), CpG Islands (yellow), erythroid-specific DHS (red dots), capture oligos (red bars), transgenic region (grey bar), mouse syntenic region (purple bar)

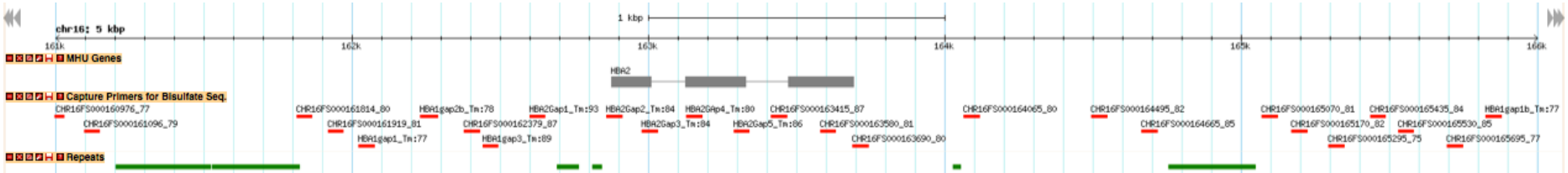

Figure 44: Close-up view of the capture area showing a 5kb window around *α2* (grey) illustrating the distance between the capture oligos (red bars).

The oligos were designed to match unique sequences only; repetitive elements (green bars) were only pulled down via adjacent oligos matching to the unique sequences.

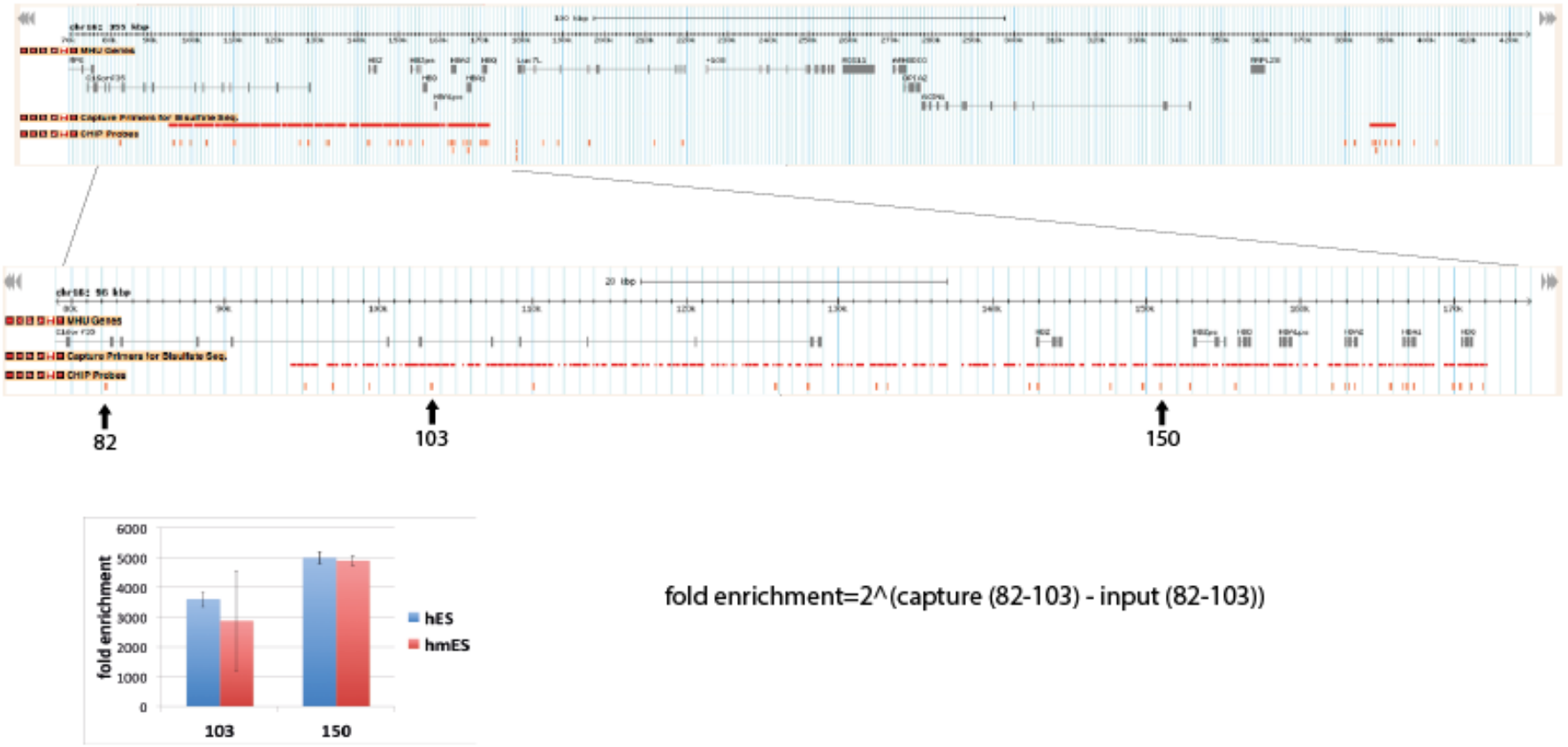

Figure 45: Enrichment of captured over uncaptured material.

Location of oligo library is shown in Gbrowse. Location of primers used for qPCR are indicated by arrows. Graph showing the fold enrichment of captured material normalised to non-captured DNA in human and humanised mouse ESCs. The delta delta Ct formula used to calculate the fold enrichment of the captured DNA is shown on the right.

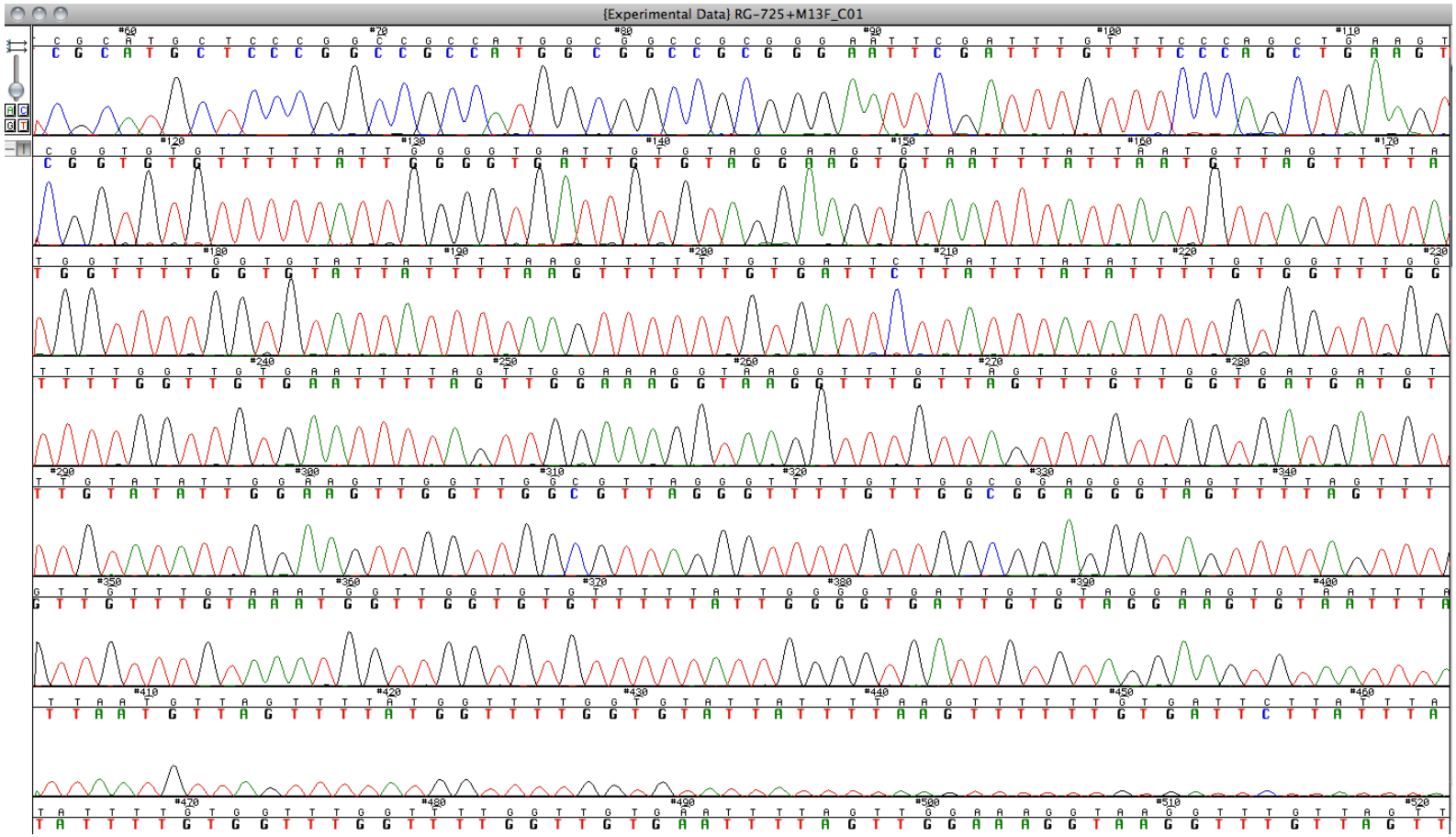

Figure 46: Verification of conversion of DNA by treatment with sodium bisulfite.

An example of the DNA sequence of a subcloned DNA fragment captured and bisulfite converted as part of the TBS protocol. The conversion efficiency was estimated by calculating the number of unconverted strands over the total number of strands analysed. In the initial **optimis**ation of the TBS protocol, the best results were obtained with the MethylEasy Xceed Kit (Human genetic signatures) (50-100% conversion). If the conversion efficiency was higher than 70%, the sample was taken through the rest of the TBS protocol.

Bioinformatic analysis

The downstream processing of the high throughput sequence data received from the WTCHG was carried out with the help of Nicki Gray (Molecular Haematology Unit). General optimisations of the method and the downstream analyses were discussed in meetings with Nicki Gray, Noemi Roy and Richard Gibbons. A summary of the downstream analyses is given in Figure 47. It involves several steps, which will be briefly discussed below.

FastQC

First, the quality of the raw sequencing data is assessed using Fast QC. FastQC is a quality control tool developed at the Babraham Institute (Cambridge) used for data coming from high-throughput sequencing platforms. It runs a series of modules and reports if there are any unusual sequences in the library. Some of the modules are not applicable to bisulfite sequencing data, such as GC content and base composition. A useful module, however, is the sequence quality control module, which plots the sequence quality per base call across the sequence reads. A quality score well above 20 is considered as good quality, anything below 20 was discarded. Typically, the sequence quality decreases towards the end of the read (this has to do with the synchronisation of sequencing by synthesis) and it will become visible in the per base sequencing plot. In the case in which parts of the sequence are of low quality (e.g. the read ends), these parts can be trimmed off to shorter reads to exclude bases with low quality scores. It is also possible to identify whether a subset of sequences have a low quality, which can for example arise from poor imaging due to their position in the flow cell. Sequences of inadequate quality have been discarded and only high

quality sequences (with a quality score well above 20) were be used for mapping. FastQC also has a tool to identify non-matching sequences by mapping them to bacterial genomes (to test for contamination) and to the adapters used in the run.

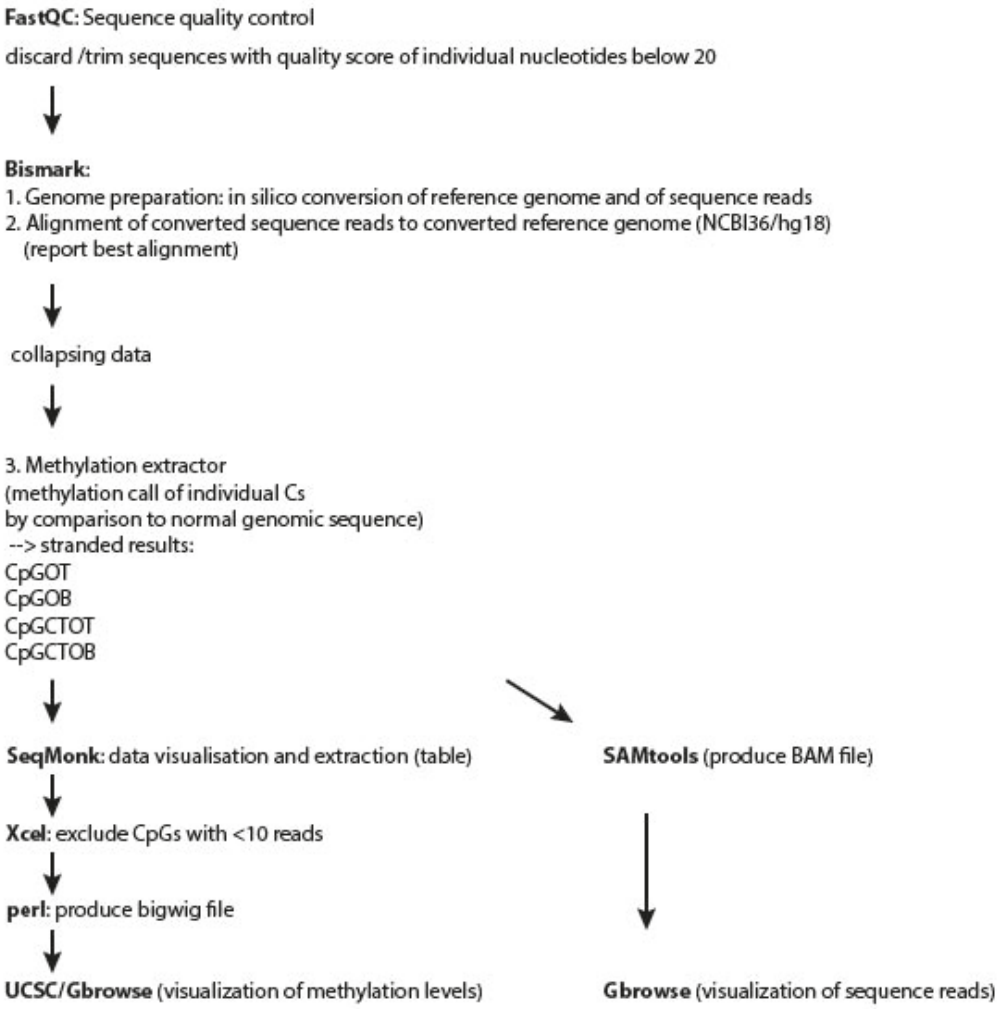

Figure 47: Bioinformatics pipeline.
Methylation extractor produces 4 text files containing the information for CpG sites on the original top strand (CpGOT), complementary to original top (CpGCTOT), original bottom (CpGOB), or complementary to original bottom (CpGCTOB). The capture technique was designed to capture the top strand and as expected the majority of sequences reads matched to CpGOT and CpGCTOT strands.

Bismark

Mapping bisulfite-converted sequence reads to the genome is an intricate challenge for bioinformaticians. There are several reasons for this. Firstly,

conversion of Cs to Ts reduces the complexity of the genome (from a 4 base to a 3 base genome). Secondly, after bisulfite conversion the two strands that were originally complementary are now non-complementary and subsequent PCR amplification gives rise to four possible strands of any given sequence essentially doubling the size of genome to be searched against. Thirdly, a T in a given read sequence can originate from a T or a C in the original genome further complicating the mapping procedure.

Various algorithms have been developed to circumvent the problems that arise from bisulfite-converted material. Earlier generations of bisulfite sequencing mapping programs, such as BISMA, BSMAP, SOAP2, BRAT were relatively slow and inaccurate. Newer methods, such as Bismark and BS-Seeker, are faster and more accurate and integrate the determination of the methylation state in the mapping process, making a post-processing unnecessary (Krueger & Andrews, 2011). The way the Bismark program solves some of the problems that arise from bisulfite sequencing data is by converting residual cytosines *in silico* into a fully bisulfite-converted form before alignment takes place. In comparison to BS-Seeker, Bismark performs better at discarding non-unique alignments, although this comes at the prize of a longer run time of the program (Krueger & Andrews, 2011). Here, Bismark has been chosen for the analysis of bisulfite sequencing data.

Bismark was developed at the Babraham Institute by Felix Krueger and integrates a set of three tools which will be described in more detail below: 1) Genome preparation, (2) Sequence alignment, (3) Methylation Extractor.

In the first step, the genome is converted *in silico* (Genome Preparation). This converts all Cs in top and bottom strands into Ts, which after *in silico* conversion are no longer complementary to each other. The complementary strands to the converted top and bottom strands are also produced. In this case Gs are converted to As (matching to the Ts in the original strands). To allow subsequent bowtie alignments the converted genome is indexed[19]. The Genome Preparation step needs to be done only once for each genome.

Once the Genome Preparation step has been carried out for the required reference genome, the actual Bismark program aligns the fully converted sequence reads to the converted genome using Bowtie[20]. Bowtie is an ultra-fast memory efficient short read aligner (Langmead, 2010). Running Bismark produces two files: a SAM (Sequence Alignment/Map) file containing the information on alignment and on methylation calls as well as a run statistics report.

To verify that the sequences have mapped correctly to the genome the sequences of the mapped reads were compared to the sequence in the reference genome; this was done by eye for selected regions. To do this, the SAM output file from Bismark was converted into a bam file (binary SAM file) using SAMtools. This allows viewing the DNA sequence of the mapped reads in a genome browser, such as UCSC or Gbrowse (an internal genome browser developed by Nicki Gray and Simon McGowan for the MRC Molecular Haematology Unit). A 20 kb region was screened by eye for correct

[19] Indexing means that the alignments are assigned to a position in the reference genome. It enables to quickly retrieve alignments without going through the whole alignments.
[20] Fully converted sequence reads means that all residual Cs (in the forward reads) and As (in the reverse strands) are *in silico* converted.

mapping of the sequence reads against the original genome and for correct methylation calling at individual CpG sites (Figure 48).

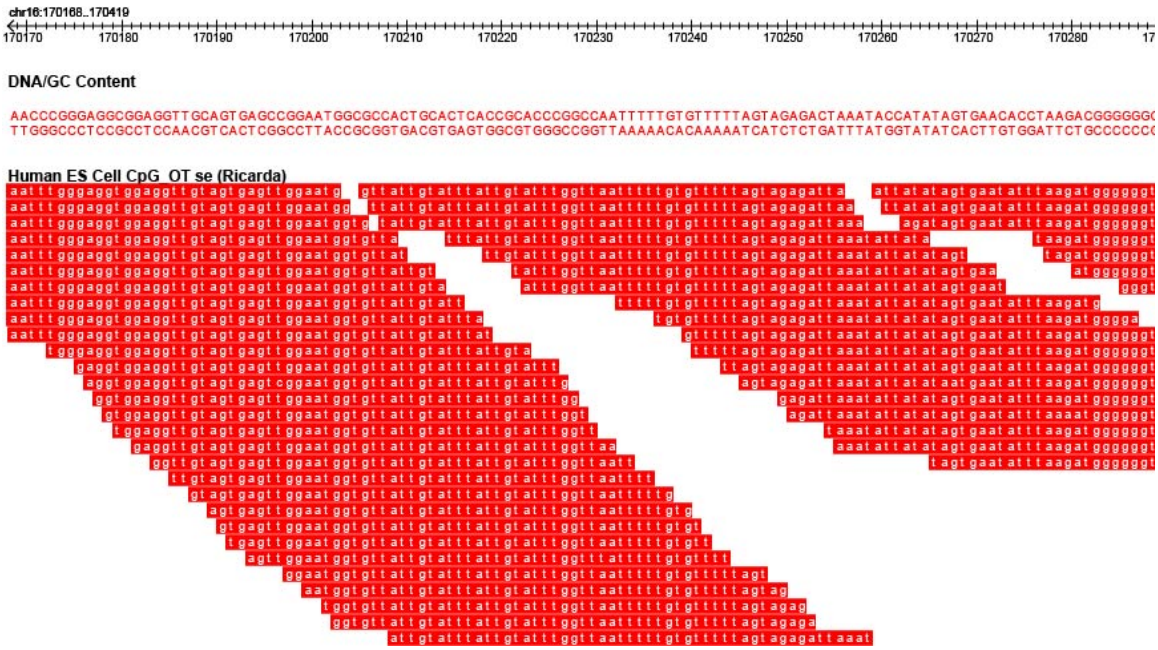

Figure 48: Close-up view of the sequence reads obtained by TBS and mapped to the human genome.
A 20kb region was screened by eye to test whether sequence reads have mapped correctly.

PCR amplification of the bisulfite-converted DNA can result in the generation of identical sequences originating from the same DNA template, so-called clonal variants. To exclude any biases resulting from PCR amplification of the template DNA identical sequence reads (reads with the same start and end position, identical sequence and methylation profile) were considered as clonal variants and only considered once (collapsing of data, Figure 49).

In the final step of Bismark, methylation states are called at individual CpG sites using the Methylation Extractor. This is done by comparing the sequence of each read to the normal (unconverted) genomic sequence and calling the methylation state at individual cytosines. This step returns stranded results for all 4 possible strands, named original top (OT), original bottom (OB), complementary to original top (CTOT) and complementary to original bottom (CTOB).

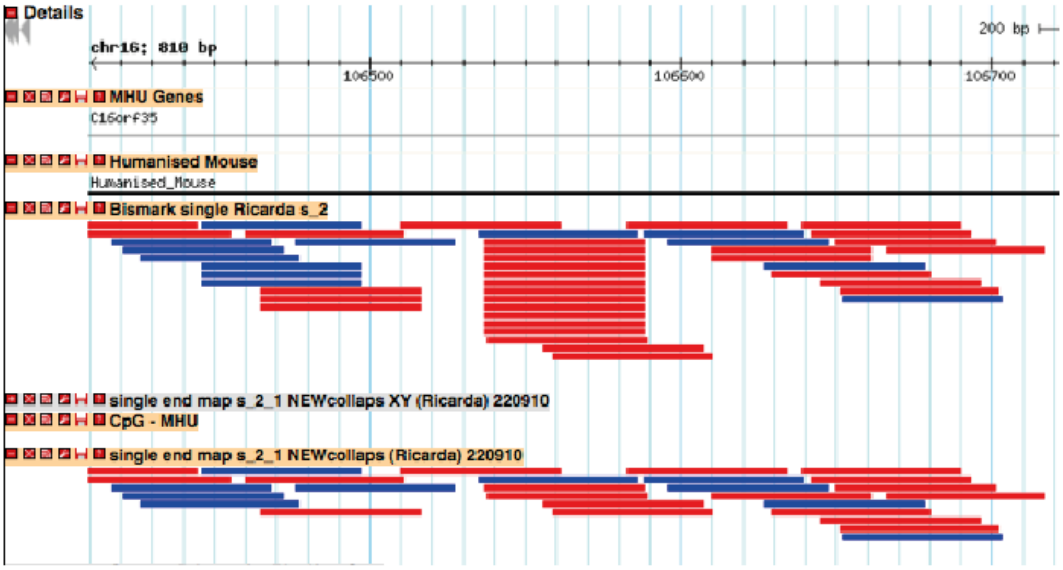

Figure 49: Example locus of read sequences before and after collapsing of data.
Forward reads (red) and reverse reads (blue) are shown before and after collapsing of the data. Identical sequence reads that start and end at the same position were considered as clonal variants and only considered once.

The stranded results coming from Methylation Extractor were visualised using SeqMonk. SeqMonk was also used to create a table with the DNA methylation level for individual CpG sites. The DNA methylation levels were calculated as the percentage of Cs of all reads (sum of Cs and Ts) at a given CpG site:

$$number\ of\ Cs * 100/(number\ of\ C's + number\ of\ Ts)$$

7.1.1 Coverage of CpGs in ESCs and brain using TBS

TBS was used to interrogate DNA methylation levels across the human transgenic region in humanised mouse ESCs and brain and compared to human ESCs and human brain, respectively. The DNA fragments were sequenced from both ends (paired-end sequencing) 100 nucleotides into the fragments yielding approximately 60 million reads per lane (and sample) (Table 10).

Table 10: Mapping efficiency of the TBS data.

	hES	hmES	hbrain	hmbrain
Total no of sequence reads	57,023,574	58,629,652	38,312,907	38,452,225
No of reads that map to the human genome	1,308,345 (2.3%)	921,683 (1.6%)	2,306,876 (6%)	7,411,573 (19.3%)
No of reads that map to the captured region	348,545	429,603	130,521	75,392
No of reads that map to the captured region (collapsed)	n.d.	30,090	16,858	10,865
Size of the captured region	78kb			
Total no of CpG sites within capture region	2,233			
No of CpGs within capture region that are covered by >1 read	1,486	1,486	1,099	1,078
No of CpGs within capture region that are covered by >10 read	949	981	540	327
Common CpG sites within two samples	785		247	
Common CpG sites within all four samples	236			

7.1.1.1 ESCs

Of the 57,023,574 and 58,629,652 reads generated for human and humanised mouse ESCs, respectively, 1,308,345 reads have mapped to the human genome in the case of human ESCs (2.3% mapping efficiency) and 921,683 reads have mapped to the human genome in the case of humanised mouse ESCs (1.6% mapping efficiency). Of these about 300'000 reads (348,545 and 429,603 for human and humanised mouse ESCs) map to the captured region. To exclude biases arising from clonal variation, the data was collapsed (as explained in the previous section), resulting in approximately 30'000 reads that map uniquely in the captured area.

The area of capture is estimated to 78kb harbouring approximately 2'233 CpG sites[21]. To get an estimation of how many CpG sites are covered using this mapping approach, the number of CpG sites that are covered by at least one read was calculated. In both human and humanised mouse ESCs, 1'486 CpG

[21]The capture area was estimated to range from (give position). This corresponds to the start position and end position of the first and last oligo in the library, respectively, and extending it by 500 bp on either side (because the DNA fragments captured with this library were on average 500 bp in length).

sites of the 2'233 CpG sites within the captured area are covered by at least one read (Table 10). CpGs with ten or more reads were allowed for further DNA methylation analysis. In human ESCs, 949 CpG sites fulfilled this criterion; in humanised mouse ESCs, 981 sites were obtained. 788 CpGs are shared between the two datasets allowing direct comparison of the methylation status of these sites. Hence, TBS yielded in methylation data in roughly half of all CpG sites in the capture area in ESCs and methylation levels of 35% of all CpG sites within the captured region could be directly compared to each other in human and humanised mouse ESCs.

7.1.1.2 Brain

In the case of brain samples, out of the approximately 60 million reads generated in each sample only 38,312,907 and 38,452,225 million reads were used for downstream mapping in human and humanised mouse brain, respectively. This is because the reverse reads did not satisfy the FastQC parameters due to low sequence quality and had to be discarded. Hence, all only forward reads were used for downstream analysis. In human brain, 2,306,876 reads (6% of all analysed reads) and in humanised mouse brain 7,411,573 reads (19% of all analysed reads) mapped uniquely to the human genome. In human and humanised mouse brain 130,521 and 75,392 reads, respectively, mapped uniquely to the captured area. About 10% of these remain after collapsing of the data (16,858 and 10,865 reads non-identical sequence reads in human and humanised mouse brain, respectively).

Due to the lower number of sequences analysed for human and humanised mouse brain, the mapping resulted in 1099 and 1078 CpG sites covered by at

least one sequence read, and 540 and 327 that are covered by at least 10 reads in human and humanised mouse brain, respectively. 247 CpG sites overlap in brain and could be directly compared to each other. In all four samples, human and humanised mouse ES and brain, only 236 are overlapping (Table 10). Hence, only 10% of all 2'233 CpG sites within the captured area could be directly compared to each other in all four samples.

Due to a relatively low coverage of CpG sites in human and humanised mouse brain, this data set is ignored in the following analysis and only the ESC data set will be used to assess any biases of TBS towards capturing certain DNA sequences.

7.1.2 Testing TBS for biases in the efficiency to capture different types of DNA sequences

The TBS technique involves several steps, any of which might introduce biases. The efficiency of capture of different types of DNA sequences was evaluated to examine if there are sequences which are underrepresented rendering biases of the technique.

7.1.2.1 Capture of DNA with high GC content

The DNA sequence capture method relies on a DNA denaturing step. The efficiency of DNA denaturation is dependent on the GC content of the underlying sequence, because GC base pairs with three hydrogen bridges form a stronger bond than AT base pairs, which are held together via only two hydrogen bonds. Hence, the GC content of DNA might have an effect on the

capture efficiency of TBS with DNA sequences of low GC content being pulled down more efficiently than DNA with a high GC content.

To assess the efficiency of the DNA sequence capture method to pull down DNA of increasing GC content the sequence read counts were plotted over the GC content (Figure 50, top panel). In the scatter plot, each dot represents an individual CpG site. The GC content of 200 bp of DNA surrounding the CpG site is shown on the x-axis and the read count of each CpG site is shown on the y-axis. Red dots represent CpG sites that are located within CGIs, whereas blue dots are CpG sites outside of CGIs (Figure 50, top panel). From this plot it becomes apparent that most CpG sites that lie within regions of very high GC content (above 80% GCs in 200 bp) are represented by less than 10 reads. Hence, sequences with a GC content of more than 80% are not efficiently captured using this method.

To better visualise how many of the CpG sites within DNA regions of differing GC content are covered by more than 10 reads, the data was stratified according to CpG sites within DNA of differing GC content (Figure 50, bottom panel). Most of the CpG sites within GC-rich DNA (80-100%) is not captured (0 % and 2% of CpG sites in human and humanised mouse ESCs are covered by less than 10 reads). On the other hand, about half of the CpG sites within areas of 20 to 80% GC content are covered by sufficient reads to calculate DNA methylation levels (i.e. >10 reads). This is comparable to the overall capture efficiency across the captured area (949 and 981 of 2233 CpG sites for human and humanised mouse, respectively). The capture efficiency for GC poor regions (20-40%) is slightly lower (23% and 40% for human and

humanised mouse ESCs). In conclusion, the capture method is not suitable for analysis of regions of very high GC content (80-95 % GC in 200 bp, Figure 50).

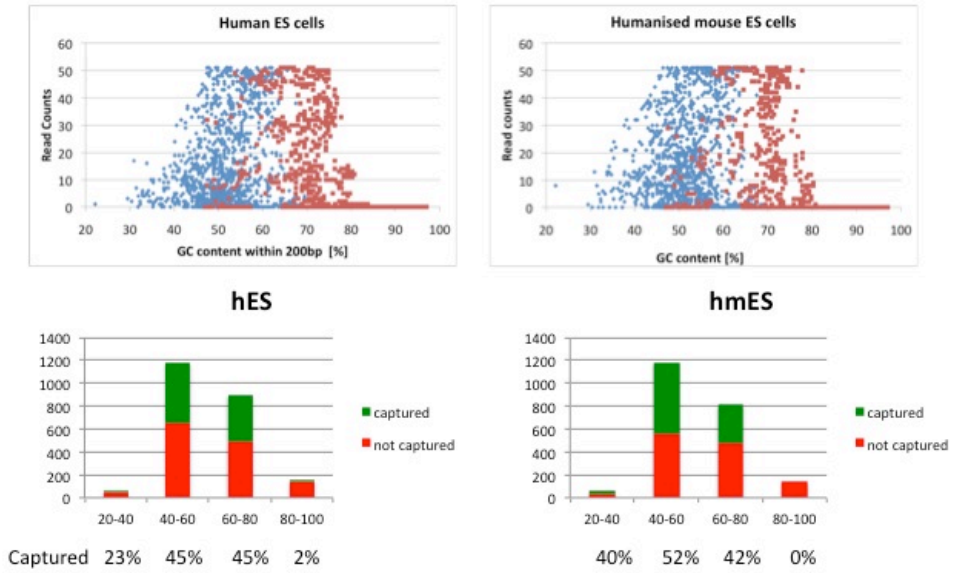

Figure 50: Efficiency of the method to capture CpG sites of high GC content.

The top panel shows a scatterplot with each dot representing a single CpG site. The number of reads of each CpG site is shown on the y-axis, and the GC content/200 bp on the x-axis. CpG sites that are located in **CGIs** are shown in red, those that are located outside of CGIs are shown in blue. It becomes apparent that CpGs that lie within regions of very high GC content (80-100%) are not efficiently captured with this method. The lower panel shows the same data set as a bar graph and the data stratified according to GC content. The bar graph helps to better visualise the fraction of non-captured DNA. The percentage of non-captured CpG sites is indicated below the graph.

7.2 Gel loading controls

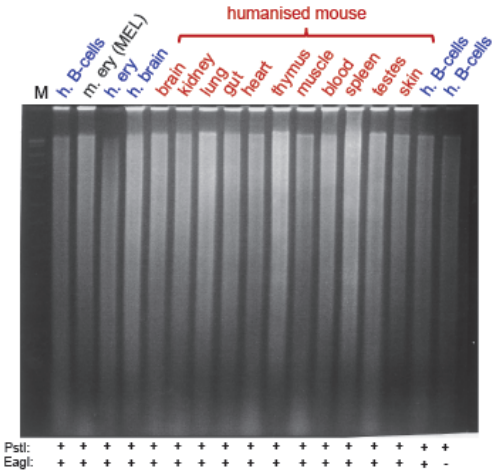

Figure 51: Ethidium bromide-stained gel corresponding to Southern shown in Figure 5.

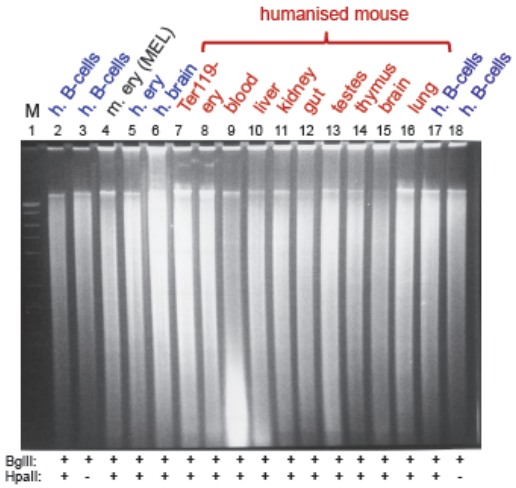

Figure 52: Ethidium bromide-stained gel corresponding to Southern shown in Figure 6.

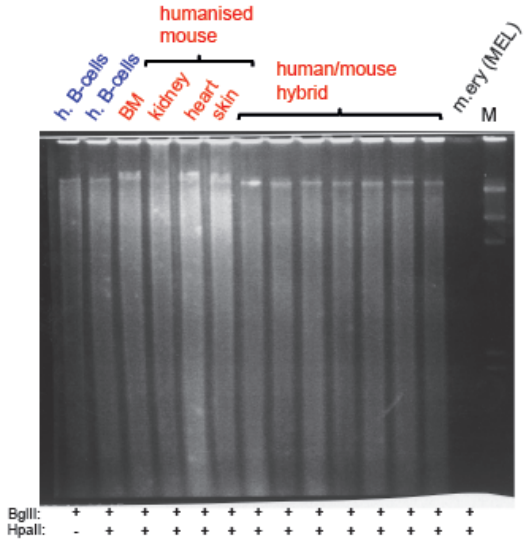

Figure 53: Ethidium bromide-stained gel corresponding to Southern shown in Figure 7.

7.3 Differentiation stage of human and mouse erythroblasts

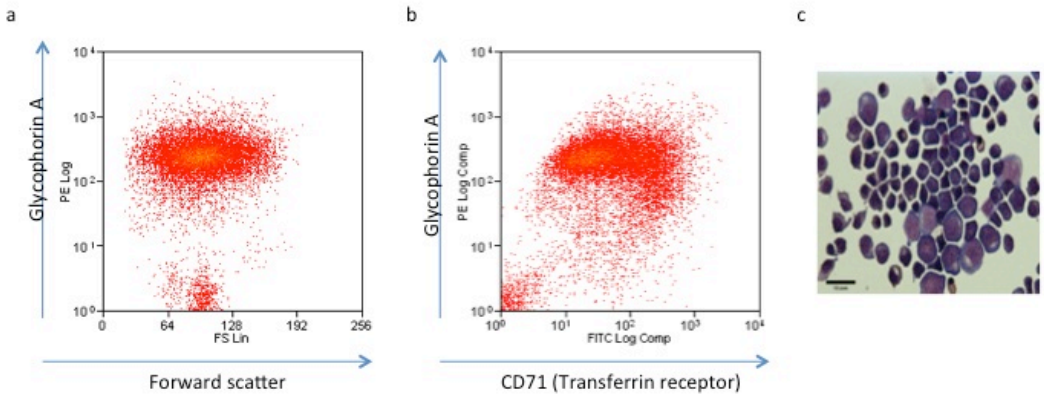

Figure 54: Differentiation stage of human erythroblasts.
a, b) Flow cytometry analysis of human Fibach cultures (phase II day 12) used in this study. c) Giemsa-stained touch preparations (c) is shown for representative human Fibach culture (phase II day 12).

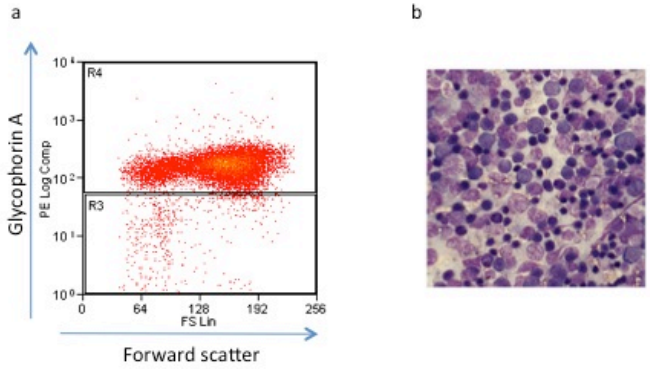

Figure 55: Differentiation stage of humanised mouse erythroblasts.
a) Flow cytometry analysis of humanized mouse Ter119+ erythroid cells used in this study. (b) Giemsa-stained touch preparations of representative spleens of mice from which erythroid cells with a similar Ter119+ expression level have been isolated (taken from (Vernimmen et al, 2009)).

7.4 List of primers

Primer sequences are given in 5' to 3' direction. All primers were purchased from Sigma or Thermo.

Table 11: List of primers used to amplify across the human *α*-globin genes (CGI G, H, I, J).
These primers are designed for bisuflite converted DNA. The location of the primer sequences refers to human chromosome 16 (hg18).

primer name		location	sequence
G1	F	152687	AGTTTTAGTGAAGTGAGGAATGGA
	R	152943	AAACCCTCAAATAAAACTAAAAACTA
G4	F	153840	GGAGGGTAGGGTAAGGAGG
	R	154077	CTCCCTAACCCCTCCCA
G5	F	154060	TGGGAGGGGTTAGGGAG
	R	154297	CCCAAACCTCTACAACTCCC
G6	F	154320	GGAGGGGTAGTTGTTTGG
	R	154580	ACCACCACCTTAAAACC
G9	F	155039	TTGTTTGTTGGTTATTTTGGT
	R	155295	CTCCTTCCCTCCAATACTAA
G10	F	155295	GGAGAAGTTGAGTGATGGGT
	R	155478	ACTCAAAAACCCCAAAAAAC
G11	F	155459	GTTTTTTGGGGTTTTTGAGT
	R	155673	ACAAAAAACCCCCTACTCTCT
G12	F	155653	AGAGAGTAGGGGGTTTTTTGT
	R	155816	TCCCAAAATTTCTAACATTAACC
G13	F	155795	GGTTAATGTTAGAAATTTTGGG
	R	156049	CAATCAAATCCCAAACCTAC
G14	F	156049	GTAGGTTTGGGATTTGATTG
	R	156265	CCTACAAAACCAAATCAATAACC
G15	F	156245	GGTTATTGATTTGGTTTTGTAGG
	R	156530	CCAAAACCCCCCCAC
G16	F	156522	GGGTTTTGGGGGTTTTTAG
	R	156782	CCACACAAACAAAAACAAATAATTC
G17	F	156773	GTTTGTGTGGTTTTTGGGGAG
	R	156958	ACCACCCTCCCTACCCCC
G18	F	156911	GATGGGGTAGAGGGAGTATTG
	R	157087	CCAAAATTAACATCTACAAATCCC
G19	F	157096	GTAGAGAGATGTAGAAATTGAGG
	R	157297	CCACTAAAAACTTCCTCCC
G20	F	157258	TGTAGTTTTGGGGATGAGGT
	R	157476	CCCTACAAACACCCACCCTA
G21	F	157457	TAGGGTGGGTGTTTGTAGGG
	R	157656	CCCACTCTTTCCTCAACCC
G22	F	157715	GTGTGGGGAGTATTGGATTAGG
	R	157900	CCAAATTCTCCAAACTACC
G23	F	157778	GAAGAGAGTTGTTGGAAGTGGAG
	R	158029	CCAAAACTATACCACTACACTCCA
G24	F	158007	GGAGTGTAGTGGTATAGTTTTGG
	R	158237	CCTATCATCCCAACACTTTAA
G25	F	158217	TTAAAGTGTTGGGATGATAGG
	R	158476	CTCTCCACATAATAACCTTCCC
G26	F	158455	GGGAAGGTTATTATGTGGAGAG
	R	158734	ACCAACTTTCTCCCAAACC
G27	F	158736	GGTTTGGGAGAAAGTTGGT

	R	158946	AACTAAACCATAACTCAAAACCA
G28	F	158924	TGGTTTTGAGTTATGGTTTAGTT
	R	159182	CCACCTATACTCCACCTCCC
G29	F	159303	TGTTTTTGTGAGTGTTGTGTTGATT
	R	159507	AACACCTCCACTATTAACACATTCC
G30	F	159636	GAGGTGTTAGGTTTGTAGGTAGTGG
	R	159888	CACTCCAACCACCTACCCTCTAAC
G31	F	159934	GAGAGTAGTTTAGTTAGTGGGGAAGG
	R	160219	CCCAACATAAATTCCAAAACTTC
G32	F	160197	GAAGTTTTGGAATTTATGTTGGG
	R	160352	CCTCTCCACCCTACCCA
G33	F	160336	TGGGTAGGGTGGAGAGG
	R	160593	AATCAAACTAATCTCAAACTCCC
G34	F	160578	TGAGATTAGTTTGATTAATATGGAGA
	R	160735	TAATCTCAACTCACCTCAACCTC
G35	F	160666	GTTATTAGGAAGGTTGAGGTAGGAG
	R	160894	AACACTCTCCTACTTTAAATAACAC
G37	F	161253	GTTATTTAGGTTGGAGTGTAGTGG
	R	161501	CCTATAATCCCAACACATTAAAAAAC
G38	F	161344	GTAGTTGAGTAGTTGGGATTATAGG
	R	161583	TACCACAACACTCTAACCTAAC
G39	F	161562	GTTAGGTTAGAGTGTTGTGGTA
	R	161793	CCCAACACTTTAAAAAACC
G40	F	161711	GGGGTTTTTTTATGTTGG
	R	161861	TACCAATAAATACATCCCTACTC
G41	F	161839	GAGTAGGGATGTATTTATTGGTA
	R	162103	TAACTAACTCTCCCACTAACCT
G42	F	162046	GGAGAGTTGGGGGTGTTATAGTG
	R	162308	CTTCCTACAACTCCCTTTCCCTC
H43	F	162175	GTAGAGGAGTTTGAATTTGGAG
	R	162328	CCTCCTACTCTCCAACCTC
G44	F	162310	GAGGTTGGAGAGTAGGAGG
	R	162557	CCACCCTTTCCTTTC
G45	F	162543	GAAAGGAAAGGGTGG
	R	162724	ACCTCCCCTAAAAATACAC
G46	F	162670	GGGTGGAGGGTGGAGA
	R	162920	CAACACCATAATAAATTCTCTCTAAATC
G47	F	162893	GATTTAGAGAGAATTTATTATGGTGTTG
	R	163012	CCTCACCTCTCCAAAACCTCC
G48	F	163182	GGTTTTGTTTAGGTTAAGGGTTA
	R	163333	CTCACCTTAAAATTAACC
G49	F	163316	GGTTAATTTTAAGGTGAG
	R	163495	TCACCAACAAACAATAACTTA
G50	F	163483	TGTTTGTTGGTGATTTTGG
	R	163710	CTACTACCCACTCAAACTTTATTCA
G51	F	163702	GGGTAGTAGTTTGTGTGTGTTTGGG
	R	163844	CTTCCCCTCCCATACTCCCTAC
G53	F	164130	GTAGGTGGTTGGAGTGAGG
	R	164342	CACACAATAAATTCTAAATACACC
G54	F	164320	GGTGTATTTAGAATTTATTGTGTG
	R	164482	CAACTTCCCAACATAAATTCC
G56	F	164802	GAGGTGGGAGGATTATTTGAGGT
	R	164986	CTTAACTCACTACAACCTCTACCTCCT
G58	F	165518	TTTTAATTATAGGGTTTTTGTTT
	R	165708	CAAACCTCATCACAAACCC
G59	F	165726	GTGTGATGGTTAGAGGATTG
	R	165921	CTCACAATTTATATAACTAACTCTCC

G60	**F**	165921	GTTTATGATTTGGGGTTTAGTTAG
	R	166133	CCTCCTACTCTCCAACCTC
G62	**F**	166348	GAAAGGAAAGGGTGG
	R	166529	ACCTCCCCTAAAAATACAC
G63	**F**	166511	GTGTATTTTTAGGGGAGGT
	R	166722	CACCATAATAAATTCTCTCTAAATC
G64	**F**	166698	GATTTAGAGAGAATTTATTATGGTG
	R	166814	CACCTCTCCAAAACCTCC
G65	**F**	166797	GGAGGTTTTGGAGAGGTGA
	R	167009	TAACCCTTAACCTAAACAAAACC
G66	**F**	166987	GGTTTTGTTTAGGTTAAGGG
	R	167137	CTCACCTTAAAATTAACC
G67	**F**	167121	GGTTAATTTTAAGGTGAG
	R	167313	CCAAAATCACCAACAAAC
G68	**F**	167289	GTTATTGTTTGTTGGTGATTTTGG
	R	167515	CCACTCAAACTTTATTCAAAAACCAC
G69	**F**	167490	GTGGTTTTTGAATAAAGTTTGAGTGG
	R	167662	CCTCCTCCATCCCCTCCTC
G70	**F**	167644	GAGGAGGGGATGGAGGAGG
	R	167842	CCAACCAATTCTTAACACAAAAATACC
G71	**F**	167823	TTGTGTTAAGAATTGGTTGG
	R	168033	ACCTAAAACCTACAATTACACCC
G72	**F**	168011	GGGTGTAATTGTAGGTTTTAGGT
	R	168241	CTTCTAACCAAACCTCTATAACCC
G73	**F**	168286	GTAGTGTTTGGGGTATTAGAGGG
	R	168520	AAATCAAACCTCCATTCTTCC
G74	**F**	168542	GTATAGTAGAGGGTTGGTTGGAGTG
	R	168747	CCTAACTTCAAATAATCCACCTACC
G76	**F**	168965	GGTGGTTGGGTATAGTGGTTTA
	R	169165	ATTCTCCTACCTCAACCTCCC
G77	**F**	169173	ATTTAGGAGGTAGAGGTTGTAGTG
	R	169433	CATACAATTTATAACCTTTAATATCC
G78	**F**	169408	GGATATTAAAGGTTATAAATTGTATG
	R	169582	CAACCATCACCACTATCTAAATC
G79	**F**	169762	GTTTATTGTTTTGAAGAAATATTTG
	R	170027	CCTCAACCTCCCAAAATTC
G80	**F**	170106	GGTATGGTGGTATGTATTTGTG
	R	170301	CCCTAAATCCAAAACCCC
G82	**F**	170285	GGGTTTTGGATTTAGGG
	R	170544	CTACCCAACTTCTTCCACA

Table 12: List of primers used to amplify across CGI E (*NPRL3*).

These primers are designed for bisuflite converted DNA. The location of the primer sequences refers to human chromosome 16 (hg18).

name		location	sequence
NP_1	Fa	126349	GGTAGAATAGTGTGTATGGTGTGA
	Ra	126607	CAATTCTCCTACCTCAACCTC
NP_2	Fa	126591	TGAGGTAGGAGAATTGTTTGATAGG
	Ra	126770	CTATTCCCATTTTTCTATATCTTTCCA
NP_3	Fa	126797	GTTGGGGGTTAATAGTAGGAAAG
	Ra	127030	TTACCATATTAACCAAACTAATCTC
NP_4	Fa	127103	TAGTTATTTGGGAGGTTGAGGTAGGAG
	Ra	127334	TCAACCTCACCCTTAACTACAAACTACTAC

NP_5	Fa	127304	GTAGTAGTTTGTAGTTAAGGGTGAGGTTG
	Ra	127525	TACCACCCCCAATTCTAACAAAC
NP_6	Fa	127512	ATTGGGGGTGGTAAAGATA
	Ra	127771	CCAACATAATTAAAACTCTATCTCTAC
NP_7	Fa	127753	GAGTTTTAATTATGTTGGTTAGGTTGG
	Ra	127962	CAACTACTACTAACATTTCATCAATCTCC
NP_8	Fa	127960	TGTTTTGGTGTGGTGAATAATGTATG
	Ra	128211	CAATAAACTACTATTCAAATACCCCTTCC
NP_9	Fb	128188	GGTATTTGAATAGTAGTTTATTGTTTTTG
	Rb	128405	CTCCTCAAACCCCCTAC
NP_11	Fa	128586	GGTTTGAGGAGGATAGGG
	Rb	128782	ATAACTTCCAAAAAACCCC
NP_12	Fa	128763	GGGGTTTTTTGGAAGTTAT
	Ra	128970	ACCTCCCAAAATACTAAAATTAC
NP_13	Fb	128947	GTAATTTTAGTATTTTGGGAG
	Rb	129142	TAATCTTAACTCAATACAACCTC
NP_14	Fb	129118	GGAGGTTGTATTGAGTTAAG
	Rd	129335	CACCATATTAACCAAAATAATCTC
NP_15	Fc	129311	GAGATTATTTTGGTTAATATGGTG
	Rb	129571	ATCTCAAAAATCTCCTATC
NP_16	Fa	129554	TAGGAGATTTTTGAGATTTTGGG
	Ra	129735	ATCTAAATCTCCTAATCTCATAATCTACC
NP_17	Fa	129725	AGATTTAGATTATTTTGGTTAATATGG
	Ra	129911	CCAAACCAAAATTCAATAACAC
NP_18	Fa	129889	GTGTTATTGAATTTTGGTTTGG
	Ra	130056	TCCATTCTAAATACAATTCCCTTATC
NP_19	Fa	130152	GAAGATGTTTAGTTTAATTGGTAGAGG
	Ra	130367	TTTCTCCACACCCCTACTAAC
NP_20	Fa	130346	GTTAGTAGGGGTGTGGAGAA
	Ra	130584	ATACCCTAAAATTACTTCCACTTTAAAC
NP_21	Fa	130600	GTGAATATGGTATAGTTTGTTTAAGG
	Ra	130833	AACATTTTCATCACTCCCCTTC
NP_22	Fa	130800	ATGGGTTTTTTGAAGGG
	Ra	131001	TCACTATCTTAACCAAAATAATCTC
NP_23	Fa	130976	GAGATTATTTTGGTTAAGATAGTGA
	Ra	131220	ATCCACTTCAAATATACACTTCC

Table 13: List of primers used to amplify across CGI E (*NPRL3*).

These primers are designed for bisuflite converted DNA. The location of the primer sequences refers to human chromosome 16 (hg18).

name		location	sequence
102-104k_1	Fb	102057	GGTTGAAAGTAAAGAATTGGAAATTG
	Rb	102227	CTTTCTTCCCACACACTACTAC
102-104k_2	Fb	102196	GGTGATTATGTAGTAGTGTGTGGG
	Ra	102432	ATAAAATCCTCACAACAACTTC

102-104k_3	Fa	102410	GAAGTTGTTGTGAGGATTTTAT
	Ra	102633	CCATAACTAATAATAAATACAACAAAAC
102-104k_4	Fa	102605	GTTTTGTTGTATTTATTATTAGTTATGG
	Ra	102835	CTATATCCAACTATATCCCTATACACAC
102-104k_5	Fa	102807	GTGTGTATAGGGATATAGTTGGATATAG
	Ra	103026	TACTATTCAAACAACCACTCCCA
102-104k_6	Fb	103112	GGATTGAATGGAAGAAATTAG
	Rb	103318	CAATTAACCTAATACCTCTCTTTTTATACC

Table 14: PCR primers used for nested PCR.

These primers are designed for bisuflite converted DNA. The location of the primer sequences refers to human chromosome 16 (hg18). Primer pairs F1/R1 and F2/R2 were used in first and second rounds of PCRs.

Primer name	location	F/R	sequence
1	150917-151399	F1	TTGGTGTTGTGGGTTTGTT
		R1	AACTCAATATAAATACCCTTCCTAC
		F2	GTTTTGGGGTTGGTGGT
		R2	ATATAAATACCCTTCCTACCCTC
2	159054-159482	F1	AGATTAAGAAATAATGTAAGTAGGTGG
		R1	TACAATTCTCCCTCCCCAAC
		F2	TGATATGTTTAATGATGTGTTTGAG
		R2	CCCCAACCCCATACAACTAC
3	161285-161619	F1	TTTTAGTGAAGTGTTTTTTTTGTA
		R1	AAAAATAAAAACTAAAAAATCCTT
		F2	TTAGGTTGGAGTGTAGTGGTAT
		R2	AACCAAAAAATTAAAAATTACAA
	161609-162057	F1	ATTGTAATTTTTAATTTTTTGGTTTA
		R1	TCACTATAACACCCCCAACTCT
		F2	ATTTTTTGGTTTAAAGGATTTTTTA
		R2	CCCCAACTCTCCAAATACTATCT
4	169357-169577	F1	GGTTGAGGTAGGAGAATTATTTG
		R1	AACACTCACTCTACTATACCCCTTA
		F2	ATTGATAGGATGTAGTTGTATTTTG
		R2	ATCACCACTATCTAAATCAAAACTT
	169732-170160 **(4b)**	F1	TTATTTTTGGGTTATATTTTAGGTAT
		R1	AATCTCTACTAAAAACACAAAAATTAAC
		F2	TTTGGGTTATATTTTAGGTATTTAAT
		R2	TTCTCCTACCTCAACCTCCC
	170141-170542	F1	TGGTATGGTGGTATGTATTTGTG
		R1	CACCTTTCCAAAACCTCTATC
		F2	GGGAGGTTGAGGTAGGAGAA
		R2	ACCCAACTTCTTCCACAAAAC

Table 15: PCR primers used for amplification of mouse DNA.

These primers are designed for bisuflite converted DNA.

primer name		sequence
5'mBP_1	F	TTGTGTAGTTTTGGTTGTTTTGGA
	R	CTCTAATTCAAAAATCTACCTATCTCTACC
5'mBP_2	F	GAAGGAAGAAGTTGAAGTTTATTTAGGT

3'mBP	R	TCTTCTCCCTATCCCTCCTACACTCCTTACC
	F	TTGATATGGGAAAGTAGATTTAGTT
	R	CAAACCTCTATCAATCCCACCC

Table 16: List of primers used for amplification of non-converted DNA.

name	sequence
M13 F	GTTTTCCCAGTCACGACG
M13 R	CAGGAAACAGCTATGACCAT
Colony PCR F	GCTATTACGCCAGCTGGCGAAAGGGGGATGTG
Colony PCR R	CCCCAGGCTTTACACTTTATGCTTCCGGCTCG
RG3A	GTTTCCCAGCTGAAGTCNNNNNN
RG4	GTTTCCCAGCTGAAGTC

7.5 Bibliography

Aapola U, Kawasaki K, Scott HS, Ollila J, Vihinen M, Heino M, Shintani A, Kawasaki K, Minoshima S, Krohn K, Antonarakis SE, Shimizu N, Kudoh J, Peterson P (2000) Isolation and initial characterization of a novel zinc finger gene, DNMT3L, on 21q22.3, related to the cytosine-5-methyltransferase 3 gene family. *Genomics* **65:** 293-298

Alami R, Greally JM, Tanimoto K, Hwang S, Feng YQ, Engel JD, Fiering S, Bouhassira EE (2000) Beta-globin YAC transgenes exhibit uniform expression levels but position effect variegation in mice. *Human molecular genetics* **9:** 631-636

Allen MD, Grummitt CG, Hilcenko C, Min SY, Tonkin LM, Johnson CM, Freund SM, Bycroft M, Warren AJ (2006) Solution structure of the nonmethyl-CpG-binding CXXC domain of the leukaemia-associated MLL histone methyltransferase. *The EMBO journal* **25:** 4503-4512

Anguita E, Hughes J, Heyworth C, Blobel GA, Wood WG, Higgs DR (2004) Globin gene activation during haemopoiesis is driven by protein complexes nucleated by GATA-1 and GATA-2. *The EMBO journal* **23:** 2841-2852

Anguita E, Sharpe JA, Sloane-Stanley JA, Tufarelli C, Higgs DR, Wood WG (2002) Deletion of the mouse alpha-globin regulatory element (HS -26) has an unexpectedly mild phenotype. *Blood* **100:** 3450-3456

Antequera F, Bird A (1993) Number of CpG islands and genes in human and mouse. *Proceedings of the National Academy of Sciences of the United States of America* **90:** 11995-11999

Arima T, Hata K, Tanaka S, Kusumi M, Li E, Kato K, Shiota K, Sasaki H, Wake N (2006) Loss of the maternal imprint in Dnmt3Lmat-/- mice leads to a differentiation defect in the extraembryonic tissue. *Developmental biology* **297:** 361-373

Arita K, Ariyoshi M, Tochio H, Nakamura Y, Shirakawa M (2008) Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. *Nature* **455:** 818-821

Augui S, Nora EP, Heard E (2011) Regulation of X-chromosome inactivation by the X-inactivation centre. *Nature reviews Genetics* **12:** 429-442

Avvakumov GV, Walker JR, Xue S, Li Y, Duan S, Bronner C, Arrowsmith CH, Dhe-Paganon S (2008) Structural basis for recognition of hemi-methylated DNA by the SRA domain of human UHRF1. *Nature* **455:** 822-825

Azuara V, Perry P, Sauer S, Spivakov M, Jorgensen HF, John RM, Gouti M, Casanova M, Warnes G, Merkenschlager M, Fisher AG (2006) Chromatin signatures of pluripotent cell lines. *Nature cell biology* **8:** 532-538

Ball MP, Li JB, Gao Y, Lee JH, LeProust EM, Park IH, Xie B, Daley GQ, Church GM (2009) Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nature biotechnology* **27:** 361-368

Bannister AJ, Kouzarides T (2011) Regulation of chromatin by histone modifications. *Cell research* **21:** 381-395

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K (2007) High-resolution profiling of histone methylations in the human genome. *Cell* **129:** 823-837

Bedell MA, Jenkins NA, Copeland NG (1997) Mouse models of human disease. Part I: techniques and resources for genetic analysis in mice. *Genes & development* **11:** 1-10

Bell AC, Felsenfeld G (2000) Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature* **405:** 482-485

Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R, Schreiber SL, Lander ES (2006a) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125:** 315-326

Bernstein E, Duncan EM, Masui O, Gil J, Heard E, Allis CD (2006b) Mouse polycomb proteins bind differentially to methylated histone H3 and RNA and are enriched in facultative heterochromatin. *Molecular and cellular biology* **26:** 2560-2569

Bestor TH (2000) The DNA methyltransferases of mammals. *Human molecular genetics* **9:** 2395-2402

Bestor TH, Ingram VM (1983) Two DNA methyltransferases from murine erythroleukemia cells: purification, sequence specificity, and mode of interaction with DNA. *Proceedings of the National Academy of Sciences of the United States of America* **80:** 5559-5563

Bi X, Broach JR (2001) Chromosomal boundaries in S. cerevisiae. *Current opinion in genetics & development* **11:** 199-204

Bird A (2002) DNA methylation patterns and epigenetic memory. *Genes & development* **16:** 6-21

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

Bird AP (1980) DNA methylation and the frequency of CpG in animal DNA. *Nucleic acids research* **8:** 1499-1504

Bird AP, Taggart MH, Nicholls RD, Higgs DR (1987) Non-methylated CpG-rich islands at the human alpha-globin locus: implications for evolution of the alpha-globin pseudogene. *The EMBO journal* **6:** 999-1004

Blackledge NP, Long HK, Zhou JC, Kriaucionis S, Patient R, Klose RJ (2012) Bio-CAP: a versatile and highly sensitive technique to purify and characterise regions of non-methylated DNA. *Nucleic acids research* **40:** e32

Blackledge NP, Zhou JC, Tolstorukov MY, Farcas AM, Park PJ, Klose RJ (2010) CpG islands recruit a histone H3 lysine 36 demethylase. *Molecular cell* **38:** 179-190

Blaschke K, Ebata KT, Karimi MM, Zepeda-Martinez JA, Goyal P, Mahapatra S, Tam A, Laird DJ, Hirst M, Rao A, Lorincz MC, Ramalho-Santos M (2013) Vitamin C induces Tet-dependent DNA demethylation and a blastocyst-like state in ES cells. *Nature* **500:** 222-226

Bourc'his D, Bestor TH (2004) Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature* **431:** 96-99

Bourc'his D, Xu GL, Lin CS, Bollman B, Bestor TH (2001) Dnmt3L and the establishment of maternal genomic imprints. *Science* **294:** 2536-2539

Brandeis M, Frank D, Keshet I, Siegfried Z, Mendelsohn M, Nemes A, Temper V, Razin A, Cedar H (1994) Sp1 elements protect a CpG island from de novo methylation. *Nature* **371:** 435-438

Brown JM, Leach J, Reittie JE, Atzberger A, Lee-Prudhoe J, Wood WG, Higgs DR, Iborra FJ, Buckle VJ (2006) Coregulated human globin genes are frequently in spatial proximity when active. *The Journal of cell biology* **172:** 177-187

Callebaut I, Courvalin JC, Mornon JP (1999) The BAH (bromo-adjacent homology) domain: a link between DNA methylation, replication and transcriptional regulation. *FEBS letters* **446:** 189-193

Carrozza MJ, Li B, Florens L, Suganuma T, Swanson SK, Lee KK, Shia WJ, Anderson S, Yates J, Washburn MP, Workman JL (2005) Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell* **123:** 581-592

Chedin F, Lieber MR, Hsieh CL (2002) The DNA methyltransferase-like protein DNMT3L stimulates de novo methylation by Dnmt3a. *Proceedings of the National Academy of Sciences of the United States of America* **99:** 16916-16921

Chen T, Tsujimoto N, Li E (2004) The PWWP domain of Dnmt3a and Dnmt3b is required for directing DNA methylation to the major satellite repeats at pericentric heterochromatin. *Molecular and cellular biology* **24:** 9048-9058

Chen T, Ueda Y, Dodge JE, Wang Z, Li E (2003) Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b. *Molecular and cellular biology* **23:** 5594-5605

Chen T, Ueda Y, Xie S, Li E (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:** 38746-38754

Chen ZX, Mann JR, Hsieh CL, Riggs AD, Chedin F (2005) Physical and functional interactions between the human DNMT3L protein and members of the de novo methyltransferase family. *Journal of cellular biochemistry* **95:** 902-917

Chen ZX, Riggs AD (2005) Maintenance and regulation of DNA methylation patterns in mammals. *Biochemistry and cell biology = Biochimie et biologie cellulaire* **83:** 438-448

Chuang LS, Ian HI, Koh TW, Ng HH, Xu G, Li BF (1997) Human DNA-(cytosine-5) methyltransferase-PCNA complex as a target for p21WAF1. *Science* **277:** 1996-2000

Clouaire T, Webb S, Skene P, Illingworth R, Kerr A, Andrews R, Lee JH, Skalnik D, Bird A (2012) Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. *Genes & development* **26:** 1714-1728

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452:** 215-219

Cooper DN, Taggart MH, Bird AP (1983) Unmethylated domains in vertebrate DNA. *Nucleic acids research* **11:** 647-658

Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, Lovci MT, Morell M, O'Shea KS, Moran JV, Gage FH (2009) L1 retrotransposition in human neural progenitor cells. *Nature* **460:** 1127-1131

Coulondre C, Miller JH, Farabaugh PJ, Gilbert W (1978) Molecular basis of base substitution hotspots in Escherichia coli. *Nature* **274:** 775-780

De Gobbi M, Anguita E, Hughes J, Sloane-Stanley JA, Sharpe JA, Koch CM, Dunham I, Gibbons RJ, Wood WG, Higgs DR (2007) Tissue-specific histone modification and transcription factor binding in alpha globin gene expression. *Blood* **110:** 4503-4510

De Gobbi M, Garrick D, Lynch M, Vernimmen D, Hughes JR, Goardon N, Luc S, Lower KM, Sloane-Stanley JA, Pina C, Soneji S, Renella R, Enver T, Taylor S, Jacobsen SE, Vyas P, Gibbons RJ, Higgs DR (2011) Generation of bivalent chromatin domains during cell fate decisions. *Epigenetics & chromatin* **4:** 9

Deaton AM, Bird A (2011) CpG islands and the regulation of transcription. *Genes & development* **25:** 1010-1022

Deng J, Shoemaker R, Xie B, Gore A, LeProust EM, Antosiewicz-Bourget J, Egli D, Maherali N, Park IH, Yu J, Daley GQ, Eggan K, Hochedlinger K, Thomson J, Wang W, Gao Y, Zhang K (2009) Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nature biotechnology* **27:** 353-360

Dennis K, Fan T, Geiman T, Yan Q, Muegge K (2001) Lsh, a member of the SNF2 family, is required for genome-wide methylation. *Genes & development* **15:** 2940-2944

Dhayalan A, Rajavelu A, Rathert P, Tamas R, Jurkowska RZ, Ragozin S, Jeltsch A (2010) The Dnmt3a PWWP domain reads histone 3 lysine 36 trimethylation and guides DNA methylation. *The Journal of biological chemistry* **285:** 26114-26120

Dodge JE, Okano M, Dick F, Tsujimoto N, Chen T, Wang S, Ueda Y, Dyson N, Li E (2005) Inactivation of Dnmt3b in mouse embryonic fibroblasts results in DNA hypomethylation, chromosomal instability, and spontaneous immortalization. *The Journal of biological chemistry* **280:** 17986-17991

Doi A, Park IH, Wen B, Murakami P, Aryee MJ, Irizarry R, Herb B, Ladd-Acosta C, Rho J, Loewer S, Miller J, Schlaeger T, Daley GQ, Feinberg AP (2009) Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nature genetics* **41:** 1350-1353

Dong KB, Maksakova IA, Mohn F, Leung D, Appanah R, Lee S, Yang HW, Lam LL, Mager DL, Schubeler D, Tachibana M, Shinkai Y, Lorincz MC (2008) DNA methylation in ES cells requires the lysine methyltransferase G9a but not its catalytic activity. *The EMBO journal* **27:** 2691-2701

Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA, Haefliger C, Horton R, Howe K, Jackson DK, Kunde J, Koenig C, Liddle J, Niblett D, Otto T, Pettett R, Seemann S, Thompson C, West T, Rogers J, Olek A, Berlin K, Beck S (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature genetics* **38:** 1378-1385

Eden A, Gaudet F, Waghmare A, Jaenisch R (2003) Chromosomal instability and tumors promoted by DNA hypomethylation. *Science* **300:** 455

Egger G, Liang G, Aparicio A, Jones PA (2004) Epigenetics in human disease and prospects for epigenetic therapy. *Nature* **429:** 457-463

Ehrlich M, Gama-Sosa MA, Huang LH, Midgett RM, Kuo KC, McCune RA, Gehrke C (1982) Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic acids research* **10:** 2709-2721

Ekwattanakit S Thesis.

Engle SJ, Womer DE, Davies PM, Boivin G, Sahota A, Simmonds HA, Stambrook PJ, Tischfield JA (1996) HPRT-APRT-deficient mice are not a model for lesch-nyhan syndrome. *Human molecular genetics* **5:** 1607-1610

Erfurth FE, Popovic R, Grembecka J, Cierpicki T, Theisler C, Xia ZB, Stuart T, Diaz MO, Bushweller JH, Zeleznik-Le NJ (2008) MLL protects CpG clusters from methylation within the Hoxa9 gene, maintaining transcript expression. *Proceedings of the National Academy of Sciences of the United States of America* **105:** 7517-7522

Esteller M (2002) CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. *Oncogene* **21:** 5427-5440

Esteller M (2005) Aberrant DNA methylation as a cancer-inducing mechanism. *Annual review of pharmacology and toxicology* **45:** 629-656

Esteve PO, Chin HG, Smallwood A, Feehery GR, Gangisetty O, Karpf AR, Carey MF, Pradhan S (2006) Direct interaction between DNMT1 and G9a coordinates DNA and histone methylation during replication. *Genes & development* **20:** 3089-3103

Farcas AM, Blackledge NP, Sudbery I, Long HK, McGouran JF, Rose NR, Lee S, Sims D, Cerase A, Sheahan TW, Koseki H, Brockdorff N, Ponting CP, Kessler BM, Klose RJ (2012) KDM2B links the Polycomb Repressive Complex 1 (PRC1) to recognition of CpG islands. *eLife* **1:** e00205

Feinberg AP, Tycko B (2004) The history of cancer epigenetics. *Nature reviews Cancer* **4:** 143-153

Feldman N, Gerson A, Fang J, Li E, Zhang Y, Shinkai Y, Cedar H, Bergman Y (2006) G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis. *Nature cell biology* **8:** 188-194

Feng S, Jacobsen SE, Reik W (2010) Epigenetic reprogramming in plant and animal development. *Science* **330:** 622-627

Ferguson-Smith AC (2011) Genomic imprinting: the emergence of an epigenetic paradigm. *Nature reviews Genetics* **12:** 565-575

Ficz G, Branco MR, Seisenberger S, Santos F, Krueger F, Hore TA, Marques CJ, Andrews S, Reik W (2011) Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473:** 398-402

Ficz G, Hore TA, Santos F, Lee HJ, Dean W, Arand J, Krueger F, Oxley D, Paul YL, Walter J, Cook SJ, Andrews S, Branco MR, Reik W (2013) FGF Signaling Inhibition in ESCs Drives Rapid Genome-wide Demethylation to the Epigenetic Ground State of Pluripotency. *Cell stem cell* **13:** 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. *Nucleic acids research* **32:** 287-297

Flint J, Thomas K, Micklem G, Raynham H, Clark K, Doggett NA, King A, Higgs DR (1997) The relationship between chromosome structure and function at a human telomeric region. *Nature genetics* **15:** 252-257

Francis NJ, Kingston RE, Woodcock CL (2004) Chromatin compaction by a polycomb group protein complex. *Science* **306:** 1574-1577

Fuks F, Hurd PJ, Deplus R, Kouzarides T (2003) The DNA methyltransferases associate with HP1 and the SUV39H1 histone methyltransferase. *Nucleic acids research* **31:** 2305-2312

Futscher BW, Oshiro MM, Wozniak RJ, Holtan N, Hanigan CL, Duan H, Domann FE (2002) Role for DNA methylation in the control of cell type specific maspin expression. *Nature genetics* **31:** 175-179

Gama-Sosa MA, Wang RY, Kuo KC, Gehrke CW, Ehrlich M (1983) The 5-methylcytosine content of highly repeated sequences in human DNA. *Nucleic acids research* **11:** 3087-3095

Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. *Journal of molecular biology* **196:** 261-282

Garrick D, De Gobbi M, Lynch M, Higgs DR (2008) Switching genes on and off in haemopoiesis. *Biochemical Society transactions* **36:** 613-618

Garrick D, Fiering S, Martin DI, Whitelaw E (1998) Repeat-induced gene silencing in mammals. *Nature genetics* **18:** 56-59

Garrick D, Sharpe JA, Arkell R, Dobbie L, Smith AJ, Wood WG, Higgs DR, Gibbons RJ (2006) Loss of Atrx affects trophoblast development and the pattern of X-inactivation in extraembryonic tissues. *PLoS genetics* **2:** e58

Garrick D, Sutherland H, Robertson G, Whitelaw E (1996) Variegated expression of a globin transgene correlates with chromatin accessibility but not methylation status. *Nucleic acids research* **24:** 4902-4909

Gaudet F, Hodgson JG, Eden A, Jackson-Grusby L, Dausman J, Gray JW, Leonhardt H, Jaenisch R (2003) Induction of tumors in mice by genomic hypomethylation. *Science* **300:** 489-492

Ge YZ, Pu MT, Gowher H, Wu HP, Ding JP, Jeltsch A, Xu GL (2004) Chromatin targeting of de novo DNA methyltransferases by the PWWP domain. *The Journal of biological chemistry* **279:** 25447-25454

Geiman TM, Sankpal UT, Robertson AK, Zhao Y, Zhao Y, Robertson KD (2004) DNMT3B interacts with hSNF2H chromatin remodeling enzyme, HDACs 1 and 2, and components of the histone methylation system. *Biochemical and biophysical research communications* **318:** 544-555

Gibbons RJ, McDowell TL, Raman S, O'Rourke DM, Garrick D, Ayyub H, Higgs DR (2000) Mutations in ATRX, encoding a SWI/SNF-like protein, cause diverse changes in the pattern of DNA methylation. *Nature genetics* **24:** 368-371

Ginno PA, Lott PL, Christensen HC, Korf I, Chedin F (2012) R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. *Molecular cell* **45:** 814-825

Glass JL, Fazzari MJ, Ferguson-Smith AC, Greally JM (2009) CG dinucleotide periodicities recognized by the Dnmt3a-Dnmt3L complex are distinctive at retroelements and imprinted domains. *Mammalian genome : official journal of the International Mammalian Genome Society* **20:** 633-643

Goll MG, Bestor TH (2005) Eukaryotic cytosine methyltransferases. *Annual review of biochemistry* **74:** 481-514

Goll MG, Kirpekar F, Maggert KA, Yoder JA, Hsieh CL, Zhang X, Golic KG, Jacobsen SE, Bestor TH (2006) Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science* **311:** 395-398

Gowher H, Liebert K, Hermann A, Xu G, Jeltsch A (2005) Mechanism of stimulation of catalytic activity of Dnmt3A and Dnmt3B DNA-(cytosine-C5)-methyltransferases by Dnmt3L. *The Journal of biological chemistry* **280:** 13341-13348

Goyal R, Reinhardt R, Jeltsch A (2006) Accuracy of DNA methylation pattern preservation by the Dnmt1 methyltransferase. *Nucleic acids research* **34:** 1182-1188

Gu TP, Guo F, Yang H, Wu HP, Xu GF, Liu W, Xie ZG, Shi L, He X, Jin SG, Iqbal K, Shi YG, Deng Z, Szabo PE, Pfeifer GP, Li J, Xu GL (2011) The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. *Nature* **477:** 606-610

Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald OG, Wen B, Wu H, Liu Y, Diep D, Briem E, Zhang K, Irizarry RA, Feinberg AP (2011)

Increased methylation variation in epigenetic domains across cancer types. *Nature genetics* **43:** 768-775

Hark AT, Schoenherr CJ, Katz DJ, Ingram RS, Levorse JM, Tilghman SM (2000) CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* **405:** 486-489

Hashimoto H, Horton JR, Zhang X, Bostick M, Jacobsen SE, Cheng X (2008) The SRA domain of UHRF1 flips 5-methylcytosine out of the DNA helix. *Nature* **455:** 826-829

Hata K, Kusumi M, Yokomine T, Li E, Sasaki H (2006) Meiotic and epigenetic aberrations in Dnmt3L-deficient male germ cells. *Molecular reproduction and development* **73:** 116-122

Hata K, Okano M, Lei H, Li E (2002) Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. *Development* **129:** 1983-1993

Havecker ER, Gao X, Voytas DF (2004) The diversity of LTR retrotransposons. *Genome biology* **5:** 225

He YF, Li BZ, Li Z, Liu P, Wang Y, Tang Q, Ding J, Jia Y, Chen Z, Li L, Sun Y, Li X, Dai Q, Song CX, Zhang K, He C, Xu GL (2011) Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333:** 1303-1307

Heard E (2004) Recent advances in X-chromosome inactivation. *Current opinion in cell biology* **16:** 247-255

Heintzman ND, Ren B (2009) Finding distal regulatory elements in the human genome. *Current opinion in genetics & development* **19:** 541-549

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics* **39:** 311-318

Hellman A, Chess A (2007) Gene body-specific methylation on the active X chromosome. *Science* **315:** 1141-1143

Herman JG, Baylin SB (2003) Gene silencing in cancer in association with promoter hypermethylation. *The New England journal of medicine* **349:** 2042-2054

Hermann A, Goyal R, Jeltsch A (2004) The Dnmt1 DNA-(cytosine-C5)-methyltransferase methylates DNA processively with high preference for hemimethylated target sites. *The Journal of biological chemistry* **279:** 48350-48359

Hermann A, Schmitt S, Jeltsch A (2003) The human Dnmt2 has residual DNA-(cytosine-C5) methyltransferase activity. *The Journal of biological chemistry* **278:** 31717-31721

Higgs DR, Wood WG, Jarman AP, Sharpe J, Lida J, Pretorius IM, Ayyub H (1990) A major positive regulatory region located far upstream of the human alpha-globin gene locus. *Genes & development* **4:** 1588-1601

Hodges E, Molaro A, Dos Santos CO, Thekkat P, Song Q, Uren PJ, Park J, Butler J, Rafii S, McCombie WR, Smith AD, Hannon GJ (2011) Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. *Molecular cell* **44:** 17-28

Holliday R, Pugh JE (1975) DNA modification mechanisms and gene activity during development. *Science* **187:** 226-232

Horsley SW, Daniels RJ, Anguita E, Raynham HA, Peden JF, Villegas A, Vickers MA, Green S, Waye JS, Chui DH, Ayyub H, MacCarthy AB, Buckle VJ, Gibbons RJ, Kearney L, Higgs DR (2001) Monosomy for the most telomeric, gene-rich region of the short arm of human chromosome 16 causes minimal phenotypic effects. *European journal of human genetics : EJHG* **9:** 217-225

Huang Y, Pastor WA, Shen Y, Tahiliani M, Liu DR, Rao A (2010) The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PloS one* **5:** e8888

Hughes JR, Cheng JF, Ventress N, Prabhakar S, Clark K, Anguita E, De Gobbi M, de Jong P, Rubin E, Higgs DR (2005) Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences. *Proceedings of the National Academy of Sciences of the United States of America* **102:** 9830-9835

Hung MS, Karthikeyan N, Huang B, Koo HC, Kiger J, Shen CJ (1999) Drosophila proteins related to vertebrate DNA (5-cytosine) methyltransferases. *Proceedings of the National Academy of Sciences of the United States of America* **96:** 11940-11945

Ichiyanagi K (2012) Inhibition of MspI cleavage activity by hydroxymethylation of the CpG site: a concern for DNA modification studies using restriction endonucleases. *Epigenetics : official journal of the DNA Methylation Society* **7:** 131-136

Illingworth R, Kerr A, Desousa D, Jorgensen H, Ellis P, Stalker J, Jackson D, Clee C, Plumb R, Rogers J, Humphray S, Cox T, Langford C, Bird A (2008) A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS biology* **6:** e22

Illingworth RS, Bird AP (2009) CpG islands--'a rough guide'. *FEBS letters* **583:** 1713-1720

Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr AR, James KD, Turner DJ, Smith C, Harrison DJ, Andrews R, Bird AP (2010) Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS genetics* **6:** e1001134

Inoue A, Shen L, Dai Q, He C, Zhang Y (2011) Generation and replication-dependent dilution of 5fC and 5caC during mouse preimplantation development. *Cell research* **21:** 1670-1676

Inoue A, Zhang Y (2011) Replication-dependent loss of 5-hydroxymethylcytosine in mouse preimplantation embryos. *Science* **334:** 194

Iqbal K, Jin SG, Pfeifer GP, Szabo PE (2011) Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. *Proceedings of the National Academy of Sciences of the United States of America* **108:** 3642-3647

Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, Cui H, Gabo K, Rongione M, Webster M, Ji H, Potash JB, Sabunciyan S, Feinberg AP (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature genetics* **41:** 178-186

Ito S, D'Alessio AC, Taranova OV, Hong K, Sowers LC, Zhang Y (2010) Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **466:** 1129-1133

Jeltsch A (2010) Molecular biology. Phylogeny of methylomes. *Science* **328:** 837-838

Jenuwein T, Allis CD (2001) Translating the histone code. *Science* **293:** 1074-1080

Jeong S, Liang G, Sharma S, Lin JC, Choi SH, Han H, Yoo CB, Egger G, Yang AS, Jones PA (2009) Selective anchoring of DNA methyltransferases 3A and 3B to nucleosomes containing methylated DNA. *Molecular and cellular biology* **29:** 5366-5376

Ji H, Ehrlich LI, Seita J, Murakami P, Doi A, Lindau P, Lee H, Aryee MJ, Irizarry RA, Kim K, Rossi DJ, Inlay MA, Serwold T, Karsunky H, Ho L, Daley GQ, Weissman IL, Feinberg AP (2010) Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature* **467:** 338-342

Jia D, Jurkowska RZ, Zhang X, Jeltsch A, Cheng X (2007) Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. *Nature* **449:** 248-251

Jones PA, Baylin SB (2002) The fundamental role of epigenetic events in cancer. *Nature reviews Genetics* **3:** 415-428

Jones PA, Liang G (2009) Rethinking how DNA methylation patterns are maintained. *Nature reviews Genetics* **10:** 805-811

Joshi AA, Struhl K (2005) Eaf3 chromodomain interaction with methylated H3-K36 links histone deacetylation to Pol II elongation. *Molecular cell* **20:** 971-978

Kaneda M, Okano M, Hata K, Sado T, Tsujimoto N, Li E, Sasaki H (2004) Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. *Nature* **429:** 900-903

Karimi MM, Goyal P, Maksakova IA, Bilenky M, Leung D, Tang JX, Shinkai Y, Mager DL, Jones S, Hirst M, Lorincz MC (2011) DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. *Cell stem cell* **8:** 676-687

Kaufman RM, Pham CT, Ley TJ (1999) Transgenic analysis of a 100-kb human beta-globin cluster-containing DNA fragment propagated as a bacterial artificial chromosome. *Blood* **94:** 3178-3184

Keogh MC, Kurdistani SK, Morris SA, Ahn SH, Podolny V, Collins SR, Schuldiner M, Chin K, Punna T, Thompson NJ, Boone C, Emili A, Weissman JS, Hughes TR, Strahl BD, Grunstein M, Greenblatt JF, Buratowski S, Krogan NJ (2005) Cotranscriptional set2 methylation of histone H3 lysine 36 recruits a repressive Rpd3 complex. *Cell* **123:** 593-605

Kouzarides T (2007) Chromatin modifications and their function. *Cell* **128:** 693-705

Kriaucionis S, Heintz N (2009) The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324:** 929-930

Krueger F, Andrews SR (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27:** 1571-1572

Kuhlmann M, Borisova BE, Kaller M, Larsson P, Stach D, Na J, Eichinger L, Lyko F, Ambros V, Soderbom F, Hammann C, Nellen W (2005) Silencing of retrotransposons in Dictyostelium by DNA methylation and RNAi. *Nucleic acids research* **33:** 6405-6417

Kunert N, Marhold J, Stanke J, Stach D, Lyko F (2003) A Dnmt2-like protein mediates DNA methylation in Drosophila. *Development* **130:** 5083-5090

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R,

Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ (2001) Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921

Langmead B (2010) Aligning short sequencing reads with Bowtie. *Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]* **Chapter 11:** Unit 11 17

Latham T, Gilbert N, Ramsahoye B (2008) DNA methylation in mouse embryonic stem cells and development. *Cell and tissue research* **331:** 31-55

Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, Low HM, Kin Sung KW, Rigoutsos I, Loring J, Wei CL (2010) Dynamic changes in the human methylome during differentiation. *Genome research* **20:** 320-331

Lee JH, Voo KS, Skalnik DG (2001) Identification and characterization of the DNA binding domain of CpG-binding protein. *The Journal of biological chemistry* **276:** 44669-44676

Lehnertz B, Ueda Y, Derijck AA, Braunschweig U, Perez-Burgos L, Kubicek S, Chen T, Li E, Jenuwein T, Peters AH (2003) Suv39h-mediated histone H3 lysine 9

methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. *Current biology : CB* **13:** 1192-1200

Leonhardt H, Page AW, Weier HU, Bestor TH (1992) A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. *Cell* **71:** 865-873

Li E (2002) Chromatin modification and epigenetic reprogramming in mammalian development. *Nature reviews Genetics* **3:** 662-673

Li E, Beard C, Jaenisch R (1993) Role for DNA methylation in genomic imprinting. *Nature* **366:** 362-365

Li E, Bestor TH, Jaenisch R (1992) Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69:** 915-926

Liang G, Chan MF, Tomigahara Y, Tsai YC, Gonzales FA, Li E, Laird PW, Jones PA (2002) Cooperativity between DNA methyltransferases in the maintenance methylation of repetitive elements. *Molecular and cellular biology* **22:** 480-491

Lienert F, Wirbelauer C, Som I, Dean A, Mohn F, Schubeler D (2011) Identification of genetic elements that autonomously determine DNA methylation states. *Nature genetics* **43:** 1091-1097

Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, Lucero J, Huang Y, Dwork AJ, Schultz MD, Yu M, Tonti-Filippini J, Heyn H, Hu S, Wu JC, Rao A, Esteller M, He C, Haghighi FG, Sejnowski TJ, Behrens MM, Ecker JR (2013) Global epigenomic reconfiguration during mammalian brain development. *Science* **341:** 1237905

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462:** 315-322

Liu Y, Oakeley EJ, Sun L, Jost JP (1998) Multiple domains are involved in the targeting of the mouse DNA methyltransferase to the DNA replication foci. *Nucleic acids research* **26:** 1038-1045

Lock LF, Takagi N, Martin GR (1987) Methylation of the Hprt gene on the inactive X occurs after chromosome inactivation. *Cell* **48:** 39-46

Long HK, Blackledge NP, Klose RJ (2013a) ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection. *Biochemical Society transactions* **41:** 727-740

Long HK, Sims D, Heger A, Blackledge NP, Kutter C, Wright ML, Grutzner F, Odom DT, Patient R, Ponting CP, Klose RJ (2013b) Epigenetic conservation at gene

regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. *eLife* **2:** e00348

Lorincz MC, Dickerson DR, Schmitt M, Groudine M (2004) Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. *Nature structural & molecular biology* **11:** 1068-1075

Lorsbach RB, Moore J, Mathew S, Raimondi SC, Mukatira ST, Downing JR (2003) TET1, a member of a novel protein family, is fused to MLL in acute myeloid leukemia containing the t(10;11)(q22;q23). *Leukemia* **17:** 637-641

Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389:** 251-260

Lyko F, Ramsahoye BH, Jaenisch R (2000) DNA methylation in Drosophila melanogaster. *Nature* **408:** 538-540

Lyle R, Watanabe D, te Vruchte D, Lerchner W, Smrzka OW, Wutz A, Schageman J, Hahner L, Davies C, Barlow DP (2000) The imprinted antisense RNA at the Igf2r locus overlaps but does not imprint Mas1. *Nature genetics* **25:** 19-21

Lynch M The role of primary DNA sequence in templating chromatin states.

Lynch MD, Smith AJ, De Gobbi M, Flenley M, Hughes JR, Vernimmen D, Ayyub H, Sharpe JA, Sloane-Stanley JA, Sutherland L, Meek S, Burdon T, Gibbons RJ, Garrick D, Higgs DR (2012) An interspecies analysis reveals a key role for unmethylated CpG dinucleotides in vertebrate Polycomb complex recruitment. *The EMBO journal* **31:** 317-329

Macleod D, Ali RR, Bird A (1998) An alternative promoter in the mouse major histocompatibility complex class II I-Abeta gene: implications for the origin of CpG islands. *Molecular and cellular biology* **18:** 4433-4443

Macleod D, Charlton J, Mullins J, Bird AP (1994) Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. *Genes & development* **8:** 2282-2292

Maeshima K, Hihara S, Eltsov M (2010) Chromatin structure: does the 30-nm fibre exist in vivo? *Current opinion in cell biology* **22:** 291-297

Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y, Turecki G, Delaney A, Varhol R, Thiessen N, Shchors K, Heine VM, Rowitch DH, Xing X, Fiore C, Schillebeeckx M, Jones SJ, Haussler D, Marra MA, Hirst M, Wang T, Costello JF (2010) Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **466:** 253-257

Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, Gnirke A, Jaenisch R, Lander ES (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454:** 766-770

Mellen M, Ayata P, Dewell S, Kriaucionis S, Heintz N (2012) MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. *Cell* **151:** 1417-1430

Mertineit C, Yoder JA, Taketo T, Laird DW, Trasler JM, Bestor TH (1998) Sex-specific exons control DNA methyltransferase in mammalian germ cells. *Development* **125:** 889-897

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448:** 553-560

Mohn F, Weber M, Rebhan M, Roloff TC, Richter J, Stadler MB, Bibel M, Schubeler D (2008) Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. *Molecular cell* **30:** 755-766

Molaro A, Hodges E, Fang F, Song Q, McCombie WR, Hannon GJ, Smith AD (2011) Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell* **146:** 1029-1041

Muotri AR, Chu VT, Marchetto MC, Deng W, Moran JV, Gage FH (2005) Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* **435:** 903-910

Muotri AR, Marchetto MC, Coufal NG, Oefner R, Yeo G, Nakashima K, Gage FH (2010) L1 retrotransposition in neurons is modulated by MeCP2. *Nature* **468:** 443-446

Nady N, Lemak A, Walker JR, Avvakumov GV, Kareta MS, Achour M, Xue S, Duan S, Allali-Hassani A, Zuo X, Wang YX, Bronner C, Chedin F, Arrowsmith CH, Dhe-Paganon S (2011) Recognition of multivalent histone states associated with heterochromatin by UHRF1 protein. *The Journal of biological chemistry* **286:** 24300-24311

Nakamura T, Liu YJ, Nakashima H, Umehara H, Inoue K, Matoba S, Tachibana M, Ogura A, Shinkai Y, Nakano T (2012) PGC7 binds histone H3K9me2 to protect against conversion of 5mC to 5hmC in early embryos. *Nature* **486:** 415-419

O'Doherty A, Ruf S, Mulligan C, Hildreth V, Errington ML, Cooke S, Sesay A, Modino S, Vanes L, Hernandez D, Linehan JM, Sharpe PT, Brandner S, Bliss TV, Henderson DJ, Nizetic D, Tybulewicz VL, Fisher EM (2005) An aneuploid mouse

strain carrying human chromosome 21 with Down syndrome phenotypes. *Science* **309:** 2033-2037

Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E (2007) Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nature genetics* **39:** 730-732

Okano M, Bell DW, Haber DA, Li E (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99:** 247-257

Okano M, Xie S, Li E (1998a) Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nature genetics* **19:** 219-220

Okano M, Xie S, Li E (1998b) Dnmt2 is not required for de novo and maintenance methylation of viral DNA in embryonic stem cells. *Nucleic acids research* **26:** 2536-2540

Ono R, Taki T, Taketani T, Taniwaki M, Kobayashi H, Hayashi Y (2002) LCX, leukemia-associated protein with a CXXC domain, is fused to MLL in acute myeloid leukemia with trilineage dysplasia having t(10;11)(q22;q23). *Cancer research* **62:** 4075-4080

Ooi SK, Qiu C, Bernstein E, Li K, Jia D, Yang Z, Erdjument-Bromage H, Tempst P, Lin SP, Allis CD, Cheng X, Bestor TH (2007) DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature* **448:** 714-717

Otani J, Nankumo T, Arita K, Inamoto S, Ariyoshi M, Shirakawa M (2009) Structural basis for recognition of H3K4 methylation status by the DNA methyltransferase 3A ATRX-DNMT3-DNMT3L domain. *EMBO reports* **10:** 1235-1241

Pastor WA, Aravind L, Rao A (2013) TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. *Nature reviews Molecular cell biology* **14:** 341-356

Pastor WA, Pape UJ, Huang Y, Henderson HR, Lister R, Ko M, McLoughlin EM, Brudno Y, Mahapatra S, Kapranov P, Tahiliani M, Daley GQ, Liu XS, Ecker JR, Milos PM, Agarwal S, Rao A (2011) Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* **473:** 394-397

Peters AH, O'Carroll D, Scherthan H, Mechtler K, Sauer S, Schofer C, Weipoltshammer K, Pagani M, Lachner M, Kohlmaier A, Opravil S, Doyle M, Sibilia M, Jenuwein T (2001) Loss of the Suv39h histone methyltransferases impairs mammalian heterochromatin and genome stability. *Cell* **107:** 323-337

Ponger L, Li WH (2005) Evolutionary diversification of DNA methyltransferases in eukaryotic genomes. *Molecular biology and evolution* **22:** 1119-1128

Popp C, Dean W, Feng S, Cokus SJ, Andrews S, Pellegrini M, Jacobsen SE, Reik W (2010) Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. *Nature* **463:** 1101-1105

Pradhan M, Esteve PO, Chin HG, Samaranayke M, Kim GD, Pradhan S (2008) CXXC domain of human DNMT1 is essential for enzymatic activity. *Biochemistry* **47:** 10000-10009

Pradhan S, Bacolla A, Wells RD, Roberts RJ (1999) Recombinant human DNA (cytosine-5) methyltransferase. I. Expression, purification, and comparison of de novo and maintenance methylation. *The Journal of biological chemistry* **274:** 33002-33010

Pradhan S, Roberts RJ (2000) Hybrid mouse-prokaryotic DNA (cytosine-5) methyltransferases retain the specificity of the parental C-terminal domain. *The EMBO journal* **19:** 2103-2114

Pray-Grant MG, Daniel JA, Schieltz D, Yates JR, 3rd, Grant PA (2005) Chd1 chromodomain links histone H3 methylation with SAGA- and SLIK-dependent acetylation. *Nature* **433:** 434-438

Rai K, Chidester S, Zavala CV, Manos EJ, James SR, Karpf AR, Jones DA, Cairns BR (2007) Dnmt2 functions in the cytoplasm to promote liver, brain, and retina development in zebrafish. *Genes & development* **21:** 261-266

Rai K, Huggins IJ, James SR, Karpf AR, Jones DA, Cairns BR (2008) DNA demethylation in zebrafish involves the coupling of a deaminase, a glycosylase, and gadd45. *Cell* **135:** 1201-1212

Ramsahoye BH, Biniszkiewicz D, Lyko F, Clark V, Bird AP, Jaenisch R (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:** 5237-5242

Rauch TA, Wu X, Zhong X, Riggs AD, Pfeifer GP (2009) A human B cell methylome at 100-base pair resolution. *Proceedings of the National Academy of Sciences of the United States of America* **106:** 671-678

Rebollo R, Karimi MM, Bilenky M, Gagnier L, Miceli-Royer K, Zhang Y, Goyal P, Keane TM, Jones S, Hirst M, Lorincz MC, Mager DL (2011) Retrotransposon-induced heterochromatin spreading in the mouse revealed by insertional polymorphisms. *PLoS genetics* **7:** e1002301

Reik W, Dean W, Walter J (2001) Epigenetic reprogramming in mammalian development. *Science* **293:** 1089-1093

Riggs AD (1975) X inactivation, differentiation, and DNA methylation. *Cytogenetics and cell genetics* **14:** 9-25

Robertson G, Garrick D, Wu W, Kearns M, Martin D, Whitelaw E (1995) Position-dependent variegation of globin transgene expression in mice. *Proceedings of the National Academy of Sciences of the United States of America* **92:** 5371-5375

Robertson KD, Keyomarsi K, Gonzales FA, Velicescu M, Jones PA (2000) Differential mRNA expression of the human DNA methyltransferases (DNMTs) 1, 3a and 3b during the G(0)/G(1) to S phase transition in normal and tumor cells. *Nucleic acids research* **28:** 2108-2113

Robertson KD, Wolffe AP (2000) DNA methylation in health and disease. *Nature reviews Genetics* **1:** 11-19

Rottach A, Frauer C, Pichler G, Bonapace IM, Spada F, Leonhardt H (2010) The multi-domain protein Np95 connects DNA methylation and histone modification. *Nucleic acids research* **38:** 1796-1804

Rountree MR, Bachman KE, Baylin SB (2000) DNMT1 binds HDAC2 and a new co-repressor, DMAP1, to form a complex at replication foci. *Nature genetics* **25:** 269-277

Roy N

Sado T, Fenner MH, Tan SS, Tam P, Shioda T, Li E (2000) X inactivation in the mouse embryo deficient for Dnmt1: distinct effect of hypomethylation on imprinted and random X inactivation. *Developmental biology* **225:** 294-303

Sado T, Okano M, Li E, Sasaki H (2004) De novo DNA methylation is dispensable for the initiation and propagation of X chromosome inactivation. *Development* **131:** 975-982

Sankaran VG, Xu J, Ragoczy T, Ippolito GC, Walkley CR, Maika SD, Fujiwara Y, Ito M, Groudine M, Bender MA, Tucker PW, Orkin SH (2009) Developmental and species-divergent globin switching are driven by BCL11A. *Nature* **460:** 1093-1097

Schultz DC, Ayyanathan K, Negorev D, Maul GG, Rauscher FJ, 3rd (2002) SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. *Genes & development* **16:** 919-932

Schwartz YB, Pirrotta V (2007) Polycomb silencing mechanisms and the management of genomic programmes. *Nature reviews Genetics* **8:** 9-22

Seisenberger S, Andrews S, Krueger F, Arand J, Walter J, Santos F, Popp C, Thienpont B, Dean W, Reik W (2012) The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells. *Molecular cell* **48:** 849-862

Shann YJ, Cheng C, Chiao CH, Chen DT, Li PH, Hsu MT (2008) Genome-wide mapping and characterization of hypomethylated sites in human tissues and breast cancer cell lines. *Genome research* **18:** 791-801

Sharif J, Muto M, Takebayashi S, Suetake I, Iwamatsu A, Endo TA, Shinga J, Mizutani-Koseki Y, Toyoda T, Okamura K, Tajima S, Mitsuya K, Okano M, Koseki H (2007) The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. *Nature* **450:** 908-912

Sharma S, De Carvalho DD, Jeong S, Jones PA, Liang G (2011) Nucleosomes containing methylated DNA stabilize DNA methyltransferases 3A/3B and ensure faithful epigenetic inheritance. *PLoS genetics* **7:** e1001286

Sharp AJ, Stathaki E, Migliavacca E, Brahmachary M, Montgomery SB, Dupre Y, Antonarakis SE (2011) DNA methylation profiles of human active and inactive X chromosomes. *Genome research* **21:** 1592-1600

Shearstone JR, Pop R, Bock C, Boyle P, Meissner A, Socolovsky M (2011) Global DNA demethylation during mouse erythropoiesis in vivo. *Science* **334:** 799-802

Shen L, Kondo Y, Guo Y, Zhang J, Zhang L, Ahmed S, Shu J, Chen X, Waterland RA, Issa JP (2007) Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters. *PLoS genetics* **3:** 2023-2036

Shen L, Zhang Y (2013) 5-Hydroxymethylcytosine: generation, fate, and genomic distribution. *Current opinion in cell biology* **25:** 289-296

Smith ZD, Meissner A (2013) DNA methylation: roles in mammalian development. *Nature reviews Genetics* **14:** 204-220

Song CX, He C (2013) Potential functional roles of DNA demethylation intermediates. *Trends in biochemical sciences*

Song CX, Yi C, He C (2012a) Mapping recently identified nucleotide variants in the genome and transcriptome. *Nature biotechnology* **30:** 1107-1116

Song F, Smith JF, Kimura MT, Morrow AD, Matsuyama T, Nagase H, Held WA (2005) Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **102:** 3336-3341

Song J, Rechkoblit O, Bestor TH, Patel DJ (2011) Structure of DNMT1-DNA complex reveals a role for autoinhibition in maintenance DNA methylation. *Science* **331:** 1036-1040

Song J, Teplova M, Ishibe-Murakami S, Patel DJ (2012b) Structure-based mechanistic insights into DNMT1-mediated maintenance DNA methylation. *Science* **335:** 709-712

Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, van Nimwegen E, Wirbelauer C, Oakeley EJ, Gaidatzis D, Tiwari VK, Schubeler D (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480:** 490-495

Suzuki MM, Bird A (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nature reviews Genetics* **9:** 465-476

Syeda F, Fagan RL, Wean M, Avvakumov GV, Walker JR, Xue S, Dhe-Paganon S, Brenner C (2011) The replication focus targeting sequence (RFTS) domain is a DNA-competitive inhibitor of Dnmt1. *The Journal of biological chemistry* **286:** 15344-15351

Tachibana M, Sugimoto K, Fukushima T, Shinkai Y (2001) Set domain-containing protein, G9a, is a novel lysine-preferring mammalian histone methyltransferase with hyperactivity and specific selectivity to lysines 9 and 27 of histone H3. *The Journal of biological chemistry* **276:** 25309-25317

Tachibana M, Sugimoto K, Nozaki M, Ueda J, Ohta T, Ohki M, Fukuda M, Takeda N, Niida H, Kato H, Shinkai Y (2002) G9a histone methyltransferase plays a dominant role in euchromatic histone H3 lysine 9 methylation and is essential for early embryogenesis. *Genes & development* **16:** 1779-1791

Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, Agarwal S, Iyer LM, Liu DR, Aravind L, Rao A (2009) Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324:** 930-935

Takeshita K, Suetake I, Yamashita E, Suga M, Narita H, Nakagawa A, Tajima S (2011) Structural insight into maintenance methylation by mouse DNA methyltransferase 1 (Dnmt1). *Proceedings of the National Academy of Sciences of the United States of America* **108:** 9055-9059

Tan M, Luo H, Lee S, Jin F, Yang JS, Montellier E, Buchou T, Cheng Z, Rousseaux S, Rajagopal N, Lu Z, Ye Z, Zhu Q, Wysocka J, Ye Y, Khochbin S, Ren B, Zhao Y (2011) Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell* **146:** 1016-1028

Thomson JP, Skene PJ, Selfridge J, Clouaire T, Guy J, Webb S, Kerr AR, Deaton A, Andrews R, James KD, Turner DJ, Illingworth R, Bird A (2010) CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature* **464:** 1082-1086

Tufarelli C, Stanley JA, Garrick D, Sharpe JA, Ayyub H, Wood WG, Higgs DR (2003) Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. *Nature genetics* **34:** 157-165

Vakoc CR, Letting DL, Gheldof N, Sawado T, Bender MA, Groudine M, Weiss MJ, Dekker J, Blobel GA (2005) Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. *Molecular cell* **17:** 453-462

Vernimmen D, De Gobbi M, Sloane-Stanley JA, Wood WG, Higgs DR (2007) Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. *The EMBO journal* **26:** 2041-2051

Vernimmen D, Marques-Kranc F, Sharpe JA, Sloane-Stanley JA, Wood WG, Wallace HA, Smith AJ, Higgs DR (2009) Chromosome looping at the human alpha-globin locus is mediated via the major upstream regulatory element (HS -40). *Blood* **114:** 4253-4260

Viprakasit V, Harteveld CL, Ayyub H, Stanley JS, Giordano PC, Wood WG, Higgs DR (2006) A novel deletion causing alpha thalassemia clarifies the importance of the major human alpha globin regulatory element. *Blood* **107:** 3811-3812

Viprakasit V, Kidd AM, Ayyub H, Horsley S, Hughes J, Higgs DR (2003) De novo deletion within the telomeric region flanking the human alpha globin locus as a cause of alpha thalassaemia. *British journal of haematology* **120:** 867-875

Vyas P, Vickers MA, Simmons DL, Ayyub H, Craddock CF, Higgs DR (1992) Cis-acting sequences regulating expression of the human alpha-globin cluster lie within constitutively open chromatin. *Cell* **69:** 781-793

Wallace HA, Marques-Kranc F, Richardson M, Luna-Crespo F, Sharpe JA, Hughes J, Wood WG, Higgs DR, Smith AJ (2007) Manipulating the mouse genome to engineer precise functional syntenic replacements with human sequence. *Cell* **128:** 197-209

Walsh CP, Chaillet JR, Bestor TH (1998) Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nature genetics* **20:** 116-117

Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ, Zhao K (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature genetics* **40:** 897-903

Ward MC, Wilson MD, Barbosa-Morais NL, Schmidt D, Stark R, Pan Q, Schwalie PC, Menon S, Lukk M, Watt S, Thybert D, Kutter C, Kirschner K, Flicek P, Blencowe BJ, Odom DT (2013) Latent regulatory potential of human-specific repetitive elements. *Molecular cell* **49:** 262-272

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C,

Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520-562

Weber M, Hellmann I, Stadler MB, Ramos L, Paabo S, Rebhan M, Schubeler D (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature genetics* **39:** 457-466

Webster KE, O'Bryan MK, Fletcher S, Crewther PE, Aapola U, Craig J, Harrison DK, Aung H, Phutikanit N, Lyle R, Meachem SJ, Antonarakis SE, de Kretser DM, Hedger MP, Peterson P, Carroll BJ, Scott HS (2005) Meiotic and epigenetic defects in Dnmt3L-knockout mouse spermatogenesis. *Proceedings of the National Academy of Sciences of the United States of America* **102:** 4068-4073

Williams K, Christensen J, Pedersen MT, Johansen JV, Cloos PA, Rappsilber J, Helin K (2011) TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* **473:** 343-348

Wilson JB, Weinberg W, Johnson R, Yuspa S, Levine AJ (1990) Expression of the BNLF-1 oncogene of Epstein-Barr virus in the skin of transgenic mice induces hyperplasia and aberrant expression of keratin 6. *Cell* **61:** 1315-1327

Wilson MD, Barbosa-Morais NL, Schmidt D, Conboy CM, Vanes L, Tybulewicz VL, Fisher EM, Tavare S, Odom DT (2008) Species-specific transcription in mice carrying human chromosome 21. *Science* **322:** 434-438

Wossidlo M, Nakamura T, Lepikhov K, Marques CJ, Zakhartchenko V, Boiani M, Arand J, Nakano T, Reik W, Walter J (2011) 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. *Nature communications* **2:** 241

Wutz A, Smrzka OW, Schweifer N, Schellander K, Wagner EF, Barlow DP (1997) Imprinted expression of the Igf2r gene depends on an intronic CpG island. *Nature* **389:** 745-749

Wysocka J, Swigut T, Xiao H, Milne TA, Kwon SY, Landry J, Kauer M, Tackett AJ, Chait BT, Badenhorst P, Wu C, Allis CD (2006) A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature* **442:** 86-90

Xie S, Wang Z, Okano M, Nogami M, Li Y, He WW, Okumura K, Li E (1999) Cloning, expression and chromosome locations of the human DNMT3 gene family. *Gene* **236:** 87-95

Xu GL, Bestor TH, Bourc'his D, Hsieh CL, Tommerup N, Bugge M, Hulten M, Qu X, Russo JJ, Viegas-Pequignot E (1999) Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. *Nature* **402:** 187-191

Xu Y, Wu F, Tan L, Kong L, Xiong L, Deng J, Barbera AJ, Zheng L, Zhang H, Huang S, Min J, Nicholson T, Chen T, Xu G, Shi Y, Zhang K, Shi YG (2011) Genome-wide regulation of 5hmC, 5mC, and gene expression by Tet1 hydroxylase in mouse embryonic stem cells. *Molecular cell* **42:** 451-464

Yildirim O, Li R, Hung JH, Chen PB, Dong X, Ee LS, Weng Z, Rando OJ, Fazzio TG (2011) Mbd3/NURD complex regulates expression of 5-hydroxymethylcytosine marked genes in embryonic stem cells. *Cell* **147:** 1498-1510

Yoder JA, Bestor TH (1998) A candidate mammalian DNA methyltransferase related to pmt1p of fission yeast. *Human molecular genetics* **7:** 279-284

Zemach A, McDaniel IE, Silva P, Zilberman D (2010) Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328:** 916-919

Zentner GE, Henikoff S (2013) Regulation of nucleosome dynamics by histone modifications. *Nature structural & molecular biology* **20:** 259-266

Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW, Chen H, Henderson IR, Shinn P, Pellegrini M, Jacobsen SE, Ecker JR (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. *Cell* **126:** 1189-1201

Zhang Y, Jurkowska R, Soeroes S, Rajavelu A, Dhayalan A, Bock I, Rathert P, Brandt O, Reinhardt R, Fischle W, Jeltsch A (2010) Chromatin methylation

activity of Dnmt3a and Dnmt3a/3L is guided by interaction of the ADD domain with the histone H3 tail. *Nucleic acids research* **38:** 4246-4253

Zhang Y, Rohde C, Tierling S, Jurkowski TP, Bock C, Santacruz D, Ragozin S, Reinhardt R, Groth M, Walter J, Jeltsch A (2009) DNA methylation analysis of chromosome 21 gene promoters at single base pair and single allele resolution. *PLoS genetics* **5:** e1000438

Zhou VW, Goren A, Bernstein BE (2011) Charting histone modifications and the functional organization of mammalian genomes. *Nature reviews Genetics* **12:** 7-18

Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S (2007) Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nature genetics* **39:** 61-69

Zilberman D, Henikoff S (2007) Genome-wide analysis of DNA methylation patterns. *Development* **134:** 3959-3965

Ziller MJ, Gu H, Muller F, Donaghey J, Tsai LT, Kohlbacher O, De Jager PL, Rosen ED, Bennett DA, Bernstein BE, Gnirke A, Meissner A (2013) Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500:** 477-481

Ziller MJ, Muller F, Liao J, Zhang Y, Gu H, Bock C, Boyle P, Epstein CB, Bernstein BE, Lengauer T, Gnirke A, Meissner A (2011) Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. *PLoS genetics* **7:** e1002389

ADDIN EN.REFLIST Aapola U, Kawasaki K, Scott HS, Ollila J, Vihinen M, Heino M, Shintani A, Kawasaki K, Minoshima S, Krohn K, Antonarakis SE, Shimizu N, Kudoh J, Peterson P (2000) Isolation and initial characterization of a novel zinc finger gene, DNMT3L, on 21q22.3, related to the cytosine-5-methyltransferase 3 gene family. *Genomics* **65:** 293-298

Alami R, Greally JM, Tanimoto K, Hwang S, Feng YQ, Engel JD, Fiering S, Bouhassira EE (2000) Beta-globin YAC transgenes exhibit uniform expression levels but position effect variegation in mice. *Human molecular genetics* **9:** 631-636

Allen MD, Grummitt CG, Hilcenko C, Min SY, Tonkin LM, Johnson CM, Freund SM, Bycroft M, Warren AJ (2006) Solution structure of the nonmethyl-CpG-binding CXXC domain of the leukaemia-associated MLL histone methyltransferase. *The EMBO journal* **25:** 4503-4512

Anguita E, Hughes J, Heyworth C, Blobel GA, Wood WG, Higgs DR (2004) Globin gene activation during haemopoiesis is driven by protein complexes nucleated by GATA-1 and GATA-2. *The EMBO journal* **23:** 2841-2852

Anguita E, Sharpe JA, Sloane-Stanley JA, Tufarelli C, Higgs DR, Wood WG (2002) Deletion of the mouse alpha-globin regulatory element (HS -26) has an unexpectedly mild phenotype. *Blood* **100:** 3450-3456

Antequera F, Bird A (1993) Number of CpG islands and genes in human and mouse. *Proceedings of the National Academy of Sciences of the United States of America* **90:** 11995-11999

Arima T, Hata K, Tanaka S, Kusumi M, Li E, Kato K, Shiota K, Sasaki H, Wake N (2006) Loss of the maternal imprint in Dnmt3Lmat-/- mice leads to a differentiation defect in the extraembryonic tissue. *Developmental biology* **297:** 361-373

Arita K, Ariyoshi M, Tochio H, Nakamura Y, Shirakawa M (2008) Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. *Nature* **455:** 818-821

Augui S, Nora EP, Heard E (2011) Regulation of X-chromosome inactivation by the X-inactivation centre. *Nature reviews Genetics* **12:** 429-442

Avvakumov GV, Walker JR, Xue S, Li Y, Duan S, Bronner C, Arrowsmith CH, Dhe-Paganon S (2008) Structural basis for recognition of hemi-methylated DNA by the SRA domain of human UHRF1. *Nature* **455:** 822-825

Azuara V, Perry P, Sauer S, Spivakov M, Jorgensen HF, John RM, Gouti M, Casanova M, Warnes G, Merkenschlager M, Fisher AG (2006) Chromatin signatures of pluripotent cell lines. *Nature cell biology* **8:** 532-538

Ball MP, Li JB, Gao Y, Lee JH, LeProust EM, Park IH, Xie B, Daley GQ, Church GM (2009) Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nature biotechnology* **27:** 361-368

Bannister AJ, Kouzarides T (2011) Regulation of chromatin by histone modifications. *Cell research* **21:** 381-395

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K (2007) High-resolution profiling of histone methylations in the human genome. *Cell* **129:** 823-837

Bedell MA, Jenkins NA, Copeland NG (1997) Mouse models of human disease. Part I: techniques and resources for genetic analysis in mice. *Genes & development* **11:** 1-10

Bell AC, Felsenfeld G (2000) Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature* **405:** 482-485

Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R, Schreiber SL, Lander ES

(2006a) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125:** 315-326

Bernstein E, Duncan EM, Masui O, Gil J, Heard E, Allis CD (2006b) Mouse polycomb proteins bind differentially to methylated histone H3 and RNA and are enriched in facultative heterochromatin. *Molecular and cellular biology* **26:** 2560-2569

Bestor TH (2000) The DNA methyltransferases of mammals. *Human molecular genetics* **9:** 2395-2402

Bestor TH, Ingram VM (1983) Two DNA methyltransferases from murine erythroleukemia cells: purification, sequence specificity, and mode of interaction with DNA. *Proceedings of the National Academy of Sciences of the United States of America* **80:** 5559-5563

Bi X, Broach JR (2001) Chromosomal boundaries in S. cerevisiae. *Current opinion in genetics & development* **11:** 199-204

Bird A (2002) DNA methylation patterns and epigenetic memory. *Genes & development* **16:** 6-21

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

Bird AP (1980) DNA methylation and the frequency of CpG in animal DNA. *Nucleic acids research* **8:** 1499-1504

Bird AP, Taggart MH, Nicholls RD, Higgs DR (1987) Non-methylated CpG-rich islands at the human alpha-globin locus: implications for evolution of the alpha-globin pseudogene. *The EMBO journal* **6:** 999-1004

Blackledge NP, Long HK, Zhou JC, Kriaucionis S, Patient R, Klose RJ (2012) Bio-CAP: a versatile and highly sensitive technique to purify and characterise regions of non-methylated DNA. *Nucleic acids research* **40:** e32

Blackledge NP, Zhou JC, Tolstorukov MY, Farcas AM, Park PJ, Klose RJ (2010) CpG islands recruit a histone H3 lysine 36 demethylase. *Molecular cell* **38:** 179-190

Blaschke K, Ebata KT, Karimi MM, Zepeda-Martinez JA, Goyal P, Mahapatra S, Tam A, Laird DJ, Hirst M, Rao A, Lorincz MC, Ramalho-Santos M (2013) Vitamin C induces Tet-dependent DNA demethylation and a blastocyst-like state in ES cells. *Nature* **500:** 222-226

Bourc'his D, Bestor TH (2004) Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature* **431:** 96-99

Bourc'his D, Xu GL, Lin CS, Bollman B, Bestor TH (2001) Dnmt3L and the establishment of maternal genomic imprints. *Science* **294:** 2536-2539

Brandeis M, Frank D, Keshet I, Siegfried Z, Mendelsohn M, Nemes A, Temper V, Razin A, Cedar H (1994) Sp1 elements protect a CpG island from de novo methylation. *Nature* **371:** 435-438

Brown JM, Leach J, Reittie JE, Atzberger A, Lee-Prudhoe J, Wood WG, Higgs DR, Iborra FJ, Buckle VJ (2006) Coregulated human globin genes are frequently in spatial proximity when active. *The Journal of cell biology* **172:** 177-187

Callebaut I, Courvalin JC, Mornon JP (1999) The BAH (bromo-adjacent homology) domain: a link between DNA methylation, replication and transcriptional regulation. *FEBS letters* **446:** 189-193

Carrozza MJ, Li B, Florens L, Suganuma T, Swanson SK, Lee KK, Shia WJ, Anderson S, Yates J, Washburn MP, Workman JL (2005) Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell* **123:** 581-592

Chedin F, Lieber MR, Hsieh CL (2002) The DNA methyltransferase-like protein DNMT3L stimulates de novo methylation by Dnmt3a. *Proceedings of the National Academy of Sciences of the United States of America* **99:** 16916-16921

Chen T, Tsujimoto N, Li E (2004) The PWWP domain of Dnmt3a and Dnmt3b is required for directing DNA methylation to the major satellite repeats at pericentric heterochromatin. *Molecular and cellular biology* **24:** 9048-9058

Chen T, Ueda Y, Dodge JE, Wang Z, Li E (2003) Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b. *Molecular and cellular biology* **23:** 5594-5605

Chen T, Ueda Y, Xie S, Li E (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:** 38746-38754

Chen ZX, Mann JR, Hsieh CL, Riggs AD, Chedin F (2005) Physical and functional interactions between the human DNMT3L protein and members of the de novo methyltransferase family. *Journal of cellular biochemistry* **95:** 902-917

Chen ZX, Riggs AD (2005) Maintenance and regulation of DNA methylation patterns in mammals. *Biochemistry and cell biology = Biochimie et biologie cellulaire* **83:** 438-448

Chuang LS, Ian HI, Koh TW, Ng HH, Xu G, Li BF (1997) Human DNA-(cytosine-5) methyltransferase-PCNA complex as a target for p21WAF1. *Science* **277:** 1996-2000

Clouaire T, Webb S, Skene P, Illingworth R, Kerr A, Andrews R, Lee JH, Skalnik D, Bird A (2012) Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. *Genes & development* **26:** 1714-1728

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452:** 215-219

Cooper DN, Taggart MH, Bird AP (1983) Unmethylated domains in vertebrate DNA. *Nucleic acids research* **11:** 647-658

Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, Lovci MT, Morell M, O'Shea KS, Moran JV, Gage FH (2009) L1 retrotransposition in human neural progenitor cells. *Nature* **460:** 1127-1131

Coulondre C, Miller JH, Farabaugh PJ, Gilbert W (1978) Molecular basis of base substitution hotspots in Escherichia coli. *Nature* **274:** 775-780

De Gobbi M, Anguita E, Hughes J, Sloane-Stanley JA, Sharpe JA, Koch CM, Dunham I, Gibbons RJ, Wood WG, Higgs DR (2007) Tissue-specific histone modification and transcription factor binding in alpha globin gene expression. *Blood* **110:** 4503-4510

De Gobbi M, Garrick D, Lynch M, Vernimmen D, Hughes JR, Goardon N, Luc S, Lower KM, Sloane-Stanley JA, Pina C, Soneji S, Renella R, Enver T, Taylor S, Jacobsen SE, Vyas P, Gibbons RJ, Higgs DR (2011) Generation of bivalent chromatin domains during cell fate decisions. *Epigenetics & chromatin* **4:** 9

Deaton AM, Bird A (2011) CpG islands and the regulation of transcription. *Genes & development* **25:** 1010-1022

Deng J, Shoemaker R, Xie B, Gore A, LeProust EM, Antosiewicz-Bourget J, Egli D, Maherali N, Park IH, Yu J, Daley GQ, Eggan K, Hochedlinger K, Thomson J, Wang W, Gao Y, Zhang K (2009) Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nature biotechnology* **27:** 353-360

Dennis K, Fan T, Geiman T, Yan Q, Muegge K (2001) Lsh, a member of the SNF2 family, is required for genome-wide methylation. *Genes & development* **15:** 2940-2944

Dhayalan A, Rajavelu A, Rathert P, Tamas R, Jurkowska RZ, Ragozin S, Jeltsch A (2010) The Dnmt3a PWWP domain reads histone 3 lysine 36 trimethylation and guides DNA methylation. *The Journal of biological chemistry* **285:** 26114-26120

Dodge JE, Okano M, Dick F, Tsujimoto N, Chen T, Wang S, Ueda Y, Dyson N, Li E (2005) Inactivation of Dnmt3b in mouse embryonic fibroblasts results in DNA

hypomethylation, chromosomal instability, and spontaneous immortalization. *The Journal of biological chemistry* **280:** 17986-17991

Doi A, Park IH, Wen B, Murakami P, Aryee MJ, Irizarry R, Herb B, Ladd-Acosta C, Rho J, Loewer S, Miller J, Schlaeger T, Daley GQ, Feinberg AP (2009) Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nature genetics* **41:** 1350-1353

Dong KB, Maksakova IA, Mohn F, Leung D, Appanah R, Lee S, Yang HW, Lam LL, Mager DL, Schubeler D, Tachibana M, Shinkai Y, Lorincz MC (2008) DNA methylation in ES cells requires the lysine methyltransferase G9a but not its catalytic activity. *The EMBO journal* **27:** 2691-2701

Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA, Haefliger C, Horton R, Howe K, Jackson DK, Kunde J, Koenig C, Liddle J, Niblett D, Otto T, Pettett R, Seemann S, Thompson C, West T, Rogers J, Olek A, Berlin K, Beck S (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature genetics* **38:** 1378-1385

Eden A, Gaudet F, Waghmare A, Jaenisch R (2003) Chromosomal instability and tumors promoted by DNA hypomethylation. *Science* **300:** 455

Egger G, Liang G, Aparicio A, Jones PA (2004) Epigenetics in human disease and prospects for epigenetic therapy. *Nature* **429:** 457-463

Ehrlich M, Gama-Sosa MA, Huang LH, Midgett RM, Kuo KC, McCune RA, Gehrke C (1982) Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic acids research* **10:** 2709-2721

Ekwattanakit S Thesis.

Engle SJ, Womer DE, Davies PM, Boivin G, Sahota A, Simmonds HA, Stambrook PJ, Tischfield JA (1996) HPRT-APRT-deficient mice are not a model for lesch-nyhan syndrome. *Human molecular genetics* **5:** 1607-1610

Erfurth FE, Popovic R, Grembecka J, Cierpicki T, Theisler C, Xia ZB, Stuart T, Diaz MO, Bushweller JH, Zeleznik-Le NJ (2008) MLL protects CpG clusters from methylation within the Hoxa9 gene, maintaining transcript expression. *Proceedings of the National Academy of Sciences of the United States of America* **105:** 7517-7522

Esteller M (2002) CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. *Oncogene* **21:** 5427-5440

Esteller M (2005) Aberrant DNA methylation as a cancer-inducing mechanism. *Annual review of pharmacology and toxicology* **45:** 629-656

Esteve PO, Chin HG, Smallwood A, Feehery GR, Gangisetty O, Karpf AR, Carey MF, Pradhan S (2006) Direct interaction between DNMT1 and G9a coordinates DNA and histone methylation during replication. *Genes & development* **20:** 3089-3103

Farcas AM, Blackledge NP, Sudbery I, Long HK, McGouran JF, Rose NR, Lee S, Sims D, Cerase A, Sheahan TW, Koseki H, Brockdorff N, Ponting CP, Kessler BM, Klose RJ (2012) KDM2B links the Polycomb Repressive Complex 1 (PRC1) to recognition of CpG islands. *eLife* **1:** e00205

Feinberg AP, Tycko B (2004) The history of cancer epigenetics. *Nature reviews Cancer* **4:** 143-153

Feldman N, Gerson A, Fang J, Li E, Zhang Y, Shinkai Y, Cedar H, Bergman Y (2006) G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis. *Nature cell biology* **8:** 188-194

Feng S, Jacobsen SE, Reik W (2010) Epigenetic reprogramming in plant and animal development. *Science* **330:** 622-627

Ferguson-Smith AC (2011) Genomic imprinting: the emergence of an epigenetic paradigm. *Nature reviews Genetics* **12:** 565-575

Ficz G, Branco MR, Seisenberger S, Santos F, Krueger F, Hore TA, Marques CJ, Andrews S, Reik W (2011) Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473:** 398-402

Ficz G, Hore TA, Santos F, Lee HJ, Dean W, Arand J, Krueger F, Oxley D, Paul YL, Walter J, Cook SJ, Andrews S, Branco MR, Reik W (2013) FGF Signaling Inhibition in ESCs Drives Rapid Genome-wide Demethylation to the Epigenetic Ground State of Pluripotency. *Cell stem cell* **13:** 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. *Nucleic acids research* **32:** 287-297

Flint J, Thomas K, Micklem G, Raynham H, Clark K, Doggett NA, King A, Higgs DR (1997) The relationship between chromosome structure and function at a human telomeric region. *Nature genetics* **15:** 252-257

Francis NJ, Kingston RE, Woodcock CL (2004) Chromatin compaction by a polycomb group protein complex. *Science* **306:** 1574-1577

Fuks F, Hurd PJ, Deplus R, Kouzarides T (2003) The DNA methyltransferases associate with HP1 and the SUV39H1 histone methyltransferase. *Nucleic acids research* **31:** 2305-2312

Futscher BW, Oshiro MM, Wozniak RJ, Holtan N, Hanigan CL, Duan H, Domann FE (2002) Role for DNA methylation in the control of cell type specific maspin expression. *Nature genetics* **31:** 175-179

Gama-Sosa MA, Wang RY, Kuo KC, Gehrke CW, Ehrlich M (1983) The 5-methylcytosine content of highly repeated sequences in human DNA. *Nucleic acids research* **11:** 3087-3095

Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. *Journal of molecular biology* **196:** 261-282

Garrick D, De Gobbi M, Lynch M, Higgs DR (2008) Switching genes on and off in haemopoiesis. *Biochemical Society transactions* **36:** 613-618

Garrick D, Fiering S, Martin DI, Whitelaw E (1998) Repeat-induced gene silencing in mammals. *Nature genetics* **18:** 56-59

Garrick D, Sharpe JA, Arkell R, Dobbie L, Smith AJ, Wood WG, Higgs DR, Gibbons RJ (2006) Loss of Atrx affects trophoblast development and the pattern of X-inactivation in extraembryonic tissues. *PLoS genetics* **2:** e58

Garrick D, Sutherland H, Robertson G, Whitelaw E (1996) Variegated expression of a globin transgene correlates with chromatin accessibility but not methylation status. *Nucleic acids research* **24:** 4902-4909

Gaudet F, Hodgson JG, Eden A, Jackson-Grusby L, Dausman J, Gray JW, Leonhardt H, Jaenisch R (2003) Induction of tumors in mice by genomic hypomethylation. *Science* **300:** 489-492

Ge YZ, Pu MT, Gowher H, Wu HP, Ding JP, Jeltsch A, Xu GL (2004) Chromatin targeting of de novo DNA methyltransferases by the PWWP domain. *The Journal of biological chemistry* **279:** 25447-25454

Geiman TM, Sankpal UT, Robertson AK, Zhao Y, Zhao Y, Robertson KD (2004) DNMT3B interacts with hSNF2H chromatin remodeling enzyme, HDACs 1 and 2, and components of the histone methylation system. *Biochemical and biophysical research communications* **318:** 544-555

Gibbons RJ, McDowell TL, Raman S, O'Rourke DM, Garrick D, Ayyub H, Higgs DR (2000) Mutations in ATRX, encoding a SWI/SNF-like protein, cause diverse changes in the pattern of DNA methylation. *Nature genetics* **24:** 368-371

Ginno PA, Lott PL, Christensen HC, Korf I, Chedin F (2012) R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. *Molecular cell* **45:** 814-825

Glass JL, Fazzari MJ, Ferguson-Smith AC, Greally JM (2009) CG dinucleotide periodicities recognized by the Dnmt3a-Dnmt3L complex are distinctive at retroelements and imprinted domains. *Mammalian genome : official journal of the International Mammalian Genome Society* **20:** 633-643

Goll MG, Bestor TH (2005) Eukaryotic cytosine methyltransferases. *Annual review of biochemistry* **74:** 481-514

Goll MG, Kirpekar F, Maggert KA, Yoder JA, Hsieh CL, Zhang X, Golic KG, Jacobsen SE, Bestor TH (2006) Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science* **311:** 395-398

Gowher H, Liebert K, Hermann A, Xu G, Jeltsch A (2005) Mechanism of stimulation of catalytic activity of Dnmt3A and Dnmt3B DNA-(cytosine-C5)-methyltransferases by Dnmt3L. *The Journal of biological chemistry* **280:** 13341-13348

Goyal R, Reinhardt R, Jeltsch A (2006) Accuracy of DNA methylation pattern preservation by the Dnmt1 methyltransferase. *Nucleic acids research* **34:** 1182-1188

Gu TP, Guo F, Yang H, Wu HP, Xu GF, Liu W, Xie ZG, Shi L, He X, Jin SG, Iqbal K, Shi YG, Deng Z, Szabo PE, Pfeifer GP, Li J, Xu GL (2011) The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. *Nature* **477:** 606-610

Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald OG, Wen B, Wu H, Liu Y, Diep D, Briem E, Zhang K, Irizarry RA, Feinberg AP (2011) Increased methylation variation in epigenetic domains across cancer types. *Nature genetics* **43:** 768-775

Hark AT, Schoenherr CJ, Katz DJ, Ingram RS, Levorse JM, Tilghman SM (2000) CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* **405:** 486-489

Hashimoto H, Horton JR, Zhang X, Bostick M, Jacobsen SE, Cheng X (2008) The SRA domain of UHRF1 flips 5-methylcytosine out of the DNA helix. *Nature* **455:** 826-829

Hata K, Kusumi M, Yokomine T, Li E, Sasaki H (2006) Meiotic and epigenetic aberrations in Dnmt3L-deficient male germ cells. *Molecular reproduction and development* **73:** 116-122

Hata K, Okano M, Lei H, Li E (2002) Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. *Development* **129:** 1983-1993

Havecker ER, Gao X, Voytas DF (2004) The diversity of LTR retrotransposons. *Genome biology* **5:** 225

He YF, Li BZ, Li Z, Liu P, Wang Y, Tang Q, Ding J, Jia Y, Chen Z, Li L, Sun Y, Li X, Dai Q, Song CX, Zhang K, He C, Xu GL (2011) Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333:** 1303-1307

Heard E (2004) Recent advances in X-chromosome inactivation. *Current opinion in cell biology* **16:** 247-255

Heintzman ND, Ren B (2009) Finding distal regulatory elements in the human genome. *Current opinion in genetics & development* **19:** 541-549

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics* **39:** 311-318

Hellman A, Chess A (2007) Gene body-specific methylation on the active X chromosome. *Science* **315:** 1141-1143

Herman JG, Baylin SB (2003) Gene silencing in cancer in association with promoter hypermethylation. *The New England journal of medicine* **349:** 2042-2054

Hermann A, Goyal R, Jeltsch A (2004) The Dnmt1 DNA-(cytosine-C5)-methyltransferase methylates DNA processively with high preference for hemimethylated target sites. *The Journal of biological chemistry* **279:** 48350-48359

Hermann A, Schmitt S, Jeltsch A (2003) The human Dnmt2 has residual DNA-(cytosine-C5) methyltransferase activity. *The Journal of biological chemistry* **278:** 31717-31721

Higgs DR, Wood WG, Jarman AP, Sharpe J, Lida J, Pretorius IM, Ayyub H (1990) A major positive regulatory region located far upstream of the human alpha-globin gene locus. *Genes & development* **4:** 1588-1601

Hodges E, Molaro A, Dos Santos CO, Thekkat P, Song Q, Uren PJ, Park J, Butler J, Rafii S, McCombie WR, Smith AD, Hannon GJ (2011) Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. *Molecular cell* **44:** 17-28

Holliday R, Pugh JE (1975) DNA modification mechanisms and gene activity during development. *Science* **187:** 226-232

Horsley SW, Daniels RJ, Anguita E, Raynham HA, Peden JF, Villegas A, Vickers MA, Green S, Waye JS, Chui DH, Ayyub H, MacCarthy AB, Buckle VJ, Gibbons RJ, Kearney L, Higgs DR (2001) Monosomy for the most telomeric, gene-rich region of the short arm of human chromosome 16 causes minimal phenotypic effects. *European journal of human genetics : EJHG* **9:** 217-225

Huang Y, Pastor WA, Shen Y, Tahiliani M, Liu DR, Rao A (2010) The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PloS one* **5:** e8888

Hughes JR, Cheng JF, Ventress N, Prabhakar S, Clark K, Anguita E, De Gobbi M, de Jong P, Rubin E, Higgs DR (2005) Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences. *Proceedings of the National Academy of Sciences of the United States of America* **102:** 9830-9835

Hung MS, Karthikeyan N, Huang B, Koo HC, Kiger J, Shen CJ (1999) Drosophila proteins related to vertebrate DNA (5-cytosine) methyltransferases. *Proceedings of the National Academy of Sciences of the United States of America* **96:** 11940-11945

Ichiyanagi K (2012) Inhibition of MspI cleavage activity by hydroxymethylation of the CpG site: a concern for DNA modification studies using restriction endonucleases. *Epigenetics : official journal of the DNA Methylation Society* **7:** 131-136

Illingworth R, Kerr A, Desousa D, Jorgensen H, Ellis P, Stalker J, Jackson D, Clee C, Plumb R, Rogers J, Humphray S, Cox T, Langford C, Bird A (2008) A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS biology* **6:** e22

Illingworth RS, Bird AP (2009) CpG islands--'a rough guide'. *FEBS letters* **583:** 1713-1720

Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr AR, James KD, Turner DJ, Smith C, Harrison DJ, Andrews R, Bird AP (2010) Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS genetics* **6:** e1001134

Inoue A, Shen L, Dai Q, He C, Zhang Y (2011) Generation and replication-dependent dilution of 5fC and 5caC during mouse preimplantation development. *Cell research* **21:** 1670-1676

Inoue A, Zhang Y (2011) Replication-dependent loss of 5-hydroxymethylcytosine in mouse preimplantation embryos. *Science* **334:** 194

Iqbal K, Jin SG, Pfeifer GP, Szabo PE (2011) Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. *Proceedings of the National Academy of Sciences of the United States of America* **108:** 3642-3647

Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, Cui H, Gabo K, Rongione M, Webster M, Ji H, Potash JB, Sabunciyan S, Feinberg AP (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature genetics* **41:** 178-186

Ito S, D'Alessio AC, Taranova OV, Hong K, Sowers LC, Zhang Y (2010) Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **466:** 1129-1133

Jeltsch A (2010) Molecular biology. Phylogeny of methylomes. *Science* **328:** 837-838

Jenuwein T, Allis CD (2001) Translating the histone code. *Science* **293:** 1074-1080

Jeong S, Liang G, Sharma S, Lin JC, Choi SH, Han H, Yoo CB, Egger G, Yang AS, Jones PA (2009) Selective anchoring of DNA methyltransferases 3A and 3B to nucleosomes containing methylated DNA. *Molecular and cellular biology* **29:** 5366-5376

Ji H, Ehrlich LI, Seita J, Murakami P, Doi A, Lindau P, Lee H, Aryee MJ, Irizarry RA, Kim K, Rossi DJ, Inlay MA, Serwold T, Karsunky H, Ho L, Daley GQ, Weissman IL, Feinberg AP (2010) Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature* **467:** 338-342

Jia D, Jurkowska RZ, Zhang X, Jeltsch A, Cheng X (2007) Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. *Nature* **449:** 248-251

Jones PA, Baylin SB (2002) The fundamental role of epigenetic events in cancer. *Nature reviews Genetics* **3:** 415-428

Jones PA, Liang G (2009) Rethinking how DNA methylation patterns are maintained. *Nature reviews Genetics* **10:** 805-811

Joshi AA, Struhl K (2005) Eaf3 chromodomain interaction with methylated H3-K36 links histone deacetylation to Pol II elongation. *Molecular cell* **20:** 971-978

Kaneda M, Okano M, Hata K, Sado T, Tsujimoto N, Li E, Sasaki H (2004) Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. *Nature* **429:** 900-903

Karimi MM, Goyal P, Maksakova IA, Bilenky M, Leung D, Tang JX, Shinkai Y, Mager DL, Jones S, Hirst M, Lorincz MC (2011) DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. *Cell stem cell* **8:** 676-687

Kaufman RM, Pham CT, Ley TJ (1999) Transgenic analysis of a 100-kb human beta-globin cluster-containing DNA fragment propagated as a bacterial artificial chromosome. *Blood* **94:** 3178-3184

Keogh MC, Kurdistani SK, Morris SA, Ahn SH, Podolny V, Collins SR, Schuldiner M, Chin K, Punna T, Thompson NJ, Boone C, Emili A, Weissman JS, Hughes TR, Strahl BD, Grunstein M, Greenblatt JF, Buratowski S, Krogan NJ (2005) Cotranscriptional set2 methylation of histone H3 lysine 36 recruits a repressive Rpd3 complex. *Cell* **123:** 593-605

Kouzarides T (2007) Chromatin modifications and their function. *Cell* **128:** 693-705

Kriaucionis S, Heintz N (2009) The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324:** 929-930

Krueger F, Andrews SR (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27:** 1571-1572

Kuhlmann M, Borisova BE, Kaller M, Larsson P, Stach D, Na J, Eichinger L, Lyko F, Ambros V, Soderbom F, Hammann C, Nellen W (2005) Silencing of retrotransposons in Dictyostelium by DNA methylation and RNAi. *Nucleic acids research* **33:** 6405-6417

Kunert N, Marhold J, Stanke J, Stach D, Lyko F (2003) A Dnmt2-like protein mediates DNA methylation in Drosophila. *Development* **130:** 5083-5090

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E,

Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ (2001) Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921

Langmead B (2010) Aligning short sequencing reads with Bowtie. *Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]* **Chapter 11:** Unit 11 17

Latham T, Gilbert N, Ramsahoye B (2008) DNA methylation in mouse embryonic stem cells and development. *Cell and tissue research* **331:** 31-55

Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, Low HM, Kin Sung KW, Rigoutsos I, Loring J, Wei CL (2010) Dynamic changes in the human methylome during differentiation. *Genome research* **20:** 320-331

Lee JH, Voo KS, Skalnik DG (2001) Identification and characterization of the DNA binding domain of CpG-binding protein. *The Journal of biological chemistry* **276:** 44669-44676

Lehnertz B, Ueda Y, Derijck AA, Braunschweig U, Perez-Burgos L, Kubicek S, Chen T, Li E, Jenuwein T, Peters AH (2003) Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. *Current biology : CB* **13:** 1192-1200

Leonhardt H, Page AW, Weier HU, Bestor TH (1992) A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. *Cell* **71:** 865-873

Li E (2002) Chromatin modification and epigenetic reprogramming in mammalian development. *Nature reviews Genetics* **3:** 662-673

Li E, Beard C, Jaenisch R (1993) Role for DNA methylation in genomic imprinting. *Nature* **366:** 362-365

Li E, Bestor TH, Jaenisch R (1992) Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69:** 915-926

Liang G, Chan MF, Tomigahara Y, Tsai YC, Gonzales FA, Li E, Laird PW, Jones PA (2002) Cooperativity between DNA methyltransferases in the maintenance methylation of repetitive elements. *Molecular and cellular biology* **22:** 480-491

Lienert F, Wirbelauer C, Som I, Dean A, Mohn F, Schubeler D (2011) Identification of genetic elements that autonomously determine DNA methylation states. *Nature genetics* **43:** 1091-1097

Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, Lucero J, Huang Y, Dwork AJ, Schultz MD, Yu M, Tonti-Filippini J, Heyn H, Hu S, Wu JC, Rao

A, Esteller M, He C, Haghighi FG, Sejnowski TJ, Behrens MM, Ecker JR (2013) Global epigenomic reconfiguration during mammalian brain development. *Science* **341:** 1237905

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462:** 315-322

Liu Y, Oakeley EJ, Sun L, Jost JP (1998) Multiple domains are involved in the targeting of the mouse DNA methyltransferase to the DNA replication foci. *Nucleic acids research* **26:** 1038-1045

Lock LF, Takagi N, Martin GR (1987) Methylation of the Hprt gene on the inactive X occurs after chromosome inactivation. *Cell* **48:** 39-46

Long HK, Blackledge NP, Klose RJ (2013a) ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection. *Biochemical Society transactions* **41:** 727-740

Long HK, Sims D, Heger A, Blackledge NP, Kutter C, Wright ML, Grutzner F, Odom DT, Patient R, Ponting CP, Klose RJ (2013b) Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. *eLife* **2:** e00348

Lorincz MC, Dickerson DR, Schmitt M, Groudine M (2004) Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. *Nature structural & molecular biology* **11:** 1068-1075

Lorsbach RB, Moore J, Mathew S, Raimondi SC, Mukatira ST, Downing JR (2003) TET1, a member of a novel protein family, is fused to MLL in acute myeloid leukemia containing the t(10;11)(q22;q23). *Leukemia* **17:** 637-641

Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389:** 251-260

Lyko F, Ramsahoye BH, Jaenisch R (2000) DNA methylation in Drosophila melanogaster. *Nature* **408:** 538-540

Lyle R, Watanabe D, te Vruchte D, Lerchner W, Smrzka OW, Wutz A, Schageman J, Hahner L, Davies C, Barlow DP (2000) The imprinted antisense RNA at the Igf2r locus overlaps but does not imprint Mas1. *Nature genetics* **25:** 19-21

Lynch M The role of primary DNA sequence in templating chromatin states.

Lynch MD, Smith AJ, De Gobbi M, Flenley M, Hughes JR, Vernimmen D, Ayyub H, Sharpe JA, Sloane-Stanley JA, Sutherland L, Meek S, Burdon T, Gibbons RJ, Garrick D, Higgs DR (2012) An interspecies analysis reveals a key role for

unmethylated CpG dinucleotides in vertebrate Polycomb complex recruitment. *The EMBO journal* **31:** 317-329

Macleod D, Ali RR, Bird A (1998) An alternative promoter in the mouse major histocompatibility complex class II I-Abeta gene: implications for the origin of CpG islands. *Molecular and cellular biology* **18:** 4433-4443

Macleod D, Charlton J, Mullins J, Bird AP (1994) Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. *Genes & development* **8:** 2282-2292

Maeshima K, Hihara S, Eltsov M (2010) Chromatin structure: does the 30-nm fibre exist in vivo? *Current opinion in cell biology* **22:** 291-297

Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y, Turecki G, Delaney A, Varhol R, Thiessen N, Shchors K, Heine VM, Rowitch DH, Xing X, Fiore C, Schillebeeckx M, Jones SJ, Haussler D, Marra MA, Hirst M, Wang T, Costello JF (2010) Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **466:** 253-257

Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, Gnirke A, Jaenisch R, Lander ES (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454:** 766-770

Mellen M, Ayata P, Dewell S, Kriaucionis S, Heintz N (2012) MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. *Cell* **151:** 1417-1430

Mertineit C, Yoder JA, Taketo T, Laird DW, Trasler JM, Bestor TH (1998) Sex-specific exons control DNA methyltransferase in mammalian germ cells. *Development* **125:** 889-897

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448:** 553-560

Mohn F, Weber M, Rebhan M, Roloff TC, Richter J, Stadler MB, Bibel M, Schubeler D (2008) Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. *Molecular cell* **30:** 755-766

Molaro A, Hodges E, Fang F, Song Q, McCombie WR, Hannon GJ, Smith AD (2011) Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell* **146:** 1029-1041

Muotri AR, Chu VT, Marchetto MC, Deng W, Moran JV, Gage FH (2005) Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* **435:** 903-910

Muotri AR, Marchetto MC, Coufal NG, Oefner R, Yeo G, Nakashima K, Gage FH (2010) L1 retrotransposition in neurons is modulated by MeCP2. *Nature* **468:** 443-446

Nady N, Lemak A, Walker JR, Avvakumov GV, Kareta MS, Achour M, Xue S, Duan S, Allali-Hassani A, Zuo X, Wang YX, Bronner C, Chedin F, Arrowsmith CH, Dhe-Paganon S (2011) Recognition of multivalent histone states associated with heterochromatin by UHRF1 protein. *The Journal of biological chemistry* **286:** 24300-24311

Nakamura T, Liu YJ, Nakashima H, Umehara H, Inoue K, Matoba S, Tachibana M, Ogura A, Shinkai Y, Nakano T (2012) PGC7 binds histone H3K9me2 to protect against conversion of 5mC to 5hmC in early embryos. *Nature* **486:** 415-419

O'Doherty A, Ruf S, Mulligan C, Hildreth V, Errington ML, Cooke S, Sesay A, Modino S, Vanes L, Hernandez D, Linehan JM, Sharpe PT, Brandner S, Bliss TV, Henderson DJ, Nizetic D, Tybulewicz VL, Fisher EM (2005) An aneuploid mouse strain carrying human chromosome 21 with Down syndrome phenotypes. *Science* **309:** 2033-2037

Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E (2007) Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nature genetics* **39:** 730-732

Okano M, Bell DW, Haber DA, Li E (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99:** 247-257

Okano M, Xie S, Li E (1998a) Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nature genetics* **19:** 219-220

Okano M, Xie S, Li E (1998b) Dnmt2 is not required for de novo and maintenance methylation of viral DNA in embryonic stem cells. *Nucleic acids research* **26:** 2536-2540

Ono R, Taki T, Taketani T, Taniwaki M, Kobayashi H, Hayashi Y (2002) LCX, leukemia-associated protein with a CXXC domain, is fused to MLL in acute myeloid leukemia with trilineage dysplasia having t(10;11)(q22;q23). *Cancer research* **62:** 4075-4080

Ooi SK, Qiu C, Bernstein E, Li K, Jia D, Yang Z, Erdjument-Bromage H, Tempst P, Lin SP, Allis CD, Cheng X, Bestor TH (2007) DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature* **448:** 714-717

Otani J, Nankumo T, Arita K, Inamoto S, Ariyoshi M, Shirakawa M (2009) Structural basis for recognition of H3K4 methylation status by the DNA methyltransferase 3A ATRX-DNMT3-DNMT3L domain. *EMBO reports* **10:** 1235-1241

Pastor WA, Aravind L, Rao A (2013) TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. *Nature reviews Molecular cell biology* **14:** 341-356

Pastor WA, Pape UJ, Huang Y, Henderson HR, Lister R, Ko M, McLoughlin EM, Brudno Y, Mahapatra S, Kapranov P, Tahiliani M, Daley GQ, Liu XS, Ecker JR, Milos PM, Agarwal S, Rao A (2011) Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* **473:** 394-397

Peters AH, O'Carroll D, Scherthan H, Mechtler K, Sauer S, Schofer C, Weipoltshammer K, Pagani M, Lachner M, Kohlmaier A, Opravil S, Doyle M, Sibilia M, Jenuwein T (2001) Loss of the Suv39h histone methyltransferases impairs mammalian heterochromatin and genome stability. *Cell* **107:** 323-337

Ponger L, Li WH (2005) Evolutionary diversification of DNA methyltransferases in eukaryotic genomes. *Molecular biology and evolution* **22:** 1119-1128

Popp C, Dean W, Feng S, Cokus SJ, Andrews S, Pellegrini M, Jacobsen SE, Reik W (2010) Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. *Nature* **463:** 1101-1105

Pradhan M, Esteve PO, Chin HG, Samaranayke M, Kim GD, Pradhan S (2008) CXXC domain of human DNMT1 is essential for enzymatic activity. *Biochemistry* **47:** 10000-10009

Pradhan S, Bacolla A, Wells RD, Roberts RJ (1999) Recombinant human DNA (cytosine-5) methyltransferase. I. Expression, purification, and comparison of de novo and maintenance methylation. *The Journal of biological chemistry* **274:** 33002-33010

Pradhan S, Roberts RJ (2000) Hybrid mouse-prokaryotic DNA (cytosine-5) methyltransferases retain the specificity of the parental C-terminal domain. *The EMBO journal* **19:** 2103-2114

Pray-Grant MG, Daniel JA, Schieltz D, Yates JR, 3rd, Grant PA (2005) Chd1 chromodomain links histone H3 methylation with SAGA- and SLIK-dependent acetylation. *Nature* **433:** 434-438

Rai K, Chidester S, Zavala CV, Manos EJ, James SR, Karpf AR, Jones DA, Cairns BR (2007) Dnmt2 functions in the cytoplasm to promote liver, brain, and retina development in zebrafish. *Genes & development* **21:** 261-266

Rai K, Huggins IJ, James SR, Karpf AR, Jones DA, Cairns BR (2008) DNA demethylation in zebrafish involves the coupling of a deaminase, a glycosylase, and gadd45. *Cell* **135:** 1201-1212

Ramsahoye BH, Biniszkiewicz D, Lyko F, Clark V, Bird AP, Jaenisch R (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:** 5237-5242

Rauch TA, Wu X, Zhong X, Riggs AD, Pfeifer GP (2009) A human B cell methylome at 100-base pair resolution. *Proceedings of the National Academy of Sciences of the United States of America* **106:** 671-678

Rebollo R, Karimi MM, Bilenky M, Gagnier L, Miceli-Royer K, Zhang Y, Goyal P, Keane TM, Jones S, Hirst M, Lorincz MC, Mager DL (2011) Retrotransposon-induced heterochromatin spreading in the mouse revealed by insertional polymorphisms. *PLoS genetics* **7:** e1002301

Reik W, Dean W, Walter J (2001) Epigenetic reprogramming in mammalian development. *Science* **293:** 1089-1093

Riggs AD (1975) X inactivation, differentiation, and DNA methylation. *Cytogenetics and cell genetics* **14:** 9-25

Robertson G, Garrick D, Wu W, Kearns M, Martin D, Whitelaw E (1995) Position-dependent variegation of globin transgene expression in mice. *Proceedings of the National Academy of Sciences of the United States of America* **92:** 5371-5375

Robertson KD, Keyomarsi K, Gonzales FA, Velicescu M, Jones PA (2000) Differential mRNA expression of the human DNA methyltransferases (DNMTs) 1, 3a and 3b during the G(0)/G(1) to S phase transition in normal and tumor cells. *Nucleic acids research* **28:** 2108-2113

Robertson KD, Wolffe AP (2000) DNA methylation in health and disease. *Nature reviews Genetics* **1:** 11-19

Rottach A, Frauer C, Pichler G, Bonapace IM, Spada F, Leonhardt H (2010) The multi-domain protein Np95 connects DNA methylation and histone modification. *Nucleic acids research* **38:** 1796-1804

Rountree MR, Bachman KE, Baylin SB (2000) DNMT1 binds HDAC2 and a new co-repressor, DMAP1, to form a complex at replication foci. *Nature genetics* **25:** 269-277

Roy N

Sado T, Fenner MH, Tan SS, Tam P, Shioda T, Li E (2000) X inactivation in the mouse embryo deficient for Dnmt1: distinct effect of hypomethylation on imprinted and random X inactivation. *Developmental biology* **225:** 294-303

Sado T, Okano M, Li E, Sasaki H (2004) De novo DNA methylation is dispensable for the initiation and propagation of X chromosome inactivation. *Development* **131:** 975-982

Sankaran VG, Xu J, Ragoczy T, Ippolito GC, Walkley CR, Maika SD, Fujiwara Y, Ito M, Groudine M, Bender MA, Tucker PW, Orkin SH (2009) Developmental and species-divergent globin switching are driven by BCL11A. *Nature* **460:** 1093-1097

Schultz DC, Ayyanathan K, Negorev D, Maul GG, Rauscher FJ, 3rd (2002) SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. *Genes & development* **16:** 919-932

Schwartz YB, Pirrotta V (2007) Polycomb silencing mechanisms and the management of genomic programmes. *Nature reviews Genetics* **8:** 9-22

Seisenberger S, Andrews S, Krueger F, Arand J, Walter J, Santos F, Popp C, Thienpont B, Dean W, Reik W (2012) The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells. *Molecular cell* **48:** 849-862

Shann YJ, Cheng C, Chiao CH, Chen DT, Li PH, Hsu MT (2008) Genome-wide mapping and characterization of hypomethylated sites in human tissues and breast cancer cell lines. *Genome research* **18:** 791-801

Sharif J, Muto M, Takebayashi S, Suetake I, Iwamatsu A, Endo TA, Shinga J, Mizutani-Koseki Y, Toyoda T, Okamura K, Tajima S, Mitsuya K, Okano M, Koseki H (2007) The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. *Nature* **450:** 908-912

Sharma S, De Carvalho DD, Jeong S, Jones PA, Liang G (2011) Nucleosomes containing methylated DNA stabilize DNA methyltransferases 3A/3B and ensure faithful epigenetic inheritance. *PLoS genetics* **7:** e1001286

Sharp AJ, Stathaki E, Migliavacca E, Brahmachary M, Montgomery SB, Dupre Y, Antonarakis SE (2011) DNA methylation profiles of human active and inactive X chromosomes. *Genome research* **21:** 1592-1600

Shearstone JR, Pop R, Bock C, Boyle P, Meissner A, Socolovsky M (2011) Global DNA demethylation during mouse erythropoiesis in vivo. *Science* **334:** 799-802

Shen L, Kondo Y, Guo Y, Zhang J, Zhang L, Ahmed S, Shu J, Chen X, Waterland RA, Issa JP (2007) Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters. *PLoS genetics* **3:** 2023-2036

Shen L, Zhang Y (2013) 5-Hydroxymethylcytosine: generation, fate, and genomic distribution. *Current opinion in cell biology* **25:** 289-296

Smith ZD, Meissner A (2013) DNA methylation: roles in mammalian development. *Nature reviews Genetics* **14:** 204-220

Song CX, He C (2013) Potential functional roles of DNA demethylation intermediates. *Trends in biochemical sciences*

Song CX, Yi C, He C (2012a) Mapping recently identified nucleotide variants in the genome and transcriptome. *Nature biotechnology* **30:** 1107-1116

Song F, Smith JF, Kimura MT, Morrow AD, Matsuyama T, Nagase H, Held WA (2005) Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **102:** 3336-3341

Song J, Rechkoblit O, Bestor TH, Patel DJ (2011) Structure of DNMT1-DNA complex reveals a role for autoinhibition in maintenance DNA methylation. *Science* **331:** 1036-1040

Song J, Teplova M, Ishibe-Murakami S, Patel DJ (2012b) Structure-based mechanistic insights into DNMT1-mediated maintenance DNA methylation. *Science* **335:** 709-712

Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, van Nimwegen E, Wirbelauer C, Oakeley EJ, Gaidatzis D, Tiwari VK, Schubeler D (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480:** 490-495

Suzuki MM, Bird A (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nature reviews Genetics* **9:** 465-476

Syeda F, Fagan RL, Wean M, Avvakumov GV, Walker JR, Xue S, Dhe-Paganon S, Brenner C (2011) The replication focus targeting sequence (RFTS) domain is a DNA-competitive inhibitor of Dnmt1. *The Journal of biological chemistry* **286:** 15344-15351

Tachibana M, Sugimoto K, Fukushima T, Shinkai Y (2001) Set domain-containing protein, G9a, is a novel lysine-preferring mammalian histone methyltransferase with hyperactivity and specific selectivity to lysines 9 and 27 of histone H3. *The Journal of biological chemistry* **276:** 25309-25317

Tachibana M, Sugimoto K, Nozaki M, Ueda J, Ohta T, Ohki M, Fukuda M, Takeda N, Niida H, Kato H, Shinkai Y (2002) G9a histone methyltransferase plays a dominant role in euchromatic histone H3 lysine 9 methylation and is essential for early embryogenesis. *Genes & development* **16:** 1779-1791

Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, Agarwal S, Iyer LM, Liu DR, Aravind L, Rao A (2009) Conversion of 5-methylcytosine to 5-

hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324:** 930-935

Takeshita K, Suetake I, Yamashita E, Suga M, Narita H, Nakagawa A, Tajima S (2011) Structural insight into maintenance methylation by mouse DNA methyltransferase 1 (Dnmt1). *Proceedings of the National Academy of Sciences of the United States of America* **108:** 9055-9059

Tan M, Luo H, Lee S, Jin F, Yang JS, Montellier E, Buchou T, Cheng Z, Rousseaux S, Rajagopal N, Lu Z, Ye Z, Zhu Q, Wysocka J, Ye Y, Khochbin S, Ren B, Zhao Y (2011) Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell* **146:** 1016-1028

Thomson JP, Skene PJ, Selfridge J, Clouaire T, Guy J, Webb S, Kerr AR, Deaton A, Andrews R, James KD, Turner DJ, Illingworth R, Bird A (2010) CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature* **464:** 1082-1086

Tufarelli C, Stanley JA, Garrick D, Sharpe JA, Ayyub H, Wood WG, Higgs DR (2003) Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. *Nature genetics* **34:** 157-165

Vakoc CR, Letting DL, Gheldof N, Sawado T, Bender MA, Groudine M, Weiss MJ, Dekker J, Blobel GA (2005) Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. *Molecular cell* **17:** 453-462

Vernimmen D, De Gobbi M, Sloane-Stanley JA, Wood WG, Higgs DR (2007) Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. *The EMBO journal* **26:** 2041-2051

Vernimmen D, Marques-Kranc F, Sharpe JA, Sloane-Stanley JA, Wood WG, Wallace HA, Smith AJ, Higgs DR (2009) Chromosome looping at the human alpha-globin locus is mediated via the major upstream regulatory element (HS -40). *Blood* **114:** 4253-4260

Viprakasit V, Harteveld CL, Ayyub H, Stanley JS, Giordano PC, Wood WG, Higgs DR (2006) A novel deletion causing alpha thalassemia clarifies the importance of the major human alpha globin regulatory element. *Blood* **107:** 3811-3812

Viprakasit V, Kidd AM, Ayyub H, Horsley S, Hughes J, Higgs DR (2003) De novo deletion within the telomeric region flanking the human alpha globin locus as a cause of alpha thalassaemia. *British journal of haematology* **120:** 867-875

Vyas P, Vickers MA, Simmons DL, Ayyub H, Craddock CF, Higgs DR (1992) Cis-acting sequences regulating expression of the human alpha-globin cluster lie within constitutively open chromatin. *Cell* **69:** 781-793

Wallace HA, Marques-Kranc F, Richardson M, Luna-Crespo F, Sharpe JA, Hughes J, Wood WG, Higgs DR, Smith AJ (2007) Manipulating the mouse genome to engineer precise functional syntenic replacements with human sequence. *Cell* **128:** 197-209

Walsh CP, Chaillet JR, Bestor TH (1998) Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nature genetics* **20:** 116-117

Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ, Zhao K (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature genetics* **40:** 897-903

Ward MC, Wilson MD, Barbosa-Morais NL, Schmidt D, Stark R, Pan Q, Schwalie PC, Menon S, Lukk M, Watt S, Thybert D, Kutter C, Kirschner K, Flicek P, Blencowe BJ, Odom DT (2013) Latent regulatory potential of human-specific repetitive elements. *Molecular cell* **49:** 262-272

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang

SP, Zdobnov EM, Zody MC, Lander ES (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520-562

Weber M, Hellmann I, Stadler MB, Ramos L, Paabo S, Rebhan M, Schubeler D (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature genetics* **39:** 457-466

Webster KE, O'Bryan MK, Fletcher S, Crewther PE, Aapola U, Craig J, Harrison DK, Aung H, Phutikanit N, Lyle R, Meachem SJ, Antonarakis SE, de Kretser DM, Hedger MP, Peterson P, Carroll BJ, Scott HS (2005) Meiotic and epigenetic defects in Dnmt3L-knockout mouse spermatogenesis. *Proceedings of the National Academy of Sciences of the United States of America* **102:** 4068-4073

Williams K, Christensen J, Pedersen MT, Johansen JV, Cloos PA, Rappsilber J, Helin K (2011) TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* **473:** 343-348

Wilson JB, Weinberg W, Johnson R, Yuspa S, Levine AJ (1990) Expression of the BNLF-1 oncogene of Epstein-Barr virus in the skin of transgenic mice induces hyperplasia and aberrant expression of keratin 6. *Cell* **61:** 1315-1327

Wilson MD, Barbosa-Morais NL, Schmidt D, Conboy CM, Vanes L, Tybulewicz VL, Fisher EM, Tavare S, Odom DT (2008) Species-specific transcription in mice carrying human chromosome 21. *Science* **322:** 434-438

Wossidlo M, Nakamura T, Lepikhov K, Marques CJ, Zakhartchenko V, Boiani M, Arand J, Nakano T, Reik W, Walter J (2011) 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. *Nature communications* **2:** 241

Wutz A, Smrzka OW, Schweifer N, Schellander K, Wagner EF, Barlow DP (1997) Imprinted expression of the Igf2r gene depends on an intronic CpG island. *Nature* **389:** 745-749

Wysocka J, Swigut T, Xiao H, Milne TA, Kwon SY, Landry J, Kauer M, Tackett AJ, Chait BT, Badenhorst P, Wu C, Allis CD (2006) A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature* **442:** 86-90

Xie S, Wang Z, Okano M, Nogami M, Li Y, He WW, Okumura K, Li E (1999) Cloning, expression and chromosome locations of the human DNMT3 gene family. *Gene* **236:** 87-95

Xu GL, Bestor TH, Bourc'his D, Hsieh CL, Tommerup N, Bugge M, Hulten M, Qu X, Russo JJ, Viegas-Pequignot E (1999) Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. *Nature* **402:** 187-191

Xu Y, Wu F, Tan L, Kong L, Xiong L, Deng J, Barbera AJ, Zheng L, Zhang H, Huang S, Min J, Nicholson T, Chen T, Xu G, Shi Y, Zhang K, Shi YG (2011) Genome-wide regulation of 5hmC, 5mC, and gene expression by Tet1 hydroxylase in mouse embryonic stem cells. *Molecular cell* **42:** 451-464

Yildirim O, Li R, Hung JH, Chen PB, Dong X, Ee LS, Weng Z, Rando OJ, Fazzio TG (2011) Mbd3/NURD complex regulates expression of 5-hydroxymethylcytosine marked genes in embryonic stem cells. *Cell* **147:** 1498-1510

Yoder JA, Bestor TH (1998) A candidate mammalian DNA methyltransferase related to pmt1p of fission yeast. *Human molecular genetics* **7:** 279-284

Zemach A, McDaniel IE, Silva P, Zilberman D (2010) Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328:** 916-919

Zentner GE, Henikoff S (2013) Regulation of nucleosome dynamics by histone modifications. *Nature structural & molecular biology* **20:** 259-266

Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW, Chen H, Henderson IR, Shinn P, Pellegrini M, Jacobsen SE, Ecker JR (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. *Cell* **126:** 1189-1201

Zhang Y, Jurkowska R, Soeroes S, Rajavelu A, Dhayalan A, Bock I, Rathert P, Brandt O, Reinhardt R, Fischle W, Jeltsch A (2010) Chromatin methylation activity of Dnmt3a and Dnmt3a/3L is guided by interaction of the ADD domain with the histone H3 tail. *Nucleic acids research* **38:** 4246-4253

Zhang Y, Rohde C, Tierling S, Jurkowski TP, Bock C, Santacruz D, Ragozin S, Reinhardt R, Groth M, Walter J, Jeltsch A (2009) DNA methylation analysis of chromosome 21 gene promoters at single base pair and single allele resolution. *PLoS genetics* **5:** e1000438

Zhou VW, Goren A, Bernstein BE (2011) Charting histone modifications and the functional organization of mammalian genomes. *Nature reviews Genetics* **12:** 7-18

Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S (2007) Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nature genetics* **39:** 61-69

Zilberman D, Henikoff S (2007) Genome-wide analysis of DNA methylation patterns. *Development* **134:** 3959-3965

Ziller MJ, Gu H, Muller F, Donaghey J, Tsai LT, Kohlbacher O, De Jager PL, Rosen ED, Bennett DA, Bernstein BE, Gnirke A, Meissner A (2013) Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500:** 477-481

Ziller MJ, Muller F, Liao J, Zhang Y, Gu H, Bock C, Boyle P, Epstein CB, Bernstein BE, Lengauer T, Gnirke A, Meissner A (2011) Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. *PLoS genetics* **7:** e1002389